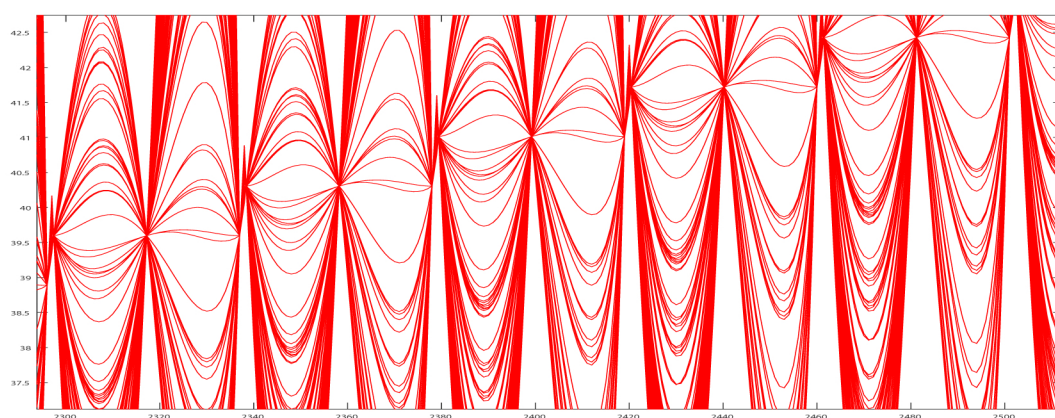


Μάριος Πούλος

Σημασιολογική Επεξεργασία της Πληροφορίας



Ανάκτηση Πληροφορίας- Συντακτική & Στατιστική Επεξεργασία
της Πληροφορίας- Υπολογιστική & Ποσοτική Γλωσσολογία
Τεχνολογίες Διαδικτύου - Οντολογίες - Εφαρμογές

Περί Σημασιολογίας



Ελληνικά Ακαδημαϊκά Ηλεκτρονικά
Συγγράμματα και Βοηθήματα
www.kallipos.gr

HEALLINK
Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών



Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο



ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ ΚΑΙ ΘΡΗΣΚΕΥΜΑΤΩΝ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης



ΕΣΠΑ
2007-2013
Ευρωπαϊκό Κοινωνικό Ταμείο

ΜΑΡΙΟΣ ΠΟΥΛΟΣ

Σημασιολογική Επεξεργασία της Πληροφορίας

Ανάκτηση πληροφορίας – Συντακτική και Στατιστική Επεξεργασία
της Πληροφορίας – Υπολογιστική και Ποσοτική Γλωσσολογία
Τεχνολογίες Διαδικτύου – Οντολογίες – Εφαρμογές

Περί Σημασιολογίας



Ελληνικά Ακαδημαϊκά Ηλεκτρονικά
Συγγράμματα και Βοηθήματα
www.kallipos.gr

Σημασιολογική Επεξεργασία της Πληροφορίας

Συγγραφή

Μάριος Πούλος

Κριτικός αναγνώστης

Σώζων Παπαβλασόπουλος

Συντελεστές έκδοσης

Γλωσσική Επιμέλεια: Μάριος Πούλος

Γραφιστική Επιμέλεια: Μάριος Πούλος

Τεχνική Επεξεργασία: Ευφροσύνη Αποστόλου

ISBN: 978-960-603-237-0

Copyright © ΣΕΑΒ, 2015



Το παρόν έργο αδειοδοτείται υπό τους όρους της άδειας Creative Commons Αναφορά Δημιουργού - Μη Εμπορική Χρήση - Όχι Παράγωγα Έργα 3.0. Για να δείτε ένα αντίγραφο της άδειας αυτής επισκεφτείτε τον ιστότοπο <https://creativecommons.org/licenses/by-nc-nd/3.0/gr/>

ΣΥΝΔΕΣΜΟΣ ΕΛΛΗΝΙΚΩΝ ΑΚΑΔΗΜΑΪΚΩΝ ΒΙΒΛΙΟΘΗΚΩΝ

Εθνικό Μετσόβιο Πολυτεχνείο

Ηρώων Πολυτεχνείου 9, 15780 Ζωγράφου

www.kallipos.gr

Πίνακας περιεχομένων

Πίνακας περιεχομένων.....	3
Πίνακας συντομεύσεων-ακρωνύμια	7
Σύνοψη	9
Εισαγωγή	10
Σημασιολογία & Επεξεργασία της Πληροφορίας	10
Εντροπία	10
Η εντροπία στη Θερμοδυναμική.....	10
Η εντροπία στη Θεωρία της Πληροφορίας.....	10
Η εντροπία του Shannon	11
Ορισμός του Shannon.....	11
Συντακτική και Στατιστική προσέγγιση της Σημασιολογικής Επεξεργασίας της Πληροφορίας.....	12
Συντακτική Προσέγγιση.....	12
Στατιστική Προσέγγιση	13
Ανάκτηση Πληροφορίας και σύγχρονα προβλήματα.....	15
Βιβλιογραφία	19
1. Ζητήματα Ανάκτησης της Πληροφορίας	21
1.1 Εισαγωγή στην διαδικασία της Ανάκτησης της Πληροφορίας	21
1.2 Μοντέλα ΑΠ και κατάταξη	23
1.2.1 VSM	23
1.2.2 Όροι στάθμισης (βάρη).....	25
1.2.3 Αξιολόγηση αποτελεσμάτων	26
1.2.4 Αξιολόγηση ΣΑΠ	27
1.2.4.1 Αποτελεσματικότητα και απόδοση ανάκτησης	27
1.2.5 Ευαισθησία – Ειδίκευση	28
1.2.6 Παρουσίαση Ανακτηθέντων Εγγράφων.....	29
1.2.7 Αρχιτεκτονική ΣΑΠ.....	30
1.2.8 Αρχιτεκτονική μηχανής αναζήτησης.....	31
1.2.9 Αρχιτεκτονική μηχανών αναζήτησης - ευρετηρίαση	31
1.2.10 Αρχιτεκτονική μηχανών αναζήτησης – διαδικασία ερωτήματος	34
Βιβλιογραφία	37
2. Συντακτική Επεξεργασία της Πληροφορίας	42
2.1 Σημασιολογικά Μοντέλα.....	42
2.2 Συλλογιστική και συμπερασματική των μηχανών	44
2.3 Αναπαράσταση Γνώσης.....	46

2.4 Μηχανισμοί εξαγωγής συμπερασμάτων (InferenceMechanisms).....	49
2.5 Γιατί το μοντέλο της RDF είναι διαφορετικό από το μοντέλο της XML	50
Το γράφημα της XML	50
Το γράφημα της RDF	52
2.6 Σημασιολογικός Ιστός	53
2.7 Υποδομή Σημασιολογικού Ιστού	54
2.8 Υπηρεσίες Σημασιολογικού Ιστού.....	56
2.8.1 Αυτόματη ανακάλυψη Υπηρεσιών Ιστού.....	56
2.8.2 Το παράδειγμα του κατανεμημένου διαμεσολαβητή βασισμένου σε εκτέλεση υπολογισμών (Agent-baseddistributedcomputingparadigm).....	57
2.8.3 Μηχάνες αναζήτησης Ιστού που βασίζονται στη σημασιολογία.....	58
2.8.4 Ψηφιακές Βιβλιοθήκες βασισμένες στη σημασιολογία.....	59
2.8.5 Νέες προκλήσεις και προοπτικές εφαρμογών για τον Σημασιολογικό Ιστό.....	59
2.8.6 Η τυπική σημασιολογία των Σημασιολογικών Γλωσσών του Ιστού	60
Βιβλιογραφία	62
3. Ποσοτική & Υπολογιστική Γλωσσολογία	64
3.1 Κείμενο και Γλωσσολογία	64
3.1.1 Το κείμενο και τα χαρακτηριστικά του.....	64
3.1.2 Αρχές κειμενικότητας	65
3.1.3 Περικείμενο	66
3.1.4 Επεξεργασία κειμένου για ανάκτηση και εξαγωγή πληροφορίας	66
3.1.5 Ποιότητα κειμένου (κατανόηση και αναγνωσιμότητα)	66
3.2 Εισαγωγή στην ΥΓ και τις βασικές της έννοιες.....	68
3.2.1 Μορφολογία.....	70
3.2.2 Σύνταξη.....	71
3.2.3 Μέρη του λόγου και επισημειωτές	72
3.2.4 Σημασιολογία	73
3.2.4.1 Προβλήματα στη Σημασιολογία	73
3.2.4.2 Μέτρα σημασιολογικής εγγύτητας (ομοιότητας)	74
3.3 Ποσοτική Γλωσσολογία.....	74
3.3.1 Στατιστικοί νόμοι στην ΠΓ	75
3.3.2 Νόμος των Menzerath-Altmann	76
3.3.3 Νόμος του Zipf	77
3.3.4 Θεωρία της Πληροφορίας και ΠΓ	78
Βιβλιογραφία	80
4. Οντολογικές Προσεγγίσεις	85
4.1 Οντολογίες	85
4.2 Διαμεσολαβητές (Agents)	88

4.3 Οντολογικές Γλώσσες.....	92
4.4 Οντολογικός Μοντελισμός	98
Βιβλιογραφία	103
5. Στατιστική Επεξεργασία της Πληροφορίας.....	105
5.1 Στοιχεία Στατιστικής.....	105
5.2 Σχεδιασμός Πειραμάτων	107
5.2.1 Περιγραφική Στατιστική	107
5.2.2 Κατανομή συχνότητας και παρουσίαση δεδομένων.....	108
5.2.3 Βασικά στατιστικά μέτρα.....	109
5.2.4 Κανονική κατανομή και τυποποιημένη τιμή Z	110
5.2.5 Μη κανονική κατανομή	110
5.2.6 Επαγωγική Στατιστική και θεωρία πιθανοτήτων.....	111
5.2.7 Έλεγχος Στατιστικών Υποθέσεων	112
5.3 Συστήματα Ομαδοποίησης & Κατηγοριοποίησης.....	114
5.3.1 Δέντρα αποφάσεων SVM - BinaryTrees (BTS)	114
5.3.2 Μέθοδοι με επίβλεψη και χωρίς επίβλεψη	117
5.4 Τεχνητά Νευρωνικά Δίκτυα.....	118
5.4.1 Ο Ανθρώπινος Εγκέφαλος.....	118
5.4.2 Μοντέλα Νευρώνων	121
5.4.1 LearningVectorQuantizer (LVQ).....	122
5.4.2 Νευρωνικό δίκτυο Τύπου Radial Basis Function (RBF)	123
Βιβλιογραφία	126
6. Παραδείγματα εξαγωγής Γνώσης.....	128
6.1 Εμπειρικές Μέθοδοι Κατηγοριοποίησης/Ταυτοποίησης.....	128
6.1.1 Η θεμελιώδης αρχή της Μεθόδου.....	128
6.1.2 Ανάπτυξη αλγορίθμου κατασκευής κυρτού πολυγώνου	129
6.1.2.1 Τομή δύο κυρτών πολυγώνων.....	131
6.1.3 Υλοποίηση αλγορίθμων υπολογιστικής γεωμετρίας.....	133
6.2 Ανάλυση Κειμένων με Στατιστικά Εργαλεία	135
6.2.1 Μεθοδολογία	136
6.2.2 Η βάση του Αλγόριθμου	137
6.2.3 Κριτήριο ομοιότητας	138
6.2.4 Μετατροπή Μεταβλητών	139
6.2.5 Στατιστική θεμελίωση.....	139
6.2.6 Πειραματικό Μέρος.....	140
6.2.7 Εφαρμογή του Αλγόριθμου και Στατιστική Επεξεργασία	140
6.2.8 Επαναληπτική Διαδικασία	142
6.2.9 Επαναληπτική Διαδικασία στο Σύνολο των Δεδομένων.....	142

6.2.10 Συμπεράσματα	144
Μελλοντικά Σχέδια:.....	144
6.3 Επεξεργασία κειμένου για ανάκτηση και εξαγωγή πληροφορίας	145
6.4 Αξιολόγηση Βιβλιοθήκης με χρήση Νευρωνικού Δικτύου	147
Βιβλιογραφία	151
ΣΥΝΟΛΙΚΗ ΒΙΒΛΙΟΓΡΑΦΙΑ	153
ΕΥΡΕΤΗΡΙΟ	159

Πίνακας συντομεύσεων-ακρωνύμια

ACL	Agent Communication Language
ANSI	American National Standards Institute
APA	American Psychological Association (style <i>body_first paragraph</i>)
CAS	Chemical Abstracts Service
CGI	Common Gateway Interface
CORBA	Common Object Request Broker Architecture
CSA	Constrained Spreading Activation
CWTS	Centre for Science and Technology Studies
DAML-S	Darpa Agent Markup Language - Services
DL	Description Logics
DOI's	Digital Object Identifier Ψηφιακά αναγνωριστικά αντικειμένου
HITS	Hypertext-Induced topic search
HTML	HyperText Markup Language Γλώσσα Σήμανσης Υπερκειμένου
IDIS	Institut für Dokumentation und Information über Sozial medizin und öffentliches Gesundheitswesen
INIST	Ινστιτούτο Επιστημονικών και Τεχνικών Πληροφοριών Digital Bibliography & Library Project
IoP	Institute of Physics
ISBN	International Standard Book Number Διεθνής μοναδικός αριθμός βιβλίου
ISI	Thomson Institute for Scientific Information Ινστιτούτο Thomson για την επιστημονική πληροφορία,
ISSI	International Society for Scientometrics and Informetrics Διεθνής εταιρία της Επιστημομετρίας και της Πληροφοριομετρίας
JASIS	Journal of the American Society for Information Science
JCI	Journal of Clinical Investigation Περιοδικό κλινικών Ερευνών
JCR	Journal Citation Reports
JCR	Journal Citation Reports
KIF	Knowledge Interchange Format
LVQ	Learning Vector Quantization
NSAPI	Network Service Access Point Identifier
OAA	Open Agent Architecture
OASIS	Organization for the Advancement of Structured Information Standards
OKBS	Open Knowledge Base Connectivity
OPAL	Object, Process, Actor Language
OWL	Web Ontology Language
PBRF	Performance-Based Research Fund Απόδοση χρηματοδότηση της Έρευνας
PLoS	Public Library of Science
RDF	Radial-Basis Function

RDF	Resource Description Framework
RDFS	Resource Description Framework Schema
RMI	Remote Method Invocation
SCC	Strongly Connected Component
SCI	Science Citations Indexes Δείκτης αναφορών επιστημών
SEI	Science and Engineering Indicators Δείκτης Επιστήμης και Τεχνολογίας
SGML	Standard Generalized Markup Language
SOAP	Simple Object Access Protocol
SPIN	Searchable Physics Information Notices
UDDI	Universal Description, Discovery, and Integration
URI	Uniform Resource Identifier
URL	Uniform Resource Locator Ενιαίος Εντοπιστής Πόρων
VQR	Valutazione Quinquennale della Ricerca
VSM	Vector Space Model
W3C	World Wide Web Consortium
WWW	World Wide Web
XML	eXtensible Markup Language
XSD	XML Schema Definition
ΣΑΠ	Συστήματα Ανάκτησης της Πληροφορίας

Σύνοψη

Το παρόν σύγγραμμα δημιουργήθηκε ώστε να καλυφθούν διδακτικές ανάγκες των φοιτητών προπτυχιακού κύκλου με προσανατολισμό την Επιστήμη της Πληροφορίας. Η στοχοθεσία των ενοτήτων ήταν να δοθεί ένας ολοκληρωμένος χάρτης περί σημασιολογίας της πληροφορίας. Σύμφωνα με τα παραπάνω δόθηκε έμφαση σε (3) τρεις τομείς.

Στο πρώτο τομέα στην εισαγωγή, περιγράφεται η έννοια της πληροφορίας και η μεθοδολογική προσέγγιση της επεξεργασίας της δηλαδή ο συντακτικός και ο στατιστικός, ενώ στο πρώτο κεφάλαιο αναπτύσσονται οι ισχύουσες πρακτικές ανάκτησής της.

Στο δεύτερο τομέα και ειδικότερα στο κεφάλαιο (2) γίνεται μια εκτενής αναφορά στη συντακτική επεξεργασία της πληροφορίας και ειδικότερα επί θεμάτων τεχνολογιών της πληροφορίας όπως σχημάτων μεταδεδομένων, μηχανισμού εξαγωγής συμπερασμάτων, μηχανών αναζήτησης με χρήση σημασιολογικών γλωσσών και τη θεμελιώδη ανάπτυξη συλλογιστικής και συμπερασματικής των μηχανών. Επίσης στο τέταρτο (4ο) κεφάλαιο αναπτύσσεται εκτενώς η έννοια της οντολογίας ως επίσης και οι γλώσσες γραφής της οντολογίας.

Στον τρίτο τομέα και συγκεκριμένα στο τρίτο (3ο) κεφάλαιο αναπτύσσονται τομείς που a priori επιλύουν γλωσσικά διαλειτουργικά προβλήματα στις μεθοδολογίες της συντακτικής επεξεργασίας της πληροφορίας και αυτοί είναι οι η υπολογιστική και η ποσοτική γλωσσολογία.

Τέλος στον τρίτο τομέα και ειδικότερα στα κεφάλαια (5) πέντε και (6) έξι αναπτύσσονται αντιπροσωπευτικές ταξινομικές μεθοδολογίες της πληροφορίας με στατιστική προσέγγιση ακολουθούμενες από τις αντίστοιχες εφαρμογές.

Εισαγωγή

Περίληψη

Στόχος κεφαλαίου: Παρουσίαση και ανάλυση τι είναι σημασιολογία μέσω της επεξεργασίας της πληροφορίας. Τα παραπάνω προϋποθέτουν μία εμπειριστατωμένη ανάλυση τι είναι εντροπία πληροφορίας και πως αυτή η έννοια εισήχθη από τον Κλωντ Σάνον το 1948 ως επίσης και ο καθορισμός των δομημένων και αδόμητων πηγών πληροφόρησης. Επίσης θίγεται το πρόβλημα πρόβλεψης τιμών μίας τυχαίας μεταβλητής μέσω ποσοτικοποίησης της πληροφορίας. Στη συνέχεια περιγράφονται εισαγωγικά αλγόριθμοι και πρακτικές ποσοτικοποίησης και μετα-περιγραφής της πληροφορίας σε διάφορους επιστημονικούς τομείς όπως στα δεδομένα του διαδικτύου (Internet), την επιστήμη της γλωσσολογίας, της ανθρώπινης αντίληψης και της στατιστικής. Τέλος, γίνεται μνεία πώς αυτές οι πρακτικές μπορούν να ενσωματωθούν στις επιστήμες της Πληροφορίας όπως της Βιβλιοθηκονομίας και της Αρχειονομίας.

Προαπαιτούμενες γνώσεις

Στατιστική Ανάλυση, Θεωρία Πιθανοτήτων, Γλώσσες Σήμανσης

Σημασιολογία & Επεξεργασία της Πληροφορίας

Η σημασιολογία, στον προπτυχιακό κύκλο, πρέπει να θεωρείται ευρύ διεπιστημονικό πεδίο που έχει τις πηγές του κυρίως στην ψυχολογία, στην τεχνητή νοημοσύνη, στη γλωσσολογία, στη φιλοσοφία και στη νευροεπιστήμη. Η διάκριση ως ένα βαθμό, ό ένας από τον άλλο, άφορα τις μεθόδους, τις θεωρίες και τα αποτελέσματα, ωστόσο τούς χαρακτηρίζει και εντυπωσιακή ενότητα, επειδή συγκλίνουν και ως προς τα κεντρικά ερωτήματα και κατά το ότι σε όλους έχει αναδυθεί ως κοινό θεωρητικό υπόβαθρο ή υπολογιστική αντίληψη, αυτή δηλαδή που εστιάζεται στην επεξεργασία πληροφοριών. Στο παρόν σύγγραμμα επιχειρείται να τηρηθεί μια συνεπής υπολογιστική γραμμή, τονίζοντας όμως δεόντως και την ιδιαίτερη συμβολή του κάθε κλάδου εστιασμένη κυρίως στον κλάδο της τεχνητής νοημοσύνης. Η επεξεργασία της πληροφορίας αναδύει χαρακτηριστικά ασαφή που προέρχονται κατά κύριο λόγο από την ευρεία διάχυση της πληροφορίας. Η ευρεία διάχυση προϋποθέτει επεξεργασία της πληροφορίας χωρίς φορμαλισμούς που οδηγεί στο πρόβλημα της διαλειτουργικότητας. Το πρόβλημα αυτό εγείρει ζητήματα ενοποίησης (Integration) και της προτυποποίησης (Standardization). Γεννάται το εξής ερώτημα πώς είναι δυνατόν να αναλυθεί η σημασιολογική επεξεργασία της πληροφορίας αν πρώτα δεν αναλυθούν ζητήματα ερμηνείας της διάχυσής σε επίπεδο εντροπίας. Στις ενότητες που ακολουθούν επιχειρείται να δοθεί μία τέτοια προσέγγιση.

Εντροπία

Η έννοια της εντροπίας χρησιμοποιήθηκε αρχικά στη Θερμοδυναμική ως επιχείρημα της ορθότητας του δεύτερου θερμοδυναμικού νόμου. Σύμφωνα με το δεύτερο θερμοδυναμικό νόμο, η κατασκευή αεικίνητου σώματος είναι αδύνατη καθώς αδύνατη είναι η παραγωγή ενέργειας εκ του μηδενός. Σε οποιαδήποτε διαδικασία, ένα μέρος της διαθέσιμης ενέργειας μετατρέπεται σε απώλειες λόγω θερμότητας ή αναδιάταξης των χημικών ενώσεων ή σε άλλες αλλαγές εντροπίας, και συνεπώς δεν μπορεί να χρησιμοποιηθεί εκ νέου για την παραγωγή μηχανικού έργου [1, 2].

Η εντροπία στη Θερμοδυναμική

Το 1865, ο Γερμανός φυσικός Rudolf Clausius παρατήρησε κατά την διάρκεια μιας αντιστρεπτής μεταβολής υπό σταθερή θερμοκρασία, ότι το πηλίκο της θερμότητας που μεταφέρεται μεταξύ του υπό μελέτη συστήματος και του περιβάλλοντος προς την θερμοκρασία κάτω από την οποία πραγματοποιείτε η μεταβολή παραμένει σταθερό. Ο Clausius θεώρησε πως το παραπάνω πηλίκο αναπαριστά πραγματικό φυσικό μέγεθος και το ονόμασε μεταβολή της εντροπίας.

Η εντροπία στη Θεωρία της Πληροφορίας

Στη θεωρία της πληροφορίας [3-6], της επιστήμης δηλαδή που ασχολείται με την ποσοτικοποίηση της πληροφορίας, η έννοια της εντροπίας εισήχθη για πρώτη φορά το 1948. Συγκεκριμένα, ο Claude Shannon χρησιμοποίησε την έννοια στην δημοσίευσή του με τίτλο “A Mathematical Theory of Communication” [7]

για να περιγράψει τη μέση ποσότητα πληροφορίας που εμπεριέχεται σε ένα μήνυμα, ενός δηλαδή γεγονότος ή ενός δείγματος ή ενός χαρακτήρα που προκύπτει από μια κατανομή ή από ένα σύνολο δεδομένων. Η εντροπία του Shannon είναι μέτρο της αβεβαιότητας που χαρακτηρίζει μια πηγή μετάδοσης πληροφοριών. Εν προκειμένω, η πηγή χαρακτηρίζεται από μια συνάρτηση πιθανότητας εμφάνισης των δειγμάτων που μπορεί να προκύψουν από αυτή. Η συνάρτηση πιθανότητας σε συνδυασμό με την ποσότητα πληροφορίας κάθε μηνύματος σχηματίζουν μια μεταβλητή της οποίας η μέση τιμή συμβολίζει την μέση τιμή της πληροφορίας που εμπεριέχει αυτή η κατανομή ή αλλιώς την εντροπία της [8].

Η εντροπία του Shannon

Στη θεωρία της πληροφορίας, της επιστήμης δηλαδή που ασχολείται με την ποσοτικοποίηση της πληροφορίας, η έννοια της εντροπίας εισήχθη για πρώτη φορά το 1948. Συγκεκριμένα, ο Claude Shannon χρησιμοποίησε την έννοια στην δημοσίευσή του με τίτλο “A Mathematical Theory of Communication” για να περιγράψει τη μέση ποσότητα πληροφορίας που εμπεριέχεται σε ένα μήνυμα, ενός δηλαδή γεγονότος ή ενός δείγματος ή ενός χαρακτήρα που προκύπτει από μια κατανομή ή από ένα σύνολο δεδομένων. Η εντροπία του Shannon είναι μέτρο της αβεβαιότητας που χαρακτηρίζει μια πηγή μετάδοσης πληροφοριών. Εν προκειμένω, η πηγή χαρακτηρίζεται από μια συνάρτηση πιθανότητας εμφάνισης των δειγμάτων που μπορεί να προκύψουν από αυτή. Η συνάρτηση πιθανότητας σε συνδυασμό με την ποσότητα πληροφορίας κάθε μηνύματος σχηματίζουν μια μεταβλητή της οποίας η μέση τιμή συμβολίζει την μέση τιμή της πληροφορίας που εμπεριέχει αυτή η κατανομή ή αλλιώς την εντροπία της .

Ορισμός του Shannon

Έστω μια τυχαία διακριτή μεταβλητή X ενός συνόλου A και η συνάρτηση πιθανότητας $p(x) = \Pr\{X = x\}, x \in A$. Η εντροπία $H(x)$ μιας διακριτής μεταβλητής X ορίζεται ως εξής:

$$H(x) = - \sum_{x \in A} p(x) * \log_2 p(x) \quad (1)$$

Με αυτό τον τρόπο ορίζονται δύο είδη εντροπίας η από κοινού και η δεσμευμένη. Τότε, επεκτείνοντας τον ορισμό της εντροπίας σε ένα ζευγάρι διακριτών τυχαίων μεταβλητών (X, Y) προκύπτει η από κοινού εντροπία (joint entropy) ως ακολούθως [4]:

$$H(X, Y) = - \sum_{x \in A} \sum_{y \in B} p(x, y) * \log p(x, y) \quad (2)$$

όπου $p(x, y)$ είναι η από κοινού κατανομή πιθανότητας των μεταβλητών X και Y . Η δεσμευμένη εντροπία (conditional entropy) ορίζεται ως ακολούθως [4]:

$$H(X | Y) = - \sum_{x \in A} \sum_{y \in B} p(x, y) * \log p(x | y) \quad (3)$$

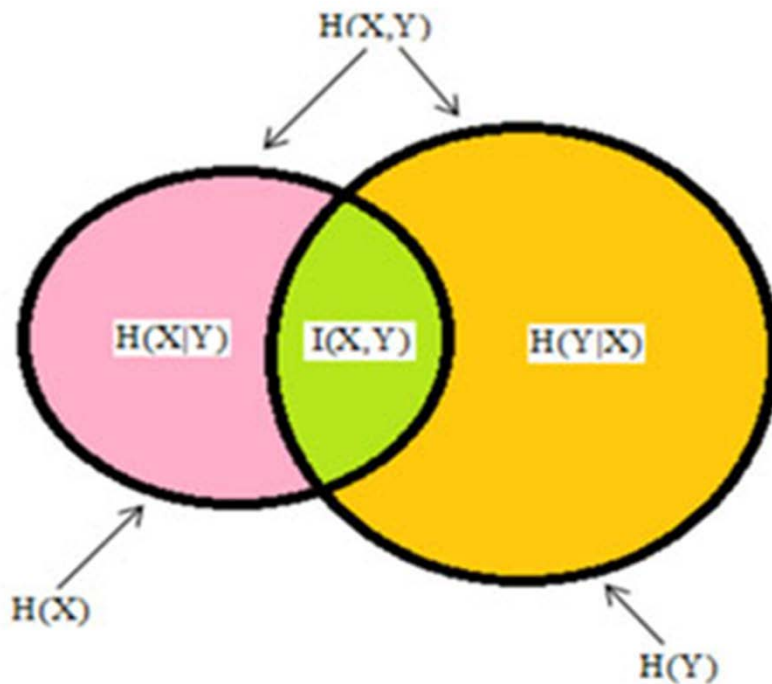
Από τους ορισμούς της από κοινού και δεσμευμένης εντροπίας προκύπτει ότι $H(X, Y) = H(X) + H(Y | X)$. Όπως σημειώθηκε προηγουμένως, η εντροπία μιας τυχαίας μεταβλητής είναι η μέση ποσότητα πληροφορίας που απαιτείται για να περιγράψει τη μεταβλητή. Από την άλλη πλευρά, η σχετική εντροπία είναι το μέτρο της απόστασης μεταξύ δύο κατανομών. Ως κατανομή μπορεί να θεωρηθεί μια τυχαία συνάρτηση πιθανότητας. Η σχετική εντροπία, $D(p||q)$, εκφράζει την ανεπάρκεια της υπόθεσης ότι μια κατανομή είναι q ενώ η πραγματική κατανομή είναι p . Για παράδειγμα, αν γνωρίζαμε την αληθινή κατανομή p μίας τυχαίας μεταβλητής, τότε θα μπορούσαμε να κατασκευάσουμε ένα κώδικα με μέσο μήκος περιγραφής $H(p)$ bits. Ωστόσο, θεωρώντας τη κατανομή q , χρησιμοποιούμε κώδικα με μέσο μήκος $H(p) + D(p||q)$ bits για να περιγράψουμε τη τυχαία μεταβλητή. Η σχετική εντροπία (relative entropy) ή απόσταση Kullback (Kullback Leibler distance) μεταξύ δύο κατανομών πιθανοτήτων $p(x)$ και $q(x)$ ορίζεται ως ακολούθως:

$$I(X, Y) = \sum_{x \in A} \sum_{y \in B} p(x, y) * \log \frac{p(x, y)}{p(x)p(y)} = D(p(x, y) || p(x)p(y)) \quad (4)$$

Από τον ορισμό της αμοιβαίας πληροφορίας προκύπτει ότι

$$I(X, Y) = H(X) - H(X | Y) \quad (5)$$

Στο διάγραμμα Venn του Σχήματος 1 φαίνεται η σχέση μεταξύ $H(X)$, $H(Y)$, $H(X, Y)$, $H(X|Y)$, $H(Y|X)$ και $I(X, Y)$. Από το Σχήμα 1 προκύπτει ότι η αμοιβαία πληροφορία αντιστοιχεί στη τομή της πληροφορίας που περιέχεται στη μεταβλητή X και της πληροφορίας που περιέχεται στη μεταβλητή Y .



Σχήμα 1. Σχέση μεταξύ Εντροπίας και Αμοιβαίας Πληροφορίας

Η συνάρτηση πιθανότητας $p(x)$ όπως δίνεται στην εξίσωση 2 ορίζει την πιθανότητα πυκνότητας της πληροφορίας η οποία είναι εξίσου σημαντική συνάρτηση στον τομέα των σημασιολογικών δικτύων.

Συντακτική και Στατιστική προσέγγιση της Σημασιολογικής Επεξεργασίας της Πληροφορίας

Συντακτική Προσέγγιση

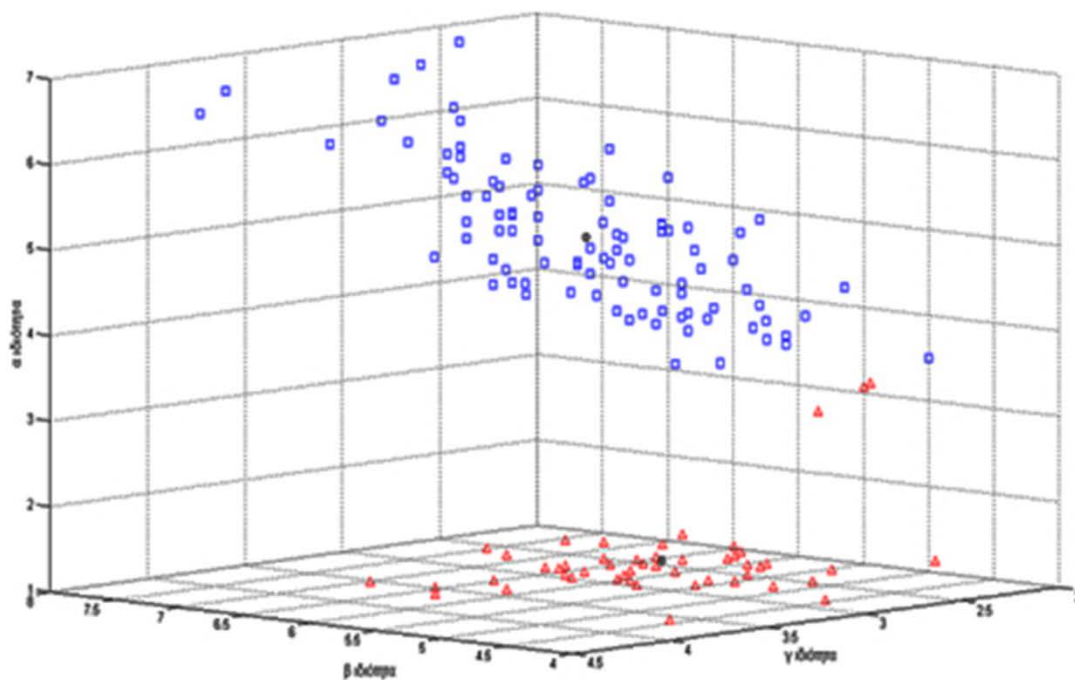
Επανερχόμενοι στο πρόβλημα της διαλειτουργικότητας, σε αυτό το στάδιο μπορεί να τεθεί ως ορισμός της «Η ικανότητα ενός συστήματος, να συνεργάζεται με άλλα συστήματα χωρίς την ανάγκη ιδιαίτερης προσπάθειας». Σε αυτό το σημείο ορίζεται η έννοια της ετερογένειας ως σημαντικό χαρακτηριστικό της διαλειτουργικότητας και αποτελεί μια άλλη εκδοχή της εντροπίας. Στο σημείο αυτό θα πρέπει να εισαχθεί η έννοια των μεταδεδομένων τα οποία δηλωτικά είναι δεδομένα τα οποία περιγράφουν άλλα δεδομένα τα οποία αποτελούν μια πηγή πληροφορίας. Επί της ουσίας δίδεται η δυνατότητα η μορφή αυτής της ετερογένειας να δεικτοδοτηθεί. Ο καθορισμός της ετερογένειας αν αυτή εξετασθεί σε επίπεδο διαμορφωμένης πληροφορίας που ονομάζεται τεκμήριο, αυτή δύναται να εκφρασθεί σε δύο συστατικά τη συντακτική και τη σημασιολογική. Η συντακτική ετερογένεια καθορίζεται στο ότι τα ίδια τα δεδομένα αναπαρίστανται με διαφορετικό τρόπο και αυτό παρουσιάζεται λόγω της αυξημένης εκφραστικότητας που χαρακτηρίζει την ανθρώπινη γλώσσα η οποία οδηγεί σε ένα πλουραλισμό κωδικοποιήσεων δεδομένων πχ διαφορετικές κλίμακες, διαφορετικά λειτουργικά υπολογιστικά συστήματα, διαφορετικές πλατφόρμες κλπ). Η σημασιολογική ετερογένεια κατά Goh σχετίζεται σε ζητήματα όπως: σύγκρουσης αντιθέσεων, κλιμάκωσης αντιθέσεων και διαφορά ονομάτων. Στην περίπτωση της σύγκρουσης αντιθέσεων τα αντικείμενα πληροφορίας δείχνουν να έχουν την ίδια έννοια αλλά ουσιαστικά διαφέρουν. (π.χ. χρονικό πλαίσιο). Ενώ στην κλιμάκωση αντιθέσεων μπορεί να επεξηγηθεί μέσω του καθορισμού μιας αξίας από διαφορετικά

συστήματα αναφοράς. (π.χ. αξία νομίσματος). Τέλος η περίπτωση διαφοράς ονομάτων αναφέρεται στη χρησιμοποίηση διαφορετικών σχημάτων ονομάτων πληροφοριών. (π.χ. ομώνυμα / συνώνυμα). Επίλυση των παραπάνω μορφών ετερογένειας δόθηκε από τη γλώσσα σημειοθέτησης γενικής χρήσης την XML (Extensible Markup Language - Επεκτάσιμη γλώσσα σημειοθέτησης). Η γλώσσα αυτή θεωρείται επεκτάσιμη γλώσσα διότι δίνεται η δυνατότητα στους χρήστες να καθορίσουν τις δικές τους ετικέτες. Ο πρωταρχικός της στόχος είναι να διευκολύνει το διαμοιρασμό δομημένων δεδομένων ιδίως μέσω του διαδικτύου. Χρησιμοποιείται τόσο για την κωδικοποίηση εγγράφων όσο και για να μετατραπούν τα δεδομένα σε δυαδική μορφή, προκειμένου να αποθηκευθούν τοπικά ή να μεταφερθούν μέσω δικτύου. Σχεδιάστηκε ως ένα υποσύνολο της Standard Generalized Markup Language (SGML), και δημιουργήθηκε ώστε να είναι σχετικά ευανάγνωστη για το χρήστη. Στη συνέχεια η δυνατότητα καθοδηγούμενης περιγραφής του εγγράφου μέσω του (Document Type Definition) οδήγησε σε διαλειτουργικές λύσεις αλλά από την άλλη πλευρά αύξησε την πολυπλοκότητα αφού σε θέματα πολύπλοκης και εξειδικευμένης περιγραφής εγγράφου όπως η ιατρική πληροφορία το βάθος του παραγόμενου XML δένδρου θεωρείται πολύ μεγάλο. Το ζήτημα αυτό δημιούργησε μία δεύτερη μετρική εντροπίας που βασίζεται στη μέτρηση σε ένα σταθμισμένο καταμετρητή των στοιχείων που περιέχονται στα έγγραφα που ονομάζεται βάρος το οποίο σχετίζεται με το βάθος του μέσα στο δέντρο XML. Η χαρτογράφηση αυτή των βαρών του κάθε στοιχείου του εγγράφου αποσκοπεί στο καθορισμό τριών μεταβλητών που σχετίζονται με το μέγεθος του εγγράφου, τη πυκνότητα των πληροφοριών του και τη δομή της κανονικότητας. Η πυκνότητα πληροφορίας αποτελεί σημαντικό παράγοντα καθορισμού του είδους του εγγράφου για παράδειγμα μιας χαμηλής πυκνότητας πληροφοριών έγγραφο σημαίνει ότι το έγγραφο περιέχει κυρίως διαρθρωτικά πληροφορίες, ενώ μια υψηλή τιμή σημαίνει ότι το έγγραφο περιέχει κυρίως πραγματικό περιεχόμενο (μη-στοιχείου και το περιεχόμενο ορισμός μη-αξία). Οι μεθοδολογίες και οι τεχνικές ταξινόμησης ψηφιακών πληροφοριών από το χώρο του πολιτισμού (υλικό από αρχεία, βιβλιοθήκες, μουσεία) αποτελούν ένα σημαντικό αντικείμενο έρευνας της Επιστήμης της Πληροφορίας. Οι σημερινές μεθοδολογίες κυρίως εστιάζονται σε τεχνικές παραγωγής μεταδεδομένων. Αντικειμενικός στόχος σε αυτή την περίπτωση είναι η μονοσήμαντη δομική περιγραφή εγγράφων, η οποία βελτιώνει την έκφραση, διάθεση, μεταφορά και αποθήκευση των πληροφοριών. Παράλληλα, η εφαρμογή της XML ως γλώσσας έκφρασης πληροφοριών επιλύει προβλήματα ετερογένειας στην κωδικοποίηση δεδομένων όπως αναφέρθηκε και προηγούμενα. Σύμφωνα με τη διεθνή βιβλιογραφία, η σύγχρονη πρακτική οργάνωσης και δημιουργίας κανόνων συσχέτισης ανάμεσα στη γλώσσα XML και σε γλώσσες οντολογιών, όπως είναι η RDF και η OWL, στοχεύει στη σημασιολογική περιγραφή του περιεχομένου δομημένων εγγράφων και αποτελεί την κορωνίδα της παραγωγής σημασιολογικών δικτύων ακολουθώντας τη πρακτική επίλυσης της συντακτικής ετερογένειας.

Στατιστική Προσέγγιση

Κυρίαρχο σημείο στη σημασιολογική επεξεργασία με συντακτικό τρόπο ορίζεται μέσω της μνήμης. Η δημιουργία εικονικής μνήμης αποτελεί αντικείμενο μελέτης της Επαγωγικής Στατιστικής (ΕΣ). Η επαγωγική στατιστική αποτελεί τη διαδικασία μία σειρά μεθόδων που χρησιμοποιείται για την εξαγωγή συμπερασμάτων σχετικά με τα χαρακτηριστικά του πληθυσμού που βασίζονται σε δεδομένα από το δείγμα. Σε αντίθεση στη Περιγραφική Στατιστική (ΠΣ) περιγράφονται τα στατιστικά δείγματα κατά ομάδες τα οποία στη συνέχεια παρουσιάζονται υπό μορφή ανάλυσης σε πίνακες, διαγράμματα με χαρακτηριστικές τιμές, ή ιδιότητες. Η ανάλυση των δειγμάτων μας οδηγεί σε δύο μεθοδολογίες ανάλυσης την μονοπαραμετρική και την πολυπαραμετρική. Η μονοπαραμετρική ανάλυση περιλαμβάνει την περιγραφή της κατανομής μιας μόνο μεταβλητής, και συμπεριλαμβάνει την κεντρική τάση και τη διασπορά. Το σχήμα της κατανομής μπορεί επίσης να περιγραφεί μέσω δεικτών όπως η ασυμμετρία και η κύρτωση. Χαρακτηριστικά της κατανομής μιας μεταβλητής περιλαμβάνει επίσης την απεικόνιση σε μορφή γραφικών ή σε μορφή πίνακα των ιδιοτήτων των δειγμάτων. Η πολυπαραμετρική ανάλυση χρησιμοποιείται όταν ένα δείγμα αποτελείται από περισσότερες από μία μεταβλητές, η (ΠΣ) πρέπει να χρησιμοποιείται για να περιγράψει τη σχέση μεταξύ ζευγών μεταβλητών. Οι παραπάνω διαδικασίες μπορούν να καθορίσουν τεχνικές ομαδοποίησης των δειγμάτων που αυτές μπορούν να ορίσουν νέες σημασιολογικές προσεγγίσεις ανάλυσης των δεδομένων με βάση την αρχή της ομαδοποίησης. Η Ανάλυση Ομάδων (Clusters) ή Ομαδοποίηση είναι η κατάταξη ενός συνόλου δεδομένων, έτσι ώστε τα δεδομένα της ίδιας ομάδας να ομοιάζουν όσο το δυνατόν περισσότερο μεταξύ τους, από ότι άλλων ομάδων (clusters), τα συνδέει δηλαδή μια τουλάχιστον κοινή ιδιότητα. Η μεθοδολογία αυτή αποτελεί το βασικό εργαλείο του επιστημονικού κλάδου εξόρυξης δεδομένων, αλλά και κοινή τεχνική ανάλυσης στατιστικών δεδομένων που εφαρμόζεται σε πολλούς επιστημονικούς τομείς, συμπεριλαμβανομένων των Μηχανικής Μάθησης, Αναγνώρισης Μορφών (patterns, motives), ανάλυσης εικόνας, ανάκτησης

πληροφοριών. Η Ανάλυση Ομάδων επιτελείται με πολλές διαφορετικές αλγοριθμικές τεχνικές. Αυτό μπορεί να επιτευχθεί με διάφορους αλγόριθμους, που διαφέρουν σημαντικά μεταξύ τους, τόσο όσον αφορά στην κατανόηση του περιεχομένου της έννοιας ομαδοποίησης, αλλά και στη διαδικασία εντοπισμού και αποτελεσματικού διαχωρισμού των δεδομένων σε ομάδες κοινών ιδιοτήτων. Υψηλής ομογενοποίησης ομάδες είναι για παράδειγμα αυτές που χαρακτηρίζονται από μικρές ευκλείδειες «αποστάσεις» μεταξύ των μελών τους, αυτές με περιοχές υψηλής πυκνότητας στο χώρο των δεδομένων, διαστήματα κλπ. Η «Ομαδοποίηση», κατά συνέπεια, αποτελεί συνηθισμένο πρόβλημα πολυπαραμετρικής (multi-objective) βελτιστοποίησης. Οι κατάλληλοι αλγόριθμοι ομαδοποίησης και οι παράμετροι ρυθμίσεων (συμπεριλαμβανομένων των τιμών απόστασης, των κατωφλίων πυκνότητας ή αριθμού των αναμενόμενων συμπλεγμάτων), εξαρτώνται από το κάθε σύνολο δεδομένων και της σκοπούμενης χρήσης των αποτελεσμάτων. Η Ανάλυση Clusters συνεπώς, δεν είναι μία «αυτόματη εργασία», αλλά μια επαναληπτική διαδικασία «εξόρυξης γνώσης» ή διαδραστικής πολυπαραμετρικής βελτιστοποίησης, που βασίζεται στη γενική πρακτική «δοκιμής και λάθους». Συχνά θεωρείται αναγκαίο να τροποποιηθούν η προκαταρκτική επεξεργασία ή/και κάποιες ή όλες οι ιδιότητες ή οι παράμετροι, μέχρι να επιτευχθεί το προσδοκώμενο (ή το επιθυμητό) αποτέλεσμα (βλ. σχήμα 2).



Σχήμα 2. Ομαδοποίηση δεδομένων πολυπαραμετρικών χαρακτηριστικών (α , β , γ) με κανόνα τη συνοχή

Ομαδοποιήσεις (συμπλέγματα) που βρέθηκαν με διαφορετικούς αλγόριθμους, μπορεί να διαφέρουν σημαντικά στις ιδιότητές τους, και τα μοντέλα αυτά ομαδοποίησης είναι το κλειδί για την κατανόηση των διαφορών μεταξύ των διαφόρων αλγορίθμων. Τα Κεντροειδή Μοντέλα: Οι αλγόριθμοι αυτοί (k-means algorithms) αναπαριστούν κάθε ομάδα με ένα Διάνυσμα. Τα Μοντέλα Συνδεσιμότητας: Αποτελούν περιπτώσεις Ιεραρχικής Ομαδοποίησης και συνθέτουν μοντέλα, που στηρίζονται σε απόσταση συνδεσιμότητας. Μοντέλα Πυκνότητας: Τα μοντέλα OPTICS και DBSCAN για παράδειγμα καθορίζουν συμπλέγματα σύνδεσης πυκνών περιοχών στο χώρο των δεδομένων. Μοντέλα Κατανομής: Παρέχεται η δυνατότητα να μοντελοποιηθούν με τη χρήση στατιστικών κατανομών. Μοντέλα Υποχώρων (Subspace) μοντελοποιούνται και με τα δύο μέλη του συμπλέγματος και με πρόσθετες ενδεχομένως κοινές ιδιότητες. Δεν υπάρχει κανένας αντικειμενικά «σωστός» αλγόριθμος ομαδοποίησης, αλλά συχνά χρειάζεται και επιλέγεται πειραματικά, εκτός και αν υπάρχει ένας ιδιαίτερος μαθηματικός λόγος προτίμησης ενός μοντέλου. Πρέπει να σημειωθεί ότι ένας αλγόριθμος που έχει σχεδιαστεί για συγκεκριμένο είδος μοντέλου, δεν έχει πιθανότητες επιτυχίας σε ένα γενικό σύνολο δεδομένων που στηρίζεται σε διαφορετικού είδους μοντέλα.

Ανάκτηση Πληροφορίας και σύγχρονα προβλήματα

Η στατιστική και η συντακτική προσέγγιση επεξεργασία της πληροφορίας οδήγησαν στην ανάπτυξη νέων επιστημονικών οριζόντων που σχετίζονται με θέματα διαχείρισης της πληροφορίας όπως η ανάκτηση. Επιπροσθέτως και κλάδοι επιστήμων όπως η ποσοτική και η υπολογιστική γλωσσολογία αλληλεπιδράσαν με την παραπάνω προσέγγιση καθορίζοντας ένα νέο μοτίβο τεχνολογίας που ονομάζεται γλωσσική τεχνολογία. Είναι γνωστό καθώς ο όγκος της πληροφορίας στο διαδίκτυο ολοένα και αυξάνεται τόσο μεγαλώνει η ανάγκη για σωστή διαχείριση της αποθηκευμένης πληροφορίας ώστε να είναι προσβάσιμη με τον καλύτερο δυνατό τρόπο ανάλογα με τις εκάστοτε πληροφοριακές ανάγκες του χρήστη. Η παραγωγή πληροφοριών αυξάνεται όλο και περισσότερο (Moens 2006) αγγίζοντας τα 167 terabytes πληροφορίας στον επιφανειακό ιστό (surface web) και στο βαθύ ιστό (deep web) υπολογίζονται 400 φορές μεγαλύτερα μεγέθη. Με τον όρο βαθύ ιστό γίνεται αναφορά στο τμήμα των πληροφοριών που βρίσκονται στο διαδίκτυο και ενώ είναι προσβάσιμες για τον χρήστη, είναι μη ανιχνεύσιμες για τις μηχανές αναζήτησης.

Λαμβάνοντας υπόψη την αδόμητη φύση των πληροφοριών, η αναζήτηση συναφούς πληροφορίας αποτελεί μεγάλο εγχείρημα για τον χρήστη. Η καταγραφή της πληροφορίας και η κατάλληλη επεξεργασία της αποτελεί βασικό προαπαιτούμενο προκειμένου να μπορεί να ανακτηθεί. Ουσιαστικά η πληροφορία που δεν έχει επεξεργαστεί κατάλληλα ώστε να καθίσταται ανακτήσιμη μπορεί να θεωρηθεί πραγματικά χαμένη, καθώς χωρίς την επεξεργασία αυτή είναι αδύνατη η αναγνώριση της μέσα σε τεράστιες συλλογές ή όγκο πληροφοριών.

Η ανάγκη αυτή παρουσιάστηκε σχεδόν απαρχής γέννησης της γραφής και παραγωγής του γραπτού λόγου. Σύμφωνα με την βιβλιογραφία (Singhal 2001), οι Σουμέριοι ήδη από το 3000 π. Χ. είχαν αντιληφθεί την σημασία οργάνωσης και πρόσβασης των αρχείων τους, χρησιμοποιούσαν ειδικές περιοχές για την αποθήκευση πήλινων πινακίδων με σφηνοειδή γραφή και ανέπτυξαν ειδικούς τρόπους ταξινόμησης ώστε να μπορούν να αναγνωρίσουν τις πινακίδες και αντίστοιχα το περιεχόμενό τους. Αργότερα οι αρχαίοι Έλληνες και οι Ρωμαίοι (Ceri et al. 2013) κατέγραφαν πληροφορίες σε πάπυρους καθώς επίσης και περιλήψεις για το περιεχόμενό τους, ενώ από τον 2^ο αιώνα π.Χ. για πρώτη φορά εμφανίζονται σε ελληνικούς πάπυρους οι πίνακες περιεχομένων. Οι ανάγκες για κατάλληλη οργάνωση και διαχείριση της πληροφορίας αυξάνεται καθώς η ιστορία προχωρά και ειδικότερα καθώς αναπτύσσεται η παραγωγή του γραπτού λόγου. Κομβικά σημεία αποτελούν η ανακάλυψη του χαρτιού και η ανακάλυψη της τυπογραφίας με αποκορύφωμα την εφεύρεση του ηλεκτρονικού υπολογιστή και του διαδικτύου, όπου πλέον η παραγωγή πληροφοριών είναι τεράστια.

Η ανάκτηση της πληροφορίας αποτελεί πολύ βασικό ζήτημα ειδικότερα για τον χώρο των βιβλιοθηκών, που βασική τους λειτουργία είναι η συγκέντρωση πνευματικής παραγωγής, οργάνωση του υλικού και επεξεργασία του ώστε να είναι προσβάσιμο. Έτσι όπως αναφέρεται και στη βιβλιογραφία (Onwuchekwa 2011) η ανάκτηση της πληροφορίας κατά της δεκαετία του '50 θεωρήθηκε πολύ σημαντική πρόοδος για την οργάνωση. Σύμφωνα με τον Van Rijsbergen (1979), το ζήτημα οργάνωσης της πληροφορίας και ανάκτησής της, στα περιβάλλοντα των βιβλιοθηκών, αντιμετωπίστηκε με την ενσωμάτωση βασικών βιβλιοθηκονομικών καθηκόντων στους ηλεκτρονικούς υπολογιστές, μέσω Συστημάτων Ανάκτησης Πληροφορίας (ΣΑΠ).

Όμως, όπως αναφέρεται στη βιβλιογραφία (Blair και Kimbrough, 2002) η ανάκτηση πληροφοριών δεν περιορίζεται μόνο στο χώρο των βιβλιοθηκών, αλλά με την εμφάνιση του διαδικτύου και της διαρκούς αύξησης της πληροφορίας, ο σχεδιασμός ΣΑΠ αντιμετωπίζει πολλές προκλήσεις για την βελτίωση της αποτελεσματικότητας της αναζήτησης, όπως την εύρεση κατάλληλου τρόπου αναπαράστασης των εγγράφων προκειμένου να μπορεί ο χρήστης να έχει πρόσβαση στο σημασιολογικό περιεχόμενό τους. Έτσι ο Van Rijsbergen (1979) αναφέρει πως στην αποθήκευση και ανάκτηση της πληροφορίας δόθηκε ιδιαίτερη έμφαση κατά την δεκαετία του '40, πέρα από τον χώρο των βιβλιοθηκών. Έτσι δημιουργούνται τα συστήματα ανάκτησης ηλεκτρονικών υπολογιστών. Πιο συγκεκριμένα η ιδέα για αποθήκευση πληροφοριών και η αυτόματη πρόσβαση σε αυτές (Singhal 2001) γεννήθηκε το 1945 από ένα άρθρο του Vannevar Bush και υλοποιήθηκε κατά τη διάρκεια της δεκαετίας του '50, ενώ το '60 πραγματοποιείται η ανάπτυξη του συστήματος SMART από τον Gerard Salton και την επιστημονική ομάδα του στο Cornell University το οποίο είχε σκοπό την βελτίωση της ποιότητας αναζήτησης. Ταυτόχρονα όπως αναφέρεται στη βιβλιογραφία (Singhal 2001) πραγματοποιούνται τα τεστ Cranfield, από τον Cyril Cleverdon και την επιστημονική του ομάδα στο College of Aeronautics από τα οποία προέκυψε η μεθοδολογία αξιολόγησης ΣΑΠ.

Κατά τις δεκαετίες '70 – '80, δίνεται έμφαση στην έρευνα των ΣΑΠ και πραγματοποιούνται απόπειρες βελτίωσης των ήδη υπαρχόντων συστημάτων. Όπως αναφέρεται στη βιβλιογραφία (Singhal 2001),

η έρευνα και οι βελτιώσεις που πραγματοποιούνται αφορούν στην ανάκτηση εγγράφων και περιορίζονται σε συλλογές μικρής κλίμακας καθώς δεν υπήρχαν ιδιαίτερα μεγάλες συλλογές κειμένων, μέχρις ότου ξεκίνησαν το 1992 τα Text Retrieval Conferences (TREC), υπό την χορηγία του National Institute of Standards and Technology (NIST) από την κυβέρνηση των ΗΠΑ για έρευνα σχετική με την ανάκτηση της πληροφορίας σε μεγάλες συλλογές. Τελειώνοντας την ιστορική αναδρομή της ανάκτησης της πληροφορίας, έχει ήδη αναφερθεί πως η εφεύρεση των υπολογιστών και ο παγκόσμιος ιστός έπαιξε καταλυτικό ρόλο για την αποθήκευση, πρόσβαση και αναζήτηση πληροφορίας σε συλλογές εγγράφων (Ceri et al. 2013).

Έτσι λοιπόν με την βασική προϋπόθεση της αποθήκευσης και κατάλληλης επεξεργασίας της κάθε πληροφορίας ώστε να είναι ανακτήσιμη για να μπορέσει να καλύψει της πληροφοριακές ανάγκες του χρήστη κατά τη διάρκεια μιας αναζήτησης του για συναφή πληροφορία, δημιουργήθηκε ο επιστημονικός κλάδος της Ανάκτησης Πληροφορίας (ΑΠ). Πιο συγκεκριμένα σύμφωνα με τους Baeza-Yates και Ribeiro-Neto (2011) *«η ανάκτηση πληροφοριών ασχολείται με την αναπαράσταση, αποθήκευση, οργάνωση και πρόσβαση πληροφοριακών τεκμηρίων (items), όπως έγγραφα, ιστοσελίδες, online κατάλογοι, δομημένα και ημι-δομημένα αρχεία, αντικείμενα πολυμέσων. Η αναπαράσταση και οργάνωση των πληροφοριακών τεκμηρίων θα πρέπει να είναι τέτοια ώστε να παρέχει στους χρήστες εύκολη πρόσβαση στην πληροφορία που τους ενδιαφέρει»*. Οι Manning, Raghavan και Schutze (2008) ορίζουν την ΑΠ ως εξής: *«η αναζήτηση και εύρεση υλικού (συνήθως εγγράφων) αδόμητης φύσης (συνήθως κείμενο) που ικανοποιεί μια πληροφοριακή ανάγκη μέσα από μεγάλες συλλογές (συνήθως αποθηκευμένες σε υπολογιστές)»*. Από τον παραπάνω ορισμό, ο όρος «έγγραφο» αφορά το ίδιο το υλικό ενώ ο όρος «κείμενο» χαρακτηρίζει τη φύση του υλικού. Συνεπώς οι όροι αυτοί θα χρησιμοποιούνται με τις παραπάνω σημασίες στην διατριβή.

Έναν ακόμη ορισμό ΑΠ βρίσκουμε στους Grossman και Frieder (2000): *«ως η αρχή εύρεσης συναφών εγγράφων σε αντίθεση με την απλή αντιστοίχιση λεξιλογικών μοτίβων σε ένα ερώτημα»*, όπου εισάγεται και η έννοια της συνάφειας (relevance). Συναφές (Croft et al. 2009) θεωρείται ένα έγγραφο το οποίο περιέχει την πληροφορία την οποία αναζητούσε ο χρήστης κατά την υποβολή του ερωτήματος του στο ΣΑΠ όπως π.χ. συχνά είναι μια μηχανή αναζήτησης. Τα ΣΑΠ έχουν ως στόχο την ανάκτηση εγγράφων από μια συλλογή, τα οποία έγγραφα θα ικανοποιούν τις ανάγκες του χρήστη. Η συνάφεια των αποτελεσμάτων σε σχέση με το ερώτημα του χρήστη θα καθορίσει το πόσο επιτυχημένο και αποτελεσματικό είναι το ΣΑΠ. Χαρακτηριστικά της συνάφειας (Ceri 2013) είναι η υποκειμενικότητά της, καθώς έχει να κάνει με την κρίση του χρήστη ως προς το αν ικανοποιείται η πληροφοριακή του ανάγκη από τα αποτελέσματα που επιστρέφει το ΣΑΠ και η δυναμική της φύση στο χωροχρόνο και το ότι είναι πολύπλευρη καθώς καθορίζεται τόσο από το περιεχόμενο των ανακτηθέντων αποτελεσμάτων όσο και από ζητήματα που έχουν να κάνουν με την πηγή της πληροφορίας όπως η αυθεντικότητα ή η ειδικότητα. Σύμφωνα με τον Ingwersen (1992) το πρόβλημα της συνάφειας επικεντρώνεται στην εύρεση της κατάλληλης πληροφορίας σε σχέση με το ερώτημα του χρήστη. Έτσι η πληροφορία που βρίσκεται σε ένα κείμενο συγκρίνοντας το με άλλα κείμενα μπορεί να είναι πιο συναφής ως προς συγκεκριμένη απαίτηση πληροφορίας του χρήστη. Ακόμη η σημασία που έχει ένα κείμενο εξαρτάται από τις πληροφοριακές απαιτήσεις και μπορεί να αλλάζει καθώς αλλάζουν και αυτές. Οι Croft et al. (2009) διακρίνουν το φαινόμενο της συνάφειας σε τοπική συνάφεια και συνάφεια χρήστη. Ως τοπική συνάφεια περιγράφουν το φαινόμενο κατά το οποίο ένα έγγραφο και το ερώτημα του χρήστη έχουν το ίδιο θέμα ενώ η συνάφεια χρήστη έγκειται στις προσωπικές απαιτήσεις του χρήστη, οι οποίες καθιστούν το έγγραφο συναφές ή όχι.

Η ΑΠ αποτελείται από δύο μεγάλες διαδικασίες: την ευρετηρίαση και την αναζήτηση ερωτήματος χρήστη, όπου και οι δύο διαδικασίες αντιμετωπίζουν τα ίδια επιστημονικά προβλήματα, τα οποία προκύπτουν από το πολύπλευρο ζήτημα της συνάφειας. Οι Baeza-Yates και Ribeiro-Neto (2011) αναλύουν περισσότερο τα προβληματικά σημεία της ΑΠ εστιάζοντας στο ζήτημα «μετάφρασης» των πληροφοριακών αναγκών του χρήστη στο πλαίσιο μιας αναζήτησης. Υποδεικνύουν τις επιμέρους επεξεργασίες νοητικά για τον χρήστη και αυτοματοποιημένα για τον υπολογιστή όπου αντίστοιχα ο χρήστης πρέπει να περιγράψει την πληροφοριακή του ανάγκη και να την συμπίεσει σε ορισμένες λέξεις-κλειδιά που θα αποτελέσουν το ερώτημα που θα εισάγει μέσω της διεπαφής ενώ από την άλλη το ΣΑΠ θα πρέπει να ανατρέξει στο ευρετήριο των εγγράφων μιας συλλογής, να «κατανοήσει» το περιεχόμενο τους, να κρίνει τα συναφή έγγραφα και να τα επιστρέψει ως αποτελέσματα σε μια λίστα καταταγμένη ανάλογα με το βαθμό συνάφειας σε σχέση με το ερώτημα. Για την παραπάνω διαδικασία είναι απαραίτητη η γλωσσολογική γνώση, η οποία αφορά την κατανόηση του κειμένου σε συντακτικό και σημασιολογικό επίπεδο.

Σε σχέση με το χρήστη (Woods et al. 2000), η αποτυχία εύρεσης κατάλληλης πληροφορίας εξαρτάται από τις λέξεις-κλειδιά που εισάγει ως ερώτημα και από το αν και κατά πόσον αυτές χρησιμοποιούνται από το αντίστοιχο συναφές υλικό, ως προς την πληροφοριακή του ανάγκη. Ακόμη (Woods et al. 2000) προβληματικό

σημείο είναι και ο απαιτούμενος χρόνος που χρειάζεται να δαπανήσει ένας χρήστης διαβάζοντας τα ανακτηθέντα αποτελέσματα της αναζήτησης ώστε να ανακαλύψει εάν η πληροφορία που αναζητά περιέχεται σε κάποιο από αυτά. Σύμφωνα με τον Lewis και την Sparck Jones (1996) κατά τη διαδικασία της ανάκτησης εγγράφων, ο χρήστης έχει άγνοια ως προς την ακριβή πληροφορία που αναζητά και αυτό επιφορτίζει το ΣΑΠ με το σημαντικό καθήκον της σύνδεσης της πληροφοριακής ανάγκης του χρήστη με αντίστοιχα συναφή έγγραφα. Για να πραγματοποιηθεί όμως αυτό θα πρέπει το υπολογιστικό σύστημα να αναζητήσει τις σχέσεις των δύο παραπάνω πλευρών. Η δυσκολία της σύνδεσης αυτών των δύο έχει να κάνει στο ότι δεν υπάρχει κάποια συγκεκριμένη συνθήκη με βάση την οποία να ικανοποιείται η πληροφοριακή ανάγκη του χρήστη, καθώς το ερώτημα του αποτελεί μια συντομευμένη περιγραφή, θεωρητικά αρκετά ασαφή.

Από την πλευρά των ΣΑΠ (Woods et al. 2000) υπάρχει ανάγκη βελτίωσης της αποτελεσματικότητας της αναζήτησης, μέσω γλωσσικών συσχετίσεων στα κείμενα όπως μορφολογικές σχέσεις μεταξύ των λέξεων και ταξινομικές σχέσεις μεταξύ εννοιών. Ο Van Rijsbergen (1979) αναφέρει πως το πρόβλημα εστιάζεται στην δυσκολία ερμηνείας της φυσικής γλώσσας με αυτόματο τρόπο από τα υπολογιστικά συστήματα και της κατανόησης του περιεχομένου των εγγράφων και πως η δυσκολία αυτή έχει δύο άξονες: την διαδικασία ανάγνωσης (με σκοπό την εύρεση συναφούς πληροφορίας), η οποία περιλαμβάνει συντακτική και σημασιολογική ανάλυση και την απόφαση της συνάφειας ή μη σε σχέση με μεμονωμένο έγγραφο.

Η ευρετηρίαση αφορά τον καθορισμό του περιεχομένου των εγγράφων, σύμφωνα με τον Lewis και την Sparck Jones (1996) και είναι υπεύθυνη για την αύξηση της ακρίβειας και ανάκλησης σε ένα ΣΑΠ. Τα προβλήματα που εμφανίζονται εσωτερικά κατά την ευρετηρίαση αφορούν διάφορους περιορισμούς όπως: γλωσσικοί περιορισμοί με παράδειγμα το φαινόμενο της πολυσημίας, περιορισμοί ως προς την έκφραση του ερωτήματος (ασαφή ερωτήματα λόγω της άγνοιας χρήστη και ατελή ερωτήματα, τα οποία δεν περιλαμβάνουν αρκετές λεπτομέρειες) και περιορισμοί στην αναπαράσταση εγγράφων, με αποτέλεσμα την απώλεια πληροφορίας από τα πρωτότυπα έγγραφα. Όπως αναφέρει ο Van Rijsbergen (1979) μέσω των κατάλληλων αναπαραστάσεων εγγράφων και ερωτημάτων θα κριθεί η συνάφεια ή μη ενός εγγράφου και ειδικότερα για την ευρετηρίαση τονίζει ότι η πληροφορία που αποθηκεύεται στα ευρετήρια σε σχέση με τα έγγραφα αποτελεί συμκρυμμένες αναπαραστάσεις εγγράφων και όχι αυτούσια όλες τις πληροφορίες που περιέχουν αυτά με αποτέλεσμα να υπάρχει απώλεια στην πληροφορία. Ειδικότερα ως προς τα μοντέλα ΑΠ και τον καθορισμό συνάφειας ο Sproerri (1995) διακρίνει διάφορες προσεγγίσεις ανάκτησης, εκ των οποίων και η στατιστική προσέγγιση η οποία χρησιμοποιεί στατιστικές πληροφορίες. Στην προσέγγιση αυτή ανήκει και το Vector Space Model (VSM), το οποίο χρησιμοποιεί τις πληροφορίες στατιστικής εμφάνισης όρων στα έγγραφα, για την παραγωγή των συναφών αποτελεσμάτων σε μορφή λίστας κατάταξης. Στην στατιστική προσέγγιση ένα από τα μειονεκτήματα των μοντέλων αυτών που απαντώνται σύμφωνα με Sproerri (1995) αφορά την έλλειψη σχετικής δομής ώστε να μπορούν να εκφραστούν γλωσσολογικά χαρακτηριστικά όπως φράσεις, καθώς και περιορισμοί εγγύτητας.

Οι Galvez, de Moya-Anegon και Solana (2005) συμπληρώνουν ειδικότερα για το ζήτημα της ανεπαρκούς αναπαράστασης εγγράφων και ερωτημάτων πώς επηρεάζει την απόδοση των ΣΑΠ και τονίζουν τη σημαντικότερη συμβολή των τεχνικών Επεξεργασίας Φυσικής Γλώσσας (Natural Language Processing/NLP), προκειμένου να αντιμετωπιστεί το ζήτημα της γλωσσικής ανάλυσης των όρων που περιέχουν τα έγγραφα, το οποίο αφορά σε ποικιλία λεξιλογικών, συντακτικών και μορφολογικών χαρακτηριστικών για κάθε μεμονωμένο όρο. Μέσω της ενσωμάτωσης τεχνικών Επεξεργασίας Φυσικής Γλώσσας στα ΣΑΠ γίνεται μια απόπειρα εφαρμογής των πρακτικών της Υπολογιστικής Γλωσσολογίας (ΥΓ). Τα βασικά προβλήματα έχουν να κάνουν με την πολυσημία ενός όρου (ένας όρος/μία λέξη με πολλές σημασίες) και με τους όρους που αποτελούνται από πολλές φράσεις έχοντας έτσι δομή φράσεων, η οποία δομή επιδέχεται πολλές διαφορετικές παραλλαγές. Μάλιστα όπως συμπληρώνουν οι Galvez, de Moya-Anegon και Solana (2005), οι γλωσσικές παραλλαγές που παρατηρούνται στους όρους χωρίζονται σε τρεις κατηγορίες:

- Μορφολογικές παραλλαγές, αφορά τις διαφορετικές μορφές εμφάνισης ενός όρου (όπου εφαρμόζεται η διαδικασία περιστολής ώστε να επιλεγεί μια ενιαία μορφολογική παραλλαγή).
- Λεξιλογικές - σημασιολογικές παραλλαγές, αφορούν τη σημασιολογική εγγύτητα των λέξεων (μία έννοια - πολλοί όροι ή πολλές έννοιες – ένας όρος).
- Συντακτικές παραλλαγές, αφορούν τις διαφορετικές δομές όρων πολλών λέξεων (μείωση/ διαίρεση κανονικών/κανονικοποιημένων συντακτικών δομών (canonical syntactic structure)).

Οι παραπάνω παραλλαγές θεωρούνται σημαντικές στον κλάδο της Βιβλιοθηκονομίας και της Αρχειονομίας επικεντρωμένες κυρίως στον τομέα της συλλογής. Η συλλογή στον τομέα της

Βιβλιοθηκονομίας θεωρεί ως θεμέλιο λίθο το τεκμήριο αναφοράς. Τεκμήριο αναφοράς θεωρείται το ιδεατό τεκμήριο που περιλαμβάνει το σύνολο των κοινών ιδιοτήτων (χαρακτηριστικών) των τεκμηρίων που περιλαμβάνει η συλλογή. Αυτές άλλοτε επικεντρώνονται στο δωρητή (ή πρόσωπο) άλλοτε γίνεται αναφορά σε μία εποχή ή σε ένα συγκεκριμένο κλάδο επιστήμης ή ειδικά σε ένα θέμα. Ενώ η συλλογή στον κλάδο της Αρχειονομίας θεωρείται η οργάνωση των τεκμηρίων ενός αρχειακού συνόλου ανά κατηγορίες οι οποίες διαρθρώνονται ιεραρχικά. Σε αμφότερες τις περιπτώσεις η οργάνωση αυτή επιβάλλει μια πρωτόλεια προσέγγιση της γλωσσικής επεξεργασίας που προϋποθέτει να τεθούν πρωτίστως οι τεχνολογικές βάσεις οι οποίες δανείζονται από τους επιμέρους κλάδους της τεχνικής νοημοσύνης, οι οποίοι αναπτύσσονται στα περεταίρω κεφάλαια.

Βιβλιογραφία

- Aghagolzadeh, M., Soltanian-Zadeh, H., Araabi, B., & Aghagolzadeh, A. (2007). A hierarchical clustering based on mutual information maximization. In *Image Processing, 2007. ICIIP 2007. IEEE International Conference on* (Vol. 1, pp. 1–277). IEEE. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4378945
- Bifet, A., Holmes, G., Kirkby, R., & Pfahringer, B. (2010). Moa: Massive online analysis. *The Journal of Machine Learning Research*, 11, 1601–1604.
- Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC press.
- Carter, T. (2007). An introduction to information theory and entropy. *Complex Systems Summer School, Santa Fe*. Retrieved from <http://csustan.csustan.edu/~tom/Lecture-Notes/Information-Theory/info-lec.pdf>
- Cha, S.-H., & Tappert, C. (2009). A genetic algorithm for constructing compact binary decision trees. *Journal of Pattern Recognition Research*, 4(1), 1–13.
- Clausius, R. (1879). *The mechanical theory of heat*. MacMillan.
- Cover, T. M., & Thomas, J. A. (2012). *Elements of information theory*. John Wiley & Sons. Retrieved from <https://www.google.com/books?hl=en&lr=&id=VWq5GG6ycxMC&oi=fnd&pg=PT10&dq=+Elements+of+information+theory&ots=bX7iL3X5RN&sig=Wf-b3G6yznIcm98eW5PzkEpngmg>
- Deng, H., Runger, G., & Tuv, E. (2011). Bias of importance measures for multi-valued attributes and solutions. *Artificial Neural Networks and Machine Learning–ICANN 2011*, 293–300.
- Ding, Y. (2008). *Treating missing data in classification trees*. ProQuest. Retrieved from <https://www.google.com/books?hl=en&lr=&id=XP3IgeZF2X4C&oi=fnd&pg=PA1&ots=dzbJQv2Rq6&sig=NzIjcvv7BWvGSp2CVSC13eFNbxM>
- Eden, B. L. (2005). *Digital library usability studies* (Vol. 21). Emerald Group Publishing. Retrieved from <https://www.google.com/books?hl=en&lr=&id=jU5RHEgfnlcC&oi=fnd&pg=PT1&dq=Digital+Library+Usability+Studies&ots=L4cS42wdHm&sig=rMg15hGBIszvTUdPUCYLfWCaNtA>
- Feinman, R. D., & Fine, E. J. (2004). A calorie is a calorie" violates the second law of thermodynamics. *Nutr J*, 3(9), 10–1186.
- Gorunescu, F. (2011). *Data Mining: Concepts, models and techniques* (Vol. 12). Springer Science & Business Media. Retrieved from <https://www.google.com/books?hl=en&lr=&id=yJvKY-sB6zkC&oi=fnd&pg=PP2&dq=%22Data+Mining:+Concepts,+Models+and+Techniques%E2%80%9D+&ots=psMwtavTtG&sig=zpejNv87ez08vdTeZdWPHOT5RMU>
- Holmes, G., Donkin, A., & Witten, I. H. (1994). Weka: A machine learning workbench. In *Intelligent Information Systems, 1994. Proceedings of the 1994 Second Australian and New Zealand Conference on* (pp. 357–361). IEEE. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=396988
- Hyafil, L., & Rivest, R. L. (1976). Constructing optimal binary decision trees is NP-complete. *Information Processing Letters*, 5(1), 15–17.
- Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical Review*, 106(4), 620.
- Kargupta, H., Han, J., Philip, S. Y., Motwani, R., & Kumar, V. (2008). *Next generation of data mining*. CRC Press. Retrieved from <https://www.google.com/books>
- Karmeshu. (2003). *Entropy measures, maximum entropy principle and emerging applications* (Vol. 119). Springer Science & Business Media. Retrieved from <https://www.google.com/books>
- Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer. Retrieved from <http://link.springer.com/book/10.1007/978-1-4614-6849-3?no-access=true>

- Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 79–86.
- Lee, J.-Y. (2008). *A new approach of top-down induction of decision trees for knowledge discovery*. ProQuest. Retrieved from <https://www.google.com/books>
- Manning, C. D., Raghavan, P., Schütze, H., & others. (2008). *Introduction to information retrieval* (Vol. 1). Cambridge university press Cambridge. Retrieved from http://www.langtoninfo.co.uk/web_content
- Ville, B. D. (2006). Decision Tree for Business Intelligence and Data Mining. Retrieved from <http://dl.acm.org/citation.cfm?id=1204951>

1. Ζητήματα Ανάκτησης της Πληροφορίας

Περίληψη

Στο κεφάλαιο αυτό αναπτύσσονται ζητήματα ανάκτησης της πληροφορίας τα οποία αποτελούν θεμέλιο λίθο της σημασιολογίας από την άποψη της εξέτασης θεμάτων που αποσκοπούν την σημασιολογική πλευρά της ανάκτησης. Προς το σκοπό αυτό αναπτύσσονται οι μεθοδολογίες που αφορούν την πληροφορία κειμένου και συγκεκριμένα τη μετατροπή τόσο του ερωτήματος χρήστη όσο και των εγγράφων σε λογικές αναπαραστάσεις.

Προαπαιτούμενες γνώσεις

Αναπαράσταση εγγράφων, μοντέλο VSM, TF-IDF, Αξιολόγηση Αποτελεσμάτων.

1.1 Εισαγωγή στην διαδικασία της Ανάκτησης της Πληροφορίας

Στο κεφάλαιο αυτό αναπτύσσεται το θεωρητικό υπόβαθρο στο οποίο εντάσσονται οι διαδικασίες εξόρυξης γνώσης μέσω των πληροφοριών που διαμοιράζει το διαδίκτυο. Προκειμένου αυτή η διαδικασία να περιγραφεί με ένα φορμαλιστικό τρόπο επιλέχθηκε ως μοντέλο σχεδιασμού το σύστημα ταξινόμησης της ACM (ανακτήθηκε 20/02/2015 από <http://www.acm.org/about/class/2012>). Σύμφωνα με την ταξινόμηση αυτή, δίνεται έμφαση σε έναν επιστημονικό υπο-κλάδο των πληροφοριακών συστημάτων αυτόν της Ανάκτησης της Πληροφορίας (ΑΠ), ο οποίος βέβαια είναι άμεσα συνυφασμένος με αυτόν του Παγκόσμιου Ιστού όπως φαίνεται και στην Σχήμα. 1.1



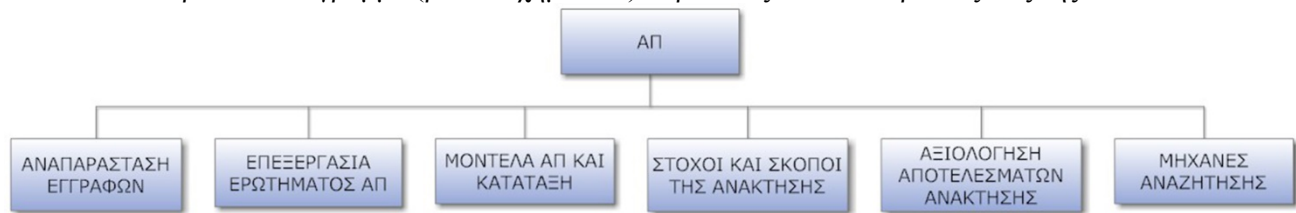
Σχήμα 1.1. Υπο-κλάδοι πληροφοριακών συστημάτων σύμφωνα με ACMClassification

Ο παραπάνω κλάδος της ΑΠ επιλέχθηκε για τον εξής λόγο: διότι ασχολείται με όλες εκείνες τις διαδικασίες που αφορούν την αναπαράσταση εγγράφων ως επίσης και τις διαδικασίες ανάλυσης και αξιολόγησης. Ο δε κλάδος του Παγκόσμιου Ιστού αποτελεί ένα δυναμικό μέσο παραγωγής εγγράφων και έτσι στη σημερινή πραγματικότητα αυτοί οι δύο κλάδοι αποτελούν δύο αναπόσπαστα τμήματα λειτουργίας. Είναι γνωστό, ότι στον κλάδο του Παγκόσμιου Ιστού μελετάται κυρίως το πρίσμα της διαδικασίας αναζήτησης και ανάκτησης πληροφορίας, χωρίς να περιλαμβάνει άλλες παραμέτρους του κλάδου αυτού. Στο παρόν σύγγραμμα χρησιμοποιείται ως μέσο οι πρακτικές ανάλυσης αυτών των εγγράφων με σκοπό να ευρεθεί ένα πρότυπο (pattern), το οποίο θα αποτελέσει δυνητικά ένα κυρίαρχο δομικό λίθο στον κλάδο της ΑΠ.

Σε αυτό το σημείο σκόπιμος είναι ο ορισμός της λέξης «έγγραφο», ο οποίος απαντάται πολύ συχνά στη διεθνή βιβλιογραφία της ΑΠ. Όπως αναφέρεται στην βιβλιογραφία (Baeza-Yates και Ribeiro-Neto 2011, Manning, Raghavan και Schutze 2008), οι τύποι πληροφοριακών αντικειμένων στην ΑΠ ποικίλουν, αν και η ΑΠ εστιάζει κυρίως στην ανάκτηση εγγράφων (documents). Ο όρος έγγραφο στην ΑΠ χρησιμοποιείται υπό μια ευρεία έννοια και υποδηλώνει την βασική μονάδα που απαρτίζει ένα Σύστημα Ανάκτησης Πληροφορίας (ΣΑΠ), είτε αυτή είναι κείμενο είτε οποιαδήποτε άλλη μορφή καθώς η πληροφορία που αναζητά ο χρήστης μπορεί να βρίσκεται σε οποιαδήποτε μορφή, όπως σε ένα αρχείο ήχου ή βίντεο. Τα έγγραφα απαρτίζονται από δεδομένα τα οποία και αυτά διακρίνονται σε 2 κατηγορίες: τα δομημένα και αδόμητα δεδομένα. Με βάση τους Manning, Raghavan και Schutze (2008), αδόμητα ονομάζονται τα δεδομένα εκείνα τα οποία για έναν υπολογιστή δεν θεωρούνται σαφή και σημασιολογικά εμφανή ως προς τη δομή τους. Δομημένα δεδομένα είναι το αντίθετο των αδόμητων δεδομένων. Παρόλα αυτά πάντοτε υπάρχει μια τυπική δομή στα εκάστοτε κείμενα, η οποία προέρχεται από τη χρήση της γλώσσας, καθώς απαιτεί συγκεκριμένη σύνταξη. Επιπλέον τα

περισσότερα κείμενα διαθέτουν δομή, η οποία πηγάζει από το πως τα έχει διαμορφώσει ο δημιουργός τους τοποθετώντας επικεφαλίδες, παραγράφους, υποσημειώσεις. Π.χ. στο διαδίκτυο για τα έγγραφα η δομή αποδίδεται μέσω των γλωσσών σήμανσης (markup).

Στο παρακάτω διάγραμμα (βλέπε Σχήμα. 1.2) παρουσιάζονται οι κύριοι άξονες της **ΑΠ**:

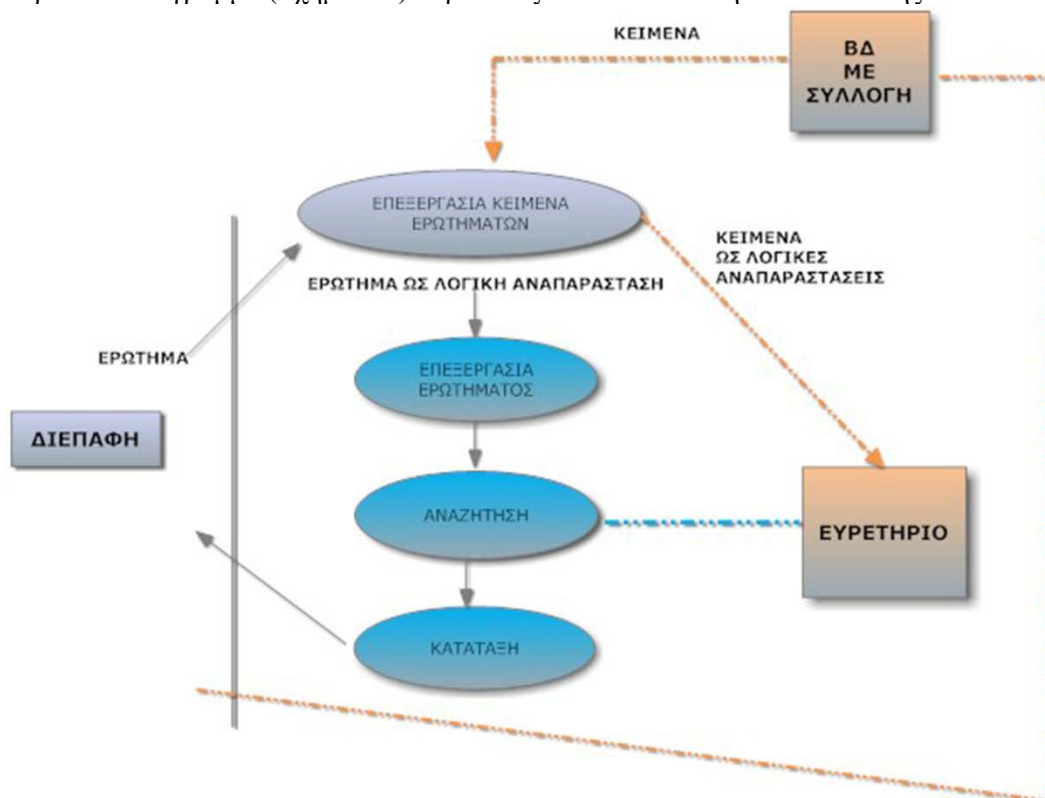


Σχήμα 1.2. Επισκόπηση ΑΠ σύμφωνα με ACMClassification

Στο παρόν κεφάλαιο θα αναπτυχθούν κυρίως οι ενότητες «μοντέλα ΑΠ και κατάταξη», δίνοντας έμφαση στο μοντέλο VSM καθώς και «μηχανές αναζήτησης», εστιάζοντας στην αρχιτεκτονική τους, η οποία αφορά στα ζητήματα ανάκτησης πληροφορίας. Προκειμένου να υπάρχει μια ολοκληρωμένη παρουσίαση είναι απαραίτητη η παρουσίαση της διαδικασίας ΑΠ συνοπτικά. Οι στόχοι της ΑΠ παρουσιάστηκαν στο προηγούμενο κεφάλαιο.

Συνοπτικά η διαδικασία ΑΠ έγκειται στο ερώτημα που εισάγει ένας χρήστης σε μια διεπαφή υπολογιστή προκειμένου να αναζητήσει - ανακτήσει την πληροφορία που χρειάζεται. Στην συνέχεια το ερώτημα υφίσταται ένα σύνολο επεξεργασιών (κατάλληλες μετατροπές ώστε να είναι επεξεργάσιμο από τον υπολογιστή) ώστε τέλος να πραγματοποιηθεί αναζήτηση σε μια πηγή πληροφοριών, όπως μια βάση δεδομένων και τέλος να επιστραφούν τα σχετικά αποτελέσματα στον χρήστη.

Μάλιστα σύμφωνα με τους Cerietal. (2013) καθώς η ΑΠ ασχολείται κυρίως με δεδομένα κειμένου (έγγραφα και ερωτήματα σε μορφή κειμένου), ένα σύνολο κειμενικών διεργασιών είναι απαραίτητο. Στο παρακάτω διάγραμμα (Σχήμα 1.3) παρουσιάζεται αναλυτικά η διαδικασία της ΑΠ.



Σχήμα 1.3. Διαδικασία ΑΠ

Σύμφωνα με τους Ceri et al. (2013) η διαδικασία ΑΠ συνοψίζεται στο σχήμα 1.3 και αναλύεται ως ακολούθως: Ο χρήστης εισάγει στη διεπαφή (μπορεί να είναι κάποιος φυλλομετρητής) κάποιες λέξεις κλειδιά που αποτελούν το ερώτημα του, που εκφράζουν την πληροφοριακή ανάγκη. Εν συνεχεία το ερώτημα αυτό υφίσταται ένα σύνολο διεργασιών για την επεξεργασία του ως κείμενο καθώς βρίσκεται σε φυσική γλώσσα, διεργασίες όμοιες με αυτές που πραγματοποιούνται στα αποθηκευμένα έγγραφα του συστήματος κατά την ευρετηρίαση. Ειδικότερα για την ΑΠ που αφορά πληροφορία κειμένου είναι αναγκαία η μετατροπή τόσο του ερωτήματος χρήστη όσο και των εγγράφων σε λογικές αναπαραστάσεις. Στην συνέχεια το ερώτημα υποβάλλεται σε επιμέρους διεργασίες, πραγματοποιείται αναζήτηση στη βάση (η ταχύτητα της αναζήτησης εξαρτάται από το ευρετήριο) και επιστρέφονται συναφή αποτελέσματα στο χρήστη σε μια λίστα κατάταξης σύμφωνα με τη συνάφεια ως προς την πληροφοριακή ανάγκη του χρήστη. Καθώς ο χρήστης περιηγείται στη λίστα μπορεί να εντοπίσει κάποιο υποσύνολο που θεωρεί πολύ συναφές, γνωστό ως ανάδραση (feedback) που καταχωρείται στο σύστημα.

1.2 Μοντέλα ΑΠ και κατάταξη

Η αναπαράσταση εγγράφων και η διαδικασία επεξεργασίας ερωτήματος χρήστη έχουν επεξηγηθεί συνοπτικά στην παραπάνω ενότητα. Έτσι σε αυτήν την ενότητα θα παρουσιαστούν τα μοντέλα ΑΠ (IR models) δίνοντας έμφαση στο VectorSpaceModel (VSM).

Τα μοντέλα ΑΠ αφορούν τη διευκρίνηση της συνάφειας εγγράφων σε σχέση με ένα ερώτημα χρήστη. Μάλιστα διαρκώς αξιολογούνται, ώστε να αναβαθμίζονται ικανοποιητικά προκειμένου να λειτουργούν όσο το δυνατόν αποτελεσματικότερα. Σύμφωνα με τους Baeza-Yates και Ribeiro-Neto (2011), η διάκριση των συναφών εγγράφων από τα υπόλοιπα της συλλογής, εξαρτάται κυρίως από τον αλγόριθμο κατάταξης (θεωρείται πυρήνας των ΣΑΠ) και από τις βασικές προϋποθέσεις που χρησιμοποιεί αυτός. Έτσι οι διαφορετικές παραλλαγές συνόλων προϋποθέσεων αποτελούν διαφορετικά μοντέλα ΑΠ, τα οποία χωρίζονται σε κάποιες επιμέρους κατηγορίες.

Οι Baeza-Yates και Ribeiro-Neto (2011) αναφέρουν τα **κλασσικά μοντέλα ΑΠ**:

1. **Λογικό (Boolean)**: βασίζεται στη θεωρία Συνόλων και τα έγγραφα και τα ερωτήματα αναπαρίστανται ως σύνολα όρων ευρετηρίου.
2. **Διανυσματικό (Vector)**: βασίζεται στη θεωρία Διανυσμάτων (ευρύτερα θεωρείται αλγεβρικό μοντέλο), στο οποίο τα αντίστοιχα έγγραφα και ερωτήματα αναπαρίστανται ως διανύσματα σε πολυδιάστατο διανυσματικό χώρο.
3. **Πιθανοτικό (Probabilistic)**: βασίζεται στη θεωρία Πιθανοτήτων.

Στα κλασσικά μοντέλα ΑΠ παραπάνω αναφέρθηκαν οι μαθηματικές βάσεις. Όσον αφορά τις εξαρτήσεις όρων οι Bendersky και Croft (2012) παρατηρούν μια αλλαγή ενδιαφέροντος προς μοντέλα ΑΠ που ενσωματώνουν εξαρτήσεις όρων την τελευταία δεκαετία, με παραδείγματα τα μοντέλα *Markovrandomfields* και *lineardiscriminantmodel*. Σύμφωνα με τους Metzler και Croft (2005) οι εξαρτήσεις όρων σε μια συλλογή εγγράφων υπάρχουν και εστιάζουν κυρίως σε ζεύγη όρων. Μάλιστα αναφέρουν ότι η εξάρτηση όρων μελετάται είτε στο πλαίσιο των φράσεων, όσον αφορά την εγγύτητα μεταξύ των όρων ή εστιάζει στις εμφανίσεις των όρων που εξετάζονται από κοινού σε έγγραφα. Όπως φαίνεται και από το παραπάνω διάγραμμα (Βλ. σχήμα 1.4) οι παραλλαγές μοντέλων ΑΠ είναι πάρα πολλές ώστε να επεξηγηθούν αναλυτικά. Παρακάτω θα αναπτυχθεί το διανυσματικό μοντέλο VSM το οποίο παρουσιάζει ιδιαίτερο ενδιαφέρον σε θέματα ανάκτησης πληροφορίας.

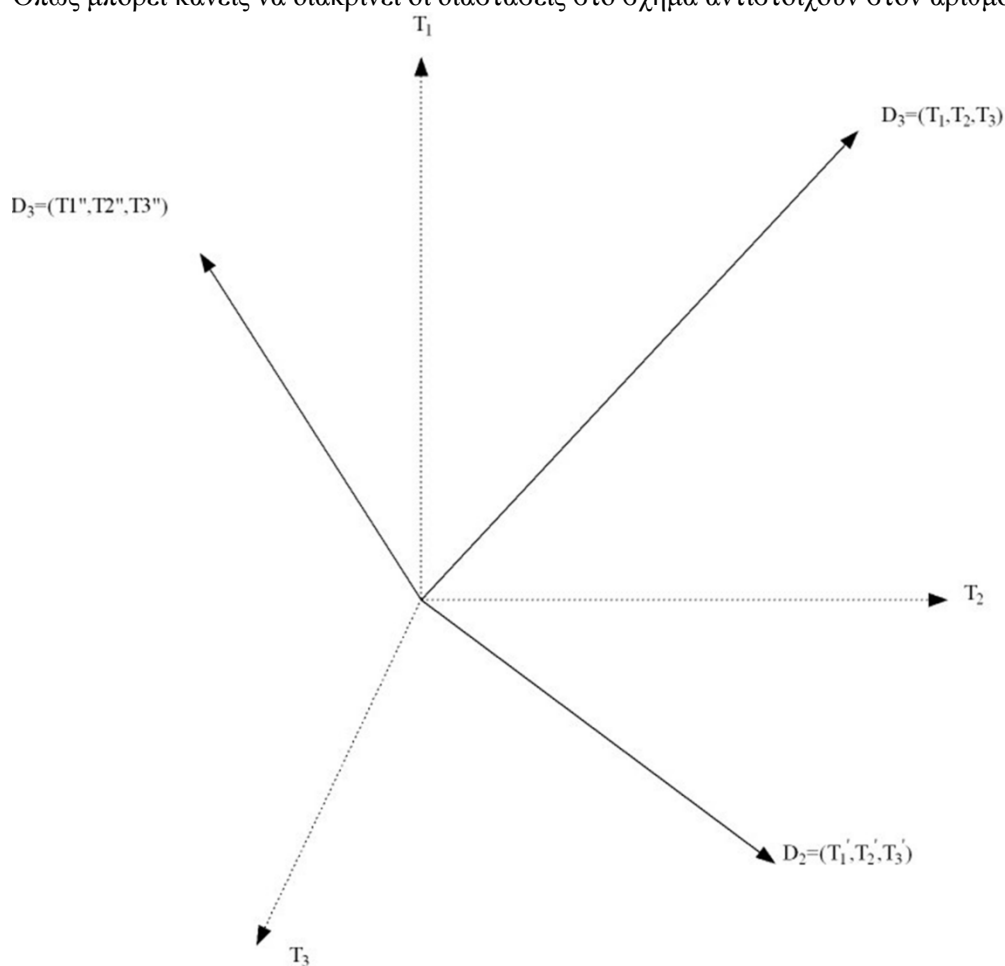
1.2.1 VSM

Το VSM απαντάται στη βιβλιογραφία (Dubin 2004) ως ένα μαθηματικό μοντέλο ιδιαίτερος πολυχρηστικό καθώς μπορεί να εφαρμοστεί σε πολλούς επιστημονικούς τομείς αλλά και ειδικότερα στην ΑΠ αναπαριστώντας έγγραφα ως διανύσματα. Το VSM είναι αποτέλεσμα της έρευνας κατά την ανάπτυξη του ΣΑΠ *System for the Mechanical Analysis and Retrieval of Text (SMART)* στο Πανεπιστήμιο Cornell, με επικεφαλής της ερευνητικής ομάδας τον Gerard Salton

Ειδικότερα, το VSM ανήκει στα στατιστικά μοντέλα (Spoerri 1995), τα οποία για να καθορίσουν τη συνάφεια μεταξύ των εγγράφων μιας συλλογής και του ερωτήματος του χρήστη χρησιμοποιούν τη συχνότητα εμφάνισης των όρων στα έγγραφα και επιστρέφουν ως αποτελέσματα λίστες κατάταξης εγγράφων ανάλογα με τη συνάφεια τους ως προς το ερώτημα του χρήστη.

Όπως αναφέρουν οι Raghavan και Wong (1986) η κεντρική ιδέα του VSM στηρίζεται στην αναπαράσταση κάθε αντικειμένου ΑΠ (λέξη, έγγραφο κλπ.) ως διάνυσμα ενώ οι Turney και Pantel (2010) συμπληρώνουν πως κάθε έγγραφο μιας συλλογής αντικατοπτρίζεται ως διάνυσμα σε ένα διανυσματικό χώρο. Μάλιστα όπως αναφέρεται στη βιβλιογραφία (Lan, Tan και Su 2009) κάθε όρος καθορίζεται από ένα δείκτη σημασίας, ο οποίος αντικατοπτρίζει τη σημαντικότητα του σε σχέση με το περιεχόμενο του εκάστοτε εγγράφου και αναπαριστά πόσο ο όρος αυτός συνεισφέρει στη σημασιολογία του εγγράφου.

Μια από τις βασικές παραδοχές του VSM σύμφωνα με τους Raghavan και Wong (1986) είναι πως το σύστημα στο οποίο υπάρχει ο διανυσματικός χώρος είναι γραμμικό. Σε ένα γραμμικό σύστημα ισχύουν βασικοί κανόνες και αξιώματα της άλγεβρας όπως η ιδιότητα πρόσθεσης δύο συστατικών του συστήματος από τα οποία προκύπτει ένα άλλο συστατικό του συστήματος και η δυνατότητα πολλαπλασιασμού κάθε συστατικού του συστήματος με πραγματικό αριθμό. Χρησιμοποιώντας τη θεωρία διανυσμάτων τα έγγραφα συγκρίνονται ως προς την ομοιότητα τους. Όπως αναφέρεται στην βιβλιογραφία (Sproerri 1995) τα έγγραφα μετατρέπονται σε διανύσματα και τοποθετούνται σε πολυδιάστατο χώρο, τις διαστάσεις του οποίου αποτελούν οι όροι που περιέχονται στα έγγραφα. Στο Σχήμα 1.4 παρατίθεται ένα παράδειγμα αναπαράστασης διανυσμάτων εγγράφων σε τρισδιάστατο χώρο, με D να αντιστοιχεί στο έγγραφο και T στον όρο εγγράφου. Όπως μπορεί κανείς να διακρίνει οι διαστάσεις στο σχήμα αντιστοιχούν στον αριθμό των όρων εγγράφων.



Σχήμα 1. 4. Αναπαράσταση διανυσμάτων εγγράφων στον πολυδιάστατο χώρο από Salton, Wong and Yang (1975)

Οι όροι που περιέχονται στα έγγραφα χρησιμοποιούνται προκειμένου να δημιουργηθεί ένα ευρετήριο για την αναπαράσταση των εγγράφων της συλλογής και μέσω λεξιλογικής ανάλυσης πρέπει να καθοριστούν οι σημαντικοί όροι ως προς το σημασιολογικό περιεχόμενο του κάθε εγγράφου. Όσο μεγαλύτερη είναι η απόσταση μεταξύ των διανυσμάτων τόσο λιγότερη ομοιότητα έχουν ενώ όσο πιο κοντά βρίσκονται στο διανυσματικό χώρο τόσο πιο κοντά είναι και ως προς το περιεχόμενο τους. Τα ερωτήματα του χρήστη και αυτά μετατρέπονται σε διανύσματα ώστε να είναι εφικτή η σύγκριση μεταξύ των εγγράφων της συλλογής και

του ερωτήματος και τελικά να επιστραφούν τα σχετικότερα αποτελέσματα στον χρήστη σε μορφή μιας λίστας κατάταξης.

Αναλυτικότερα σύμφωνα με τους Salton και Buckley (1988) η αναπαράσταση των εγγράφων από όρους - διανύσματα (termvectors) περιγράφεται από το μαθηματικό τύπο (1.1):

$$D = (t_1, t_j, \dots, t_p) \quad (1.1)$$

Κάθε t_k αντιστοιχεί σε έναν όρο που ανατίθεται σε κάποιο έγγραφο D. Καθώς τα ερωτήματα των χρηστών αναπαρίστανται και αυτά είτε σε διανυσματική μορφή είτε ως λογικές δηλώσεις (booleanstatements), ο παραπάνω τύπος μετατρέπεται ως ακολούθως όπως φαίνεται στους μαθηματικούς τύπους (1.2) και (1.3):

$$Q = (q_a, q_b, \dots, q_r) \quad (1.2)$$

$$Q = (q_a \text{ and } q_b) \text{ ή } (q_c \text{ and } q_d \text{ and } \dots) \text{ or} \quad (1.3)$$

Όπου q_k αναπαριστά έναν όρο ο οποίος έχει ανατεθεί σε ένα ερώτημα Q.

Σύμφωνα με τους Salton και Buckley (1988) αν w_{dk} (ή w_{qk}) αναπαριστά το βάρος του όρου t_k στο έγγραφο D (ή το ερώτημα Q) και για την αναπαράσταση του περιεχομένου διατίθενται t όροι, τότε οι τύποι (1) και (2) μετατρέπονται ως ακολούθως στους μαθηματικούς τύπους (4) και (5) αντίστοιχα:

$$D = (t_0 w_{d_0}; t_1 w_{d_1}; \dots; t_t w_{d_t}) \quad (1.4)$$

$$Q = (q_0 w_{q_0}; q_1 w_{q_1}; \dots; q_t w_{q_t}) \quad (1.5)$$

Όταν ο όρος k δεν έχει ανατεθεί στο έγγραφο D ή στο ερώτημα Q τότε το βάρος του όρου w_{dk} ή w_{qk} είναι ίσο με 0 ενώ στην αντίστροφη περίπτωση είναι ίσο με το 1.

1.2.2 Όροι στάθμισης (βάρη)

Η απόδοση βάρους στους όρους αποτελεί ένα πολύ σημαντικό βήμα για την βελτίωση της αποτελεσματικότητας στην ΑΠ. Όπως αναφέρεται στη βιβλιογραφία (Lan, Tan και Su 2009) οι μέθοδοι απόδοσης βάρους που αναφέρονται είναι οι παρακάτω: (α) δυαδική (binary) αναπαράσταση, (β) πιο ευρέως χρησιμοποιούμενη tf-idf και παραλλαγές της. Όπως αναφέρουν και οι Salton και Buckley (1988) στόχος της απόδοσης βαρών σε όρους είναι η βελτίωση της αποτελεσματικότητας (effectiveness) της ανάκτησης, δηλαδή η βελτίωση της διαδικασίας ανάκτησης συναφών εγγράφων με βάση τις πληροφοριακές ανάγκες του χρήστη αλλά ταυτόχρονα και στην εκκαθάριση της λίστας κατάταξης από μη συναφή έγγραφα. Τα δύο βασικά μέτρα για την επίτευξη των παραπάνω επιμέρους στόχων είναι η ανάκληση και η ακρίβεια, μέσω των οποίων επιτυγχάνεται η αξιολόγηση των ΣΑΠ.

Επεξηγώντας αναλυτικότερα τις έννοιες της ανάκλησης και ακρίβειας οι Salton και Buckley (1988) τονίζουν ότι είναι αναγκαίοι οι συμβιβασμοί στις απαιτήσεις ως προς τις έννοιες αυτές σε σχέση με ένα ΣΑΠ. Ιδανικά ένα ΣΑΠ θα έπρεπε να έχει υψηλή ανάκληση και υψηλή ακρίβεια, μα καθώς η ανάκληση σχετίζεται με την υψηλή συχνότητα όρων στα έγγραφα της συλλογής ενώ η υψηλή ακρίβεια από τη χρήση σπανιότερων όρων, το ακριβώς αντίθετο δηλαδή είναι κατανοητή η ανάγκη για συμβιβασμό μεταξύ των επιπέδων των δύο μέτρων. Προκειμένου να επιτευχθεί ο συμβιβασμός αυτός στη βιβλιογραφία (Lan, Tan και Su 2009) αναφέρονται οι παράγοντες απόδοσης βάρους συνοπτικά και χωρίζονται σε τρεις κατηγορίες:

- Παράγοντες συχνότητας εμφάνισης όρου (termfrequencyfactors)
- Δυαδική αναπαράσταση (0 απουσία όρου, 1 παρουσία όρου), tf κ.α.
- Παράγοντες συχνότητας εμφάνισης όρου στην συλλογή (collection Frequency Factor).
- Παράγοντες κανονικοποίησης (normalization Factor)

Για την εξίσωση του διαφορετικού μήκους των εγγράφων ώστε να περιοριστεί το βάρος όρου στο διάστημα (0,1). Είναι προφανές ότι στη δυαδική αναπαράσταση η κανονικοποίηση δεν έχει χρησιμότητα αφού οι τιμές είναι ήδη 0 ή 1. Αναλυτικότερα οι Salton και Buckley (1988) παρουσιάζουν τους πιο δημοφιλείς παράγοντες για την απόδοση βάρους tf-idf. Ο παράγοντας συχνότητας εμφάνισης όρων tf (termfrequency) μετρά τη συχνότητα εμφάνισης όρων στα έγγραφα ή τα ερωτήματα και αφορά την ανάκληση. Όμως η συχνότητα εμφάνισης των όρων δεν αρκεί διότι αν οι συχνότεροι όροι δεν βρίσκονται μονάχα σε κάποια

μεμονωμένα έγγραφα αλλά είναι διάσπαρτοι σε όλη την συλλογή τότε είναι πιθανό να ανακτηθούν όλα τα έγγραφα με αποτέλεσμα τα επίπεδα της ακρίβειας να είναι ιδιαιτέρως χαμηλά.

Έτσι ο παράγοντας *idf* (inversedocumentfrequency), ο οποίος εξαρτάται από τη συλλογή έρχεται για να εξισορροπήσει τα επίπεδα ανάκλησης - ακρίβειας. Ένας *idf* παράγοντας υπολογίζεται από $\log N/n$, όπου *n* τα έγγραφα στα οποία έχει ανατεθεί ένας όρος σε μια συλλογή και *N* το σύνολο των εγγράφων της συλλογής. Σύμφωνα πάντα με τους Salton και Buckley (1988) οι καλύτεροι όροι θα πρέπει από την μια πλευρά να έχουν υψηλή συχνότητα εμφάνισης όρων και από την άλλη χαμηλές συνολικά συχνότητες στη συλλογή, ώστε να απομονώσουν τα συναφή έγγραφα στο πλαίσιο μιας συλλογής. Έτσι ο υπολογισμός της σημαντικότητας των όρων πραγματοποιείται μέσω του γινόμενου: $tf \times idf$.

Τέλος, οι Salton και Buckley (1988) αναφέρονται στον παράγοντα κανονικοποίησης που καθιστά εφικτή την αντιμετώπιση όλων των συναφών εγγράφων ως ίσης σημασίας στην ανάκτηση. Καθώς όλα τα έγγραφα δεν έχουν το ίδιο μήκος, η σύγκριση τους χωρίς κανονικοποίηση δεν θεωρείται έγκυρη αφού ανάλογα με το μήκος των εγγράφων διαμορφώνεται και το μήκος των διανυσμάτων τους. Έτσι τα μικρού μήκους έγγραφα (shortdocuments) τείνουν να αναπαρίστανται από μικρά διανύσματα όρων και αντίστοιχα τα μεγάλου μήκους έγγραφα αναπαρίστανται από μεγάλα διανύσματα όρων. Συνεπώς τα μεγαλύτερου μήκους έγγραφα έχουν καλύτερες πιθανότητες ανάκτησης.

Σύμφωνα με τους Salton και Buckley (1988) από τις εξισώσεις (4) και (5) μπορεί να εξαχθεί μια τιμή ομοιότητας μεταξύ εγγράφων και ερωτήματος, συγκρίνοντας τα διανύσματα τους, όπως φαίνεται στον παρακάτω μαθηματικό τύπο (6):

$$similarity(Q, D) = \sum_{k=1}^l w_{qk} \cdot w_{dk} \quad (1.6)$$

Μάλιστα σύμφωνα με τους Salton και Buckley (1988) κάθε βάρος όρου, εξαρτάται σε κάποιο βαθμό από τα βάρη των υπόλοιπων όρων στο ίδιο διάνυσμα. Ένα τυπικό βάρος όρου που χρησιμοποιεί παράγοντα κανονικοποίησης μήκους διανύσματος φαίνεται στο μαθηματικό τύπο (7):

$$\frac{w_{dk}}{\sqrt{\sum_{vector} (w_{dk})^2}} \quad (1.7)$$

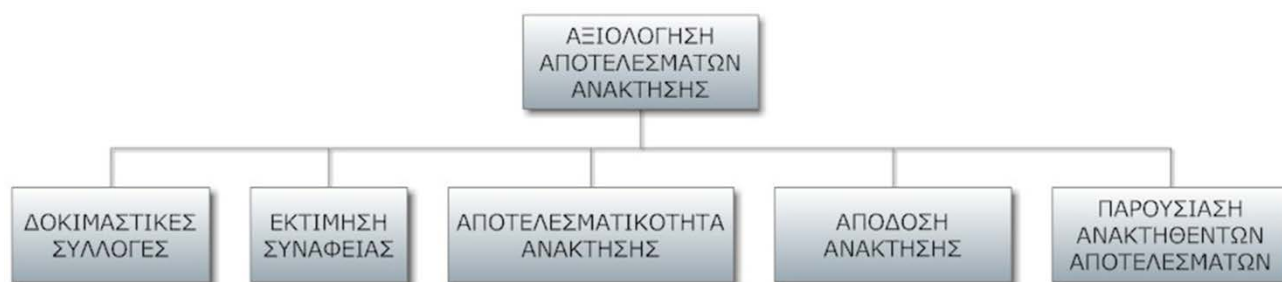
Τότε ο τύπος (6) σε συνδυασμό με τον παράγοντα κανονικοποίησης (7) δίνουν ως αποτέλεσμα τον τύπο *cosinevectorsimilarity* (8) :

$$similarity(Q, D) = \frac{\sum_{k=1}^l w_{qk} \cdot w_{dk}}{\sqrt{\sum_{k=1}^l (w_{qk})^2 \cdot \sum_{k=1}^l (w_{dk})^2}} \quad (1.8)$$

1.2.3 Αξιολόγηση αποτελεσμάτων

Ένα σημαντικό κεφάλαιο στην ΑΠ αφορά στην αξιολόγηση των ΣΑΠ. Όπως έχει ήδη αναφερθεί το επίκεντρο των συστημάτων αυτών είναι τα μοντέλα ΑΠ και πιο συγκεκριμένα ο αλγόριθμος κατάταξης, με βάση τον οποίο παρουσιάζονται τα αποτελέσματα στο χρήστη. Εξ αρχής ο απώτερος στόχος της ΑΠ είναι η ικανοποίηση των πληροφοριακών αναγκών του χρήστη. Αναφερόμαστε λοιπόν για χρηστο-κεντρική αξιολόγηση της αναζήτησης που εκτελεί το ΣΑΠ σύμφωνα με τους Croft et al. (2010), καθώς ο τελικός κριτής της ποιότητας αναζήτησης είναι ο χρήστης. Όμως τα ερωτήματα που εισάγει ο χρήστης αποτελούν περιγραφές της πληροφοριακής του ανάγκης, μη συγκεκριμένες. Έτσι αποτελεί ιδιαίτερη πρόκληση το να καθοριστεί ένα έγγραφο συναφές ως προς αυτές καθώς η υποκειμενική φύση των ερωτημάτων και η δυναμική τους ιδιότητα αποτελούν τα μεγαλύτερα εμπόδια.

Έτσι η μέθοδος που ακολουθείται είναι η διαρκής αξιολόγηση των αποτελεσμάτων της ανάκτησης, ώστε τα συστήματα να βελτιώνονται όλο και περισσότερο. Παρακάτω φαίνονται σχηματικά (βλέπε Σχήμα. 1.5) τα κύρια θέματα που αφορούν στην αξιολόγηση:



Σχήμα 1.5. Αξιολόγηση ΑΠ σύμφωνα με ταξινόμηση ACM

Για να πραγματοποιηθεί αξιολόγηση ενός ΣΑΠ θα πρέπει να υπάρχουν κάποιες μετρήσιμες ιδιότητες του, ως προς τις οποίες θα εξεταστεί. Σύμφωνα με Cerietal. (2013) πρέπει να είναι γνωστή η ύπαρξη μιας συλλογή D και ενός συνόλου ερωτημάτων Q ως προς τα οποία να μπορεί να γίνει αναφορά και τέλος να υπάρχει μια πλειάδα $t_{jk} = (d_j, q_k, r^*)$ όπου για κάθε $q_k \in Q$ και $d_j \in D$ να υπάρχει μια αντίστοιχη δυαδική κρίση (binary judgment) r^* της μεταξύ τους συνάφειας (του d_j σε σχέση με το q_k δηλαδή), όπως εκτιμήθηκε από μια αρχή αναφοράς.

1.2.4 Αξιολόγηση ΣΑΠ

Προκειμένου να είναι εφικτή η σύγκριση και αξιολόγηση του ΣΑΠ αναπτύχθηκαν διάφορες τέτοιες συλλογές εγγράφων (standard test collections), όπως αναφέρθηκαν παραπάνω, οι οποίες χρησιμοποιούνται ήδη από τη δεκαετία του 1960. Ορισμένα ευρέως χρησιμοποιούμενα παραδείγματα αποτελούν η συλλογή Cranfield, TREC, Reuters και TRC2. Οι συλλογές αυτές περιλαμβάνουν, όπως αναφέρθηκε παραπάνω, σύνολα εγγράφων, ερωτημάτων και σύνολα κριτικών συνάφειας. Ο λόγος ύπαρξης των παραπάνω συλλογών είναι ο υπολογισμός της εκτίμησης συνάφειας με τις πληροφοριακές ανάγκες χρηστών, με σκοπό την αξιολόγηση. Συγκεκριμένα, σύμφωνα με τους Manning, Raghavan και Schutze (2008) οι πληροφοριακές αυτές ανάγκες σχεδιάζονται από ειδικούς του τομέα. Η συγκέντρωση εκτιμήσεων συνάφειας έγκειται σε κάποιο ανθρώπινο δυναμικό το οποίο θα δαπανήσει ιδιαίτερο χρόνο ώστε να εφαρμόσει όλα τα ερωτήματα στη συλλογή, μια ιδιαίτερα δαπανηρή διαδικασία. Όπως αναφέρουν οι Manning, Raghavan και Schutze (2008) τέτοιες διαδικασίες είναι εφικτές για μικρές συλλογές ενώ για μεγάλες συλλογές η εκτίμηση συνάφειας συνηθίζεται να εφαρμόζεται σε περιορισμένο υποσύνολο εγγράφων για κάθε ερώτημα και όχι για όλη τη συλλογή.

1.2.4.1 Αποτελεσματικότητα και απόδοση ανάκτησης

Στην ενότητα αυτή θα γίνει αναφορά σε δύο βασικές έννοιες που αφορούν την αξιολόγηση των ΣΑΠ: την αποτελεσματικότητα και την ανάκληση. Όπως λοιπόν αναφέρεται στη βιβλιογραφία (Croft et al. 2009) η **αποτελεσματικότητα** ορίζεται ως η **ανάκτηση των περισσότερο συναφών εγγράφων** σε σχέση με το ερώτημα χρήστη ενώ η **απόδοση** ως η όσο το δυνατόν **ταχύτερη επεξεργασία των ερωτημάτων** του χρήστη.

Σύμφωνα με τους Manning et al. (2008) για την αξιολόγηση της **αποτελεσματικότητας** ενός ΣΑΠ πρέπει να υπολογιστούν η ακρίβεια και η ανάκληση, μέτρα τα οποία χρησιμοποίησε πρώτος την δεκαετία του 1960 ο Cyril Cleverdon, πρωτοπόρος των ΣΑΠ. Στα μέτρα ανάκλησης και ακρίβειας έγινε σύντομη αναφορά σε παραπάνω ενότητα. Όπως αναφέρεται στη βιβλιογραφία (Kowalski 2011) μέσω της ανάκλησης και της ακρίβειας αξιολογείται η ποιότητα της αποτελεσματικότητας της αναζήτησης. Παρακάτω αναλύονται οι μετρικές συνάφειας, **ακρίβεια** και **ανάκληση**, σύμφωνα με τον Ingwersen (1992) οι οποίες υπολογίζονται με τους παρακάτω τύπους και έχουν μια **αντιστρόφως ανάλογη σχέση μεταξύ τους**:

- Ακρίβεια: αριθμός συναφών ανακτηθέντων εγγράφων R προς τον αριθμό των ανακτηθέντων εγγράφων L .
- Ανάκληση: αριθμός συναφών εγγράφων που έχουν ανακτηθεί R προς τον συνολικό αριθμό συναφών εγγράφων στην συλλογή C .

Σύμφωνα με Ingwersen (1992) υπάρχει ακόμα ένα μέτρο, το οποίο μπορεί σε κάποιες περιπτώσεις να αντικαταστήσει την ακρίβεια και ονομάζεται αστοχία. Η αστοχία αντιπροσωπεύει την σχέση μεταξύ του

αριθμού μη συναφών ανακτηθέντων εγγράφων και όλων των μη συναφών εγγράφων (όπου N ο συνολικός αριθμός συλλογής).

Σχετικά με την απόδοση των ΣΑΠ, όπως αναφέρεται στην βιβλιογραφία (Kowalski 2011), το αιτούμενο είναι ο καθορισμός του χρόνου απόκρισης, δηλαδή του χρονικού διαστήματος στο οποίο εκτελείται η διαδικασία της αναζήτησης. Η απόδοση του συστήματος είναι ιδιαίτερος σημαντική καθώς όπως αναφέρεται στη βιβλιογραφία (Buettcher, Clarke και Cormack 2010) επηρεάζει το κόστος λειτουργίας του συστήματος αλλά και την ικανοποίηση των χρηστών. Σύμφωνα με Kowalski (2011) κάποιες βασικές έννοιες στον χρόνο απόκρισης είναι η αρχή και το τέλος του χρονικού διαστήματος αυτού, δηλαδή το πότε αρχίζει η αναζήτηση (με την εντολή έναρξης του χρήστη) και το πότε τελειώνει (ολοκλήρωση της αναζήτησης). Ακόμη ο Kowalski (2011) συμπληρώνει πως η ολοκλήρωση της αναζήτησης μπορεί να είναι ένα ζήτημα με διαφορετική προσέγγιση για τον χρήστη σε σχέση με το ΣΑΠ. Για το ΣΑΠ η αναζήτηση ολοκληρώνεται μονάχα όταν έχουν βρεθεί όλα τα αποτελέσματα, ενώ κατά τη χρήση μια αναζήτηση έχει ολοκληρωθεί ακόμη και αν έχει επιστραφεί έστω ένα αποτέλεσμα. Οι Croftetal. (2009) από την άλλη τονίζουν ότι η απόδοση εξαρτάται από τα ευρήματα.

Ένα άλλο ζήτημα που επηρεάζει την απόδοση μιας μηχανής αναζήτησης είναι αυτό της εξυπηρέτησης πληθώρας χρηστών από τη μηχανή ταυτόχρονα. Όπως αναφέρεται στη βιβλιογραφία (Buettcher, Clarke και Cormack 2010) μια μηχανή αναζήτησης μπορεί να πρέπει να εξυπηρετήσει ταυτόχρονα χιλιάδες ερωτήματα για μεμονωμένους χρήστες και για να θεωρείται αποδοτική ώστε να μη χάνει τους χρήστες της, ο χρόνος απόκρισης που εξυπηρετεί τον καθένα ξεχωριστά θα πρέπει να μην επηρεάζεται σε τέτοιο βαθμό ώστε να τον αφήσει δυσαρεστημένο. Ο μέσος αριθμός ερωτημάτων που επεξεργάζεται η μηχανή σε συγκεκριμένο χρονικό διάστημα ονομάζεται ρυθμός διαμεταγωγής (throughput).

1.2.5 Ευαισθησία – Ειδικότητα

Οι όροι ευαισθησία και ειδικότητα εμφανίζονται κυρίως για την αξιολόγηση διαγνωστικών τεστ στην ιατρική αλλά και σε πιο εξειδικευμένες περιπτώσεις αξιολόγησης σε διάφορους επιστημονικούς κλάδους, όπως π.χ. στην αξιολόγηση και εκπαίδευση νευρωνικών δικτύων (Poulos et al. 2007), στην αξιολόγηση απόδοσης ταξινομητών σε σχέση με δυαδική ταξινόμηση (Powers 2011, Fawcett 2006) και γενικότερα στην αξιολόγηση αλγορίθμων και μέτρηση απόδοσης μοντέλων. Σύμφωνα με τους Manning, Raghavan και Schutze (2008) οι μετρικές της ευαισθησίας και της ειδικότητας στο πλαίσιο της ΑΠ συμπίπτουν με την ανάκληση και την ακρίβεια αντίστοιχα και συχνά στο πλαίσιο της αξιολόγησης αναπαρίστανται με γραφικό τρόπο μέσω μιας Receiver Operating Characteristics curve ή αλλιώς ROC curve.

Ειδικότερα για τον έλεγχο υποθέσεων στα διαγνωστικά τεστ, οι Wong και Gek (2011) τονίζουν πως τα διαγνωστικά – κλινικά τεστ, τα οποία καθορίζουν την ύπαρξη ή μη μιας ασθένειας στα αντικείμενα της μελέτης, στον πληθυσμό δηλαδή που έχει υποβληθεί στο τεστ, δεν είναι πάντοτε έγκυρα και υπάρχει η πιθανότητα κάποιου ποσοστού του αποτελέσματος να είναι λάθος. Γι' αυτό τον λόγο, οι Wong και Gek (2011) τονίζουν πως πρέπει να γίνεται αξιολόγηση των τεστ αυτών (testvalidation) και αυτό επιτυγχάνεται μέσω της μέτρησης 4 παραμέτρων: ευαισθησία, ειδικότητα, θετική προγνωστική αξία (positivepredictivevalue, PPV) και αρνητική προγνωστική αξία (NPV). Συνεπώς χρησιμοποιώντας τις παραπάνω παραμέτρους μπορεί να γίνει έλεγχος της εγκυρότητας ή μη μιας υπόθεσης, όχι μόνο στην ιατρική και τα διαγνωστικά τεστ αλλά και γενικότερα για στατιστικούς λόγους.

Οι Lalkhen και McCluskey (2008), συμπληρώνουν τους παρακάτω ορισμούς προκειμένου να κατανοηθεί στη συνέχεια η ανάλυση των όρων ευαισθησίας και ειδικότητας. Οι πιθανές τιμές που μπορεί να έχει ως αποτέλεσμα ένα διαγνωστικό τεστ είναι οι ακόλουθες:

- Αληθινό θετικό: όταν ο ασθενής έχει την ασθένεια και το τεστ βγαίνει θετικό.
- Ψεύτικο θετικό: όταν ο ασθενής δεν έχει την ασθένεια και το τεστ είναι θετικό
- Αληθινό αρνητικό: όταν ο ασθενής δεν έχει την ασθένεια και το τεστ είναι αρνητικό
- Ψεύτικο αρνητικό: όταν ο ασθενής έχει την ασθένεια αλλά το τεστ είναι αρνητικό.

Σύμφωνα με τον Akobeng (2007) αρκετά χρήσιμο στην κατανόηση των όρων ευαισθησία – ειδικότητα είναι η τοποθέτηση των παραπάνω πιθανών αποτελεσμάτων, όπως φαίνονται στον Πίνακα 1:

Πίνακας 1. Πιθανά αποτελέσματα για υπολογισμό ευαισθησίας - ειδικότητας

	ΑΣΘΕΝΗΣ	ΜΗ ΑΣΘΕΝΗΣ
ΘΕΤΙΚΟ	A	B
ΑΡΝΗΤΙΚΟ	C	D

Ο Akobeng (2007) στο άρθρο του αναφέρει τον υπολογισμό των όρων μέσω των (9), (10), (11) και (12):

$$\text{ευαισθησία} = \frac{a}{a + c} \quad (1.9)$$

$$\text{εξειδίκευση} = \frac{b}{b + d} \quad (1.10)$$

$$PPV = \frac{a}{a + b} \quad (1.11)$$

$$NPV = \frac{d}{c + d} \quad (1.12)$$

Οι Zhu, Zeng και Wang (2010) αναφέρουν ότι η ευαισθησία αφορά το πόσο καλό είναι το τεστ στην ανίχνευση των πραγματικών ασθενών (δηλαδή θετική απάντηση στο αποτέλεσμα του τεστ) ενώ η ειδικότητα αφορά την ανίχνευση των μη πασχόντων (δηλαδή αρνητική απάντηση στο αποτέλεσμα του τεστ).

1.2.6 Παρουσίαση Ανακτηθέντων Εγγράφων

Σύμφωνα με τη βιβλιογραφία (Wu, Bi και Zeng 2010, Kamps 2009), η παρουσίαση ανακτηθέντων αποτελεσμάτων συνδέεται άμεσα με την αξιολόγηση και ο τρόπος παρουσίασης των ανακτηθέντων αποτελεσμάτων στους χρήστες μπορεί να καθορίσει το αν το ΣΑΠ θα θεωρηθεί επιτυχημένο ή όχι. Ειδικότερα στην βιβλιογραφία αναφέρεται (Joho και Jose 2006) πως έχει αποδειχτεί ότι οι χρήστες όταν ψάχνουν για μια πληροφορία στο διαδίκτυο μέσω μηχανών αναζήτησης εξετάζουν όσο το δυνατόν μικρότερο αριθμό εγγράφων και συνήθως περιορίζονται στην πρώτη σελίδα αποτελεσμάτων από τη λίστα. Έτσι έγκειται στον τρόπο παρουσίασης των αποτελεσμάτων, στο σύνολο τους αλλά και σε κάθε μεμονωμένο αποτέλεσμα, να δελεάσουν και να προτρέψουν τον χρήστη να επιλέξει κάποια από αυτά τα αποτελέσματα για περιήγηση. Οι πληροφορίες που παρουσιάζονται συνήθως στα αποτελέσματα είναι τίτλος, περίληψη, διεύθυνση url, αποσπάσματα εγγράφου. Μέσω των στοιχείων που θα εμφανιστούν στο χρήστη μόλις δει τα αποτελέσματα για πρώτη φορά θα ληφθεί η απόφαση από τον ίδιο για τη συνάφεια τους και για τον αν θα ενδιαφερθεί να διαβάσει περαιτέρω κάποιο έγγραφο ώστε να βρει την πληροφορία που χρειάζεται. Τέλος από το αν θα είναι ευχαριστημένος από την όλη αυτή διαδικασία επιλογής και προβολής αποτελεσμάτων θα γίνει αξιολόγηση του συστήματος και θα προτιμάται ή όχι από τους χρήστες.

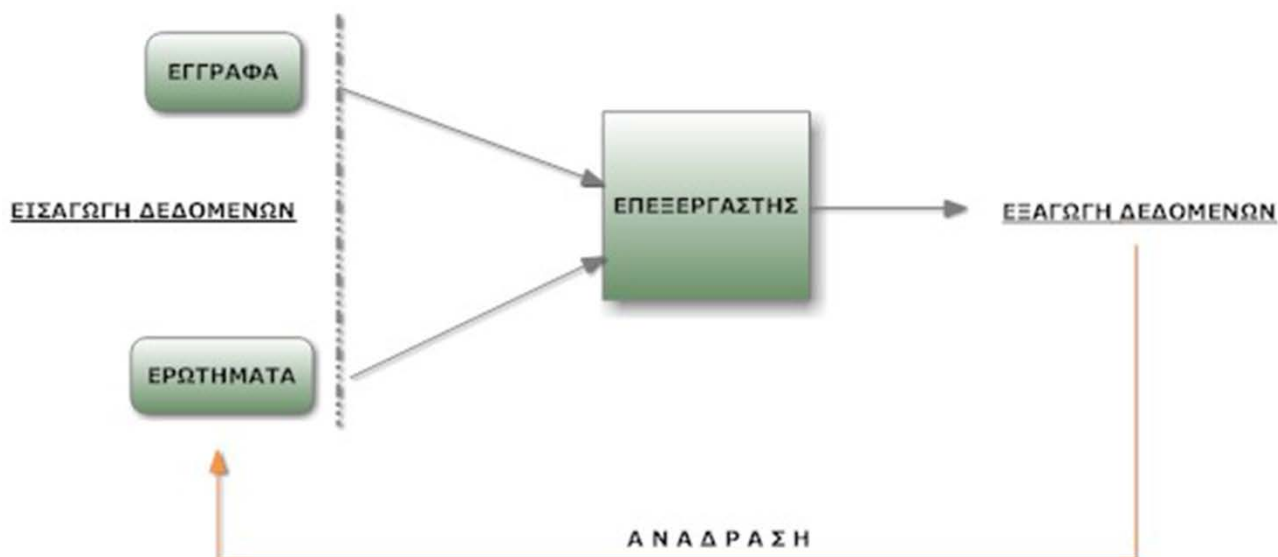
Κατά την κατάταξη των εγγράφων για την παρουσίαση των αποτελεσμάτων, σε κάθε έγγραφο ανατίθεται ένας βαθμός συνάφειας (Wu, Bi και Zeng 2010) μέσα στη συλλογή. Από τις βαθμολογίες αυτές γίνεται η κατάταξη τους με πιο συνηθισμένη μορφή παρουσίασης ανακτηθέντων αποτελεσμάτων αυτή της λίστας καταταγμένων εγγράφων. Όσα βρίσκονται ψηλότερα στην κατάταξη είναι πιο συναφή ενώ προχωρώντας προς τα κάτω στη λίστα ακολουθείται μια φθίνουσα πορεία σε σχέση με τη συνάφεια της πληροφοριακής ανάγκης του χρήστη. Εναλλακτική παρουσίαση είναι αυτή που χρησιμοποιεί ιεραρχικές δομές. Στη βιβλιογραφία αναφέρεται (Crestani και Wu 2006) πως σύμφωνα με πληθώρα ερευνών και πειραμάτων, η παρουσίαση ανακτηθέντων αποτελεσμάτων με συσταδοποίηση, δηλαδή ομαδοποίηση αποτελεσμάτων χωρίς να υπάρχουν προκαθορισμένες κλάσεις είναι αποτελεσματικότερη από μια απλή λίστα

κατάταξης, καθώς βασίζεται στην υπόθεση ότι τα έγγραφα που ομαδοποιούνται μαζί έχουν παρόμοια συνάφεια για συγκεκριμένο ερώτημα. Όπως αναφέρουν οι Wu, Bi και Zeng (2010), η σχέση μεταξύ των βαθμολογιών εγγράφων για την κατάταξη αποτελεί τις πιθανότητες συνάφειας, οι οποίες μπορεί να είναι ιδιαίτερα χρήσιμες σε προχωρημένα ΣΑΠ.

1.2.7 Αρχιτεκτονική ΣΑΠ

Σύμφωνα με τη βιβλιογραφία (Canfora και Cerulo 2004), οι μηχανές αναζήτησης θεωρούνται συνώνυμες με τα ΣΑΠ στο περιβάλλον του διαδικτύου. Κατά συνέπεια, όπως είναι αναμενόμενο και σύμφωνα με τη βιβλιογραφία (Croft et al. 2009) οι βασικοί στόχοι που πρέπει να επιτευχθούν για μια μηχανή αναζήτησης ως προς τον σχεδιασμό και την αρχιτεκτονική της, είναι ταυτόσημοι με αυτούς των ΣΑΠ (αποτελεσματικότητα και απόδοση). Καθώς η αρχιτεκτονική μιας μηχανής αναζήτησης είναι ιδιαιτέρως πολύπλοκη, αρχικά θα γίνει παρουσίαση μιας απλουστευμένης μορφής ενός ΣΑΠ. Έτσι σύμφωνα με τον Van Rijsbergen (1979) ένα ΣΑΠ παρουσιάζεται στην απλούστερη και πιο γενικευμένη μορφή του ως ένα σύστημα αποτελούμενο από τρία βασικά μέρη: εισαγωγή δεδομένων (input), επεξεργαστή και εξαγωγή δεδομένων (output), όπως φαίνεται σχηματικά στην Σχήμα 1.5

- **Εισαγωγή δεδομένων:** το ΣΑΠ δέχεται ως δεδομένα τα έγγραφα από την μια πλευρά και από την άλλη τα ερωτήματα του χρήστη και τα δύο πρέπει να αναπαρίστανται με κατάλληλο τρόπο για τον υπολογιστή. Καθώς ο χρήστης αλλάζει τις απαιτήσεις του κατά την διάρκεια μιας ενότητας διαδικτυακής αναζήτησης, μπορεί να βελτιωθεί η αναζήτηση μέσω της διαδικασίας ανάδρασης.
- **Επεξεργαστής:** ασχολείται με τη διαδικασία ανάκτησης. Ορισμένες από τις διεργασίες που περιλαμβάνει είναι η δόμηση της πληροφορίας με κατάλληλο τρόπο και η διαδικασία αυτή καθαυτή της ανάκτησης του συνόλου εγγράφων.
- **Εξαγωγή δεδομένων:** περιλαμβάνει τα αποτελέσματα της διαδικασίας ανάκτησης.



Σχήμα 1.6. Αναπαράσταση αρχιτεκτονικής ΣΑΠ

Εμβαθύνοντας περαιτέρω στην αρχιτεκτονική ενός ΣΑΠ, μια αρκετά πιο λεπτομερής παρουσίαση (βλέπε Σχήμα 1.6) προέρχεται από τους Baeza-Yates και Ribeiro-Neto (2011), στην οποία ένα ΣΑΠ υλοποιείται από δύο μεγάλες ενότητες διαδικασιών. Η πρώτη ενότητα αφορά τη (α) συλλογή των εγγράφων και όλες τις διαδικασίες που την αφορούν και η δεύτερη την (β) ανάκτηση των εγγράφων και όλων των διαδικασιών που περιλαμβάνονται για την υλοποίηση της. Αναλυτικότερα:

1. **Συλλογή εγγράφων:** είναι αποθηκευμένη σε ένα κεντρικό αποθετήριο και πρέπει να ευρετηριαστεί, δηλαδή να δημιουργηθεί ένα ευρετήριο από τα έγγραφα που την απαρτίζουν. Όλα τα βήματα για τη δημιουργία του ευρετηρίου αυτού αποτελούν τη διαδικασία ευρετηρίασης, η οποία πραγματοποιείται εκτός διαδικτύου.

2. **Ανάκτηση εγγράφων:** περιλαμβάνει όλα τα βήματα που ακολουθούνται ώστε να παραχθεί ένα σύνολο ανακτηθέντων εγγράφων, τα οποία θα είναι συναφή ως προς το ερώτημα χρήστη. Η διαδικασία ανάκτησης εγγράφων αφορά την αναζήτηση του χρήστη για πληροφορία με βάση την πληροφοριακή του ανάγκη ή σε περιήγηση πληροφορίας όταν ο χρήστης περιηγείται μέσω υπερσυνδέσμων. Εστιάζοντας στην αναζήτηση πληροφοριών το ερώτημα του χρήστη αναλύεται, επεκτείνεται και επεξεργάζεται προκειμένου να επιστραφεί ένα υποσύνολο εγγράφων από το ευρετήριο. Πιο συγκεκριμένα, το ερώτημα αναλύεται και τροποποιείται εκτελώντας κάποιες βασικές διεργασίες, όπως διόρθωση ορθογραφικών λαθών, αποκλεισμός όρων ως διακόπτουσες λέξεις (λέξεις που κατέχουν λειτουργικό ρόλο στο σχηματισμό δομής της πρότασης ενώ ταυτόχρονα δεν συνεισφέρουν σημασιολογικά στο έγγραφο) και τέλος τροποποιείται ξανά μέσω προτάσεων από το σύστημα. Το τελικό τροποποιημένο ερώτημα επεξεργάζεται και σύμφωνα με αυτό επιστρέφεται ένα σύνολο ανακτηθέντων εγγράφων με έγγραφα που περιέχουν τους όρους του ερωτήματος χρήστη. Τα ανακτηθέντα έγγραφα κατατάσσονται με βάση τη συνάφεια σε σχέση με το ερώτημα του χρήστη και επιστρέφονται αυτά που κατά την αξιολόγηση κρίνονται πιο συναφή.

Όπως ήδη έχει αναφερθεί στην αρχή της ενότητας αυτής, οι μηχανές αναζήτησης αποτελούν την πιο διαδεδομένη εφαρμογή για την υλοποίηση τεχνικών ΑΠ και θεωρούνται σχεδόν ταυτόσημες με την έννοια της ΑΠ. Μάλιστα στη βιβλιογραφία (Croft et al. 2010) αναφέρεται ο παρακάτω ορισμός της μηχανής αναζήτησης: «μια μηχανή αναζήτησης αποτελεί την πρακτική εφαρμογή τεχνικών ανάκτησης πληροφοριών σε μεγάλης κλίμακας συλλογές κειμένων». Κατά συνέπεια τα βασικά ζητήματα ΑΠ αποτελούν βασικά ζητήματα και για τη σχεδίαση μηχανών αναζήτησης.

Φυσικά η ΑΠ εφαρμόζεται σε πληθώρα άλλων πεδίων εκτός των μηχανών αναζήτησης. Ορισμένα παραδείγματα χρήσης της στη βιβλιογραφία (Ceri et al. 2013) αφορούν συστήματα φιλτραρίσματος πληροφορίας, εξαγωγή περιλήψεων εγγράφων, συσταδοποίηση και ταξινόμηση εγγράφων, συστήματα απάντησης ερωτήσεων για την διαγλωσσική ανάκτηση.

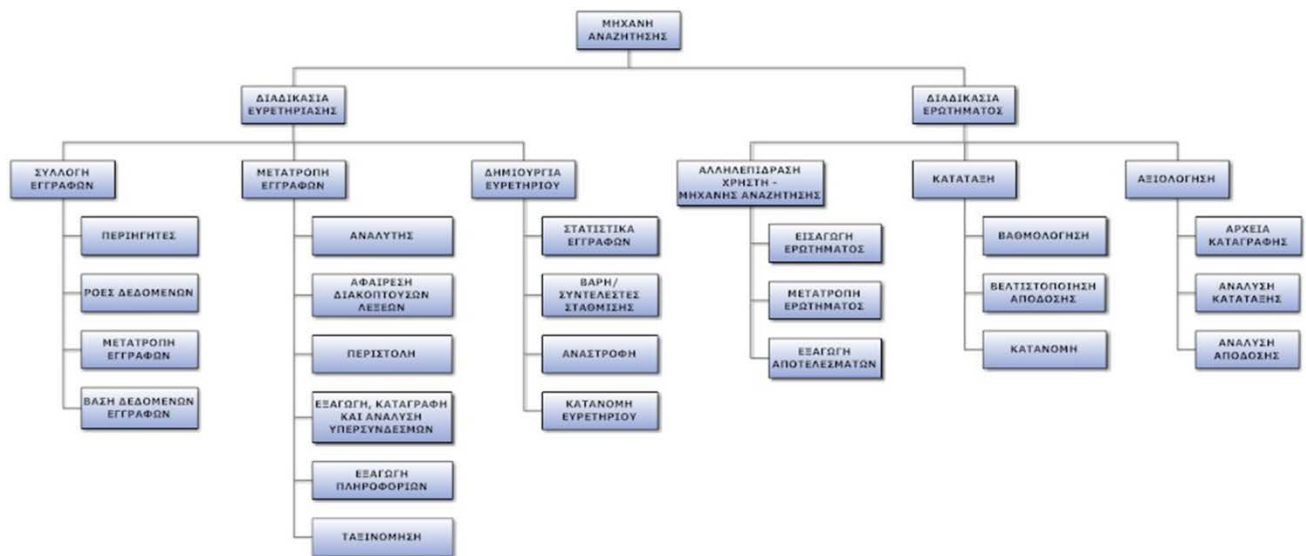
Σε αυτό το σημείο κατόπιν ολοκλήρωσης της παρουσίασης της αρχιτεκτονικής ενός απλοποιημένου ΣΑΠ, στην ενότητα που ακολουθεί θα αναλυθεί η αρχιτεκτονική των μηχανών αναζήτησης ακόμη εκτενέστερα. Έτσι θα γίνει περισσότερο κατανοητός όχι μόνον ο τρόπος λειτουργίας τους αλλά και ο ρόλος των μοντέλων ΑΠ και ειδικότερα του αλγόριθμου κατάταξης.

1.2.8 Αρχιτεκτονική μηχανής αναζήτησης

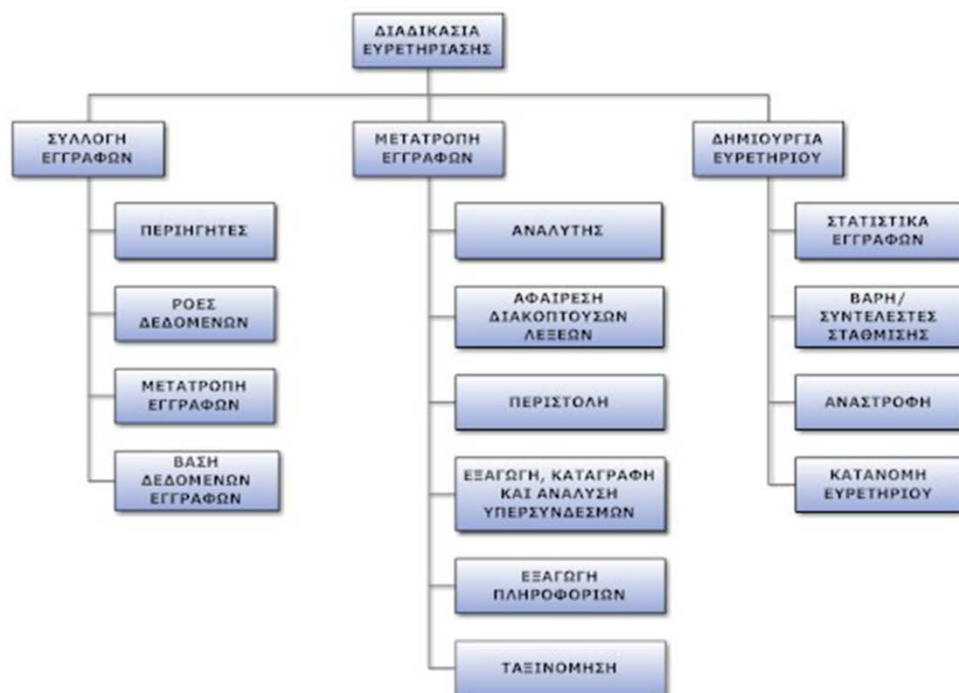
Η αρχιτεκτονική των μηχανών αναζήτησης αναλύεται σε βάθος από τους CroftB. et al. (2009), όπου παρουσιάζονται με ιδιαίτερη λεπτομέρεια όλα τα λειτουργικά τμήματα που αλληλεπιδρούν και κατηγοριοποιούνται σε δύο κεντρικές λειτουργίες: **τη διαδικασία ευρετηρίασης (indexing process) και τη διαδικασία ερωτήματος (query process)**. Σε αυτές έχει ήδη γίνει γενική αναφορά στην προηγούμενη ενότητα, εδώ θα παρουσιαστούν πολύ συγκεκριμένα. Σχηματικά η αρχιτεκτονική μιας μηχανής αναζήτησης συνοψίζεται στο Σχήμα 1.7.

1.2.9 Αρχιτεκτονική μηχανών αναζήτησης - ευρετηρίαση

Ξεκινώντας με τη **διαδικασία ευρετηρίασης** στη βιβλιογραφία (Croft et al. 2009) στη συνέχεια ακολουθούν τρεις επιμέρους βασικές διαδικασίες: η δημιουργία **συλλογής εγγράφων** (text acquisition), η **τροποποίηση εγγράφων** (text transformation) και η **κατασκευή ευρετηρίου** (index creation), όπως φαίνεται σχηματικά στο Σχήμα 1.8:



Σχήμα 1.7. Αρχιτεκτονική μηχανής αναζήτησης



Σχήμα 1.8. Διαδικασία ευρετηρίασης

Συλλογή εγγράφων: κατά τη διαδικασία αυτή αναγνωρίζονται και συλλέγονται τα έγγραφα της συλλογής (η οποία μπορεί να είναι ήδη αποθηκευμένη ή να δημιουργείται μέσω webcrawling) και πραγματοποιούνται διαδικασίες ώστε τα έγγραφα αυτά να μπορούν να είναι προσβάσιμα κατά την αναζήτηση. Ειδικότερα περιλαμβάνει:

1. Τους **περιηγητές (crawlers)**, λογισμικά τα οποία αναγνωρίζουν και επιστρέφουν έγγραφα στη μηχανή αναζήτησης, όπως π.χ. η κατηγορία generalwebcrawler. Το πρόγραμμα αυτό ακολουθεί τους υπερσυνδέσμους των ιστοσελίδων, εντοπίζει ιστοσελίδες και τις αποθηκεύει.

2. Τις **διαδρομές που ακολουθούν τα δεδομένα (feeds)**, μέσω των οποίων είναι δυνατή η πρόσβαση σε πραγματικό χρόνο ακολουθίας εγγράφων μέσω εγγραφής των χρηστών σε ιστοτόπους. Παράδειγμα αποτελεί το πρότυπο RSS.
3. Τη **τροποποίηση(conversion)** των εγγράφων, σε κατάλληλο μορφότυπο αρχείου κειμένου, καθώς συνήθως ανακτώνται σε αρχεία xml, html, pdf κ.α.
4. Τη **βάση δεδομένων εγγράφων (documentdatastore)**, η οποία διαχειρίζεται ένα σημαντικό όγκο των εγγράφων και των δεδομένων τους όπως είναι τα μεταδεδομένα ή άλλες πληροφορίες που πηγάζουν από το έγγραφο (π.χ. υπερσύνδεσμοι).

Σημαντικό σε αυτό το σημείο είναι να καθορισθούν οι διάφορες σειρές χαρακτήρων των εγγράφων και οι διαφορές τους. Μια σειρά χαρακτήρων σύμφωνα με τους Manning, Raghavan και Schutze (2008) μπορεί να κατηγοριοποιηθεί σε σύμβολα (tokens), τύπους (types) ή όρους (terms). Σύμβολο ονομάζεται μια ακολουθία χαρακτήρων που παρουσιάζεται σε συγκεκριμένο έγγραφο και αποτελεί μια σημασιολογική μονάδα. Τύπος ονομάζεται μια κλάση που αποτελείται από σύμβολα με ίδια ακολουθία χαρακτήρων. Τέλος, όρος ονομάζεται ένας τύπος ο οποίος εντάσσεται στο λεξικό του ΣΑΠ. Οι τύποι και οι όροι είναι συνήθως κανονικοποιημένοι, κάτι που θα ήδη έχει παρουσιαστεί σε προηγούμενη ενότητα. Θα παρουσιαστούν οι διαφορές των παραπάνω σε ένα σύντομο παράδειγμα:

- Έστω ότι υπάρχει μια πρόταση με 7 λέξεις από τις οποίες οι 2 θεωρούνται διακόπτουσες και πως οι 2 διακόπτουσες αυτές λέξεις είναι η ίδια λέξη “of” (δηλαδή η λέξη “of” επαναλαμβάνεται δύο φορές). Τότε:
 - ο Οι 7 λέξεις θεωρούνται σύμβολα, δηλαδή όσες είναι οι μονάδες της πρότασης τόσα και τα σύμβολα,
 - ο Οι διαφορετικοί τύποι που εμφανίζονται είναι 6, καθώς η μια λέξη “of” εμφανίζεται 2 φορές
 - ο Οι όροι που θα ενταχθούν στο ευρετήριο τελικά είναι 5, καθώς οι διακόπτουσες λέξεις αφαιρούνται.

Μετατροπή εγγράφων: συνεχίζοντας την παρουσίαση σύμφωνα με τη βιβλιογραφία (Croft et al. 2009), η διαδικασία αυτή περιέχει την τροποποίηση των **εγγράφων σε όρους ευρετηρίου (indexterms)**, έχοντας ως αποτέλεσμα να δημιουργείται ένα λεξιλόγιο όρων (indexvocabulary). Το οποίο ειδικότερα περιλαμβάνει:

1. Τον **αναλυτή (parser)**, ο οποίος **διαίρει την σειρά χαρακτήρων του εγγράφου σε σύμβολα (tokenization)**. Όταν υποβάλλεται το ερώτημα του χρήστη τότε αυτό διαιρείται σε σύμβολα ώστε να συγκριθεί με τα έγγραφα της συλλογής. Στη διαδικασία της διαίρεσης σε σύμβολα σημαντικά ζητήματα που πρέπει να αντιμετωπιστούν είναι οι ειδικοί χαρακτήρες, το ενωτικό και η απόστροφος. Ακόμη ο αναλυτής ασχολείται και με την λήψη της δομής τους εγγράφου από τη γλώσσα σήμανσης στην οποία έχει συνταχθεί (HTML ή XML), χρησιμοποιώντας το συντακτικό της γλώσσας και έτσι αντιλαμβάνεται τα μέρη από τα οποία αποτελείται ένα έγγραφο (τίτλος, επικεφαλίδα κλπ.).
2. Τις **διακόπτουσες λέξεις (stopwords)**. Βασική εργασία είναι η **αφαίρεση** των λέξεων αυτών από την ακολουθία συμβόλων του εγγράφου, κατά συνέπεια και από το λεξιλόγιο όρων ώστε να μειωθεί το μέγεθος του ευρετηρίου. Οι διακόπτουσες λέξεις έχουν ήδη οριστεί σε προηγούμενη ενότητα. Μάλιστα όλες οι διακόπτουσες λέξεις συγκεντρώνονται σε μια κοινή λίστα διακοπτουσών λέξεων (stopwordlist).
3. Την **περιστολή (stemming)** είναι η διαδικασία κατά την οποία ένας αλγόριθμος συγκεντρώνει λέξεις με κοινό τμήμα, τις ομαδοποιεί και τέλος τις αντικαθιστά με μια προκαθορισμένη λέξη προκειμένου να επιτύχει ομοιογένεια και κατά συνέπεια να βελτιωθεί η διαδικασία κατάταξης αποτελεσμάτων.
4. Την **εξαγωγή, καταγραφή και ανάλυση των υπερσυνδέσμων και του κειμένου αγκύρωσης (anchortext/** αποτελεί το κείμενο, το οποίο κάνει κλικ πάνω του ο χρήστης και μεταφέρεται στον υπερσύνδεσμο) των ιστοσελίδων που είναι αποθηκευμένες στη βάση δεδομένων εγγράφων, προκειμένου να χρησιμοποιηθούν από αλγόριθμους ανάλυσης υπερσυνδέσμων (όπως PageRank) με σκοπό να παρέχουν στην μηχανή αναζήτησης είτε μια κατάταξη του πόσο δημοφιλείς είναι οι ιστοσελίδες αυτές στον ιστό είτε για το **authority** της κάθε ιστοσελίδας, το οποίο αφορά το κύρος των υπερσυνδέσμων που φέρει αυτή.
5. Την **Εξαγωγή Πληροφοριών**, ώστε να αναγνωριστούν όροι ευρετηρίου οι οποίοι έχουν πιο πολύπλοκη μορφή ως λέξεις. Ορισμένα παραδείγματα αποτελούν οι λέξεις που βρίσκονται σε πλάγια ή έντονη γραφή, λέξεις ως επικεφαλίδες κ.α. Ακόμη όσον αφορά την εξαγωγή πληροφοριών σχετικών

με τη σύνταξη του κειμένου χρησιμοποιούνται προγράμματα επισημείωσης μερών του λόγου (part-of-speechtaggers).

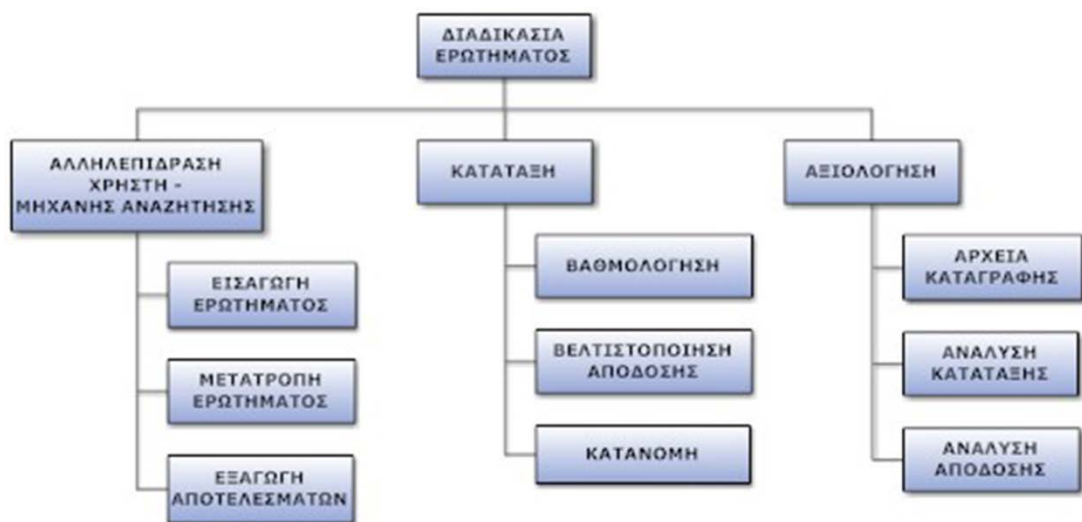
6. Την **ανάθεση εγγράφων σε ομάδες**. Αυτό μπορεί να γίνει είτε μέσω **τεχνικών ταξινόμησης με προκαθορισμένες κλάσεις** όπου ο ταξινομητής (classifier) πραγματοποιεί αναγνώριση σύμφωνα με τα μεταδεδομένα εγγράφων ή μερών τους και έτσι τα αναθέτει στις κατάλληλες προκαθορισμένες ετικέτες κλάσεων (θεματικές ή κατάταξη εγγράφων ως κακόβουλα (spam) ή αναγνώριση αντικειμένων που δεν αποτελούν μέρος του εγγράφου όπως διαφημίσεις), είτε μέσω **τεχνικών συσταδοποίησης** (clustering) όπου τα έγγραφα ομαδοποιούνται χωρίς να υπάρχουν προκαθορισμένες κλάσεις. Η ομαδοποίηση γίνεται με τέτοιο τρόπο ώστε να χρησιμεύει για την κατάταξη εγγράφων ή για την αλληλεπίδραση με το χρήστη.

Όπως αναφέρουν οι Baeza-Yates και Ribeiro-Neto (2011), η δημιουργία του ευρετηρίου αφορά κυρίως τη μείωση του χρόνου απόκρισης κατά τη διάρκεια της αναζήτησης. Όπως αναφέρουν η καλύτερη επιλογή ευρετηρίασης είναι τα ανεστραμμένα ευρετήρια αλλά υπάρχουν και άλλες τεχνικές ευρετηρίασης (πίνακες επιθεμάτων, αρχεία υπογραφών). Η διαδικασία ευρετηρίασης ολοκληρώνεται σύμφωνα με τη βιβλιογραφία (Croft et al. 2009) με το τελευταίο στάδιο της **δημιουργίας ευρετηρίου**, το οποίο αποτελείται από τα παρακάτω βήματα:

1. **Συγκέντρωση στατιστικών πληροφοριών σχετικές με τις λέξεις, τα έγγραφα και τα χαρακτηριστικά τους**. Παραδείγματα τέτοιων στατιστικών στοιχείων αποτελούν ο αριθμός εμφάνισης όρων ευρετηρίου σε κάθε έγγραφο της συλλογής ή η θέση κάθε όρου του ευρετηρίου στο έγγραφο, τα οποία χρησιμοποιούνται στην κατάταξη και στον υπολογισμό βαθμολογίας (score) εγγράφων, στα πλαίσια μιας αναζήτησης. Το ποιες στατιστικές πληροφορίες θα συλλεχθούν καθορίζεται από το μοντέλο ΑΠ που χρησιμοποιεί η μηχανή αναζήτησης και ειδικότερα από τον αλγόριθμο κατάταξης που χρησιμοποιεί. Τέλος οι στατιστικές πληροφορίες αποθηκεύονται σε πίνακες αναζήτησης (look-up tables).
2. **Προσδιορισμός συντελεστών στάθμισης/βάρη (weighting)** όρων ευρετηρίου ώστε να διακρίνονται οι όροι μεταξύ τους με βάση τη διαφορετική σημασία που έχουν για κάθε έγγραφο, η οποία μεταβάλλεται από έγγραφο σε έγγραφο και βοηθά στον υπολογισμό βαθμολογίας εγγράφων στην κατάταξη. Οι συντελεστές στάθμισης/βάρη όρων αποθηκεύονται επίσης σε πίνακες αναζήτησης και η μορφή τους καθορίζεται από το μοντέλο ΑΠ. Ένα από τα πιο συνηθισμένα μοντέλα προσδιορισμού συντελεστών στάθμισης είναι το **termfrequency - invertedocumentfrequency (tf-idf)**, το οποίο ήδη έχει επεξηγηθεί σε προηγούμενη ενότητα (βλέπε ενότητα 2.2).
3. **Αναστροφή (inversion)**, η οποία ουσιαστικά αφορά στην διαδικασία αντιστοίχισης όρων σε έγγραφα, ώστε όταν ζητείται ένας όρος κατά την αναζήτηση, να μπορεί να οδηγήσει στα έγγραφα της συλλογής που τον περιέχουν. Τα λεγόμενα **αντεστραμμένα ευρετήρια** βελτιώνουν την απόδοση της μηχανής αναζήτησης καθώς η αναζήτηση πραγματοποιείται πιο γρήγορα. Έτσι από την αποδόμηση και ανάλυση των εγγράφων της συλλογής με όλες τις παραπάνω διαδικασίες, πραγματοποιείται η αναστροφή, η οποία αποτελεί είδος συντομεύσεων από όρους προς έγγραφα της συλλογής.
4. **Κατανομή ευρετηρίων (indexdistribution)**, πρόκειται για τη διανομή του ευρετηρίου σε πολλαπλούς υπολογιστές και σε πολλαπλές ιστοσελίδες στο διαδίκτυο, ώστε τόσο η ευρετηρίαση όσο και η διαδικασία υποβολής ερωτήματος του χρήστη να μπορούν να πραγματοποιηθούν παράλληλα, για τη συλλογή.

1.2.10 Αρχιτεκτονική μηχανών αναζήτησης – διαδικασία ερωτήματος

Η δεύτερη μεγάλη διαδικασία είναι αυτή της διαδικασίας ερωτήματος. Σύμφωνα με τη βιβλιογραφία (Croft et al. 2009) περιγράφεται ως μια διαδικασία παραγωγής λίστας αποτελεσμάτων τα οποία θεωρούνται συναφή ως προς το ερώτημα του χρήστη (βλέπε Σχήμα 1.9). Η διαδικασία αυτή χωρίζεται ξανά σε τρεις επιμέρους βασικές διαδικασίες: την αλληλεπίδραση χρήστη με την μηχανή αναζήτησης (user interaction), την κατάταξη (ranking) και την αξιολόγηση (evaluation).



Σχήμα 1.9. Διαδικασία ερωτήματος

Αναλύονται οι τρεις προαναφερόμενες υπό-διαδικασίες σύμφωνα με την βιβλιογραφία (Croft et al. 2009) ως εξής:

Αλληλεπίδραση χρήστη με τη μηχανή αναζήτησης: Η διαδικασία αυτή αφορά στο συστατικό της μηχανής που (α) παίρνει το ερώτημα που υποβάλλει ο χρήστης, το οποίο μετατρέπεται σε όρους ευρετηρίου, (β) οργανώνει τη λίστα κατάταξης εγγράφων και εμφανίζει τα αποτελέσματα στο χρήστη. Πιο συγκεκριμένα:

1. **Εισαγωγή δεδομένων ερωτήματος (queryinput):** αποτελείται από την διεπαφή για τον χρήστη και έναν αναλυτή γλώσσας ερωτημάτων (querylanguageparser). Οι γλώσσες ερωτημάτων αποτελούνται από τους τελεστές (operators), δηλαδή σύνολο εντολών οι οποίες χρησιμοποιούνται για να υποδείξουν κείμενο το οποίο χρήζει ιδιαίτερης μεταχείρισης, ώστε να αποσαφηνιστεί το τι ψάχνει ο χρήστης μέσω του ερωτήματος που έχει εισάγει. Όπως ήδη έχει αναφερθεί σε προηγούμενες ενότητες το ερώτημα του χρήστη αποτελείται από κάποιες λέξεις κλειδιά σε φυσική γλώσσα, οι οποίες πρέπει να γίνουν κατανοητές από τη μηχανή.
2. **Μετατροπή ερωτήματος (querytransformation):** πρόκειται για ένα σύνολο τεχνικών που εφαρμόζονται στο ερώτημα του χρήστη, ώστε να βελτιωθεί πριν και μετά τη διαδικασία παραγωγής αποτελεσμάτων για το χρήστη σε μορφή καταταγμένων εγγράφων. Πολλές από αυτές τις τεχνικές είναι ίδιες με αυτές που εφαρμόζονται κατά τη μετατροπή εγγράφων στην ευρετηρίαση (διαίρεση σε σύμβολα, διακόπτουσες λέξεις κλπ.). Άλλες πρόσθετες αφορούν στον ορθογραφικό έλεγχο (spellchecking) και σε εναλλακτικές ερωτήματος (querysuggestion), προκειμένου το ερώτημα να γίνει πιο συγκεκριμένο.
3. **Εξαγωγή αποτελεσμάτων (resultsoutput):** ουσιαστικά αφορά στην παρουσίαση της λίστας καταταγμένων εγγράφων, όπου περιλαμβάνονται ποικίλες εργασίες όπως:
 - ο Παραγωγή περιλήψεων ανακτηθέντων εγγράφων γνωστά ως ψήγματα αποτελεσμάτων (resultsnippets).
 - ο Τονισμό σημαντικών λέξεων μέσα στο έγγραφο.
 - ο Εμφάνιση με ομαδικό τρόπο των αποτελεσμάτων ώστε να ταυτιστούν οι επιμέρους κλάσεις (π.χ. θεματικές κλάσεις).
 - ο Αναζήτηση κατάλληλων διαφημιστικών προτάσεων για την προβολή αποτελεσμάτων.
 - ο Μετάφραση αποτελεσμάτων.

Κατάταξη εγγράφων: αποτελεί τη δεύτερη υπό-διαδικασία της αναζήτησης και σύμφωνα με τη βιβλιογραφία (Croft et al. 2009) αφορά μόνο την κατάταξη των εγγράφων. Καθώς το ερώτημα του χρήστη έχει μετατραπεί ήδη σε όρους ευρετηρίου, βαθμολογούνται τα έγγραφα με βάση έναν αλγόριθμο κατάταξης (ανάλογα το μοντέλο ΑΠ) και παράγεται μια λίστα με τα καταταγμένα έγγραφα. Πιο συγκεκριμένα σύμφωνα με τη βιβλιογραφία (Croft et al. 2009) τα συστατικά κατάταξης είναι τα ακόλουθα:

1. **Βαθμολόγηση (scoring)**, η διαδικασία κατά την οποία υπολογίζονται βαθμολογίες για τα έγγραφα χρησιμοποιώντας τον αλγόριθμο κατάταξης (rankingalgorithm).
2. **Βελτιστοποίηση απόδοσης(performanceoptimization)**, αφορά στην σχεδίαση αλγορίθμων κατάταξης και ευρετηρίων ώστε ο χρόνος απόκρισης να ελαχιστοποιηθεί.
3. **Κατανομή(distribution)** της λίστας κατάταξης.

Αξιολόγηση: στο στάδιο αυτό πραγματοποιείται μέτρηση και έλεγχος της αποτελεσματικότητας και της απόδοσης της μηχανής. Σύμφωνα με τη βιβλιογραφία (Croft et al. 2009) η αξιολόγηση αφορά στα:

1. **Αρχεία καταγραφής (logfiles ή logs)χρηστών:** πρόκειται για δεδομένα αλληλεπίδρασης μεταξύ των χρηστών με τις μηχανές αναζήτησης. Χρησιμοποιούνται σε μεγάλο αριθμού εργασιών (π.χ. ορθογραφικός έλεγχος). Ιδιαίτερα χρήσιμα για την αξιολόγηση αλλά και τους αλγόριθμους κατάταξης θεωρούνται τα αρχεία καταγραφής πλοήγησης (clickthrough logs), τα οποία περιλαμβάνουν τις επιλογές του χρήστη για περιήγηση για συγκεκριμένη λίστα κατάταξης.
2. **Ανάλυση κατάταξης (rankinganalysis):** συνίσταται στον υπολογισμό της αποτελεσματικότητας του αλγόριθμου κατάταξης. Επίσης περιλαμβάνει και τη σύγκρισή του με άλλους αλγόριθμους.
3. **Ανάλυση απόδοσης (performanceanalysis): υπολογισμός της απόδοσης της μηχανής αναζήτησης και τρόποι βελτίωσης.**

Βιβλιογραφία

- Ahlsén, E. (2006). Introduction to neurolinguistics. John Benjamins Publishing. Retrieved from https://www.google.com/books?hl=en&lr=&id=ziC_Dbl4KnIC&oi=fnd&pg=PR1&dq=Introduction+to+Neurolinguistics.+&ots=kZo4-xMj9w&sig=fbprd4XowkkCFwmDourBm7r9xjk
- Akobeng, A. K. (2007). Understanding diagnostic tests 1: sensitivity, specificity and predictive values. *Acta Paediatrica*, 96(3), 338–341.
- Altmann, G. (1980). Prolegomena to Menzerath's law. *Glottometrica*, 2, 1–10.
- Altmann, G. (2002). Zipfian linguistics. *Glottometrics*, 3, 19–26.
- Baeza-Yates, R., Ribeiro-Neto, B., & others. (1999). *Modern information retrieval* (Vol. 463). ACM press New York. Retrieved from ftp://mail.im.tku.edu.tw/seke/slide/baeza-yates/chap10_user_interfaces_and_visualization-modern_ir.pdf
- Baixeries, J., Hernandez-Fernandez, A., Forns, N., & Ferrer-i-Cancho, R. (2013). The parameters of the Menzerath-Altmann Law in genomes. *Journal of Quantitative Linguistics*, 20(2), 94–104.
- Bendersky, M., & Croft, W. B. (2012). Modeling higher-order term dependencies in information retrieval using query hypergraphs. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval* (pp. 941–950). ACM. Retrieved from <http://dl.acm.org/citation.cfm?id=2348408>
- Blair, D. C., & Kimbrough, S. O. (2002). Exemplary documents: a foundation for information retrieval design. *Information Processing & Management*, 38(3), 363–379.
- Bokos, M. P. G., Papavaslopoulos, N. K. S., & Avlonitis, M. (2007). Specific selection of FFT amplitudes from audio sports and news broadcasting for classification purposes. Retrieved from <http://www.emis.ams.org/journals/JGAA/accepted/2007/Poulos+2007.11.1.pdf>
- Bolshakov, I. A., & Gelbukh, A. (2004). Computational Linguistics Models, Resources, Applications.
- Brants, T. (2003). Natural Language Processing in Information Retrieval. In *CLIN*. Retrieved from <http://www.clips.ua.ac.be/clin2003/proc/03Brants.pdf>
- Buk, S., & Rovenchak, A. (2008). Menzerath–Altmann Law for Syntactic Structures in Ukrainian. *Glottology*, 1(1), 10–17.
- Büttcher, S., Clarke, C. L., & Cormack, G. V. (2010). *Information retrieval: Implementing and evaluating search engines*. Mit Press. Retrieved from <https://www.google.com/books>
- Carstens, W. (1999). Text-linguistics: Relevant linguistics. *School of Languages and Arts, Potchefstroom University for CHE*. Retrieved from <http://www.pala.ac.uk/uploads/2/5/1/0/25105678/carstens.pdf>
- Ceri, S., Bozzon, A., Brambilla, M., Della Valle, E., Fraternali, P., & Quarteroni, S. (2013). *Web Information Retrieval*. Springer Science & Business Media. Retrieved from <https://www.google.com/books>
- Cerulo, L., & Canfora, G. (2004). A taxonomy of information retrieval models and tools. *CIT. Journal of Computing and Information Technology*, 12(3), 175–194.
- Crestani, F., & Wu, S. (2006). Testing the cluster hypothesis in distributed information retrieval. *Information Processing & Management*, 42(5), 1137–1150.
- Croft, W. B., Metzler, D., & Strohman, T. (2010). *Search engines: Information retrieval in practice*. Addison-Wesley Reading. Retrieved from http://library.mpib-berlin.mpg.de/toc/z2009_2465.pdf
- Dale, E., & Chall, J. S. (1949). The concept of readability. *Elementary English*, 26(1), 19–26.
- Daniel, J., & James, H. M. (2000). Speech and Language processing: An introduction to natural language processing. *Computational Linguistics and Speech Recognition*. Prentice Hall, NJ, USA.

- De Beaugrande, R. (1981). Introduction to text linguistics. Retrieved from <http://www.citeulike.org/group/236/article/202549>
- Diggle, P. J., & Chetwynd, A. G. (2011). *Statistics and scientific method: an introduction for students and researchers*. Oxford University Press.
- DuBay, W. H. (2004). The Principles of Readability. *Online Submission*. Retrieved from <http://eric.ed.gov/?id=ED490073>
- Dubin, D. (2004). The most influential paper Gerard Salton never wrote. Retrieved from <https://www.ideals.illinois.edu/handle/2142/1697>
- Eroglu, S. (2013). Menzerath–Altmann law for distinct word distribution analysis in a large text. *Physica A: Statistical Mechanics and Its Applications*, 392(12), 2775–2780.
- Espunya i Prat, A. (1994). Computational linguistics: a brief introduction. In *Links & Letters* (pp. 009–23). Retrieved from <http://ddd.uab.cat/pub/lal/11337397n1/11337397n1p9.pdf>
- Everitt, B. S., & Skrondal, A. (2002). The Cambridge dictionary of statistics. *Cambridge: Cambridge*. Retrieved from <http://www.maa.org/publications/maa-reviews/the-cambridge-dictionary-of-statistics>
- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861–874.
- Fisher, R. A. (1992). Statistical methods for research workers. In *Breakthroughs in Statistics* (pp. 66–70). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-1-4612-4380-9_6
- Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32(200), 675–701.
- Galvez, C., de Moya-Anegón, F., & Solana, V. H. (2005). Term conflation methods in information retrieval: non-linguistic and linguistic approaches. *Journal of Documentation*, 61(4), 520–547.
- Gries, S. T. (2013). *Statistics for linguistics with R: a practical introduction*. Walter de Gruyter. Retrieved from <https://www.google.com/books>
- Gries, S. T. (n.d.). Quantitative methods in linguistics. Retrieved from http://www.linguistics.ucsb.edu/faculty/stgries/research/ToApp_STG_QuantMethInLing_IESBS2.pdf
- Grossman, D. A., & Frieder, O. (2012). *Information retrieval: Algorithms and heuristics* (Vol. 15). Springer Science & Business Media. Retrieved from <https://www.google.com/books>
- Harispe, S., Ranwez, S., Janaqi, S., & Montmain, J. (2013). Semantic Measures for the Comparison of Units of Language, Concepts or Instances from Text and Knowledge Base Analysis. *arXiv Preprint arXiv:1310.1285*. Retrieved from <http://arxiv.org/abs/1310.1285>
- Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *Science*, 298 (5598), 1569–1579.
- Hayes, B., Curtiss, S., Szabolcsi, A., Stowell, T., Stabler, E., Sportiche, D., ... others. (2001). *Linguistics: An introduction to linguistic theory*. Wiley-Blackwell.
- Hempel, C. G. (1952). Fundamentals of concept formation in empirical science.(Int. Encyclopedia Unified Science, Vol. II. No. 7). Retrieved from <http://psycnet.apa.org/psycinfo/1953-03119-000>
- Hoey, M. (2004). Lexical priming and the properties of text. *Corpora and Discourse*, 385–412.
- Hogan, P. C. (2011). *The Cambridge encyclopedia of the language sciences*. Cambridge University Press Cambridge, NY. Retrieved from http://www.langtoninfo.com/web_content
- Hřebíček, L. (2002). Zipf's Law and Text. *Glottometrics*, 27.
- Ingwersen, P. E. R. (1992). *Information Retrieval Interaction:(click on title for download access via 'Documents')*. Taylor Graham. Retrieved from <http://pure.iva.dk/ws>
- Johnson, K. (2011). *Quantitative methods in linguistics*. John Wiley & Sons. Retrieved from <https://www.google.com/books>

- Joho, H., & Jose, J. M. (2006). A comparative study of the effectiveness of search result presentation on the web. In *Advances in Information Retrieval* (pp. 302–313). Springer. Retrieved from http://link.springer.com/chapter/10.1007/11735106_27
- JOSEPH SORELL, C. (2012). Zipf's law and vocabulary. *The Encyclopedia of Applied Linguistics*. Retrieved from <http://onlinelibrary.wiley.com/doi/10.1002/9781405198431.wbeal1302/full>
- Kamps, J. (2009). Presenting structured text retrieval results. *Encyclopedia of Database Systems*, 2130–2134.
- Karaman, B. I. (2003). *Polysemy in Natural Language: Case Studies on the Structural Description of Polysemous Lexemes in English, German and Turkish*. University of Surrey. Retrieved from <http://eprints.surrey.ac.uk/id/eprint/2816>
- Kennedy, C. (2009). Ambiguity and vagueness: An overview. *Semantics: An International Handbook of Natural Language Meaning*.
- Kovacs, E. (2011). Polysemy in Traditional vs. Cognitive Linguistics. *EGER JOURNAL OF ENGLISH STUDIES*, 3.
- Kowalski, G. (2010). *Information retrieval architecture and algorithms*. Springer Science & Business Media. Retrieved from <https://www.google.com/books>
- Kracht, M. (n.d.). Introduction to Linguistics. Department of Linguistics, UCLA.–211 p.[Электронный ресурс].–URL: [Http://www. Linguistics. Ucla. edu/people/Kracht/courses/ling20-fall07/ling-Intro. Pdf](http://www.Linguistics.Ucla.edu/people/Kracht/courses/ling20-fall07/ling-Intro.Pdf) (дата обращения: 08.10. 13). Retrieved from http://www.researchgate.net/profile/Marcus_Kracht/publication/229026759_Introduction_to_Linguistics/links/5416d8200cf2bb7347db7772.pdf
- Ku\lacka, A., & Mačutek, J. (2007). A discrete formula for the Menzerath-Altmann law*. *Journal of Quantitative Linguistics*, 14(1), 23–32.
- Kuropka, D. (2004). Modelle zur Repräsentation natürlichsprachlicher Dokumente-Information-Filtering und-Retrieval mit relationalen Datenbanken. *Advances in Information Systems and Management Science*, 10.
- Lalkhen, A. G., & McCluskey, A. (2008). Clinical tests: sensitivity and specificity. *Continuing Education in Anaesthesia, Critical Care & Pain*, 8(6), 221–223.
- Lan, M., Tan, C. L., Su, J., & Lu, Y. (2009). Supervised and traditional term weighting methods for automatic text categorization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(4), 721–735.
- Law, J. (2015). *A dictionary of law*. Oxford University Press. Retrieved from [https://www.google.com/books?hl=en&lr=&id=3CNDBgAAQBAJ&oi=fnd&pg=PP1&dq=Power+law:+definitions+of+power+law+in+Oxford+Dictionary+\(British+and+World+English\)+\(2015\).+&ots=FscwCpA2ox&sig=awzm1O1AGkWXX6fB8FHs16Cj2yE](https://www.google.com/books?hl=en&lr=&id=3CNDBgAAQBAJ&oi=fnd&pg=PP1&dq=Power+law:+definitions+of+power+law+in+Oxford+Dictionary+(British+and+World+English)+(2015).+&ots=FscwCpA2ox&sig=awzm1O1AGkWXX6fB8FHs16Cj2yE)
- Legendre, P. (2005). Species associations: the Kendall coefficient of concordance revisited. *Journal of Agricultural, Biological, and Environmental Statistics*, 10(2), 226–245.
- Legendre, P. (2010). Coefficient of concordance. *Encyclopedia of Research Design*, SAGE Publications. [Http://dx. Doi. org/10.4135/9781412961288, \(55\).](http://dx. Doi. org/10.4135/9781412961288, (55).)
- Lewis, D. D., & Jones, K. S. (1996). Natural language processing for information retrieval. *Communications of the ACM*, 39(1), 92–101.
- Manning, C. D., Raghavan, P., Schütze, H., & others. (2008). *Introduction to information retrieval* (Vol. 1). Cambridge university press Cambridge. Retrieved from http://www.langtoninfo.co.uk/web_content/9780521865715_frontmatter.pdf
- Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT press. Retrieved from <https://www.google.com/books>

- Metzler, D., & Croft, W. B. (2005). Modeling query term dependencies in information retrieval with markov random fields. Retrieved from <http://ciir.cs.umass.edu/pubfiles/ir-484.pdf>
- Mikk, J. (2005). Text comprehensibility. *Quantitative Linguistics: An International Handbook*, 909–921.
- Mikros, G., & Milička, J. (2014). *Distribution of the Menzerath's Law on the Syllable Level in Greek texts*. Forthcoming. Retrieved from http://www.researchgate.net/profile/Jii_Milicka/publication
- Miller, G. A. (2003). The cognitive revolution: a historical perspective. *Trends in Cognitive Sciences*, 7(3), 141–144.
- Moens, M.-F. (2006). *Information extraction: algorithms and prospects in a retrieval context* (Vol. 21). Springer Science & Business Media. Retrieved from <https://www.google.com/books>
- Nadel, L., & others. (2003). *Encyclopedia of cognitive science* (Vol. 3). nature publishing group London,, United Kingdom. Retrieved from http://web.stanford.edu/class/linguist156/Boroditsky_2003.pdf
- Neideen, T., & Brasel, K. (2007). Understanding statistical tests. *Journal of Surgical Education*, 64(2), 93–96.
- Onwuchekwa, E. O., & Jegede, O. R. (2011). Information Retrieval Methods in Libraries and Information Centers. *African Research Review*, 5(6), 108–120.
- Panik, M. J. (2012). *Statistical Inference: A Short Course*. John Wiley & Sons. Retrieved from <https://www.google.com/books>
- Poulimenou, S., Stamou, S., Papavlasopoulos, S., & Poulos, M. (n.d.-a). Keywords Extraction from Articles' Title for Ontological Purposes. Retrieved from <http://www.inase.org/library/2014/santorini/bypaper/MATH/MATH-16.pdf>
- Poulimenou, S., Stamou, S., Papavlasopoulos, S., & Poulos, M. (n.d.-b). Short Text Coherence Hypothesis. Retrieved from http://www.researchgate.net/profile/Marios_Poulos/publication/271167995_Short_Text_Coherence_Hypothesis/links/54bf5d570cf28ce68e6b4c71.pdf
- Powers, D. M. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. Retrieved from <http://dspace2.flinders.edu.au/xmlui/handle/2328/27165>
- Pulvermüller, F. (2002). *The neuroscience of language: on brain circuits of words and serial order*. Cambridge University Press. Retrieved from <https://www.google.com/books>
- Raghavan, V. V., & Wong, S. M. (1986). A critical analysis of vector space model for information retrieval. *Journal of the American Society for Information Science*, 37(5), 279–287.
- Richards, J. C., & Schmidt, R. W. (2013). *Longman dictionary of language teaching and applied linguistics*. Routledge. Retrieved from <https://www.google.com/books>
- Robertson, S. (2004). Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of Documentation*, 60(5), 503–520.
- Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5), 513–523.
- Saussure, F. de. (1966). *Course in General Linguistics*, eds. Charles Bally and Albert Sechehaye, trans., Wade Baskin. New York: McGraw-Hill Book Co.
- Singhal, A. (2001). Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, 24(4), 35–43.
- Spoerri, A. (1995). *InfoCrystal, a visual tool for information retrieval*. Massachusetts Institute of Technology. Retrieved from <http://dspace.mit.edu/handle/1721.1/36946>
- Squire, L., Berg, D., Bloom, F. E., du Lac, S., Ghosh, A., & Spitzer, N. C. (2012). *Fundamental neuroscience*. Academic Press. Retrieved from <https://www.google.com/books>
- Franco, L., Li, C., Dehghan, M., Kambuzi, A. K. Z., Xu, J., Gong, M. N., ... others. (χ.χ.). Theory and Practice in Language Studies. Ανακτήθηκε από <http://www.cultureready.org>

- Strzalkowski, T., Lin, F., Wang, J., & Perez-Carballo, J. (1999). Evaluating natural language processing techniques in information retrieval. Στο *Natural language information retrieval* (σσ 113–145). Springer. Ανακτήθηκε από http://link.springer.com/chapter/10.1007/978-94-017-2388-6_5
- Tanskanen, S.-K. (2006). *Collaborating towards coherence: Lexical cohesion in English discourse* (T. 146). John Benjamins Publishing. Ανακτήθηκε από https://www.google.com/books?hl=el&lr=&id=SYw6AAAAQBAJ&oi=fnd&pg=PR1&dq=Collaborating+towards+coherence&ots=FFJJp8UImd&sig=MhV-PUyg_2bEB1mzyFskz_geCtw
- Tesitlova, M. (1992). . Linguistic and Literary Studies in Eastern Europe 37. *Amsterdam: Quantitative linguistics*.
- Turney, P. D., Pantel, P., & others. (2010). From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37(1), 141–188.
- Van Rijsbergen, C. J. (1995). Information retrieval. Ανακτήθηκε από <http://158.69.150.236:1080/jspui/handle/961944/37613>
- Wong, H. B., & Lim, G. H. (2011). Measures of diagnostic accuracy: sensitivity, specificity, PPV and NPV. *Proceedings of Singapore Healthcare▪ Volume*, 20(4). Ανακτήθηκε από http://www.singhealthacademy.edu.sg/documents/publications/vol20no42011/11_stats_032-0911wonghb.pdf
- Woods, W. A., Bookman, L. A., Houston, A., Kuhns, R. J., Martin, P., & Green, S. (2000). Linguistic knowledge can improve information retrieval. Στο *Proceedings of the sixth conference on Applied natural language processing* (σσ 262–267). Association for Computational Linguistics. Ανακτήθηκε από <http://dl.acm.org/citation.cfm?id=974183>
- Wu, S., Bi, Y., & Zeng, X. (2010). Retrieval result presentation and evaluation. Στο *Knowledge Science, Engineering and Management* (σσ 125–136). Springer. Ανακτήθηκε από http://link.springer.com/chapter/10.1007/978-3-642-15280-1_14
- Wyllys, R. E. (1981). Empirical and theoretical bases of Zipf's law. *Library Trends*, 30(1), 53–64.
- Zamir, O., & Etzioni, O. (1999). Grouper: a dynamic clustering interface to Web search results. *Computer Networks*, 31(11), 1361–1374.
- Zar, J. H., & others. (1999). *Biostatistical analysis*. Pearson Education India.
- Zhu, W., Zeng, N., Wang, N., & others. (2010). Sensitivity, specificity, accuracy, associated confidence interval and ROC analysis with practical SAS® implementations. *NESUG proceedings: health care and life sciences, Baltimore, Maryland*, 1–9.

2. Συντακτική Επεξεργασία της Πληροφορίας

Περίληψη

Στο κεφάλαιο αυτό αναπτύσσεται η διαδικασία της συντακτικής επεξεργασίας της πληροφορίας. Συγκεκριμένα αναπτύσσεται η σημασιολογία ως επιστήμη από το χώρο της υπολογιστικής τεχνολογίας η οποία στη συνέχεια γίνεται το απαραίτητο ερευνητικό εργαλείο στην υπηρεσία του Διαδικτύου. Αναπτύσσονται επίσης τα παραδοσιακά μοντέλα ευρετηρίασης με τη χρήση της σημασιολογίας που είναι όμοια των οντολογιών ως προς τη φιλοσοφία. Επίσης γίνεται ειδική μνεία των μηχανών αναζήτησης και των ψηφιακών βιβλιοθηκών.

Προαπαιτούμενες γνώσεις

Μεταδεδομένα, Γλώσσες Σήμανσης

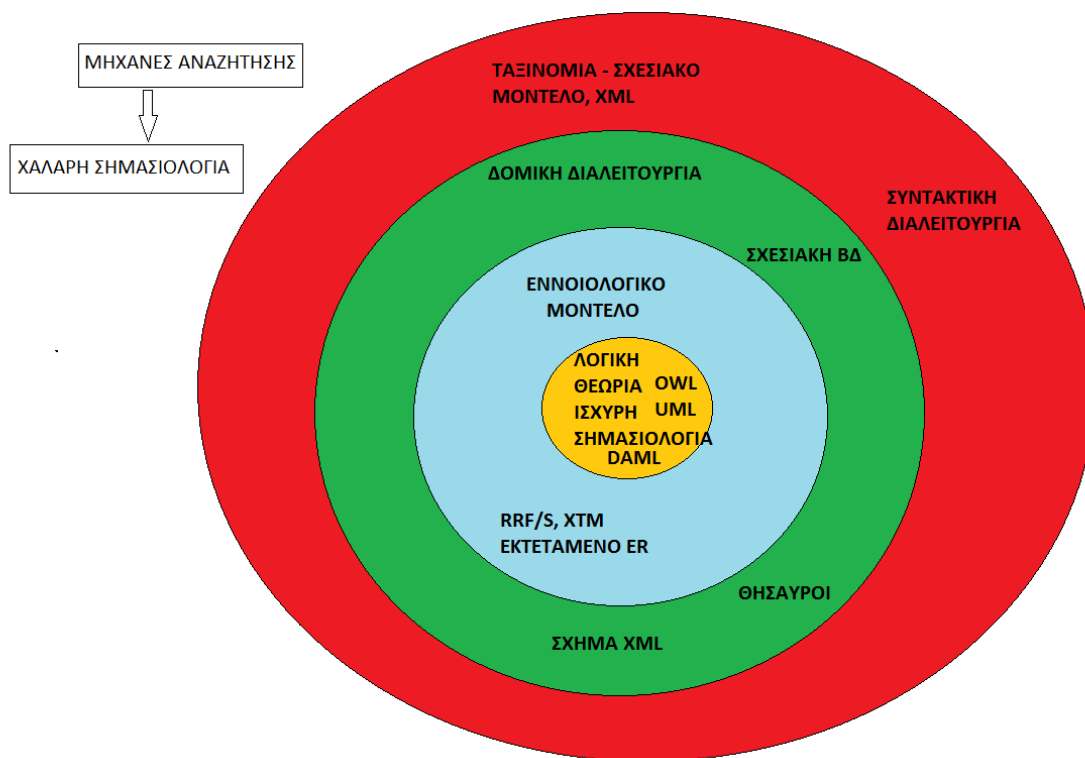
2.1 Σημασιολογικά Μοντέλα

Ενώ έχουν γίνει τεράστια άλματα στη σύνδεση εντελώς ανόμοιων πηγών δεδομένων με τη χρήση εκλεπτυσμένων λύσεων ενδιάμεσου λογισμικού, προχωρημένων πρωτοκόλλων ανταλλαγής δεδομένων και κοινών πρότυπων λεξιλογίου, υπάρχει ακόμη μια έλλειψη συσχετισμών στο επίπεδο της σημασιολογίας. Ένα από τα μεγαλύτερα κωλύματα στην πραγματική τιθάσευση της δύναμης της πληροφορίας και τον καλύτερο εξοπλισμό μας για να ανταποκριθούμε στις προκλήσεις που περιγράφηκαν πιο πάνω, είναι η έλλειψη κατανόησης του τι σημαίνει η πληροφορία, και πώς χρησιμοποιείται σε ένα σύστημα, έναντι κάποιου άλλου.

Επίσημα, η Σημασιολογία όπως αναφέρθηκε και στο προηγούμενο κεφάλαιο αποτελεί κλάδο της γλωσσολογίας που καταπίνεται με τη μελέτη της σημασίας, τις αλλαγές στη σημασία και τις αρχές που διέπουν τη σχέση ανάμεσα στις προτάσεις ή λέξεις και τις σημασίες τους (Bedford, 2004), ενώ η Σημασιολογία της Πληροφορίας είναι η σημασιολογική αναπαράσταση (σημασία) των συστημάτων, των δεδομένων, των εγγράφων ή των παραγόντων μας (Obrst, 2004).

Η Σημασιολογία της Πληροφορίας αναπαριστά τα οργανωτικά και πολιτισμικά πλαίσια ως ενσωματωμένα εντός των οργανωτικών αποστολών, ιεραρχιών, λεξιλογίων, ροής εργασίας, και προτύπων εργασίας. Η ίδια έννοια μπορεί να εκφράζεται με διαφορετικούς όρους, π.χ. σε ένα σύστημα μπορεί να εμφανίζεται «τιμή» και σε ένα άλλο «κόστος». Ο ίδιος όρος μπορεί να έχει διαφορετικές σημασίες, π.χ. ένας «Λοχαγός» του Στρατού είναι το αντίστοιχο ενός «Υποπλοίαρχου» στο Ναυτικό• ένας «Σμήναρχος» στην Αεροπορία είναι ο «Συνταγματάρχης» στο Στρατό. Παρόμοια, ένας «πληροφοριοδότης» σε ένα οργανισμό επιβολής του νόμου μπορεί να ονομάζεται «πηγή πληροφοριών» σε μια οργάνωση συλλογής πληροφοριών, και μπορεί να περιλαμβάνει και πηγές που δεν είναι άνθρωποι. Ακόμη και στον ίδιο οργανισμό, ο ίδιος όρος μπορεί να αναφέρεται σε εντελώς διαφορετικές έννοιες από διαφορετικές υπηρεσίες, π.χ. «Ασφάλεια» στο «Ασφάλεια Οικοδομής» έναντι «Ασφάλεια» στο «Ασφάλεια Δεδομένων». Ίσως πιο κρίσιμο είναι να δεχτούμε πως ο ίδιος όρος μπορεί να φέρει διαφορετικές σημασίες κατά την πάροδο του χρόνου, λόγω οργανωτικών αλλαγών.

Τα σημασιολογικά μοντέλα περιγράφουν τους σημασιολογικούς συσχετισμούς ανάμεσα σε διαφορετικούς όρους και έννοιες. Οι όροι και οι έννοιες είναι δύο διαφορετικά πράγματα. Οι όροι είναι λέξεις ή φράσεις. Οι έννοιες είναι η σημασία πίσω από τους όρους που αντιπροσωπεύουν τη σημασιολογία. Οι όροι, επομένως, δρουν ως ταμπέλες για τις έννοιες. Μπορεί να υπάρχει η έννοια του Ατόμου στην οποία όλοι οι όροι «άτομο», «κόσμος», «άνθρωπος» κτλ αναφέρονται. Υπάρχουν πολλές μορφές σημασιολογικών μοντέλων. Για να βοηθήσουμε τη συζήτησή μας για τα σημασιολογικά μοντέλα, χρησιμοποιούμε στην εικόνα «το φάσμα της οντολογίας» για να δείξουμε το «φάσμα της οντολογίας». Το φάσμα της οντολογίας (βλ. Σχήμα 2.1) δείχνει μια αλληλουχία μοντέλων, που κλιμακώνονται από κάτω αριστερά προς τα πάνω δεξιά, από μοντέλα με λιγότερο εκφραστική σημασιολογία («χαλαρή» σημασιολογία) σε μοντέλα με αυξητικά περισσότερο εκφραστική σημασιολογία («ισχυρή» σημασιολογία).



Σχήμα 2.1. Το φάσμα της οντολογίας

Εν γένει, η πρόοδος από κάτω αριστερά προς τα πάνω δεξιά αποτελεί επίσης αύξηση της ποσότητας της δομής για το μοντέλο, με τα εκφραστικότερα σημασιολογικά μοντέλα να έχουν την περισσότερη δομή. Συμπεριλαμβάνουμε παραπλεύρως του φάσματος μερικούς τύπους μοντέλων και γλωσσών που είτε ξέρετε ή έχετε ακούσει γι' αυτές: π.χ. το μοντέλο σχεσιακής βάσεως δεδομένων και η XML βρίσκονται κάτω αριστερά• μετακινούμενοι προς τα πάνω και δεξιά έχουμε το Σχήμα XML, τα μοντέλα Σχέσεως-Οντότητας, το XTM (το πρότυπο θεματικού χάρτη της XML - XML Topic Map standard), την RDF/S, την UML (Ενοποιημένη Γλώσσα Απεικόνισης - Unified Modeling Language), την OWL (γλώσσα οντολογίας του Ιστού - Web Ontology Language) και μέχρι τη λογική πρώτης τάξης (τον κατηγορηματικό λογισμό - Predicate Calculus) και πιο πάνω. Ουσιαστικά, το φάσμα επεκτείνεται πέραν εκείνων των μοντέλων που μας ενδιαφέρουν. Μια από τις απλούστερες μορφές του σημασιολογικού μοντέλου είναι η ταξινόμια. Οι ταξινομίες ορίζονται απλά ως οι δομές που χρησιμοποιούνται για να οργανωθεί η πληροφορία. Όταν ο κόσμος σκέφτεται τις ταξινομίες, συνήθως αντιλαμβάνεται ιεραρχικές δομές, όπως αυτές που φτιάχνουμε στις βιολογικές επιστήμες. Από την οπτική μιας επιστήμης της πληροφορίας, οι ταξινομίες μπορεί να εμπλέκουν ένα ή ένα συνδυασμό διάφορων τύπων δομών. Μπορούν να είναι επίπεδες δομές, ιεραρχίες, δικτυακές / πλεγματικές δομές ή πολύπλευρες ταξινομίες. Κάθε ένα από αυτά τα είδη δομών χρησιμεύει και ως ένα διαφορετικού είδους διαχείρισης της πληροφορίας και σκοπός πρόσβασης. Όλα τους είναι κρίσιμα για την υποστήριξη των περίπλοκων σημερινών λύσεων για την πληροφορία και αποτελούν ζωτικά συστατικά των περίπλοκων σημερινών συστημάτων της πληροφορίας (Bedford, 2004). Συμπεριλαμβάνουμε και το θησαυρό, το εννοιολογικό μοντέλο και τη λογική θεωρία στην εικόνα «το φάσμα της οντολογίας», επειδή αυτά δρουν ως «ενδιάμεσοι σταθμοί» της αυξανόμενης πολυπλοκότητας και του σημασιολογικού πλούτου καθώς ανεβαίνουμε το οντολογικό φάσμα. Ένας θησαυρός είναι πιο περίπλοκος από μια ταξινόμια, επειδή η σχέση γονέα-παιδιού που έχει χαρακτηρίζεται με συνέπεια ως «ευρύτερη από / στενότερη από», π.χ. ένας μητρικός κόμβος έχει μια ευρύτερη από σχέση με τους θυγατρικούς του κόμβους• ένας θυγατρικός κόμβος έχει μια στενότερη από σχέση με το μητρικό του κόμβο. Υπάρχουν και σχέσεις υπαγωγής, έτσι ένας γονέας υπάγει ένα παιδί. Ωστόσο, σε ένα θησαυρό, οι κόμβοι δεν είναι απλά ταξινομήσεις όπως είναι σε μια ταξινόμια, ούτε είναι «τάξεις» ή «έννοιες» όπως είναι σε ένα εννοιολογικό μοντέλο ή μια λογική θεωρία. Οι κόμβοι είναι μάλλον «όροι», δηλαδή λέξεις ή φράσεις, και οι όροι αυτοί έχουν στενότερες από ή ευρύτερες από σχέσεις ανάμεσά τους. Ένας θησαυρός είναι ουσιαστικά σχέσεις ανάμεσα σε όρους δομημένους κατά μία (σημασιολογικά χαλαρή) ταξινόμια. Ένας θησαυρός περιλαμβάνει και άλλες σημασιολογικές σχέσεις

ανάμεσα στους όρους, όπως τα συνώνυμα. Το σημασιολογικό μοντέλο στο οποίο οι σχέσεις (οι συσχετισμοί ανάμεσα στα αντικείμενα) ονομάζονται και διαφοροποιούνται σαφώς ονομάζεται οντολογία. Στην εικόνα «το φάσμα της οντολογίας», τόσο τα εννοιολογικά μοντέλα όσο και οι λογικές θεωρίες μπορούν να θεωρηθούν ως οντολογία (το πρώτο μια χαλαρότερη οντολογία και το δεύτερο μια ισχυρότερη οντολογία). Επειδή οι σχέσεις προσδιορίζονται, δεν υπάρχει πλέον ανάγκη αυστηρής δομής που περικλείει ή καθορίζει τις σχέσεις. Το μοντέλο ουσιαστικά γίνεται ένα δίκτυο συνδέσεων, με κάθε σύνδεση να έχει έναν ανεξάρτητο συσχετισμό από κάθε άλλη σύνδεση. Η μεταβλητότητα αυτή παρέχει φοβερή ευελιξία στην αντιμετώπιση εννοιών, επειδή πολλοί εννοιολογικοί τομείς δεν μπορούν να εκφραστούν επαρκώς με μια ταξινομία (ούτε με ένα θησαυρό, που διαμορφώνει σχέσεις όρων, σε αντίθεση με τις σχέσεις εννοιών). Σε μια ταξινομία ή ένα θησαυρό, προκύπτουν πάρα πολλές ανωμαλίες και αντιφάσεις, επιβάλλοντας έτσι αβάστακτους συμβιβασμούς. Επιπρόσθετα, η μετακίνηση ανάμεσα σε ανόμοιες έννοιες συχνά απαιτεί εύθραυστους συνδετικούς μηχανισμούς που είναι δύσκολο να συντηρούνται ή να επεκτείνονται.

2.2 Συλλογιστική και συμπερασματική των μηχανών

Η συλλογιστική και συμπερασματική των μηχανών αναφέρεται σε βασισμένη σε υπολογιστή προσομοίωση της ανθρώπινης δυνατότητας να φθάνει σε ένα πόρισμα μέσω της συλλογιστικής. Νέα δεδομένα ή συμπεράσματα προέρχονται από μια πύλη εισόδου πληροφοριών σε ένα πρόγραμμα υπολογιστή.

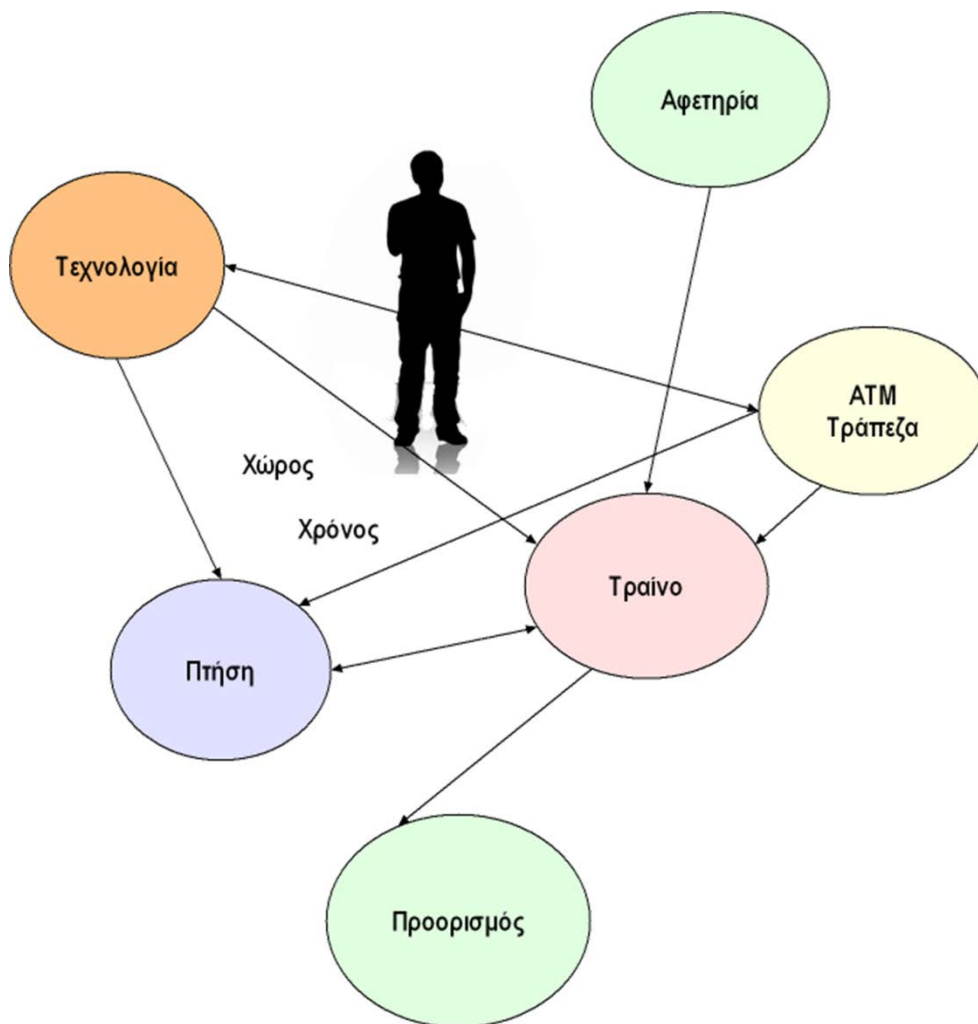
Εδώ έχουμε ένα πολύ απλό παράδειγμα του πώς εργάζεται η συλλογιστική και συμπερασματική των μηχανών, δαμάζοντας προκαθορισμένη «σημασιολογία». Έστω ότι δίνονται οι ακόλουθες πληροφορίες:

- Το Συνέδριο 2004 του ΑΒΓ θα γίνει στην αίθουσα εκδηλώσεων του Κτιρίου Χ.
- Το Κτίριο Χ έχει διεύθυνση «Δεύτερος Δρόμος, 123, Κύρια Πόλη, Βοιωτία».
- Το Κτίριο Χ είναι συγκρότημα που ανήκει στον Οργανισμό Υ.

Βασισμένοι στα ακατέργαστα αυτά στοιχεία, μπορούμε να τεκμαίρουμε ότι το Συνέδριο 2004 του ΑΒΓ έχει ως διεύθυνση το «Δεύτερος Δρόμος, 123, Κύρια Πόλη, Βοιωτία» και ότι θα γίνει στο συγκρότημα του Οργανισμού Υ. Η προκαθορισμένη σημασιολογία καθορίζει τους συσχετισμούς ανάμεσα σε διαφορετικές ιδιότητες δεδομένων, π.χ. γονέα-παιδιού και μεταβατικότητα.

Δημιουργώντας ένα μοντέλο πληροφορίας και σχέσεων, δίνουμε τη δυνατότητα στους λογικά σκεπτόμενους να εξαγάγουν λογικά συμπεράσματα βασισμένοι στο μοντέλο. Στο απλό παράδειγμα πιο πάνω, οι συνδέσεις όπως ο συσχετισμός ανάμεσα στο Κτίριο Χ και το συνέδριο ΑΒΓ καθορίζονται σαφώς, καθιστώντας δυνατό να συμπεράνουμε τη διεύθυνση για το συνέδριο ΑΒΓ.

Ο «χάρτης του μετρώ» που φαίνεται πιο κάτω είναι ένα έγκυρο διάγραμμα του Σημασιολογικού Ιστού που θα απεικονίσει περαιτέρω το πώς οι έννοιες μπορούν να συνδεθούν ή να συσχετιστούν με σχετικές ή/και μη σχετικές έννοιες (βλ. Σχήμα 2.2).



Σχήμα 2.2. Χάρτης μετρό του Σημασιολογικού Ιστού

Θεμελιώδεις έννοιες μπορούν να φανούν ως γραμμές στο διάγραμμα, με την κάθε μια να προσδιορίζει μια ξεχωριστή ατομική μορφή δεδομένων, όπως ο κόσμος, ο χρόνος ή ο χώρος. Η διασταύρωση μίας ή περισσότερων από αυτές τις θεμελιώδεις έννοιες συγκροτεί μία οντότητα με κάποιο υψηλότερο επίπεδο συσχετισμένης σημασίας. Η έννοια ενός Βιβλίου Διευθύνσεων είναι ο συνδυασμός του κόσμου, των διευθύνσεων και άλλων πληροφοριών επαφής. Η έννοια του Καταλόγου είναι μια συλλογή μερών και τιμών. Αν και απλό στην ουσία, το διάγραμμα δείχνει πως όταν η σημασιολογία καθοριστεί κατάλληλα για τα δεδομένα - μια ημερομηνία ή μια τοποθεσία, για παράδειγμα - μπορεί τότε να συσχετιστεί κατά τρόπους που είναι μεγαλύτεροι από τη συγκεκριμένη μορφή ή αναπαράσταση των δεδομένων. Με άλλα λόγια, χρησιμοποιώντας μια αναζήτηση για ένα χάρτη του Κιλκίς το 1913 ως παράδειγμα, αν το δεδομένο 1913 προσκολλάται ή αναγνωρίζεται ως ημερομηνία, τότε μπορούν να γίνουν ευφρείς αναζητήσεις, χρησιμοποιώντας ευέλικτες αναπαραστάσεις επερώτησης ημερομηνίας έναντι μιας ποικιλίας αναπαραστάσεων ημερομηνίας (Ιούνιος 1913• 1913• ή ακόμη δεκαετία του 1910, για παράδειγμα). Περαιτέρω, έχοντας συσχετισμούς από τους οποίους οι συσχετισμοί μπορούν να οριστούν ανεξάρτητα από μια ταξινομημένη δομή σχέσεως (όπως μια οντολογία), κάνει δυνατή τη συμπερίληψη ενός συσχετισμού «ημερομηνίας» ή «διακύμανσης ημερομηνίας» ανάμεσα στη «Μάχη του Κιλκίς» και το «19-22 Ιουνίου 1913». Ως αποτέλεσμα, μπορεί να βγει συμπέρασμα εντός μιας μηχανής αναζήτησης σχετικά με τη διακύμανση ημερομηνίας, αν έχει τη δυνατότητα να «διέλθει» οποιουδήποτε συσχετισμούς εντός μιας οντολογίας για μια έννοια που έχει να κάνει με ημερομηνίες.

Άλλο ένα παράδειγμα θα μπορούσε να είναι ένα σεμινάριο στον Άγιο Στέφανο της Αττικής. Η έννοια ενός τόπου μεταφέρει μαζί της συσχετισμούς που μπορούν να τοποθετήσουν πόλεις (Άγιος Στέφανος, Αττική) εντός μεγαλύτερων και πιο ευέλικτων περιοχών ορίων, όπως η μητροπολιτική περιοχή των Αθηνών ή συνδέει πόλεις, ταχυδρομικούς κωδικούς, πάρκα, μνημεία και άλλες πληροφορίες με βάση την τοποθεσία σε ένα

διασχίσιμο εννοιολογικό τομέα. Οι ευφυείς αναζητήσεις που καθίστανται εφικτές από τις σημασιολογικές προσεγγίσεις, μπορούν να τιθασεύσουν το συνδυασμό σαφώς καθορισμένων δεδομένων (όπως μια ημερομηνία ή μια τοποθεσία), παράλληλα με ευέλικτα μοντέλα συσχετισμών (το έργο γι' αυτά βρίσκεται ακόμη σε εξέλιξη) για να γεφυρώσουν τις διαφορές ανάμεσα στις συντάξεις, τις δομικές αναπαραστάσεις ή τα ευρύτερα πλαίσια. Αυτή η ιδέα του «αποκεντρωμένο αλλά συνδέσιμο» είναι θεμελιώδης για το όραμα του Σημασιολογικού Ιστού.

Ωστόσο, κανένα από αυτά τα παραδείγματα δεν υπονοεί την ύπαρξη κάποιας μαγικής μορφής τεχνητής νοημοσύνης που επιτρέπει στις μηχανές να κατανοούν τα ανθρώπινα μурμουρητά. Απλώς επισημαίνει την ικανότητα της μηχανής να επιλύσει ένα καλώς ορισμένο πρόβλημα εκτελώντας καλώς ορισμένες λειτουργίες σε καλώς ορισμένα υπάρχοντα δεδομένα (Berners-Lee, Σεπτέμβριος 1998, σ. 1).

2.3 Αναπαράσταση Γνώσης

Η αναπαράσταση γνώσης αποτελεί ένα κεντρικό πρόβλημα στον τομέα της τεχνητής νοημοσύνης (ΑΙ). Ερώτημα είναι πώς μπορεί να αποθηκεύσει και χειριστεί η γνώση μέσα σε ένα πληροφοριακό σύστημα με έναν τυπικό τρόπο έτσι ώστε να μπορεί να χρησιμοποιηθεί με μηχανισμούς για να ολοκληρώσει έναν δοσμένο έργο. Παραδείγματα εφαρμογών είναι τα έμπειρα συστήματα (expertsystems), τα συστήματα μηχανικής μετάφρασης, τα συστήματα ανάκτησης πληροφοριών και λοιπά. Στην καθημερινότητά του ο άνθρωπος αναπαριστά τη γνώση με απλή περιγραφή σε φυσική γλώσσα. Η φυσική γλώσσα θεωρείται μία από τις καλύτερες μεθόδους αναπαράστασης της γνώσης για τον άνθρωπο. Σε αντίθεση με τα υπολογιστικά συστήματα η φυσική γλώσσα είναι ακατάλληλη για να γίνει αναπαράστασή της. Το βασικό εμπόδιο είναι η πολυσημαντικότητα (ambiguity) της γλώσσας, δηλαδή οι πολλαπλές ερμηνείες που ενδέχεται μία φράση καθώς και η ερμηνεία με βάση τα συμφραζόμενα (context). Για τα συστήματα τεχνητής νοημοσύνης πρέπει να χρησιμοποιηθεί ένας μονοσήμαντος και τυποποιημένος συμβολισμός που πρέπει και να δίνει δυνατότητα ακριβούς αναπαράστασης γνώσης αλλά και να μπορεί να συνδυαστεί κατάλληλα με μηχανισμούς εξαγωγής συμπερασμάτων (inferencemechanisms) που υλοποιούνται από τη στρατηγική αναζήτησης στη γνώση ενός προβλήματος και από τη συλλογιστική (reasoning). Σύμφωνα με τον Winston, αναπαράσταση γνώσης είναι ένα σύνολο συντακτικών και σημασιολογικών παραδοχών που καθιστούν δυνατή την περιγραφή ενός κόσμου. Άρα τα βασικά δομικά στοιχεία που έχει κάθε μέθοδος αναπαράστασης γνώσης είναι :

1. το συντακτικό (syntax), ο ορισμός των συμβόλων που χρησιμοποιεί και οι κανόνες με τους οποίους τα σύμβολα αυτά μπορούν να συνδυαστούν
2. τη σημασιολογία (semantics) ο καθορισμός των εννοιών που αποδίδονται στα σύμβολα και συνδυασμούς συμβόλων που επιτρέπει το συντακτικό

Μέθοδοι αναπαράστασης γνώσης μπορούν να ταξινομηθούν σε τρεις κατηγορίες οι οποίες θεωρούνται και οι δημοφιλέστερες:

1. Λογική
 - Προτασιακή λογική (propositional logic)
 - Κατηγορηματική λογική (predicate logic) ή λογική πρώτης τάξης (first-order logic)
 - Διαζευκτική μορφή της λογικής (clausal form of logic)
2. Δομημένες αναπαραστάσεις γνώσης
 - Σημασιολογικά Δίκτυα (semantic networks)
 - Πλαίσια (frames)
 - Εννοιολογική εξάρτηση (conceptual dependency)
3. Κανόνες (if_then rules)

Η Κατηγορηματική λογική είναι προέκταση της προτασιακής λογικής εισάγοντας όρους, κατηγορήματα και ποσοδείκτες (quantifiers). Για παράδειγμα στην κατηγορηματική λογική η πρόταση «Ο Κώστας είναι προγραμματιστής» αναπαριστάται με προγραμματιστής(Κώστας) επιτρέποντας έτσι την προσπέλαση των στοιχείων του συγκεκριμένου αντικειμένου (Κώστας) από τους κανόνες διεξαγωγής συμπερασμάτων για τη δημιουργία νέων προτάσεων. Μηχανισμός εξαγωγής συμπερασμάτων στην κατηγορηματική λογική είναι η απόδειξη.

Η Γνώση Τομέα (domainknowledge) περιγράφει τα κύρια στατικά αντικείμενα πληροφορίας και γνώσης σε έναν τομέα εφαρμογής. Εντοπίζουμε και προσδιορίζουμε το βασικό είδος των συστατικών στοιχείων που χρησιμοποιούνται για την περιγραφή της γνώσης τομέα στις οντολογίες. Σύμφωνα με τον Gruber, η γνώση στις οντολογίες μπορεί να προσδιοριστεί αξιοποιώντας πέντε είδη συστατικών: έννοιες (concepts), σχέσεις (relations), λειτουργίες (functions), αξιώματα (axioms) και στιγμιότυπα (instances). Οι έννοιες στην οντολογία συνήθως οργανώνονται σε ταξινομίες. Μερικές φορές η έννοια της οντολογίας αμβλύνεται κατά κάποιο τρόπο, υπό την έννοια ότι οι ταξινομίες θεωρούνται ως πλήρεις οντολογίες. Άλλες συνιστώσες όπως διαδικασίες και κανόνες επίσης εντοπίζονται σε ορισμένες οντολογικές γλώσσες (πχ. OCML). Για κάθε μία από τις προαναφερθείσες συνιστώσες (πλην των διαδικασιών, καθώς είναι ιδιαίτερα δύσκολο το να εντοπιστούν κοινά χαρακτηριστικά για αυτές σε όλες τις γλώσσες) θα επιλέξουμε μια ομάδα χαρακτηριστικών που θεωρούμε συναφή.

Οι έννοιες (Concepts), γνωστές και ως τάξεις, χρησιμοποιούνται με ευρεία έννοια. Μπορεί να είναι αφηρημένες ή συγκεκριμένες, στοιχειώδεις ή σύνθετες, πραγματικές ή φανταστικές. Εν συντομία, μία έννοια μπορεί να είναι οτιδήποτε για το οποίο κάτι λέγεται, και, επομένως, θα μπορούσε επίσης να είναι η περιγραφή ενός έργου, μιας λειτουργίας, ενέργειας, στρατηγικής, διαδικασία συλλογισμού, κτλ. Τα ακόλουθα ερωτήματα προσδιορίζουν την εκφραστικότητα μιας γλώσσας κατά τον ορισμό τάξεων (classes):

- Είναι εφικτό να οριστούν οι μετατάξεις (metaclasses)(ίσως τάξεις ως στιγμιότυπα άλλων). Αυτές είναι σημαντικές σε περίπτωση που υφίσταται μια αναπαράσταση γνώσης (KR) οντολογία για τη γλώσσα
- Είναι δυνατό να οριστούν διαχωρισμοί (partitions)(ομάδες τεμαχισμένων τάξεων)
- Παρέχει η γλώσσα μηχανισμούς για να οριστούν θυρίδες / ιδιότητες (slots / attributes); Παραδείγματος χάριν:

τοπικές ιδιότητες (local attributes). Ιδιότητες οι οποίες ανήκουν σε μια συγκεκριμένη έννοια. Για παράδειγμα, η ιδιότητα *ηλικία* ανήκει στην έννοια *Πρόσωπο*.

ιδιότητες στιγμιότυπων οδηγών θυρίδων (instanceattributestemplateslots). Ιδιότητες των οποίων η τιμή μπορεί να διαφέρει για κάθε στιγμιότυπο (instance) της έννοιας (concept).

.θυρίδες κλάσεων (class attributes)(own slots). Ιδιότητες των οποίων η τιμή πρέπει να είναι η ίδια για όλες τις στιγμιότυπα της έννοιας (concept).

πολυμορφικές ιδιότητες (Polymorph attributes). Ιδιότητες θυρίδων (slots) με το ίδιο όνομα και διαφορετική συμπεριφορά για διαφορετικές έννοιες. Για παράδειγμα, η ιδιότητα *συγγραφέας* για την έννοια *Διατριβή* διαφέρει από την ιδιότητα *συγγραφέας* για την έννοια *Βιβλίο*. Ο τύπος της για τη *Διατριβή* είναι *Μαθητής*, ενώ ο τύπος της για το *Βιβλίο* είναι *Πρόσωπο*.

- Παρέχει η γλώσσα τις ακόλουθες προκαθορισμένες απόψεις (predefined facets)για τις ιδιότητες;
- **Εξ ορισμού τιμή κενού** (Default slot value), η οποία θα χρησιμοποιηθεί για να αποδοθεί μια τιμή στην ιδιότητα (attribute) σε περίπτωση που δεν υπάρχει ρητή προκαθορισμένη τιμή για αυτή.
- **Τύπος**, ο οποίος θα χρησιμοποιηθεί για τον περιορισμό του τύπου μιας ιδιότητας (attribute).
- **Περιορισμοί μεγέθους** (Cardinality constraints), οι οποίοι θα χρησιμοποιηθούν για να περιορίσουν τον ελάχιστο και μέγιστο αριθμό των τιμών της ιδιότητας.
- **Τεκμηρίωση** (Documentation), η οποία θα μπορούσε να περιέχει έναν ορισμό φυσικής γλώσσας.
- **Λειτουργικός Ορισμός** (Operational definition), που θα μπορούσε να περιλαμβάνει τον προσδιορισμό ή επιλογή μιας φόρμουλας, ενός κανόνα, κτλ για να χρησιμοποιηθεί, για παράδειγμα, κατά την απόδοση μιας τιμής στην ιδιότητα αυτή.
- Μπορούν να δημιουργηθούν **νέες απόψεις** (new facets)για ιδιότητες;

Ταξινομίες. Χρησιμοποιούνται ευρέως για να οργανώνουν οντολογική γνώση στον τομέα αξιοποιώντας σχέσεις γενίκευσης / προσδιορισμού συγκεκριμενοποίησης μέσω των οποίων απλή ή πολλαπλή κληρονομική διαδοχή μπορεί να εφαρμοστεί. Δεδομένου ότι υφίσταται κάποια σύγχυση αναφορικά με τους αρχέγονους τύπους που χρησιμοποιούνται για να φτιαχτούν ταξινομίες, προτείνουμε να αναλύσουμε εάν οι ακόλουθοι αρχέγονοι τύποι (που στηρίζονται στους ορισμούς που παρέχονται από το πλαίσιο οντολογίας στην Ontolingua) είναι προκαθορισμένοι ή μη στις γλώσσες.

- **Υποκατηγορία του** (Subclass of)συγκεκριμενοποιεί γενικές έννοιες σε πιο συγκεκριμένες έννοιες.
- **Ασύνδετοι διαχωρισμοί** (Disjoint decompositions)ορίζουν έναν διαχωρισμό (a partition) ως υποκατηγορία μιας κλάσης. Η κατηγοριοποίηση δεν απαιτείται να είναι ολοκληρωμένη (μπορεί να υπάρχουν στιγμιότυπα (instances) της μητρικής κλάσης που δεν συμπεριλαμβάνονται σε καμία από τις υποκατηγορίες του διαχωρισμού).

- **Εξαντλητικοί διαχωρισμοί υποκατηγοριών** (Exhaustive subclass decompositions) ορίζουν έναν διαχωρισμό ως υποκατηγορία μιας κλάσης. Η μητρική κλάση (parent class) αποτελεί την ένωση όλων των τάξεων που συνθέτουν τον διαχωρισμό (partition).
- **Μη-Υποκατηγορία του** (Not subclass of) μπορεί να χρησιμοποιηθεί για να δηλώσει ότι μια κλάση δεν αποτελεί συγκεκριμενοποίηση μιας άλλης κλάσης. Αυτό το είδος της γνώσης αποτυπώνεται συνήθως χρησιμοποιώντας την άρνηση (denial) της υποκατηγορίας του αρχέγονου τύπου (subclass of primitive). Μερικές γλώσσες διαθέτουν μια τυπική σημασιολογία για τους αρχέγονους αυτούς τύπους, ενώ κάποιες άλλες πρέπει να ορίζουν τη σημασιολογία τους χρησιμοποιώντας αξιώματα ή κανόνες.

Οι **Σχέσεις** (Relations) αντιπροσωπεύουν ένα είδος αλληλεπίδρασης μεταξύ εννοιών του τομέα. Επίσημως ορίζονται ως η εκάστοτε υποομάδα ενός προϊόντος n ομάδων (any subset of a product of n sets). Αρχικά, εξετάζουμε τη σχέση μεταξύ σχέσεων και άλλων συστατικών μερών της οντολογίας. Θα αναρωτηθούμε εάν οι έννοιες και οι ιδιότητες θεωρούνται ως μοναδικές (unary) ή δυαδικές (binary) σχέσεις αντίστοιχα..

Οι **Λειτουργίες** (Functions) θεωρούνται ένα ιδιαίτερο είδος σχέσεων όπου η τιμή του τελευταίου επιχειρήματος είναι μοναδική για μια σειρά τιμών των $n-1$ προηγούμενων επιχειρημάτων (preceding arguments). Δεύτερον, επικεντρωνόμαστε στα επιχειρήματα και στις σχέσεις και στις λειτουργίες:

- Είναι εφικτό να οριστούν οι αυθαίρετες και οι μη-αυθαίρετες σχέσεις/λειτουργίες (arbitrary n -ary relations/functions); Εάν αυτό δεν είναι δυνατό, ποιος είναι ο μέγιστος αριθμός επιχειρημάτων;
- Μπορεί να περιοριστεί ο τύπος των επιχειρημάτων;
- Είναι εφικτό να οριστούν **περιορισμοί αριότητας** (integrity constraints) προκειμένου να ελέγχουν την ορθότητα της τιμής των επιχειρημάτων;
- Γίνεται να οριστούν **λειτουργικοί ορισμοί** (operational definitions) για να συνάγονται τιμές επιχειρημάτων με διαδικασίες, φόρμουλες και κανόνες, ή για να οριστεί η σημασιολογία τους χρησιμοποιώντας αξιώματα ή κανόνες;

Τα Αξιώματα (Axioms) διαμορφώνουν προτάσεις που είναι πάντα αληθείς. Συμπεριλαμβάνονται σε μια οντολογία για διάφορους λόγους, όπως για να περιορίσουν τις πληροφορίες της, να επαληθεύσουν την ορθότητά της, ή για την εξαγωγή νέων πληροφοριών. Θα επικεντρωθούμε στα ακόλουθα χαρακτηριστικά:

- Υποστηρίζει η γλώσσα τη δημιουργία αξιωμάτων σε πρώτης τάξης λογική (first order logic);
- Και τα αξιώματα δεύτερης τάξης λογικής (second order logic);
- Τα αξιώματα ορίζονται ως ανεξάρτητα στοιχεία στην οντολογία (named axioms) ή πρέπει να συμπεριληφθούν εντός του ορισμού άλλων στοιχείων, όπως σχέσεων, εννοιών, κτλ; Το στοιχείο αυτό βελτιώνει την δυνατότητα ανάγνωσης και τη διατήρηση των οντολογιών.

Στιγμιότυπα/Άτομα/Γεγονότα/Ισχυρισμοί (Instances/Individuals/Facts/Claims).

Όλοι αυτοί οι όροι χρησιμοποιούνται για να παραστήσουν στοιχεία στον τομέα. Τα Στιγμιότυπα (Instances) αντιπροσωπεύουν στοιχεία μιας δεδομένης έννοιας. Τα γεγονότα (Facts) αντιπροσωπεύουν μια σχέση η οποία υφίσταται μεταξύ στοιχείων. Τα άτομα (Individuals) αναφέρονται σε κάθε στοιχείο στον τομέα που δεν αποτελεί κλάση (class) (και στιγμιότυπα και γεγονότα). Οι ισχυρισμοί (Claims) αναπαριστούν διαβεβαίωση ενός γεγονότος από μια περίπτωση. Είναι σημαντικό να τονιστεί η προσθήκη ισχυρισμών, καθώς οι άνθρωποι στο Διαδίκτυο μπορούν να κάνουν οποιοδήποτε ισχυρισμό επιθυμούν. Επομένως, οι διαμεσολαβητές (agents) δε θα πρέπει να ερμηνεύουν τους ισχυρισμούς ως αντικείμενα γνώσης (facts of knowledge), αλλά ως ισχυρισμούς που γίνονται από μια συγκεκριμένη περίπτωση σχετικά με τον εαυτό της ή σχετικά με άλλες στιγμιότυπα (instances) ή δεδομένα, τα οποία ίσως αποδειχτούν μη συμβατά με άλλα. Στο μέρος αυτός, θα τεθούν τα ακόλουθα ερωτήματα:

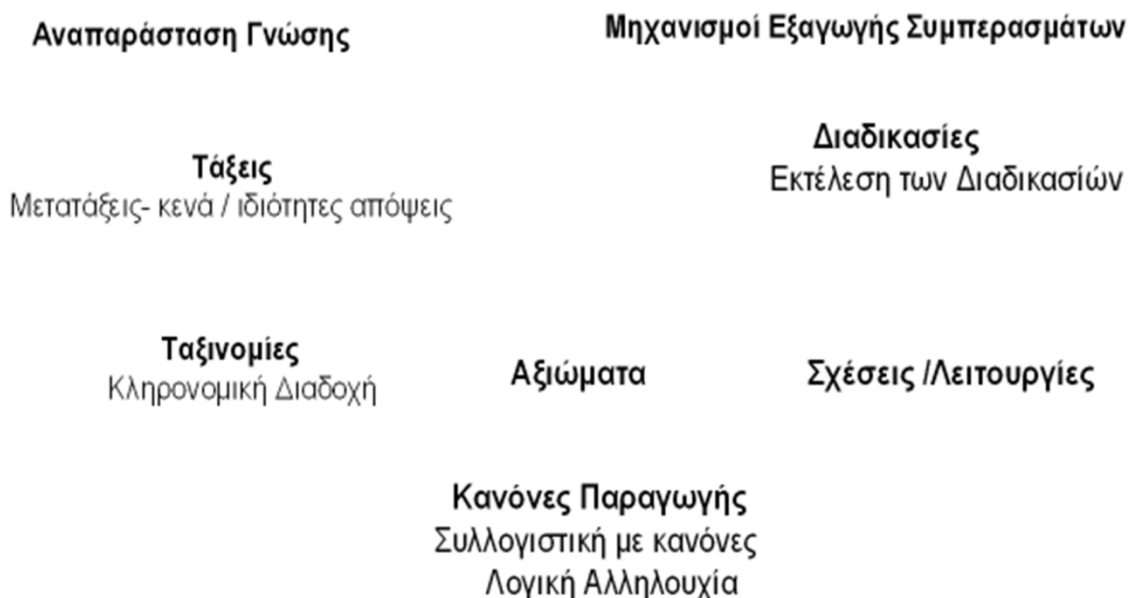
- Είναι εφικτό να οριστούν στιγμιότυπα εννοιών (instances of concepts);
- Είναι εφικτό να οριστούν στιγμιότυπα σχέσεων (γεγονότα);
- Παρέχει η γλώσσα ειδικούς μηχανισμούς για τον προσδιορισμό ισχυρισμών (claims);

Κανόνες Παραγωγής. Οι κανόνες παραγωγής (Production rules), που ακολουθούν τη δομή *Εάν... Τότε...* (If ... Then ...), χρησιμοποιούνται για να εκφράσουν ομάδες ενεργειών και ευριστικούς μηχανισμούς (heuristics) που μπορούν να παρασταθούν ανεξάρτητα από τον τρόπο που θα χρησιμοποιηθούν. Θα εξεταστεί μια ομάδα ερωτημάτων σχετικά με αυτούς τους κανόνες:

- Είναι εφικτό να οριστούν διαζευκτικά και συζευκτικά μέρη (premises);
- Μπορεί ο μηχανισμός εκτέλεσης κανόνων (chaining mechanism) να οριστεί με σαφήνεια;
- Είναι εφικτό να οριστούν οι **τιμές αληθείας** (truth values) ή οι **τιμές βεβαιότητας** (certainty values) που προσάπτονται στον κανόνα;

- Μπορούν οι **διαδικασίες**(procedures)να συμπεριληφθούν στην **ακολουθία** (consequent); Χρησιμοποιούνται συχνά για να αλλάξουν τις τιμές των ιδιοτήτων μιας έννοιας, να προσθέσουν πληροφορίες στη Γνωσιακή Βάση (KB), κτλ.
- Υποστηρίζει η γλώσσα ενημερώσεις (updates) στη Γνωσιακή Βάση (KB), οι οποίες πραγματοποιούνται προσθέτοντας ή αφαιρώντας γεγονότα (facts) ή ισχυρισμούς (βλ Σχήμα2.3);

Πλαίσιο Αξιολόγησης



Σχήμα 2.3.Το πλαίσιο αξιολόγησης

2.4 Μηχανισμοί εξαγωγής συμπερασμάτων (Inference Mechanisms)

Η διάσταση αυτή περιγράφει πώς οι στατικές δομές που αναπαρίστανται στη Γνώση Τομέα μπορούν να χρησιμοποιηθούν για τη διεξαγωγή μια συλλογιστικής διαδικασίας. Υπάρχει μια ισχυρή σχέση μεταξύ και των δύο διαστάσεων (dimensions), καθώς οι δομές που χρησιμοποιούνται για την αναπαράσταση της γνώσης αποτελούν τη βάση για τη διαδικασία συλλογισμού (reasoning process), όπως φαίνεται στο γράφημα:πλαίσιο αξιολόγησης. Αναλύουμε εάν η γλώσσα υποστηρίζει τα ακόλουθα χαρακτηριστικά ή όχι:

- Παρέχει η γλώσσα μια **μηχανή εξαγωγής συμπερασμάτων** (inference engine) που «σκέπτεται» με τη γνώση που αναπαρίσταται χρησιμοποιώντας τη γλώσσα; Είναι **λογικό** (sound); Και **ολοκλήρωμένο** (complete);
- Εκτελεί η μηχανή εξαγωγής συμπερασμάτων **αυτόματες ταξινομήσεις** (automatic classifications);
- Ασχολείται η μηχανή εξαγωγής συμπερασμάτων με **εξαιρέσεις** (exceptions); Εξαιρέσεις θεωρούνται όταν η ιδιότητα 1 ορίζεται για την έννοια E1 και την έννοια E2, όπου η E1 είναι υποκατηγορία της E2 και αναλύουμε εάν ο ορισμός της ιδιότητας1 στην έννοια E1 πρέπει να ορίσει ξανά (override) τον ορισμό της ιδιότητας1 στην έννοια E2 ή όχι.
- Είναι εφικτό να χρησιμοποιηθεί **μονοτονική, μη-μονοτονική, απλή και/ή πολλαπλή κληρονομική διαδοχή**;
- Είναι οι **διαδικασίες** (procedures)εκτελέσιμες;
- Εκτελούν τα αξιώματα κάποιου είδους **έλεγχο περιορισμών** (constraint checking);
- Κατά τη συλλογιστική με κανόνες, επιτρέπει η γλώσσα **ορθήκαι ανάστροφηεκτέλεση κανόνων** (forward and backward chaining);

Προκειμένου να λειτουργήσει ο Σημασιολογικός Ιστός, θα πρέπει οι Η/Υ να έχουν πρόσβαση σε δομημένες συλλογές πληροφοριών και σε σύνολα κανόνων εξαγωγής συμπερασμάτων ώστε να διεξάγουν

αυτοματοποιημένους συλλογισμούς. Παρόμοια συστήματα έχουν μελετηθεί πολύ πριν τη δημιουργία του Ιστού από τους ερευνητές της τεχνητής νοημοσύνης. Η αναπαράσταση της γνώσης, όπως συχνά αποκαλείται αυτή η τεχνολογία, βρίσκεται σε τέτοια φάση που μπορεί να συγκριθεί με εκείνη του υπερκειμένου (hypertext) πριν την έλευση του Ιστού. Ο σπόρος για εφαρμογές είναι εκεί αλλά για να εκμεταλλευτούμε πλήρως τις δυνατότητες του θα πρέπει να συνδεθεί πρώτα με ένα ενιαίο καθολικό σύστημα. Τα παραδοσιακά συστήματα αναπαράστασης της γνώσης είναι μέχρι στιγμής κεντριοποιημένα προϋποθέτοντας έτσι για όλους τον ίδιο ορισμό για κοινές έννοιες όπως 'πατέρας' ή 'όχημα'. Όμως, ο κεντρικός έλεγχος οδηγεί σε κορεσμό, ενώ η ολοένα αυξανόμενη έκταση και το πεδίο δράσης ενός τέτοιου συστήματος γίνεται γρήγορα ανεξέλεγκτο. Επιπλέον, αυτά τα συστήματα περιορίζουν συνήθως με προσοχή τον αριθμό των ερωτήσεων που θα υποβληθούν ώστε ο Η/Υ να μπορεί να απαντήσει με αξιοπιστία ή και να μην απαντήσει καθόλου. Προς αποφυγή προβλημάτων αξιοπιστίας, το καθένα από τα παραδοσιακά συστήματα αναπαράστασης της γνώσης είχε το δικό του περιορισμένο, ιδιόμορφο σύνολο κανόνων προκειμένου να βγάζει συμπεράσματα μόνο για τα δικά του δεδομένα. Συνεπώς, η πρόκληση του Σημασιολογικού Ιστού είναι να παράσχει μια γλώσσα πρώτο, ικανή να εκφράζει τόσο δεδομένα όσο και κανόνες για συλλογισμούς που αφορούν αυτά τα δεδομένα, και δεύτερο, που να επιτρέπει στους κανόνες από ήδη υπάρχοντα συστήματα αναπαράστασης της γνώσης να εξάγονται/ προωθούνται στον Ιστό. Η προσθήκη λογικής στον Ιστό- δηλαδή τα μέσα με τα οποία θα μπορούν να χρησιμοποιούνται κανόνες για την εξαγωγή συμπερασμάτων, για την επιλογή δράσης και απάντησης ερωτημάτων- είναι το κύριο μέλημα της κοινότητας του Σημασιολογικού Ιστού αυτή τη στιγμή. Τα πράγματα περιπλέκονται από ένα κράμα μαθηματικών αποφάσεων και αποφάσεων μηχανίκευσης (engineering). Η λογική αυτή θα πρέπει να είναι τόσο ισχυρή ώστε να περιγράφει περίπλοκες ιδιότητες των αντικειμένων, χωρίς, όμως, να προκαλεί σύγχυση στον διαμεσολαβητή ζητώντας του να αναλογιστεί ένα παράδοξο.

2.5 Γιατί το μοντέλο της RDF είναι διαφορετικό από το μοντέλο της XML

Δύο σημαντικές τεχνολογίες για την εξέλιξη του Σημασιολογικού Ιστού χρησιμοποιούνται ήδη: η eXtensibleMarkupLanguage (XML) (Γλώσσα Επεκτάσιμης Σήμανσης) και η ResourceDescriptionFramework (RDF) (Πλαίσιο Περιγραφής Πόρων). Η XML επιτρέπει στον καθένα να δημιουργήσει τα δικά του σύμβολα πεδίου (tags)- κρυμμένες ετικέτες που σχολιάζουν ιστοσελίδες ή τμήματα κειμένου σε μια σελίδα. Αυτό το σημείωμα είναι μια προσπάθεια να απαντηθεί το ερώτημα: «Γιατί θα πρέπει να χρησιμοποιήσω την RDF - γιατί όχι απλά την XML;». Αυτό το ερώτημα υπάρχει από τον καιρό που ξεκίνησε η RDF. Στο εργαστήριο Γλώσσας Επερώτησης του W3C υπήρχε μια σαφής διαφορετική άποψη ανάμεσα σε αυτούς που ήθελαν να επερωτούν έγγραφα και σε αυτούς που ήθελαν να εκμαιεύσουν το «νόημα» σε κάποια μορφή και να το επερωτήσουν αυτό. Αυτό είναι σύνηθες. Γράφθηκε το σημείωμα αυτό ως μια ανεκπλήρωτη προσπάθεια εξήγησης του τι είναι το μοντέλο της RDF για αυτούς που σκέφτονταν με βάση το μοντέλο της XML. Το σημείωμα αυτό αντιμετωπίζει το μοντέλο δεδομένων της XML σε όλη του την πολυπλοκότητα, και τη σύνταξη της RDF όπως βρίσκεται στο Μοντέλο της RDF και τη Σύνταξη, σε όλη του την πολυπλοκότητα. Δεν προσπαθεί να χαρτογραφήσει το ένα άμεσα στο άλλο - εκφράζει το μοντέλο της RDF χρησιμοποιώντας την XML. Ας πάρουμε ως παράδειγμα έναν ισχυρισμό της RDF. Ας δοκιμασθεί το «TheauthorofthepageisOra». Αυτό είναι παραδοσιακό. Στην RDF αυτό είναι τριπλό κομμάτι:

Triple (author, page, Ora) το οποίο μπορείτε να το σκεφτείτε όπως αναπαριστάται στο διάγραμμα:
Πώς αυτή η πληροφορία θα αναπαρασταθεί τυπικά σε XML;



Το γράφημα της XML

Αυτά είναι όλα απόλυτα σωστά έγγραφα της XML - και σε έναν που τα διαβάζει σημαίνουν τα ίδια πράγματα. Για μια μηχανή που τα αναλύει γραμματικά, παράγουν διαφορετικά δέντρα της XML. Ας υποθέσουμε πως απεικονίζεται τυπικά το δέντρο της XML

Δεν είναι τόσο φανερή η χρήση του. Τα ονόματα στοιχείων ήταν μια μεγάλη υπόδειξη για έναν ανθρώπινο αναγνώστη.

Χωρίς να βλέπουμε τη σχηματική παράσταση, ξέρουμε πράγματα για τη δομή του εγγράφου, αλλά τίποτα άλλο. Δεν μπορούμε να αντιληφθούμε τι θα συμπεράνουμε. Δεν ξέρουμε κατά πόσον το `rrrrr` είναι ένα `y` του `qqqq`, ή αν το `qqqq` είναι ένα `z` του `qqqq` ή άλλο. Δεν μπορούμε ούτε καν να αντιληφθούμε ποιες πραγματικές ερωτήσεις μπορούν να τεθούν. Μια πηγή μερικής σύγχυσης είναι πως στο παράδειγμα `xyz` πιο πάνω, υπάρχουν πολλές ερωτήσεις που *μπορούν* να τεθούν. Είναι ερωτήσεις όπως:

- Υπάρχει ένα στοιχείο `w` εντός ενός στοιχείου λεπτομερειών;
- Ποιο είναι το περιεχόμενο του στοιχείου `w` εντός του πρώτου στοιχείου `x`;
- Ποιο είναι το περιεχόμενο του στοιχείου `w` ακολουθώντας το πρώτο στοιχείο `y` που περιέχει ένα στοιχείο `x` του οποίου η ιδιότητα είναι `"rrrr"`;
- Και τα λοιπά.

Αυτές είναι όλες ερωτήσεις σχετικά με το *έγγραφο*. Αν γνωρίζουμε τη σχηματική παράσταση του εγγράφου (ένα μεγάλο *αν*), και αν η σχηματική παράσταση αυτή μας δίνει μόνο περιορισμένο αριθμό τρόπων έκφρασης του ίδιου πράγματος (άλλο ένα μεγάλο *αν*), τότε η ερώτηση αυτών των ερωτήσεων μπορεί να είναι στην ουσία αντίστοιχη με την ερώτηση ερωτήσεων όπως:

- Ποιος είναι ο συγγραφέας της *σελίδας*;
- Αυτό είναι δύσκολο. Είναι πιθανό, επειδή υπάρχει η χαρτογράφηση από έγγραφα της XML σε σημασιολογικά γραφήματα. Εν συντομία, είναι ζόρικο επειδή:
- Η χαρτογράφηση είναι πολλά προς ένα.
- Χρειαζόμαστε μια σχηματική παράσταση για να ξέρουμε ποια είναι η χαρτογράφηση.
- (Οι σχηματικές παραστάσεις για τις οποίες αναφερόμαστε για την XML προς το παρόν δεν το περιλαμβάνουν αυτό έτσι κι αλλιώς και θα έπρεπε να είχε προστεθεί μια ολόκληρη γλώσσα συμπερασμάτων).
- Η έκφραση που χρειαζόμαστε για την επερώτηση κάποιο πράγματος με βάση το δέντρο της XML είναι αναγκαία πιο περίπλοκη από την έκφραση που χρειαζόμαστε για την επερώτηση κάποιου πράγματος με βάση το δέντρο της RDF.

Το τελευταίο είναι μεγάλο. Αν προσπαθήσουμε να γράψουμε την έκφραση για το συγγραφέα ενός εγγράφου όπου οι πληροφορίες βρίσκονται σε κάποια αυθαίρετη σχηματική παράσταση της XML, πιθανότατα θα μπορέσουμε, αν και μπορεί να είναι ή να μην είναι ωραίο. Αν προσπαθήσουμε να συνδυάσουμε περισσότερες από μία ιδιότητες σε μια συνδυασμένη έκφραση (δώσε μου μια λίστα με βιβλία του ίδιου συγγραφέα όπως αυτό), το να το πούμε σε XML γίνεται πολύ αδέξιο για να το σκεφτούμε.

(Σκεφτείτε την προσπάθεια ορισμού της πρόσθεσης αριθμών με βάση συνήθων λειτουργιών στις συμβολοσειρές. Είναι δυνατό για την πρόσθεση. Όταν φτάσουμε στον πολλαπλασιασμό καταντά γελοίο - για να λυθεί το πρόβλημα θα καταλήγαμε να ξανά-εφεύρουμε τους αριθμούς ως ξεχωριστό τύπο).

Βλέποντας την απλή κωδικοποίηση της XML πιο πάνω,

```
<author>
<uri>page</uri>
<name>Ora</name>
</author>
```

θα μπορούσε να αναπαρασταθεί ως γράφημα.

Θα μπορούσαμε να αναπαραστήσουμε το δέντρο πιο συνοπτικά αν κάναμε μια στενογραφία, γράφοντας το όνομα κάθε στοιχείου μέσα στον κύκλο του. Φυσικά, το δέντρο της RDF που αναπαριστά αυτό (αν και δεν είναι προφανές από το δέντρο της XML, εκτός γι' αυτούς που ξέρουν) είναι: Εδώ κάναμε ένα στενογράφημα πάλι, κάνοντας την ταμπέλα του κάθε μέρους το URI του.

Η πολυπλοκότητα της επερώτησης του δέντρου της XML υφίσταται επειδή υπάρχει εν γένει μεγάλος αριθμός τρόπων κατά τους οποίους η XML χαρτογραφεί το λογικό δέντρο, και η επερώτηση που γράφουμε πρέπει να είναι ανεξάρτητη της επιλογής τους. Έτσι, μεγάλο τμήμα της επερώτησης είναι μια προσπάθεια για τη μετατροπή βασικά του συνόλου όλων των δυνατών αναπαραστάσεων ενός γεγονότος σε μία δήλωση. Αυτό είναι απλώς το τι κάνει η RDF. Μας δίνει μερικούς τυποποιημένους τρόπους εγγραφής δηλώσεων έτσι ώστε

όπως και αν προκύψουν σε ένα έγγραφο, παράγουν το ίδιο αποτέλεσμα από την άποψη της RDF. Το ίδιο δέντρο της RDF προκύπτει από πολλά δέντρα της XML.

Δεν θα ήταν ωραίο αν επιχειρείτο να ενθυλακωθεί με πλαίσια (ταμπέλες) η XML μας, έτσι ώστε όταν ο γραμματικός αναλυτής τη διαβάσει, να βρίσκει τους ισχυρισμούς (τριπλά κομμάτια) και να ξεχωρίζει τα υποκείμενα και τα αντικείμενά τους, έτσι ώστε να συμπεραίνει τους λογικούς ισχυρισμούς χωρίς να χρειάζεται την RDF; Αυτό είναι όμως απλά το τι κάνει η RDF.

Το γράφημα της RDF

Εδώ που τα λέμε, η RDF είναι πολύ ευέλικτη - μπορεί να αναπαραστήσει το τριπλό κομμάτι αυτό με πολλούς τρόπους σε XML, έτσι ώστε να μπορεί να ταιριάσει με συγκεκριμένες εφαρμογές, αλλά έτσι για να διαλέξουμε έναν, θα μπορούσαμε να γράψουμε το πιο πάνω ως:

```
<Description about="http://www.w3.org/test/page" Author ="Ora" />
```

Έχω παραλείψει το κομμάτι για τους χώρους ονομάτων. Στην ουσία, καθώς οποιοσδήποτε μπορεί να δημιουργήσει ή να κατέχει τα ρήματα, υποκείμενα και αντικείμενα σε ένα διασκορπισμένο Ιστό, οποιοσδήποτε όρος πρέπει να αναγνωρίζεται από ένα URI με κάποιο τρόπο. Αυτό το υπαρκτό αληθινό παράδειγμα, δουλεμένο στην πραγματική ζωή θα ήταν περισσότερο όπως:

```
<?xml version="1.0"?>
```

```
<Description
```

```
    xmlns="http://www.w3.org/TR/WD-rdf-syntax#"
    xmlns:s="http://docs.r.us.com/bibliography-info/"
```

```
    about="http://www.w3.org/test/page"
```

```
    s:Author ="http://www.w3.org/staff/Ora" />
```

Μπορούμε να σκεφτούμε πως η «περιγραφή» του στοιχείου της RDF δίνει την ένδειξη στο γραμματικό αναλυτή ως προς το πώς να βρει τα υποκείμενα, αντικείμενα και ρήματα σε ό,τι ακολουθεί.

Αυτός είναι πάνω κάτω ο πιο στενογραφικός τρόπος χρησιμοποιώντας τη βάση της RDF στην XML. Υπάρχουν άλλοι που είναι πιο μακροσκελείς, αλλά πιο αποδοτικοί αν έχουμε, για παράδειγμα, σύνολα από πολλές ιδιότητες του ίδιου αντικειμένου. Το χρήσιμο είναι πως φυσικά όλοι τους αποδίδουν το ίδιο τριπλό κομμάτι:

Είναι χάλια όταν χρησιμοποιούμε ερωτήσεις σχετικά με ένα έγγραφο για να προσπαθήσουμε να ρωτήσουμε ερωτήσεις σχετικά με το τι προσπαθεί να αποδώσει το έγγραφο. Θα δουλέψει. Κατά κάποιο τρόπο. Αλλά ή σαφής επισήμανση με δείκτες της γραμματικής (η σύνταξη της RDF είναι ένας τρόπος) είναι κατά πολύ καλύτερη.

Τα πράγματα που μπορούν να κάνουμε με την RDF, που δεν μπορούν να γίνουν με την XML περιλαμβάνουν:

- Μπορούμε να αναλύσουμε γραμματικά το σημασιολογικό δέντρο, που θα καταλήξει στο να μας δώσει ένα σύνολο (πιθανώς αμοιβαία αναφορικών) τριπλών κομματιών, και μετά μπορούμε να χρησιμοποιήσουμε αυτά που θέλουμε, αγνοώντας αυτά που δεν καταλαβαίνουμε.

Τα προβλήματα που προκύπτουν βασίζοντας την κατανόησή μας στη δομή περιλαμβάνουν:

- Χωρίς να μπούμε στον κόπο να πάρουμε τη σχηματική παράσταση, ή έχοντας προγραμματίσει την εφαρμογή μας στο χέρι για να αναγνωρίζει ένα συγκεκριμένο είδος εγγράφων, δεν μπορούμε να πάρουμε οποιαδήποτε σημασιολογική πληροφορία από ένα έγγραφο.
- Όταν μια σχηματική παράσταση της XML αλλάζει, θα μπορούσε τυπικά να εισαγάγει νέα ενδιάμεσα στοιχεία (όπως «details» στο δέντρο πιο πάνω ή «div» στην HTML). Αυτά μπορεί να ακυρώνουν ή να μην ακυρώνουν οποιαδήποτε επερώτηση βασισμένη στη δομή του εγγράφου.
- Αν δεν έχετε μπει στον κόπο να φτιάξετε ένα σημασιολογικό μοντέλο, τότε ίσως να μην έχετε ένα καλά καθορισμένο μοντέλο.

Σταματώντας εδώ με κάποια παραδείγματα του τελευταίου προβλήματος. Μπορούν σαφώς να αποφευχθούν με καλό σχεδιασμό, ακόμη και σε ένα σύστημα XML που δεν χρησιμοποιεί την RDF. Χρησιμοποιώντας την RDF τα πράγματα γίνονται ευκολότερα. Η σημασία εκφράζεται με τη βοήθεια της γλώσσας RDF, η οποία κωδικοποιείται σε ομάδες τριάδων, καθεμία από τις οποίες μοιάζει με το υποκείμενο, ρήμα και αντικείμενο μιας στοιχειώδους πρότασης. Αυτές οι τριάδες μπορούν να γραφούν χρησιμοποιώντας σύμβολα πεδίου XML. Στην RDF, ένα τεκμήριο διατυπώνει ισχυρισμούς ως προς το ότι συγκεκριμένα πράγματα (άνθρωποι, ιστοσελίδες, κτλ.) φέρουν ιδιότητες (π.χ. 'αδελφή του', 'συγγραφέας του') με

συγκεκριμένες αξίες (ένα άλλο άτομο, μια άλλη ιστοσελίδα). Αυτή η δομή, όπως αποδεικνύεται, είναι ο φυσικός τρόπος περιγραφής της συντριπτικής πλειοψηφίας των δεδομένων που επεξεργάζονται οι μηχανές. Το κάθε υποκείμενο και αντικείμενο αναγνωρίζεται βάσει ενός UniversalResourceIdentifier (URI), όπως συμβαίνει με μια ζεύξη σε μια ιστοσελίδα. (Οι URL, UniversalResourceLocators, αποτελούν την πιο κοινή μορφή URI). Τα ρήματα αναγνωρίζονται επίσης με URI, πράγμα που καθιστά ικανό τον οποιονδήποτε να ορίσει μια καινούρια έννοια, ένα καινούριο ρήμα, αναγνωρίζοντας έτσι απλά ένα URI που το αφορά κάπου στον Ιστό.

Αν και το μεγαλείο της γλώσσας των ανθρώπων φαίνεται στις περιπτώσεις όπου ο ίδιος όρος σημαίνει διαφορετικά πράγματα, το ίδιο δεν ισχύει και για τον αυτοματισμό. Το πρόβλημα μπορεί να λυθεί εάν για κάθε ξεχωριστή έννοια χρησιμοποιείται ένας διαφορετικός URI.

Τα τριπλά του RDF (RDFtriples) σχηματίζουν ιστούς πληροφοριών που αφορούν σχετιζόμενα πράγματα. Καθώς το RDF χρησιμοποιεί URI προκειμένου να κωδικοποιήσει αυτή την πληροφορία σε ένα τεκμήριο, οι URI διασφαλίζουν αυτές οι έννοιες να μην αποτελούν λέξεις και μόνο σε ένα τεκμήριο αλλά να συνδέονται πάντα με ένα μοναδικό ορισμό που να μπορεί ο κάθε ένας να βρει στον Ιστό. Εάν, π.χ., έχουμε πρόσβαση σε ποικίλες βάσεις δεδομένων με πληροφορίες για ανθρώπους, συμπεριλαμβανομένου και των διευθύνσεών τους, και θέλουμε να βρούμε μερικούς από αυτούς που διαμένουν σε μια περιοχή με τον ίδιο ταχυδρομικό τομέα, τότε πρέπει να ξέρουμε ποια πεδία σε κάθε βάση δεδομένων αναπαριστούν ονόματα και ποια ταχυδρομικούς τομείς. Το RDF μπορεί να διευκρινίσει ότι '(το πεδίο 5 στη βάση δεδομένων Α) (είναι πεδίο του τύπου) (ταχυδρομικός τομέας)', χρησιμοποιώντας URI παρά φράσεις για κάθε όρο. Ένα παράδειγμα τριπλών της RDF στη συνέχεια θα δοθεί για την εισαγωγή στις έννοιες δήλωσης της RDF και το πώς χρησιμοποιεί τις URI για να κάνει δηλώσεις για τους πόρους στον Ιστό. Κάθε δήλωση αποτελείται από το υποκείμενο, το κατηγορήμα και ένα αντικείμενο.

- ένα υποκείμενο <http://www.example.org/index.html>
- ένα κατηγορήμα <http://purl.org/dc/elements/1.1/creator>
- και ένα αντικείμενο <http://www.example.org/staffid/85740>

<http://www.example.org/index.html> έχει ένα **creator** που η αξία είναι **JohnSmith**

Στο επόμενο γράφημα αναφέρεται η χρήση των τριπλών

<http://www.example.org/index.html> έχει μια **creation-date** που η τιμή είναι **August 16, 1999**

<http://www.example.org/index.html> έχει μια **language** που η τιμή είναι **English**

Στην περίπτωση των τριπλών έχουμε το εξής αποτέλεσμα:

1. [<http://www.example.org/index.html>](http://www.example.org/index.html)[<http://purl.org/dc/elements/1.1/creator>](http://purl.org/dc/elements/1.1/creator)

[<http://www.example.org/staffid/85740>](http://www.example.org/staffid/85740) .

2. [<http://www.example.org/index.html>](http://www.example.org/index.html)[<http://www.example.org/terms/creation-date>](http://www.example.org/terms/creation-date)

"August 16, 1999".

3. [<http://www.example.org/index.html>](http://www.example.org/index.html)[<http://purl.org/dc/elements/1.1/language>](http://purl.org/dc/elements/1.1/language) "en" .

Στην τριπλή σημειογραφία κάθε δήλωσης στο γράφημα γράφεται ως απλή τριπλότητα του υποκειμένου, του κατηγορήματος και του αντικειμένου. Κάθε τριπλότητα αντιστοιχεί σε ένα απλό ακροδέκτη στο γράφημα. Έτσι η διεύθυνση <http://www.example.org/index.html> εμφανίζεται τρεις φορές, μια φορά σε κάθε τριπλή. Οι τριπλές εκπροσωπούν ακριβώς την ίδια πληροφορία. Η θεμελιώδης αρχή της RDF είναι αυτό το μοντέλο της γραφικής παράστασης.

Αυτό, όμως, δεν αρκεί. Δύο βάσεις δεδομένων μπορεί να χρησιμοποιούν διαφορετικά αναγνωριστικά σύμβολα (identifiers) κυριολεκτικά για την ίδια έννοια όπως ο *ταχυδρομικός τομέας*. Ένα πρόγραμμα που θέλει να συγκρίνει ή να συνδυάσει πληροφορίες από αυτές τις δύο βάσεις οφείλει να γνωρίζει ότι αυτοί οι δύο όροι χρησιμοποιούνται για το ίδιο πράγμα. Υπό ιδανικές συνθήκες, το πρόγραμμα θα πρέπει να είναι σε θέση να ανακαλύπτει αυτές τις κοινές σημασίες σε οποιαδήποτε βάση δεδομένων συναντά. Σε αυτό το πρόβλημα μπορεί να δοθεί λύση με τη βοήθεια του τρίτου βασικού συστατικού μέρους του Σημασιολογικού Ιστού: συλλογές πληροφοριών που τις αποκαλούμε οντολογίες.

2.6 Σημασιολογικός Ιστός

Ο TimBerners-Lee εμπνευστής του σημασιολογικού ιστού σε άρθρο του στο περιοδικό ScientificAmerican αναφέρεται σε ένα παράδειγμα από την καθημερινότητα του μέλλοντος (αναζήτηση ενός φυσιοθεραπευτή στην πλησιέστερη κλινική και σε ώρες και ημέρες που να βολεύουν τόσο την άρρωστη μητέρα της Lucy και του Peter όσο και τους ίδιους), όπου ο καθένας από εμάς θα μπορεί με τη βοήθεια ενός φορητού

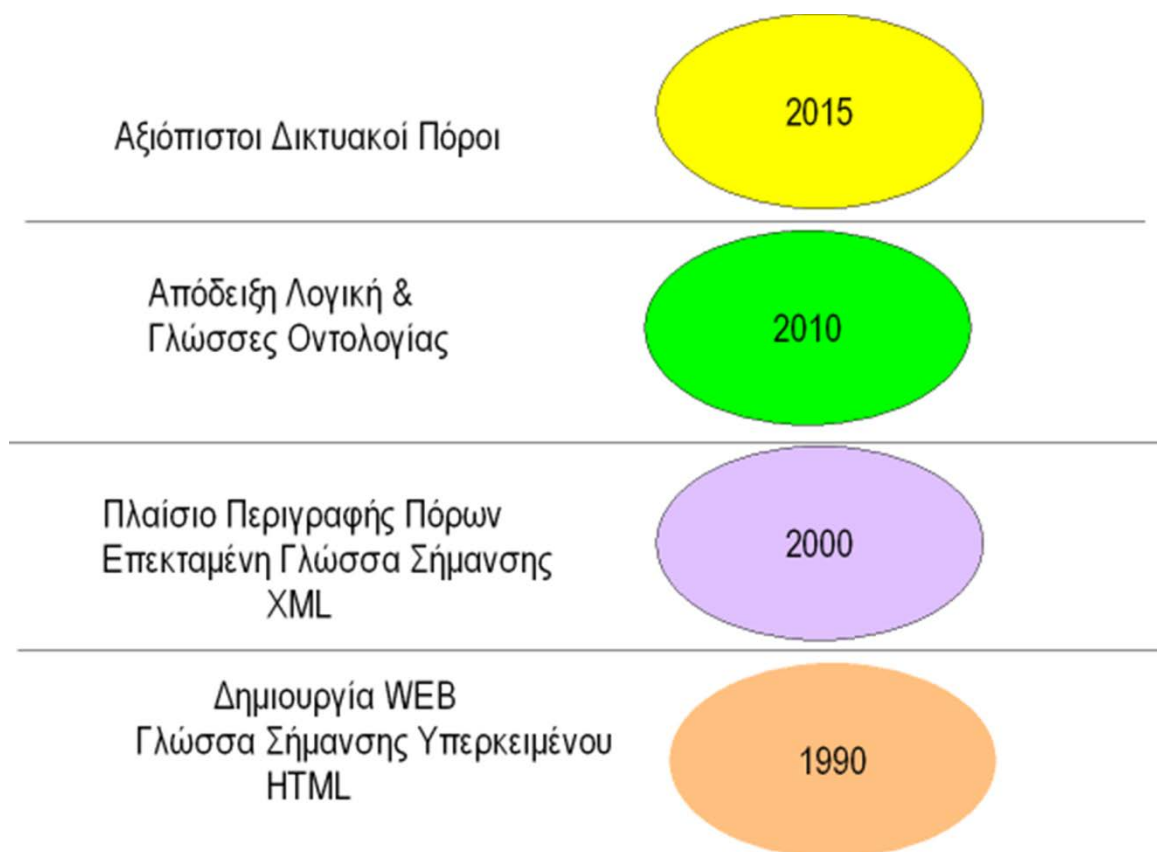
φυλλομετρητή Ιστού να κάνει χρήση των δυνατοτήτων που προσφέρει ένας διαμεσολαβητής Σημασιολογικού Ιστού (SemanticWebagent). Μέσω αυτού του διαμεσολαβητή θα επικοινωνούμε με άλλους διαμεσολαβητές Σημασιολογικού Ιστού προκειμένου να ανακτηθούν οι πληροφορίες που επιθυμούμε. Αυτές οι πληροφορίες θα διατίθενται από τους διαμεσολαβητές μεμονωμένων παροχών μέσω των ιστοσελίδων τους. Στο παραπάνω παράδειγμα, οι πληροφορίες ανακτήθηκαν όχι χάρη στον παγκόσμιο Ιστό του σήμερα αλλά χάρη στο Σημασιολογικό Ιστό του αύριο. Το μεγαλύτερο μέρος του Ιστού σήμερα έχει σχεδιαστεί έτσι ώστε να είναι αναγνώσιμο από τους ανθρώπους κι όχι για να βγάζουν νόημα και να το χειρίζονται οι ίδιοι οι Η/Υ. Ναι μεν μπορούν άνετα να αναλύουν την κατάσταση (layout) των ιστοσελίδων και να κάνουν επεξεργασίες ρουτίνας (μια κεφαλίδα εδώ και ένας σύνδεσμος εκεί) αλλά γενικά δεν μπορούν να επεξεργαστούν τις πληροφορίες σημασιολογικά.

Ο Σημασιολογικός Ιστός θα μπορέσει να δομήσει το εννοιολογικό περιεχόμενο των ιστοσελίδων, αφού δημιουργήσει ένα περιβάλλον όπου οι διαμεσολαβητές λογισμικού θα περιάγονται από σελίδα σε σελίδα και θα μπορούν με ευκολία να εκτελούν τις περίπλοκες αναζητήσεις που τους αναθέτουν οι χρήστες. Έτσι, αν θυμηθούμε το παράδειγμα ανωτέρω, θα δούμε ότι τα παιδιά βρήκαν αυτό που έψαχναν όχι επειδή χρησιμοποίησαν λέξεις κλειδιά ('θεραπευτική αγωγή, ιατρική, φυσιοθεραπεία') όπως γίνεται σήμερα, αλλά επειδή ο διαμεσολαβητής που επισκέφτηκε την ιστοσελίδα της συγκεκριμένης κλινικής ήξερε ότι ο τάδε γιατρός δουλεύει σε αυτήν την κλινική κάθε Δευτέρα, Πέμπτη και Παρασκευή και ότι το σενάριο διαθέτει το μορφότυπο έτος-μήνας-ημέρα ώστε να μπορεί να μας ενημερώνει για τις ώρες επισκέψεων του γιατρού. Όλα αυτά θα τα 'γνωρίζει' χωρίς να χρειάζεται τεχνητή νοημοσύνη στην κλίμακα του Hal 2001 ή του StarWarsC-3PO. Αντί γι' αυτό, το σημασιολογικό αυτό περιεχόμενο ενσωματώθηκε ως κώδικας στην ιστοσελίδα της κλινικής από τον υπεύθυνο (χωρίς βέβαια να έχει παρακολουθήσει ποτέ στο πανεπιστήμιο το μάθημα CompSci 101) ως εξής: το διαμόρφωσε στις ανάγκες του χρησιμοποιώντας ένα απλό λογισμικό του εμπορίου για τη σύνταξη σελίδων Σημασιολογικού Ιστού ενώ παράλληλα έκανε χρήση των πόρων που ήταν καταλογογραφημένοι στην ιστοσελίδα της Ένωσης Φυσιοθεραπευτών. Ο Σημασιολογικός Ιστός αποτελεί προέκταση του παγκόσμιου Ιστού, όπου δίνεται στις πληροφορίες μια καλά οριοθετημένη σημασία για καλύτερη συνεργασία του ανθρώπου με τους Η/Υ. Το κύριο χαρακτηριστικό του παγκόσμιου Ιστού είναι η παγκόσμια εμβέλεια του: μια υπερκειμενική ζεύξη (hypertextlink) έχει τη δύναμη να συνδέει τα πάντα μεταξύ τους. Άρα, η τεχνολογία του Ιστού θα πρέπει τώρα πια να αρχίσει να διαχωρίζει τις πληροφορίες με βάση άλλους άξονες· εάν, δηλ., οι πληροφορίες αυτές προορίζονται κυρίως για κατανάλωση από το ανθρώπινο δυναμικό ή από άλλες μηχανές. Ο Σημασιολογικός Ιστός σήμερα θα πρέπει να χρησιμοποιείται ως μέσο αυτόματης επεξεργασίας δεδομένων και πληροφοριών παρά ως μέσο πρόσβασης σε αρχεία και έγγραφα από το κοινό. Όπως το Διαδίκτυο, έτσι και ο Σημασιολογικός Ιστός θα αποκεντριοποιηθεί όσο το δυνατό περισσότερο. Παρόμοια συστήματα με αυτό του Ιστού αποτελούν πρόκληση για κάθε τομέα και ενδιαφερόμενο, προσκομίζοντας οφέλη που είναι αδύνατον να προβλεφθούν. Ωστόσο, η αποκεντριοποίηση προϋποθέτει συμβιβασμούς: ο Ιστός έπρεπε να απορρίψει το ιδανικό της απόλυτης συνοχής όλων των διασυνδέσεων του, κάτι που οδηγούσε στο περιβόητο 'Error 404: Not Found' αλλά επέτρεπε τη σταδιακή ανάπτυξη (βλ Σχήμα 2.4).

2.7 Υποδομή Σημασιολογικού Ιστού

Το Διαδίκτυο και ο Ιστός έχουν προκαλέσει επανάσταση στην τεχνολογία της πληροφορικής καθώς και στην καθημερινή ζωή των περισσότερων ανθρώπων. Ωστόσο, οι περισσότερες σύγχρονες μορφές περιεχομένου στο web έχουν σχεδιαστεί και είναι δομημένες για να χρησιμοποιούνται από ανθρώπους και μόλις που γίνονται κατανοητές από τους υπολογιστές. Ο στόχος του Σημασιολογικού Ιστού είναι, όπως το είχε οραματιστεί ο Berners-Lee, να δημιουργηθούν γλώσσες επικοινωνίας προκειμένου να περιγράφονται οι πληροφορίες με τρόπο κατανοητό στις μηχανές. Η XML (ExtensibleMark-upLanguage) έχει προσδώσει τεράστια χαρακτηριστικά γνώρισμα και ελπιδοφόρες προοπτικές για την εξέλιξη του Σημασιολογικού Ιστού. Επί του παρόντος υπάρχουν πολυάριθμες τεχνικές και εργαλεία διαθέσιμα για XML, e.g., SAX (Simple API for XML), DOM (Document Object Model), XSL (Extensible Stylesheet Language), XSLT (XSL Transformation), XPath, XLink, and XPointer, and XML κατατμητές είναι διαθέσιμα για διαφορετικές πλατφόρμες. Χρησιμοποιώντας την XML μπορεί κάποιος να περιγράψει ειδή κειμένων / άρθρων / αρχείων σε διάφορα πεδία γνώσης και για διάφορους σκοπούς. Για παράδειγμα, αρχεία XML μπορεί να ενσωματώνουν παρουσιάσεις με πολυμέσα καθώς και επαγγελματικές / εμπορικές συναλλαγές (XML/EDI-

Group). Εφαρμογές μπορούν να έχουν πρόσβαση σε αρχεία XML μέσω συγκεκριμένων διεπαφών όπως SAX and DOM. Ένας αριθμός γλώσσων διατύπωσης ερωτήσεων XML έχουν προταθεί συμπεριλαμβάνοντας τις XML-QL, X-QL, Lorel, XQuery.



Σχήμα 2.4. Η χρονολογική εξέλιξη του Ιστού

Περαιτέρω μερικοί ερευνητές προτείνουν αυτές τις γλώσσες διατύπωσης ερωτήσεων να επεκταθούν με μια αναβαθμισμένη ικανότητα ώστε ένα αποθετήριο XML αρχείων να μετατραπεί σε βάση δεδομένων XML. Η XML θα συνεχίσει να παίζει σημαντικό ρόλο στην εξέλιξη του Σημασιολογικού Ιστού. Ωστόσο δεν προσφέρει ολοκληρωμένη λύση στις απαιτήσεις του Σημασιολογικού Ιστού. Η XML μπορεί να παρουσιάζει μόνο κάποιες σημασιολογικές ιδιότητες μέσω της συντακτικής της δομής π.χ. με την εμφώλευση ή την διαδοχική διάταξη σχέσης μεταξύ των στοιχείων (XML ετικέτες). Οι αναζητήσεις μέσω XML πρέπει να λαμβάνουν υπόψη αυτή τη συντακτική δομή μέσω του είδους κειμένου το οποίο ορίζεται με ένα DTD. Παρόλο που κάποιος θα μπορούσε να εξάγει κάποιο είδος σημασίας από τη δομή των κειμένων μέσα στο πλαίσιο του είδους κειμένου, η σημασία κάθε στοιχείου (XML ετικέτας) δεν προσδιορίζεται και η ερμηνεία του στηρίζεται εξ ολοκλήρου στην υπονοούμενη γνώση σε προγράμματα εφαρμογών δουλεμένα χειροποίητα (hardcoded). Το να δημιουργήσεις έναν Ιστό που βασίζεται στη σημασιολογία, οι πόροι του οποίου χρειάζεται να παρουσιαστούν ή να σχολιαστούν με δομημένες περιγραφές και του περιεχομένου τους και των σχέσεών τους και να είναι κατανοητές από τις μηχανές, χρησιμοποιώντας λεξιλόγια και κατασκευές που να έχουν ρητά και τυπικά καθοριστεί από μια οντολογία περιοχής γνώσης. Η θεώρηση του κόσμου που αντιπροσωπεύει μια οντολογία συνήθως θεωρείται ως μια ιεραρχική περιγραφή μιας σειράς εννοιών, αρχών, ιδιοτήτων και των μεταξύ τους σχέσεων και μιας σειράς συμπερασματικών κανόνων. Ο Berners-Lee περιέγραψε επιγραμματικά την αρχιτεκτονική του Σημασιολογικού Ιστού με τα ακόλουθα 3 επίπεδα:

The metadatalayer. Το metadatalamodel περιέχει σε αυτό το επίπεδο μόνο τις έννοιες των πόρων και των ιδιοτήτων. Επί του παρόντος το RDF (ResourceDescriptionFramework) θεωρείται το πιο δημοφιλές μοντέλο δεδομένων για το επίπεδο μεταδεδομένων (metadatalayer).

The schemalayer. Οι Οντολογικές γλώσσες Ιστού εισάγονται σε αυτό το επίπεδο για να ορίσουν μια ιεραρχική περιγραφή των αρχών (is-a hierarchy) και των ιδιοτήτων. Επί του παρόντος το RDFS (RDFS Schema) αποτελεί γλώσσα σε επίπεδο Σχήμα (schemalayer).

Theologicallayer. Σε αυτό το επίπεδο εισάγονται πιο δυναμικές οντολογικές γλώσσες Ιστού. Αυτές οι γλώσσες παρέχουν ένα πιο πλούσιο σύνολο αρχικών μοντέλων που μπορούν να χαρτογραφηθούν σε Περιγραφική Λογική (DescriptionLogics) Παράδειγμα γλωσσών σ' αυτό το επίπεδο αποτελούν η OIL (Ontology Inference Layer) και η DAML-OIL (Darpa Agent Markup Language-Ontology Inference Layer) ως δημοφιλείς γλώσσες του Λογικού επιπέδου (logical layer). Με τη δημιουργία του Σημασιολογικού Ιστού διάφοροι πόροι Ιστού θα είναι προσβάσιμοι από τις μηχανές με τρόπο σημασιολογικό. Τα ερωτήματα που προκύπτουν είναι, τι ευκαιρίες θα μας φέρει η καινούρια τεχνολογία και τι προκλήσεις και δουλειά μας περιμένει για να μεταβούμε από τον Ιστό του σήμερα στον Ιστό του αύριο.

2.8 Υπηρεσίες Σημασιολογικού Ιστού

Όπως και με άλλες τεχνολογίες, το ενδιαφέρον για τη δημιουργία και ανάπτυξη του Σημασιολογικού Ιστού προέρχεται από τις ευκαιρίες που αυτό μπορεί να δημιουργήσει: μπορεί είτε να λύσει νέα προβλήματα είτε να λύσει τα παλιά με καλύτερο τρόπο. Εδώ, αντί να απαριθμήσουμε όλες τις ευκαιρίες που ανοίγονται με τον Σημασιολογικό Ιστό θα εστιάσουμε τη συζήτηση στις παρακάτω στενά συνδεδεμένες σχετικές προοπτικές όπως: υπηρεσίες Ιστού, καταναεμημένους διαμεσολαβητές βασιζόμενους σε εκτέλεση υπολογισμών (agent-based distributed computing), μηχανές αναζήτησης βασιζόμενες στη σημασιολογία Ιστού (semantics-based web), και Ψηφιακές Βιβλιοθήκες βασιζόμενες στη σημασιολογία. Μεταξύ των σημαντικότερων πόρων στον Ιστό του Σημασιολογικού Ιστού είναι οι λεγόμενες υπηρεσίες Ιστού. Εδώ, με τον όρο υπηρεσίες Ιστού αναφερόμαστε σε Ιστοτόπους που δεν παρέχουν απλά στατικές πληροφορίες αλλά επιτρέπουν σε κάποιον να ενεργήσει και να προκαλέσει αλλαγές στον κόσμο. Ο Σημασιολογικός Ιστός θα δώσει τη δυνατότητα στους χρήστες να εντοπίζουν, συνθέτουν, και ελέγχουν τις υπηρεσίες του Ιστού αυτόματα.

2.8.1 Αυτόματη ανακάλυψη Υπηρεσιών Ιστού

Καθώς οι σημασιολογικές περιγραφές των υπηρεσιών Ιστού είναι καταχωρημένες σε ορισμένους δημόσιους καταλόγους, οι ευφυείς φορητοί διαμεσολαβητές θα μπορούν να μεταφέρονται από μια υπηρεσία καταχώρησης σε άλλη προκειμένου να βρουν τις επιθυμητές υπηρεσίες στον Ιστό όπως αυτές ορίζονται από τον χρήστη. Έχει γίνει κάποια εργασία σχετικά με τη διαδικασία αναζήτησης με τη βοήθεια διαμεσολαβητών. Π.χ. το πρόγραμμα Letizia είναι ένας διαμεσολαβητής για χρήστες που βοηθά ένα χρήστη που ψάχνει στο WWW ανάλογα με το προφίλ του χρήστη. Το Letizia μπορεί να μάθει το προφίλ ενός χρήστη βασιζόμενο στο ιστορικό αναζητήσεων του χρήστη. Αυτή η ιδέα μπορεί να εφαρμοστεί και στην υπηρεσία ανακάλυψης Ιστού. Π.χ. ένας εξυπηρετητής μπορεί να παρέχει υπηρεσίες Ιστού σε διαφορετικά επίπεδα ή με διαφορετικές παραμέτρους ως προς διάφορους χρήστες και ένας διαμεσολαβητής θα μπορεί να βρει μια υπηρεσία Ιστού βασιζόμενος στο προφίλ του χρήστη.

Σήμερα οι περισσότερες υπηρεσίες Ιστού απαιτούν την ανθρώπινη παρέμβαση κατά την εφαρμογή τους. Π.χ. για να αγοραστεί ένα βιβλίο από <http://www.greekbooks.gr/> ο ιστότοπος ζητά από το χρήστη να συμπληρώσει μια φόρμα και να πατήσει ένα κουμπί για να ενεργοποιήσει την υπηρεσία. Συνήθως πολλαπλές συναλλαγές μεταξύ χρήστη και υπηρεσίας Ιστού είναι απαραίτητες προκειμένου να ενεργοποιηθεί η υπηρεσία. Η δυνατότητα αυτόματης κλήσης υπηρεσίας Ιστού συνεπάγεται ότι το λογισμικό του διαμεσολαβητή θα μπορεί να πραγματοποιήσει αυτές τις κλήσεις για λογαριασμό του χρήστη ο οποίος χρειάζεται μόνο να πει στο μεσάζοντα "πήγαινε στο greekbooks και αγόρασε ένα βιβλίο με τίτλο ο Σημασιολογικός Ιστός αξίας 50 €", και ο διαμεσολαβητής θα επικοινωνήσει με τη συγκεκριμένη υπηρεσία Ιστού με κατάλληλα εισαγόμενα δεδομένα μέσω API (Application Program Interfaces) κατανοητά από τον υπολογιστή.

Αυτόματη σύσταση υπηρεσίας Ιστού και διαλειτουργικότητας.

Δεδομένης μιας βιβλιοθήκης υπηρεσιών Ιστού και με ένα στόχο κάποιος θα μπορούσε αυτόματα να επιλέξει και να συνθέσει μια νέα υπηρεσία Ιστού προκειμένου να πετύχει ένα νέο στόχο. Οι προϋπόθεσης και οι συνέπειες των διάφορων υπηρεσιών Ιστού πρέπει να περιγράφουν με επίσημο τρόπο και η τεχνολογία για αυτόματη παραγωγή ροή εργασίας θα μπορούσε να χρησιμοποιηθεί για να παράγει υπηρεσίες Ιστού αυτόματα.

Αυτόματος έλεγχος διεξαγωγής υπηρεσιών Ιστού.

Για μακράς διάρκειας πολύπλοκες υπηρεσίες Ιστού είναι επιθυμητό να μπορεί κάποιος να παρακολουθήσει και να εξετάσει την κατάσταση της εφαρμογής κάθε υπηρεσίας καθώς και των επιμέρους μηχανισμών της. Η βιομηχανία έχει ήδη δει τη δυναμική της αγοράς που δημιουργείται από τις υπηρεσίες Ιστού και κάποιες προσπάθειες έχουν επενδυθεί στην ανάπτυξη των προδιαγραφών για το ηλεκτρονικό

εμπόριο, συγκεκριμένα για την περιγραφή των υπηρεσιών Ιστού. Για παράδειγμα η Microsoft, IBM η Ariba προτείνουν το UDDI (Universal Description, Discovery, and Integration) να περιγράψει ένα πρότυπο για online εγγραφή, επίσης η Microsoft και η IBM προτείνουν το WSDL (Web Service Definition Language) ως μια XML γλώσσα για να περιγράψουν τις διαπαφές σε καταραμμένες υπηρεσίες Ιστού σε μια βάση UDDI. Η DAML Services Coalition προτείνει την DAML-S (Darpa Agent Markup Language - Service) ως μια οντολογία για να περιγράψει τις υπηρεσίες Ιστού. Τέλος η OASIS και τα Ηνωμένα Έθνη ανέπτυξαν την ebXML (Electronic Business XML Initiative) για να περιγράψουν τις επιχειρησιακές συναλλαγές από μια άποψη ροής εργασίας. Ένας αριθμός πρωτοκόλλων επικοινωνίας έχουν αναπτυχθεί για την κλήση υπηρεσιών Ιστού ως ακολούθως:

Το Remote Procedure Call αποτελεί μια υποδομή πελάτη / εξυπηρετητή που επιτρέπει ένα τμήμα πελατών μιας εφαρμογής να χρησιμοποιήσει μια λειτουργική κλήση προκειμένου να έχει πρόσβαση σε έναν εξυπηρετητή σε ένα απομακρυσμένο σύστημα. Ο μηχανισμός CGI (Common Gateway Interface) είναι ένα μοντέλο για προγράμματα πρόσβασης σε εξυπηρετητές πληροφοριών όπως οι HTTP. Το CORBA (Common Object Request Broker Architecture) χρησιμοποιεί έναν κατάλογο για να αποθηκεύει διαπαφές των κατανεμημένων αντικειμένων έτσι ώστε ένας πελάτης να μπορεί να χρησιμοποιήσει μια μέθοδο από ένα απομακρυσμένο εξυπηρετητή χωρίς να γνωρίζει την τοποθεσία του, τη γλώσσα προγραμματισμού, το λειτουργικό σύστημα και άλλες πλευρές του συστήματος οι οποίες δεν αποτελούν μέρος της διαπαφής του αντικειμένου. Το SOAP (Simple Object Access Protocol) είναι ένα πρωτόκολλο για την ανταλλαγή πληροφοριών σε ένα κατανεμημένο περιβάλλον, συγκεκριμένα στον Ιστό. Με τη χρήση του SOAP μπορεί κάποιος να περιγράψει το περιεχόμενο ενός μηνύματος και τον τρόπο επεξεργασίας του, να ορίσει είδη δεδομένων εξαρτώμενων από εφαρμογές και να παρουσιάσει την απομακρυσμένη διαδικασία κλήσεων και απαντήσεων. Το SOAP έχει τη δυνατότητα να χρησιμοποιηθεί με μια ποικιλία άλλων πρωτοκόλλων. Τελευταία στηρίζεται η σύνδεση του με το HTTP. Το σύστημα JavaRMI (Java Remote Method Invocation) δίνει στους προγραμματιστές τη δυνατότητα να αναπτύξουν εφαρμογές Java στις οποίες ο πελάτης μπορεί να χρησιμοποιήσει μια μέθοδο από ένα απομακρυσμένο αντικείμενο Java. Το σύστημα ACL (Agent Communication Language) που δημιουργήθηκε από το SRI International στο πλαίσιο του OAA (Open Agent Architecture) επιτρέπει σε κάποιον να ορίσει και να δημοσιοποιήσει τις δυνατότητες των διαμεσολαβητών ώστε το αίτημα να ταιριάζει με έναν εξυπηρετητή διαμεσολαβητή από τους βοηθούς μεσίτες.

2.8.2 Το παράδειγμα του κατανεμημένου διαμεσολαβητή βασισμένου σε εκτέλεση υπολογισμών (Agent-based distributed computing paradigm)

Ο Σημασιολογικός Ιστός θα χρησιμοποιήσει οντολογίες για να περιγράψει διάφορους πόρους του Ιστού, έτσι η γνώση στον Ιστό θα παρουσιάζεται με ένα δομημένο, λογικό και σημασιολογικό τρόπο. Αυτό θα αλλάξει τον τρόπο που οι διαμεσολαβητές θα πλοηγούν, θα συλλέγουν και θα αξιοποιούν τις πληροφορίες στον Ιστό. Από τη μια πλευρά ο Σημασιολογικός Ιστός αποτελεί ένα δίκτυο κατανεμημένων βάσεων δεδομένων και οι διαμεσολαβητές μπορούν να διαβάζουν και να αναλύουν τις διαθέσιμες γνώσεις με την καθοδήγηση των οντολογιών. Από την άλλη, ο Σημασιολογικός Ιστός είναι μια συλλογή υπηρεσιών Ιστού που περιγράφονται από οντολογίες όπως η DAML-S (Darpa Agent Markup Language - Services) και αυτό θα διευκολύνει το ταίριασμα ετερογενών διαμεσολαβητών δηλαδή οι διαμεσολαβητές παροχής υπηρεσιών μπορούν να διαφημίζουν τις δυνατότητες τους στους μεσαίους διαμεσολαβητές, οι μεσαίοι διαμεσολαβητές αποθηκεύουν αυτές τις διαφημίσεις. Ένας διαμεσολαβητής για λογαριασμό κάποιου που ζητά υπηρεσίες μπορεί να ρωτήσει έναν ενδιάμεσο διαμεσολαβητή αν γνωρίζει κάποιον παροχέα με τις επιθυμητές δυνατότητες, ο ενδιάμεσος διαμεσολαβητής συγκρίνει την εντολή με τις αποθηκευμένες διαφημίσεις και στέλνει το αποτέλεσμα, δηλαδή ένα υποσύνολο των αποθηκευμένων διαφημίσεων. Όταν οι διαμεσολαβητές είναι εφοδιασμένοι με ευφυΐα και κινητικότητα, η συμβατική σχέση πελάτη / εξυπηρετητή θα μπορούσε να αντικατασταθεί με μια σχέση βασισμένη σε διαμεσολαβητές στην οποία οι διαμεσολαβητές θα μπορούν να μετακινούνται από τον έναν ισότοπο σε άλλο μεταφέροντας τους κωδικούς τους, δεδομένα, την τρέχουσα κατάσταση και ευφυΐα (καθορισμένη από τους χρήστες) και να εκπληρώνουν τις αποστολές τους αυτόνομα και έξυπνα. Πολλοί ερευνητές έχουν παρατηρήσει ότι οι μετακινούμενοι διαμεσολαβητές είναι αναπόφευκτοι για ένα ανοικτό και κατανεμημένο περιβάλλον όπως αυτό του Σημασιολογικού Ιστού και αναγνωρίζουν σε αυτήν τη νέα κατάσταση στον τομέα της υπολογιστικής τεχνολογίας τα ακόλουθα πλεονεκτήματα;

- Μειώνεται το φορτίο του δικτύου
- Υπερνικά τη λανθάνουσα κατάσταση δικτύων
- Υποστηρίζει την ενσωμάτωση των πρωτοκόλλων

- Οι διαμεσολαβητές διεκπεραιώνουν ασύγχρονα και αυτόνομα
- Οι διαμεσολαβητές προσαρμόζουν δυναμικά
- Οι διαμεσολαβητές είναι φυσιολογικά ετερογενείς.
- Οι διαμεσολαβητές είναι ανθεκτικοί και ανεκτικοί στα λάθη.

2.8.3 Μηχανές αναζήτησης Ιστού που βασίζονται στη σημασιολογία

Οι μηχανές αναζήτησης είναι από τους πιο χρήσιμους πόρους του Ιστού και σήμερα υπάρχουν δυο είδη:

Μηχανές αναζήτησης μεγάλης κλίμακας που βασίζονται σε ρομπότ. Αυτά τα συστήματα βασίζονται σε ρομπότ για να αντλούν και να αποθηκεύουν ιστοσελίδες σε μια κεντρική βάση δεδομένων. Το θετικό αυτού του μηχανισμού είναι ότι αυξάνει την ανάκληση (ποσότητα σχετικών αρχείων που εντοπίζονται) καθώς τα ρομπότ μπορούν να αντλήσουν σχεδόν όλες τις σελίδες του Ιστού, ενώ το αρνητικό είναι ότι η ακρίβεια του αποτελέσματος της αναζήτησης (ποσότητα ευρισκόμενων αρχείων που όντως είναι σχετικά) μπορεί να είναι χαμηλή.

Μηχανές αναζήτησης μικρής κλίμακας που βασίζονται στην διερεύνηση. Γίνεται ιεράρχηση ανά κατηγορία και κάθε κατηγορία περιγράφεται από μια σειρά λέξεων κλειδιά. Οι 'αναζητητές' θα ψάξουν κάθε ιστοσελίδα (κατατεθειμένη από δημιουργούς ιστοσελίδων) και θα τη συσχετίσουν με την ανάλογη κατηγορία. Το θετικό είναι ότι αυξάνεται η ακρίβεια αλλά το αρνητικό είναι ότι η ανεύρεση μπορεί να είναι χαμηλή καθώς είναι αδύνατο να διερευνηθούν και να συμπεριληφθούν όλες οι σχετικές ιστοσελίδες. Και οι δυο τύποι των μηχανών αναζήτησης βασίζονται σε λέξεις κλειδιά και συνεπώς υπόκεινται στα πολύ γνωστά γλωσσολογικά φαινόμενα τα οποία μειώνουν κατά πολύ την ακρίβεια αναζήτησης: πολυσημία (μια λέξη με πολλές σημασίες) και συνωνυμία (πολλοί όροι, λέξεις ή φράσεις που μπορεί να ορίζουν την ίδια έννοια). Μια σειρά αλγορίθμων θέματος λέξεων (stem) έχουν αναπτυχθεί προκειμένου να αντιμετωπιστεί το πρόβλημα της συνωνυμίας, όπως απαλοιφή καταλήξεων, αυστηρή αποκοπή της σειράς των χαρακτήρων, χωρισμός λέξεων, δι-γραμμάτων λέξεων και γλωσσολογική μορφολογία. Η λογική είναι ότι τα διάφορα παράγωγα μιας λέξης είναι ίδια μεταξύ τους ως προς τη μορφή (π.χ. έχουν ίδιο πρόθεμα) και μπορούν να αναζητηθούν στην ίδια ρίζα χρησιμοποιώντας τις μεθόδους παραγωγής / προέλευσης. Ωστόσο αυτές οι μέθοδοι υπόκεινται στα ακόλουθα λάθη: λέξεις με διαφορετική σημασία μπορεί να οδηγηθούν / αναζητηθούν στην ίδια ρίζα. Π.χ. οι λέξεις *general*, *generous*, *generation*, και *generic*. Από την άλλη διαφορετικές λέξεις με την ίδια σημασία δεν μπορούν να συνδεθούν με την ίδια ρίζα. Π.χ. *Automobile* και *car*. Η κατάσταση χειροτερεύει στις μεγάλης κλίμακας μηχανές αναζήτησης που ελέγχονται από ρομπότ. Μόνο περιορισμένη σημασιολογία μπορεί να ανακληθεί από το λεξιλογικό και συντακτικό περιεχόμενο των ιστοσελίδων. Πολλά συστήματα έχουν κατασκευαστεί για να ξεπεράσουν αυτά τα προβλήματα βασισμένα στην ιδέα του να αναφέρονται σε ιστοσελίδες με ειδικές ετικέτες HTML για να παρουσιάσουν σημασιολογία συμπεριλαμβανόμενων του SHOE (SimpleHTMLOntologyExtensions) συστήματος και του GDA συστήματος.

Ωστόσο ο περιορισμός αυτών των συστημάτων είναι ότι μπορούν να επεξεργαστούν μόνο σελίδες που αναφέρονται σε HTML ετικέτες και ως τώρα δεν υπάρχει συμφωνία για μια σειρά διεθνώς αποδεκτών HTML ετικετών. Η XML είναι μια πολλά υποσχόμενη τεχνική καθώς διατηρεί περιεχόμενο, δομή και αναπαράσταση ξεχωριστά και είναι ένα πιο επαρκές μέσο αναπαράστασης γνώσεων. Ωστόσο η XML μπορεί να αναπαριστά μόνο μερικά σημασιολογικά χαρακτηριστικά μέσω της συντακτικής της δομής. Οι XML αναζητητές πρέπει να λαμβάνουν υπόψη αυτή τη συντακτική δομή. Με την εμφάνιση του Σημασιολογικού Ιστού, οι πόροι στο διαδίκτυο θα αναπαριστώνται σημασιολογικά σε οντολογίες. Μηχανές αναζήτησης βασισμένες στη σημασιολογία μπορούν να δημιουργηθούν μέσα στις οποίες κάθε αναζήτηση διεκπεραιώνεται στο πλαίσιο κάποιας οντολογίας.

Η καθοδήγηση μέσω οντολογιών θα αυξήσει την ανάκληση και την ακρίβεια του αποτελέσματος της αναζήτησης. Π.χ. κάποιος μπορεί να θέσει το ερώτημα "βρες μου όσους διερευνούν το βιβλίο "ο Σημασιολογικός Ιστός : εισαγωγή" σε μια μηχανή αναζήτησης βασισμένη στη σημασιολογία, τότε η μηχανή θα αναζητήσει μόνο τους διερευνητές για αυτό το βιβλίο αντί για ιστοσελίδες που περιέχουν τη λέξη κλειδί "διερευνητής" και / ή τον όρο "Ο Σημασιολογικός Ιστός : εισαγωγή". Ένα άλλο παράδειγμα αν κάποιος θέσει το ερώτημα "βρες μου όλες τις καρέκλες" με την καθοδήγηση κάποιας οντολογίας επίπλων θα αναζητηθούν μόνο καρέκλες (έπιπλα), και με την καθοδήγηση μιας οντολογίας για ανθρώπους θα αναζητηθούν μόνο άτομα που κατέχουν καρέκλες (αξιώματα) κάποιων οργανισμών. Αντιθέτως, μηχανές αναζήτησης που βασίζονται σε λέξεις κλειδιά θα αναζητήσουν ιστοσελίδες που περιέχουν τη λέξη κλειδί "καρέκλες", συμπεριλαμβανόμενων και των επίπλων και των αξιωμάτων που κατέχονται από ανθρώπους. Αξίζει να σημειωθεί ότι έχουν αναπτυχθεί μερικά συστήματα που χρησιμοποιούν οντολογίες για να βελτιώσουν τις μηχανές αναζήτησης

στον Ιστό. Εφόσον οι οντολογίες είναι κατασκευασμένες με βάση κάποιο τομέα, και οι μηχανές αναζήτησης Ιστού μπορούν να κατασκευαστούν με βάση έναν τομέα, και ως εκ τούτου οι μηχανές μετά-αναζήτησης των οποίων οι δι-επαφές σχετίζονται με πολλαπλές απομακρυσμένες μηχανές αναζήτησης και επιλέγουν και κατατάσσουν τις απομακρυσμένες μηχανές αναζήτησης με έξυπνο τρόπο, μπορεί να αποδειχθούν πολύ χρήσιμες.

2.8.4 Ψηφιακές Βιβλιοθήκες βασισμένες στη σημασιολογία

Τα Ψηφιακά πολυμεσικά δεδομένα σε διάφορα μορφότυπα έχουν αυξηθεί σε πολύ μεγάλο βαθμό. Με την εξέλιξη της ψηφιακής φωτογραφίας ολοένα και περισσότεροι άνθρωποι μπορούν να αποθηκεύουν φωτογραφίες στους Η/Υ τους. Το να μοιράζεται κάποιος άλμπουμ φωτογραφιών και homevideos στο Διαδίκτυο γίνεται ολοένα και πιο δημοφιλές. Επιπλέον πολλοί οργανισμοί έχουν μεγάλες συλλογές εικόνων και video σε ψηφιακή μορφή διαθέσιμες για πρόσβαση online. Οι παραγωγοί ταινιών θέλουν να διαφημίζουν ταινίες μέσω διαδραστικών κλιπ προεπισκόπησης. Τα ταξιδιωτικά γραφεία ενδιαφέρονται για ψηφιακά αρχεία φωτογραφιών από τουριστικά θέρετρα. Τα νοσοκομεία θα ήθελαν να κατασκευάσουν βάσεις δεδομένων με εικόνες ιατρικών περιστατικών. Αυτές οι νεοεμφανιζόμενες εφαρμογές για πολυμεσικές Ψηφιακές Βιβλιοθήκες δεδομένων απαιτούν διεπιστημονική αναζήτηση στους τομείς της επεξεργασίας εικόνων, computervision, ανάκτησης πληροφοριών, και διαχείρισης βάσεων δεδομένων. Η βασισμένη στη σημασιολογία άντληση πολυμεσικού ψηφιακού περιεχομένου είναι σημαντική για επαρκή χρήση των αποθετηρίων πολυμεσικών δεδομένων. Οι παραδοσιακές βασισμένες στο περιεχόμενο πολυμεσικές τεχνικές ανάκτησης περιγράφουν εικόνες/videos με βάση χαμηλού επιπέδου χαρακτηριστικά (όπως χρώμα, υφή, μορφή και σχήμα) και υποστηρίζουν την ανάκτηση με βάση αυτά τα χαρακτηριστικά. Ωστόσο ο άνθρωπος δεν βλέπει τις εικόνες/ videos με βάση χαμηλού επιπέδου χαρακτηριστικά. Η ικανότητα αναζήτησης βασισμένη στη σημασιολογία είναι ιδιαίτερα επιθυμητή. Π.χ μπορεί κάποιος να διαμορφώσει μια αναζήτηση όπως "ανάκτησέ μου όλες τις σκηνές / στιγμιότυπα σε ένα video-clip στο οποίο ένα αγόρι κάνει ποδήλατο. Η ανάκτηση εικόνων/videos που βασίζεται σε χαρακτηριστικά χαμηλού επιπέδου δεν μπορεί να προσφέρει ικανοποιητικά αποτελέσματα. Η αποτελεσματική και ακριβής πολυμεσική ανάκτηση βασισμένη στη σημασιολογία παραμένει ένα ανοικτό και γεμάτο πρόκληση πρόβλημα. Πρόσφατα άρχισαν να χρησιμοποιούνται οντολογίες στο πλαίσιο ψηφιακών βιβλιοθηκών. Π.χ η ScholOnto είναι μια Ψηφιακή Βιβλιοθήκη βασισμένη σε οντολογία η οποία υποστηρίζει την ακαδημαϊκή ερμηνεία και εκφορά λόγου και η ARION, μια άλλη η οποία υποστηρίζει την έρευνα και την πλοήγηση συνόλων γεωκινητικών δεδομένων και περιβαλλοντικών εφαρμογών. Πιστεύουμε ότι οι διάφορες Ψηφιακές Βιβλιοθήκες θα αποτελέσουν μια ακόμα πολλή μεγάλη δικτυακή πηγή του Σημασιολογικού Ιστού. Οι προκλήσεις είναι εδώ: 1) Η ανάπτυξη ενός αποδοτικού και αποτελεσματικού μηχανισμού κατηγοριοποίησης και αρχειοθέτησης για κάθε είδος Ψηφιακή Βιβλιοθήκη και 2) Η σημασιολογική διαλειτουργικότητα μεταξύ Ψηφιακών Βιβλιοθηκών ίδιου τύπου και μεταξύ μιας Ψηφιακής Βιβλιοθήκης και του Σημασιολογικού Ιστού.

2.8.5 Νέες προκλήσεις και προοπτικές εφαρμογών για τον Σημασιολογικό Ιστό

Προκλήσεις

Νέες ευκαιρίες επιβάλλουν νέες προκλήσεις. Στα παρακάτω εστιάζουμε την αναφορά μας στις παρακάτω προκλήσεις που αντιμετωπίζουμε τώρα: την ανάπτυξη οντολογιών, και την ανάπτυξη της 'τυπικής' σημασιολογίας των Σημασιολογικών Γλωσσών του Ιστού και την ανάπτυξη μοντέλων πιστοποίησης (trust) και απόδειξης (proof).

Η ανάπτυξη οντολογιών

Είναι κοινώς αναγνωρισμένο το γεγονός ότι στο πλαίσιο της κοινότητας του Σημασιολογικού Ιστού οι οντολογίες θα παίξουν ένα σημαντικό ρόλο στην ανάπτυξη του. Ποικίλες προσπάθειες έχουν αφιερωθεί στην έρευνα διαφορετικών απόψεων των οντολογιών, συμπεριλαμβανομένων των γλωσσών οντολογικής αναπαράστασης, ανάπτυξη οντολογίας, προσεγγίσεις για εκμάθηση οντολογιών, συστήματα οντολογικών βιβλιοθηκών (ontologylibrarysystems), τα οποία διαχειρίζονται και προτυποποιούν οντολογίες.

Διαχείριση

Ο βασικός στόχος των οντολογιών είναι να καταστήσει εφικτό το μοίρασμα και την επαναχρησιμοποίηση της γνώσης και ως εκ τούτου ένα τυπικό σύστημα οντολογικής βιβλιοθήκης υποστηρίζει ανοικτή αποθήκευση και οργάνωση, αναγνώριση / ταυτοποίηση και αναπροσαρμογή (versioning). Η ανοικτή αποθήκευση και οργάνωση δεικνύουν το πώς αποθηκεύονται και οργανώνονται οι οντολογίες σε ένα σύστημα βιβλιοθήκης για να διευκολύνουν την πρόσβαση και διαχείριση των οντολογιών. Η αναγνώριση / ταυτοποίηση συσχετίζει

κάθε οντολογία με ένα μοναδικό προσδιοριστή (identifier). Η αναπροσαρμογή (versioning) είναι ένα σημαντικό χαρακτηριστικό καθώς οι οντολογίες εξελίσσονται με το χρόνο και ένας μηχανισμός ‘αναπροσαρμογής’ μπορεί να διασφαλίσει τη λογική συνέχεια μεταξύ των διαφορετικών εκδοχών των οντολογιών.

Προσαρμογή

Καθώς οι οντολογίες εξελίσσονται με το χρόνο, το πώς να επεκτείνουμε και να ενημερώνουμε τις υπάρχουσες οντολογίες είναι ένα σημαντικό θέμα. Αυτό συνεπάγεται έρευνα, διόρθωση και αιτιολόγηση των οντολογιών μέσα σε ένα σύστημα οντολογικής βιβλιοθήκης.

Προτυποποίηση

Η δυνατότητα ενσωμάτωσης και διαλειτουργικότητας είναι πάντα η μέριμνα ενός ανοιχτού συστήματος. Αυτή είναι ιδιαίτερα και η μέριμνα του Σημασιολογικού Ιστού, ενός ανοιχτού συστήματος που πρέπει να είναι κλιμακωτό σε επίπεδο διαδικτύου. Τελευταία μια σειρά γλωσσών οντολογικής αναπαράστασης έχουν προταθεί και διάφορα συστήματα οντολογικών βιβλιοθηκών έχουν κατασκευαστεί. Το ερώτημα είναι ποια θα ήταν η καθορισμένη γλώσσα οντολογικής αναπαράστασης που θα είναι το πρότυπο; Κάθε μια φαίνεται να έχει τα πλεονεκτήματα και μειονεκτήματα της και τους υποστηρικτές και τους πολέμιους της. Αυτό πρέπει να είναι το χαρακτηριστικό της ανθρώπινης κοινωνίας όπου ο καθένας μας έχει τις προτιμήσεις του. Καθώς ο Σημασιολογικός Ιστός είναι ακόμα στα αρχικά του στάδια θα ήταν μάλλον νωρίς να επιβληθεί οποιαδήποτε προτυποποίηση. Κάθε γλώσσα αναπαράστασης μπορεί να αναπτυχθεί μόνη της και η μια ή οι λίγες που θα κερδίσουν θα αποτελέσουν τα εκ των πραγμάτων πρότυπα. Οι διάφορες XML μπορεί να λειτουργήσουν ως οι μεταγλώσσες αυτών των αναπαραστάσεων για να διευκολύνουν τη μελλοντική διαλειτουργικότητα και ενσωμάτωση.

2.8.6 Η τυπική σημασιολογία των Σημασιολογικών Γλωσσών του Ιστού

Η λειτουργική αρχιτεκτονική του Σημασιολογικού Ιστού έχει 3 επίπεδα τα metadatalayer, το schemalayer και το logicalayer, όπως ήδη αναφέραμε. Τελευταία το RDF (Resource Description Framework) θεωρείται το πιο δημοφιλές μοντέλο για το επίπεδο μεταδεδομένων. Αν και πιστεύεται ότι το μοντέλο πληροφοριών RDF είναι αρκετό για τον προσδιορισμό και την χρήση μεταδεδομένων, η σημασιολογία της δήλωσης σχετικά με δήλωση (reification) παραμένει να προσδιοριστεί. Το σχήμα RDFS (RDF Schema) επεκτείνει το RDF και είναι τελευταία μια δημοφιλής σε επίπεδο γλώσσα σχήματος. Είναι διαπιστωμένο ότι το RDFS στερείται τυπικής σημασιολογίας και μια πρόταση είναι να οριστεί μια μεταμοντελική αρχιτεκτονική / διάταξη για την RDFS όμοια με αυτήν για την UML (Universal Modeling Language) και ως εκ τούτου ορίζει τυπική σημασιολογία. Αυτή η προσέγγιση αν και τυπική, είναι πολύπλοκη και αυθόρμητη. Αν και το RDFS έχει κατηγορηθεί για τη σημασιολογική του σύγχυση και για κάποιο εμφανές παράδοξο, δεν έχει αποδειχθεί ότι μια τυπική σημασιολογία είναι ανέφικτη. Για να διευκολυνθεί η ανάπτυξη του Σημασιολογικού Ιστού πιστεύουμε ότι το θέμα της σημασιολογίας του RDFS πρέπει να επιλυθεί: είτε να οριστεί τυπική σημασιολογία ή το πρόβλημα του RDFS να εντοπιστεί ώστε να επιλυθεί το θέμα της σημασιολογίας.

Πιστοποίηση (proof) και αξιοπιστία (trust)

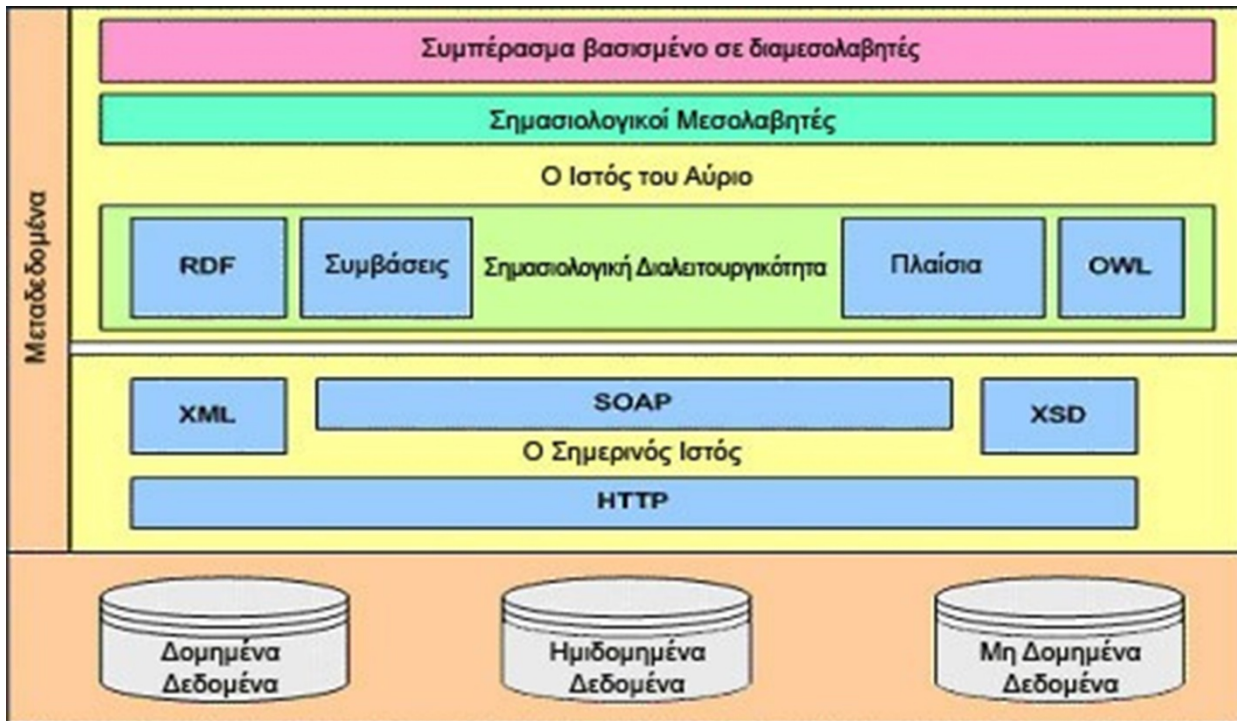
Ως ένα ανοιχτό και καταναμεμημένο σύστημα, ο Σημασιολογικός Ιστός διαπνέεται από την ιδέα ότι ο καθένας μπορεί να λέει στον καθένα οτιδήποτε. Άνθρωποι σε όλο τον κόσμο μπορεί να κάνουν κάποιες δηλώσεις που μπορεί να συγκρούονται. Ως εκ τούτου πρέπει κανείς να βεβαιώνεται ότι η αυθεντική πηγή κάνει μια δήλωση τεκμήριο (proof) και αυτή η πηγή είναι αξιόπιστη (trust).

Πιστοποίηση (proof). Η ψηφιακή υπογραφή θα παίζει σημαντικό ρόλο στην πιστοποίηση. Ο πόρος / τεκμήριο θα πρέπει να υπογράφει τη δήλωση που κάνει έτσι ώστε οι διαμεσολαβητές να μπορούν να ελέγξουν εάν η πληροφορία πραγματικά προέρχεται από τον πόρο από την οποία υποστηρίζει ότι προέρχεται. Επιπλέον, άλλες τεχνολογίες ασφάλειας όπως η κρυπτογράφηση (encryption), η απόκρυψη και ο έλεγχος πρόσβασης μπορούν να χρησιμοποιηθούν για να διασφαλίσουν την εχεμύθεια της πληροφορίας.

Αξιοπιστία (trust).

Ο καθένας θα πρέπει να είναι σε θέση να ορίζει ένα μοντέλο αξιοπιστίας για τον εαυτό του, δηλαδή μπορεί να προσδιορίζει πόση εμπιστοσύνη θα δείξει σε κάθε πόρο / τεκμήριο στο Σημασιολογικό Ιστό. Καθώς είναι ανεδαφικό να προσδιοριστεί η έκταση εμπιστοσύνης σε κάθε νέο πόρο είναι απαραίτητος ένας μηχανισμός για να εξάγει το βαθμό εμπιστοσύνης σε κάθε νέα πηγή. Μια λύση είναι η ιδέα του ‘δικτύου εμπιστοσύνης’ όπου όταν κάποιος εμπιστεύεται μια πηγή Α αυτομάτως εμπιστεύεται και όλες τις πηγές που εμπιστεύονται έναν πόρο Α αλλά σε μικρότερο βαθμό, με αυτόν τον τρόπο ένα τεράστιο δίκτυο ιεράρχησης δημιουργείται το οποίο διευκολύνει τους διαμεσολαβητές να αντλούν πληροφορίες βασισμένες σε γνώση που εμπιστεύονται.

Σήμερα, η ιδέα της απόδειξης / πιστοποίησης και εμπιστοσύνης μένει να επισημοποιηθεί και να αναπτυχθεί μια θεωρία που να ενσωματώνει σε μηχανές συμπερασμάτων του Σημασιολογικού Ιστού. Ωστόσο αυτές οι τεχνολογίες είναι πολύ σημαντικές και αποτελούν το θεμέλιο της κατασκευής αληθινών εμπορικών εφαρμογών όπως Business-to-Business, Business-to-Consumer και λοιπά συστήματα. Ο Σημασιολογικός Ιστός είναι ακόμα όραμα. Το δίκτυο θα ωριμάσει προς αυτήν την κατεύθυνση με τρόπο ανάλογο της εξέλιξης του κόσμου. Οι κοινότητες του Σημασιολογικού Ιστού θα εμφανιστούν και θα αναπτυχθούν πρώτα και η διαδραστικότητα και η διαλειτουργικότητα μεταξύ διαφορετικών κοινοτήτων θα τις καταστήσει μέρος του Σημασιολογικού Ιστού (Βλ. σχήμα 2.5).



Σχήμα 2. 5. Το όραμα του Σημασιολογικού Ιστού

Βιβλιογραφία

- Arroyo, S., Lara, R., Ding, Y., Stollberg, M., & Fensel, D. (2004). Semantic web languages. strengths and weakness. Στο *International Conference in Applied Computing*. CiteSeer. Ανακτήθηκε από <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.88.5910&rep=rep1&type=pdf>
- Bedford, D. A. (2014). Understanding and managing taxonomies as economic goods and services. *Bulletin of the American Society for Information Science and Technology*, 40(4), 15–22.
- Berners-Lee, T. (1998a). What the Semantic Web can represent. Ανακτήθηκε από <http://www.citeulike.org/group/17638/article/11988245>
- Berners-Lee, T. (1998b). *Why RDF model is different from the XML model*.
- Berners-Lee, T., & Hendler, J. (2001). *Scientific Publishing on the “Semantic Web”*. *Nature Webdebates*.
- Berners-Lee, T., Hendler, J., Lassila, O., & others. (2001). The semantic web. *Scientific american*, 284(5), 28–37.
- Berners-Lee, T., & others. (1998). *Semantic web road map*. Ανακτήθηκε από http://www.researchgate.net/profile/Tim_Berners
- Butler, M. H. (2003). Is the semantic web hype. Retrieved May, 31, 2006.
- Conklin, J. (1987). Hypertext: A survey and introduction. *IEEE Computer*, 20(9), 17–41.
- Corcho, O., & Gómez-Pérez, A. (2000). A roadmap to ontology specification languages. Στο *Knowledge Engineering and Knowledge Management Methods, Models, and Tools* (σσ 80–96). Springer. Ανακτήθηκε από http://link.springer.com/chapter/10.1007/3-540-39967-4_7
- Dacosta, M. C., Obrst, L. J., & Smith, K. T. (2003). *The Semantic Web: a guide to the future of XML, Web services, and knowledge management*. John Wiley & Sons. Ανακτήθηκε από <https://www.google.com/>
- Davies, J., Fensel, D., & Van Harmelen, F. (2003). *Towards the semantic web: ontology-driven knowledge management*. John Wiley & Sons. Ανακτήθηκε από [google.com](http://www.google.com)
- Dobson, S. A., & Burrill, V. A. (1995). Lightweight databases. *Computer Networks and ISDN Systems*, 27(6), 1009–1015.
- Gómez-Pérez, A., & Corcho, O. (2002). Ontology languages for the semantic web. *Intelligent Systems, IEEE*, 17(1), 54–60.
- Hearst, M., Elliott, A., English, J., Sinha, R., Swearingen, K., & Yee, K.-P. (2002). Finding the flow in web site search. *Communications of the ACM*, 45(9), 42–49.
- Hossain, M. A., & El Saddik, A. (2004). Creating Knowledge and Semantic Web Applications on the Web. Στο *Proc. 2nd Intl. Symposium on Innovation in Information and Communication Technology (ISIICT), Amman, Jordan, April*.
- Hyvönen, E., Mäkelä, E., Salminen, M., Valo, A., Viljanen, K., Saarela, S., ... Kettula, S. (2005). MuseumFinland—Finnish museums on the semantic web. *Web Semantics: Science, Services and Agents on the World Wide Web*, 3(2), 224–241.
- Hyvönen, E., Salminen, M., Kettula, S., & Junnila, M. (2004a). A content creation process for the semantic web. Στο *Proceeding of OntoLex* (σσ 1–7). Ανακτήθηκε από https://128.214.112.31/files/admin/sab/SAB04/material/SeCo/13_contentCreation.pdf
- Hyvönen, E., Salminen, M., Kettula, S., & Junnila, M. (2004b). A content creation process for the semantic web. Στο *Proceeding of OntoLex* (σσ 1–7). Ανακτήθηκε από https://128.214.112.31/files/admin/sab/SAB04/material/SeCo/13_contentCreation.pdf

- Hyvönen, E., Valo, A., Viljanen, K., & Holi, M. (2003). Publishing semantic web content as semantically linked HTML pages. Στο *Proceedings of XML Finland 2003, Kuopio, Finland*. Ανακτήθηκε από <https://www.seco.tkk.fi/publications/2003/hyvonen-valo-et-al-publishing-semantic-web-2003.pdf>
- Kalinichenko, L., Missikoff, M., Schiappelli, F., & Skvortsov, N. (2003). Ontological modeling. Στο *Proc. of the 5th Russian Conference on Digital Libraries RCDL2003, St.-Petersburg, Russia* (σσ 7–13). Citeseer. Ανακτήθηκε από <http://citeseerx.ist.psu.edu>.
- Kanal, L. N., Kumar, V., Kitano, H., & Suttner, C. B. (2014). Providing computationally effective knowledge representation via massive parallelism. *Parallel Processing for Artificial Intelligence* 1, 115.
- Lee, B. T., & Fischetti, M. (1999). *Web Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by Its Inventor*. Ανακτήθηκε από <http://www.citeulike.org>
- Luke, S., Spector, L., Rager, D., & Hendler, J. (1997). Ontology-based web agents. Στο *Proceedings of the first international conference on Autonomous agents* (σσ 59–66). ACM. Ανακτήθηκε από <http://dl.acm.org/citation.cfm?id=267668>
- Lu, S., Dong, M., & Fotouhi, F. (2002). The Semantic Web: opportunities and challenges for next-generation Web applications. Στο *International Journal of Information Research*. Citeseer. Ανακτήθηκε από <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.20.2117>
- Manola, F., Miller, E., & McBride, B. (2004). *Rdf primer w3c recommendation 10 february 2004*. Ανακτήθηκε από http://publishing-portal.net/download/dot/ontopage/sub_MaMi2004a.pdf
- Mizoguchi, R., & Ikeda, M. (1998). Towards ontology engineering. *Journal-Japanese Society for Artificial Intelligence*, 13, 9–10.
- Pollitt, A. S. (1998). *The key role of classification and indexing in view-based searching*. Citeseer. Ανακτήθηκε από <http://citeseerx.ist.psu.edu>
- Su, X., & Ilebrikke, L. (2002). A comparative study of ontology languages and tools. Στο *Advanced Information Systems Engineering* (σσ 761–765). Springer. Ανακτήθηκε από http://link.springer.com/chapter/10.1007/3-540-47961-9_62

3. Ποσοτική & Υπολογιστική Γλωσσολογία

Περίληψη

Στο κεφάλαιο αυτό γίνεται εκτενής αναφορά στις επιμέρους Γλωσσικές Τεχνολογίες όπως ο η Ποσοτική & η υπολογιστική Γλωσσολογία. Η συνεκτικότητα ενός κειμένου οι αρχές κειμενικότητας ως επίσης και ην μετρική που έχει αναπτυχθεί σχετικά με την ποσοτική κατανόηση του κειμένου αποτελούν ζητήματα που θα εξετασθούν σε αυτό κεφάλαιο.

Προαπαιτούμενες γνώσεις

Ευρετηρίαση, Ποσοτική & Υπολογιστή Γλωσσική Τεχνολογία

3.1 Κείμενο και Γλωσσολογία

Όπως έχει ήδη αναφερθεί στο παραπάνω κεφάλαιο, το έγγραφο ως αντικείμενο πληροφορίας έχει κεντρική θέση στην ΑΠ. Η επεξεργασία των εγγράφων που απαρτίζουν τις συλλογές για την δημιουργία ενός ευρετηρίου προκειμένου να υπάρχει αποδοτικότερη αναζήτηση, είναι αναγκαία και επεξηγήθηκε αναλυτικά σε προηγούμενο κεφάλαιο..

Η επεξεργασία αυτή μπορεί να χωριστεί σε δύο κατηγορίες: την γλωσσολογική και τη μη-γλωσσολογική. Η γλωσσολογική επεξεργασία λοιπόν αφορά στον υπό-κλάδο της Γλωσσολογίας που ονομάζεται ΥΓ και όπως έχει αναφερθεί παραπάνω αφορά στην απόπειρα ανάπτυξης λογισμικού υπολογιστών προκειμένου να μπορέσει η μηχανή να κατανοήσει τη φυσική γλώσσα.

Για την κατανόηση της τομής των επιστημών της ΑΠ και της ΥΓ, είναι απαραίτητη η ανάλυση της έννοιας του κειμένου στη Γλωσσολογία η οποία αποτελεί το περιεχόμενο του εγγράφου στην ΑΠ.

3.1.1 Το κείμενο και τα χαρακτηριστικά του

Προκειμένου να οριστεί πλήρως η έννοια του κειμένου, είναι απαραίτητη η επεξήγηση της στο πλαίσιο μιας ευρύτερης διαδικασίας, αυτής της ανθρώπινης επικοινωνίας. Προκειμένου στην περιγραφή της διαδικασίας της επικοινωνίας, όπως περιγράφεται από τους Bolshakov και Gelbukh (2004), ορίζονται οι βασικές γλωσσολογικές έννοιες σημασία (meaning), κείμενο (text), γλώσσα (language) και οι μεταξύ τους σχέσεις. Ως σημασία ορίζεται ουσιαστικά η πληροφορία που θέλει να μεταφέρει ένας άνθρωπος μέσω της επικοινωνίας, η οποία αποτελεί και το βασικό στόχο της επικοινωνιακής διαδικασίας. Ως κείμενο ορίζεται η φυσική αναπαράσταση των σκέψεων μεταξύ δύο ατόμων που επικοινωνούν, το οποίο περιέχει λέξεις, κενά, σημεία στίξης και μέσω των συνδυασμών των παραπάνω σχηματίζονται προτάσεις και παράγραφοι. Τέλος, ως φυσική γλώσσα ορίζεται ουσιαστικά ο μετατροπέας του νοήματος σε κείμενο και αντιστρόφως. Από τους παραπάνω ορισμούς συμπεραίνεται ο ρόλος του κειμένου στο πλαίσιο της επικοινωνίας και η σχέση του με τη γλώσσα.

Σύμφωνα με Bolshakov και Gelbukh (2004) υπάρχουν τρία βασικά χαρακτηριστικά τα οποία διέπουν ένα κείμενο. Το πρώτο αφορά στον σκοπό ύπαρξης ενός κειμένου, δηλαδή στο ότι ένα κείμενο γεννιέται ώστε να κωδικοποιήσει ένα σύνολο πληροφοριών, επομένως έχει μια σημασία, ένα νόημα το οποίο προορίζεται και αφορά κάποιους ανθρώπους. Για το λόγο αυτό ακριβώς πραγματοποιείται η επεξεργασία της φυσικής γλώσσας του κειμένου.

Το δεύτερο χαρακτηριστικό συνίσταται στο ότι όσες πληροφορίες και αν εμπεριέχονται σε ένα κείμενο, όσο πολύπλοκο και αν είναι η δομή του είναι πάντοτε γραμμική, αποτελεί δηλαδή ένα σύνολο από συμβολοσειρές, κενά, σημεία στίξης τα οποία σχηματίζουν μια πολύ μεγάλη γραμμή. Να σημειωθεί ότι οι πληροφορίες που αναπαρίστανται στο κείμενο με γραμμική δομή είναι μη γραμμικές.

Τέλος το τρίτο χαρακτηριστικό αφορά τη δομή του κειμένου από επιμέρους στοιχεία, τα οποία βρίσκονται «εντεθειμένα σε ομοειδή δομή» (nested structure) μέσα στο κείμενο. Για τα στοιχεία αυτά λοιπόν που απαρτίζουν το κείμενο στη βιβλιογραφία (Hoey 2003) γίνεται αναφορά στην ουσιαστική πραγματοποίηση διαχωρισμού του κειμένου στα λεγόμενα κομμάτια (chunks) με αφορμή την αλληλεπίδραση μεταξύ συγγραφέα και αναγνώστη, ώστε να μεταφερθεί η πληροφορία από τον μεν στον δε. Πιο συγκεκριμένα, ο διαχωρισμός αυτός πραγματοποιείται τόσο κατά την διάρκεια σύνταξης του κειμένου από

τον συγγραφέα (απ' αρχής δημιουργίας του κειμένου δηλαδή, προκειμένου να υφίσταται μια λογική δομή στο κείμενο), όσο και κατά την διάρκεια της ανάγνωσης από τους αναγνώστες, οι οποίοι χρησιμοποιούν τη δομή ώστε να ερμηνεύσουν καλύτερα το κείμενο. Οι Bolshakov και Gelbukh (2004) αναφέρουν ακόμη την σημασία των επιμέρους στοιχείων της δομής του κειμένου, τα οποία οργανώνονται σε λέξεις, προτάσεις, παραγράφους και όλα μαζί σχηματίζουν τον λόγο (discourse), ο οποίος έχει ως κύριο χαρακτηριστικό τη συνδετικότητα (connectivity) ή αλλιώς συνεκτικότητα (coherence).

Η συνεκτικότητα ενός κειμένου έγκειται στην κοινή σταθερότητα και συνέπεια όλων των στοιχείων του λόγου στο κείμενο ώστε να μεταφερθεί το νόημα που πρέπει μέσω αυτών. Έχοντας ένα κείμενο την οργάνωση αυτή, μέσω των παραπάνω χαρακτηριστικών τότε είναι εφικτή η ανάπτυξη μεθόδων ευφύους επεξεργασίας κειμένου. Όπως τονίζεται στη βιβλιογραφία (Hoey 2003) τα γραπτά κείμενα είναι συνεκτικά, έτσι ώστε να καθιστούν εφικτή την κατανόηση των σχέσεων των στοιχείων μέσα στο κείμενο.

3.1.2 Αρχές κειμενικότητας

Οι DeBeaugrande και Dressler (1981) στο πλαίσιο του υπο-κλάδου της Κειμενογλωσσολογίας, που μελετά τα κοινά χαρακτηριστικά που διέπουν τα κείμενα και τις μεταξύ τους διαφορές και διακρίσεις, χαρακτηρίζουν το κείμενο ως μια επικοινωνιακή εμφάνιση (communicative occurrence), η οποία πρέπει να πληροί επτά αρχές κειμενικότητας. Σε περίπτωση που κάποια από τις αρχές κειμενικότητας δεν ικανοποιείται τότε το κείμενο δε θεωρείται επικοινωνιακό και έτσι κατά συνέπεια δεν μπορεί να αντιμετωπιστεί καν ως κείμενο. Οι αρχές κειμενικότητας λοιπόν είναι άμεσα συνδεδεμένες με τη διαδικασία επικοινωνίας, όπως παρουσιάστηκε στην παραπάνω ενότητα.

Οι αρχές κειμενικότητας χωρίζονται σε **κείμενο-κεντρικές** (απευθύνονται στο υλικό του κειμένου δηλαδή), που περιλαμβάνουν τις δύο πρώτες αρχές και σε **χρηστο-κεντρικές** (αφορούν παραγωγούς και αποδέκτες στην διαδικασία επικοινωνίας), που περιλαμβάνουν τις υπόλοιπες πέντε. Σύμφωνα με τους DeBeaugrande και Dressler (1981), οι αρχές κειμενικότητας είναι οι ακόλουθες:

1. **Συνοχή** (cohesion): αφορά στους τρόπους σύνδεσης των συστατικών του επιφανειακού κειμένου (surfacetext), μέσα σε μια πρόταση, οι οποίοι έχουν να κάνουν κυρίως με γραμματικούς κανόνες. Οι De Beaugrande και Dressler (1981) τονίζουν πόσο σημαντική είναι η αλληλεπίδραση της συνοχής με τις υπόλοιπες αρχές κειμενικότητας προκειμένου να είναι αποδοτική η επικοινωνία.
2. **Συνεκτικότητα** (coherence): κάθε κείμενο για να μπορεί να είναι κατανοητό από το ανθρώπινο μυαλό πρέπει να αλληλεπιδρά η γνώση που περιέχεται στο κείμενο με τις γνώσεις που έχει ο άνθρωπος αποθηκευμένες στον εγκέφαλο του για τον κόσμο. Έτσι σε κάθε κείμενο υπάρχουν κάποιες έννοιες (concepts) και κάποιες σχέσεις (relations) οι οποίες συνδέουν τις έννοιες μεταξύ τους. Σύμφωνα με τους De Beaugrande και Dressler (1981) ο όρος **έννοια** ορίζεται ως «*το γνωσιακό περιεχόμενο το οποίο μπορεί να ανακτηθεί ή να ενεργοποιηθεί μέσα από την ενότητα και συνοχή του μυαλού*».
3. **Αποβλεπτικότητα** (intentionality): η πρόθεση του συγγραφέα-παραγωγού του κειμένου να δημιουργήσει ένα κείμενο που να πληροί τις πληροφοριακές ανάγκες του αποδέκτη.
4. **Αποδοχή** (acceptability): αφορά τη συμπεριφορά του δέκτη σε σχέση με το αν το κείμενο είναι επικοινωνιακό σε σχέση με τις πληροφοριακές του ανάγκες και αν το αποδέχεται.
5. **Πληροφοριακότητα** (informativity): σύμφωνα με τη βιβλιογραφία (Carstens 2001) η αρχή αυτή αφορά στην επικοινωνιακή αξία των μερών του κειμένου.
6. **Καταστασιακότητα** (situationality/contextuality): έχει να κάνει με τους διάφορους παράγοντες που καθιστούν ένα κείμενο σχετικό. Σύμφωνα με τη βιβλιογραφία (Carstens 2001) αφορά κυρίως το ρόλο του περικειμένου/συμφραζόμενα (context) στην ποιότητα της επικοινωνίας. Υποδηλώνει το κατά πόσο όσοι εμπλέκονται στην επικοινωνία έχουν γνώση των συμφραζομένων.
7. **Διακειμενικότητα** (intertextuality): σύμφωνα με τη βιβλιογραφία (Carstens 2001) η αρχή της διακειμενικότητας αφορά τη συμπεριφορά του παραγωγού ενός κειμένου στην πρόθεση του να παράγει πληροφορίες για έναν αποδέκτη. Μάλιστα σύμφωνα με τη βιβλιογραφία (Carstens 2001) οι αρχές κειμενικότητας αποδοχή και διακειμενικότητα θεωρούνται ζευγάρι καθώς για κάθε κείμενο υπάρχει απαραίτητα ο παραγωγός του και ο αποδέκτης του.

Τέλος, οι DeBeaugrande και Dressler (1981) προσθέτουν πως εκτός των αρχών κειμενικότητας υπάρχουν ακόμη κανονιστικές αρχές (regulative principles), προκειμένου να υπάρχει έλεγχος της επικοινωνίας

του κειμένου και αναφέρουν τις δύο βασικότερες: αποδοτικότητα (efficiency), όταν απαιτείται η ελάχιστη προσπάθεια από τους συμμετέχοντες ώστε η επικοινωνία να είναι ικανοποιητική και αποτελεσματικότητα (effectiveness), η αρχή η οποία αφορά τη δημιουργία συνθηκών για την επίτευξη ενός στόχου.

3.1.3 Περικείμενο

Στη βιβλιογραφία (Tanskanen 2006) παρουσιάζεται μια πολύ σημαντική έννοια στη Γλωσσολογία και αλληλένδετη με τη συνοχή του κειμένου, πρόκειται για την έννοια του περικείμενου. Το περικείμενο χωρίζεται στα εξής είδη: το γλωσσολογικό περικείμενο, το οποίο έχει να κάνει με το γλωσσικό υλικό που βρίσκεται γύρω από το αντικείμενο προς εξέταση, το γνωσιακό περικείμενο (cognitive) που αφορά τους γνωσιακούς παράγοντες επικοινωνίας (διανοητικές αναπαραστάσεις, γνωστική προσπάθεια που απαιτείται από τους επικοινωνούντες) και κοινωνικό περικείμενο, το οποίο αναφέρεται σε όλο το κανάλι επικοινωνίας, την κατάσταση, τους επικοινωνούντες και τους ρόλους αλληλεπίδρασης. Σχετίζεται άμεσα με την καταστασιακότητα, μια από τις αρχές κειμενικότητας που αναλύθηκαν παραπάνω.

3.1.4 Επεξεργασία κειμένου για ανάκτηση και εξαγωγή πληροφορίας

Όπως έχει ήδη αναφερθεί, η επεξεργασία της πληροφορίας είναι απαραίτητη προκειμένου να επιτευχθεί η ανάκτηση της. Πολλές φορές δε η ανάγκη ανάκτησης της πληροφορίας αφορά σε **ειδικευμένη ανάκτηση της πληροφορίας**. Σύμφωνα με τη βιβλιογραφία (Moens 2006) μέσω της Εξαγωγής Πληροφορίας (InformationExtraction) είναι εφικτή η εξειδικευμένη ανάκτηση, καθώς η **Εξαγωγή Πληροφορίας** δεν εστιάζει μονάχα στην αντιστοίχιση του ερωτήματος του χρήστη (σε μορφή λέξεων κλειδιών φυσικής γλώσσας) με κάποια συναφή έγγραφα αλλά και στην αντιστοίχιση των σημασιολογικών τάξεων των οντοτήτων (και των μεταξύ τους σχέσεων) που φέρουν την πληροφορία στα έγγραφα.

Φυσικά η παραπάνω διαδικασία προϋποθέτει την κατανόηση της φυσικής γλώσσας στο πλαίσιο ενός ΣΑΠ. Όπως αναφέρεται στη βιβλιογραφία (Moens 2006) αυτό μπορεί να πραγματοποιηθεί μέσω της Επεξεργασίας Φυσικής Γλώσσας, στόχος της οποίας είναι η ανάλυση της ανθρώπινης γλώσσας ώστε να είναι εφικτή η κατανόηση της από τους υπολογιστές. Η Επεξεργασία Φυσικής Γλώσσας εστιάζει στην επεξεργασία της δομής ενός κειμένου από γλωσσική άποψη και πιο συγκεκριμένα περιλαμβάνει μορφολογική, συντακτική, σημασιολογική ανάλυση της γλώσσας.

Με βάση λοιπόν τα παραπάνω, θα αναφερθούν επιγραμματικά οι συνηθέστερες διαδικασίες επεξεργασίας κειμένου, οι οποίες έχουν ενσωματωθεί στις μηχανές αναζήτησης για την ΑΠ σύμφωνα με τη βιβλιογραφία (Ceri et al. 2013), καθώς έχουν επεξηγηθεί αναλυτικά στο προηγούμενο κεφάλαιο. Πρόκειται για τις διαδικασίες: ανάλυσης δομής εγγράφου, λεξιλογικής ανάλυσης, αφαίρεσης διακοπτούσων λέξεων, ανίχνευσης φράσεων (μέσω μορφολογικής και συντακτικής ανάλυσης), περιστολής - λημματοποίησης, απόδοσης βαρών όρων. Στο σχήμα 3.1 φαίνονται σχηματικά όλες οι παραπάνω διαδικασίες επεξεργασίας κειμένου:

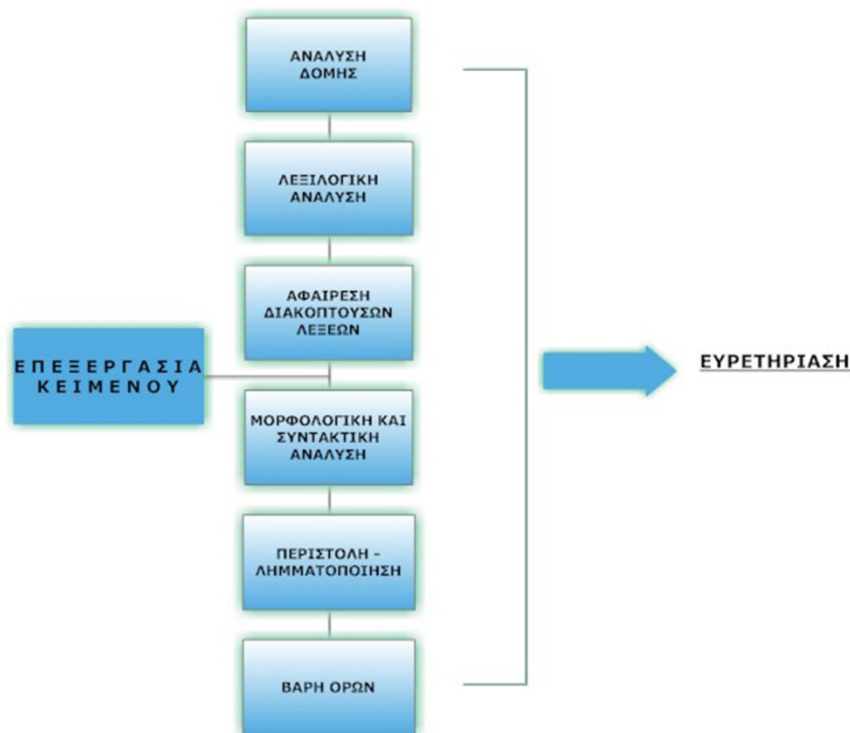
3.1.5 Ποιότητα κειμένου (κατανόηση και αναγνωσιμότητα)

Η ποιότητα ενός κειμένου εξαρτάται από το κατά πόσο αυτό ικανοποιεί το βασικό σκοπό δημιουργίας του. Για το λόγο αυτό η μέτρηση της αποτελεί σημαντικό ζήτημα στη Γλωσσολογία. Όπως ήδη έχει αναφερθεί ο βασικός στόχος ενός κειμένου έγκειται στη μεταφορά ενός νοήματος στον αποδέκτη/αναγνώστη που απευθύνεται.

Για να γίνει αυτό, το κείμενο θα πρέπει να είναι κατανοητό από τον αναγνώστη. Συνεπώς οι διάφορες παράμετροι που επηρεάζουν την κατανόησή (comprehensiveness) του θεωρούνται πολύ σημαντικές. Στη βιβλιογραφία (Mikk 2005) η κατανόηση και η αναγνωσιμότητα (readability) θεωρούνται ταυτόσημες έννοιες. Η αναγνωσιμότητα (readability) σύμφωνα με τη βιβλιογραφία (Richards και Schmidt 2002) ορίζεται ως το «*πόσο εύκολα ένα υλικό μπορεί να αναγνωστεί και να κατανοηθεί*» και αναφέρονται παράγοντες που την επηρεάζουν όπως το μέσο μήκος πρότασης ενός κειμένου ή η πολύπλοκη γραμματική που το διέπει.

Η ανάγνωση και κατανόηση συνδέεται άμεσα με το κανάλι επικοινωνίας και την ευρύτερη εγκεφαλική αντίληψη του ανθρώπου. Μάλιστα στη βιβλιογραφία (Dubay 2004) αναφέρεται η αρχή της Ελάχιστης Δυνατής Προσπάθειας στην ανθρώπινη ομιλία, όπου ο γλωσσολόγος και φιλόλογος Zipf G. K. χρησιμοποίησε τη στατιστική ανάλυση της γλώσσας για να δείξει το πώς αυτή λειτουργεί. Η ελάχιστη δυνατή προσπάθεια συνδέεται άμεσα με την εξοικονόμηση ενέργειας και τη συχνότητα εμφάνισης λέξεων. Η

στατιστική μελέτη των λέξεων που χρησιμοποιεί ο άνθρωπος για την επικοινωνία (ανάλογα την δυσκολία - ευκολία της λέξης) οδήγησε σε διάφορους στατιστικούς νόμους που αφορούν την ΠΓ



Σχήμα 3.1. Διαδικασίες επεξεργασίας κειμένου

Σύμφωνα με τους Dale και Chall (1949) η επιτυχία ενός κειμένου ως προς την αναγνωσιμότητα εξαρτάται από τους εξής αλληλένδετους παράγοντες, εστιάζοντας στο κείμενο:

1. Το περιεχόμενο του κειμένου και το πόσο ενδιαφέρει τον αναγνώστη.
2. Ο τρόπος έκφρασης.
3. Η δομή και οργάνωση του κειμένου.

Ως επιτυχία στην αναγνωσιμότητα σύμφωνα με τους Dale και Chall (1949) θεωρείται η κατανόηση και ανάγνωση του κειμένου από τους αναγνώστες στο ελάχιστο δυνατό χρόνο και με την καταβολή της ελάχιστης δυνατής προσπάθειας, η οποία βέβαια δεν **εξαρτάται** αποκλειστικά μόνο από το κείμενο όπως παρουσιάστηκε παραπάνω αλλά **και από τους ίδιους τους αναγνώστες** και πιο συγκεκριμένα από παράγοντες όπως:

1. Επιδεξιότητα στην ανάγνωση.
2. Ευφυΐα.
3. Εμπειρία.
4. Ωριμότητα.
5. Ενδιαφέροντα.
6. Σκοπός ανάγνωσης.

Στη βιβλιογραφία (Mikk 2005) οι μέθοδοι έρευνας σχετικά με την εξασφάλιση της κατανόησης και της αναγνωσιμότητας στρέφονται προς δύο διαφορετικές κατευθύνσεις: από τη μια πλευρά μελετώνται οι κανόνες προκειμένου ένα κείμενο να έχει υψηλή αναγνωσιμότητα και από την άλλη οι readability formulae, προκειμένου να μετρηθεί και να αξιολογηθεί η αναγνωσιμότητα του κειμένου. Ειδικότερα μέσω της μέτρησης αυτής στην βιβλιογραφία (Zamanian και Heydari 2012) αναφέρεται πως μπορεί να γίνει πρόβλεψη της δυσκολίας αναγνωσιμότητας κάθε κειμένου και πως η πρόβλεψη αυτή είναι ιδιαίτερος χρήσιμη σε διάφορους τομείς όπως η εκπαίδευση και η συγγραφή κειμένων. Ουσιαστικά μέσω των μετρήσεων αυτών μπορεί να

διασφαλιστεί ότι το κατάλληλο ανάγνωσμα θα δοθεί στο κατάλληλο επίπεδο αναγνώστη και μπορεί να επιτευχθεί έτσι η αποβλεπτικότητα.

Αναλυτικότερα σε σχέση με την εφαρμογή των readabilityformulae, σύμφωνα με τη βιβλιογραφία (Mikk 2005) προκειμένου να διερευνηθεί η κατανόηση κειμένου, αρχικά θα πρέπει να υπάρχει ένα σύνολο χαρακτηριστικών ως προς τα οποία θα εξεταστεί ένα αντιπροσωπευτικό δείγμα κειμένων. Τα χαρακτηριστικά αυτά συνήθως καθορίζονται από ειδικούς ή μέσα από έρευνα και ερωτηματολόγια. Εν συνεχεία, διαμορφώνονται οι readabilityformulae, οι οποίες υπολογίζουν το πόσο πολύπλοκο μπορεί να θεωρηθεί ένα κείμενο σε σχέση με την κατανόηση του, χρησιμοποιώντας έναν δείκτη αναγνωσιμότητας. Το δείγμα κειμένων υποβάλλεται σε πειράματα, εκτεταμένη ανάλυση, καθιέρωση τιμών για την κατανόηση και στατιστικές επεξεργασίες προκειμένου να υπολογιστούν οι σχέσεις μεταξύ των μεταβλητών πρόβλεψης (predictorvariables). Δύο από τις βασικότερες μεταβλητές είναι η πολυπλοκότητα περιεχομένου που αφορά το λεξιλόγιο και η πολυπλοκότητα δομής που σχετίζεται με το μήκος κειμένου.

Στη βιβλιογραφία (Zamanian και Heydari 2012), αναφέρονται τα ακόλουθα πλεονεκτήματα χρήσης των “readabilityformulae”:

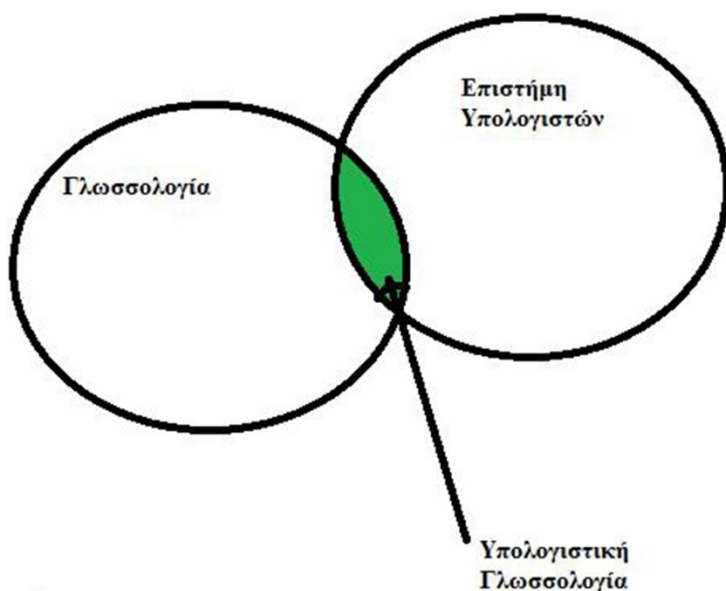
1. Μέσω των μετρήσεων ο συγγραφέας έχει στη διάθεσή του πληροφορίες ώστε να έρθει κοντά με το κοινό στο οποίο απευθύνεται και να μετατρέψει το κείμενο του σε ένα απλό κείμενο.
2. Η εφαρμογή τους πραγματοποιείται πριν το κείμενο φτάσει στον αναγνώστη μέσω υπολογιστών.

Ενώ παρουσιάζουν αντίστοιχα και τα ακόλουθα μειονεκτήματα:

1. Δεν προσδιορίζουν την κατανόηση από την πλευρά των αναγνωστών.
2. Παρουσιάζεται απόκλιση αποτελεσμάτων με χρήση διαφορετικών readabilityformulae.
3. Αδυναμία υπολογισμού διάφορων παραμέτρων όπως βαθμό ενδιαφέροντος ή συνεκτικότητας κειμένου.

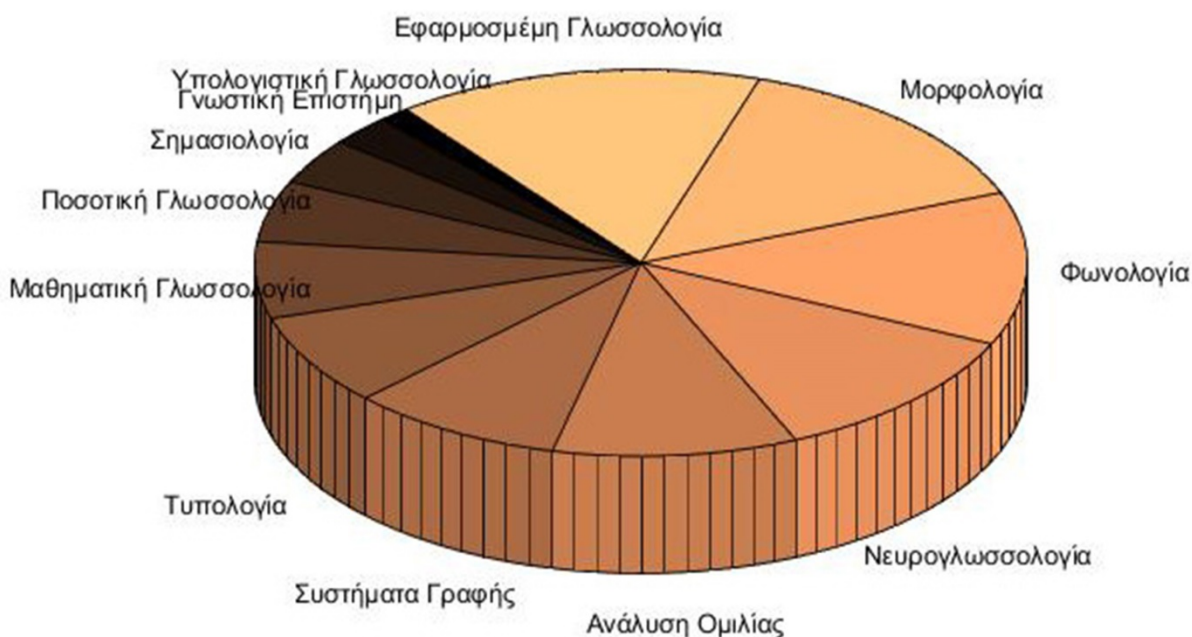
3.2 Εισαγωγή στην ΥΓ και τις βασικές της έννοιες

Σύμφωνα με τους Bolshakov και Gelbukh (2004) το κύριο μέλημα της ΥΓ είναι η δημιουργία υπολογιστικών προγραμμάτων για την αυτόματη Επεξεργασία Φυσικής Γλώσσας (π.χ. των λέξεων ή κειμένων). Η ΥΓ εφαρμόζεται σε πολλούς τομείς ορισμένοι από τους οποίους είναι ο διαχωρισμός λέξεων (hyphenation), ο γραμματικός έλεγχος (spellchecking), η ΑΠ, η μηχανική μετάφραση, η εξόρυξη δεδομένων από το κείμενο. Όπως ήδη έχει αναφερθεί, η ΥΓ είναι ένας διεπιστημονικός κλάδος, ο οποίος αφορά στο συνδυασμό πολλών κλάδων επιστημών άλλων σε μικρότερο και άλλων σε μεγαλύτερο βαθμό, με κυρίαρχους την Επιστήμη των Υπολογιστών και τη Γλωσσολογία, όπως φαίνεται στο σχήμα 3.2.



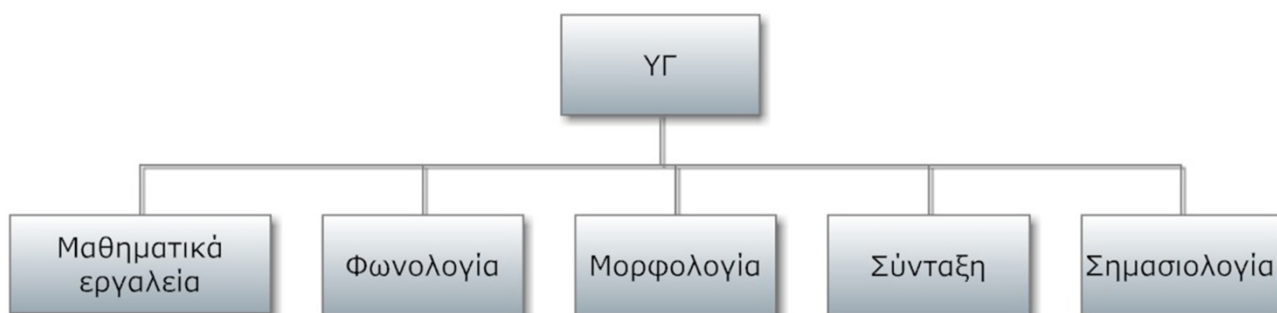
Σχήμα 3. 2. Ο υπο-κλάδος της ΥΓ

Γενικότερα, η ΥΓ σύμφωνα με την LINGUIST List αποτελεί έναν από τους πολλούς υπο-κλάδους της Γλωσσολογίας, όπως φαίνεται σχηματικά στο σχήμα 3.3. Στο διάγραμμα παρατηρούνται και άλλοι επιστημονικοί και διεπιστημονικοί κλάδοι στους οποίους έχει ήδη γίνει αναφορά όπως η Νευρογλωσσολογία και η Μορφολογία. Ο κλάδος της Μορφολογίας θα αναλυθεί παρακάτω.



Σχήμα 3. 3. Υπο-κλάδοι Γλωσσολογίας

Οι Bolshakov και Gelbukh (2004) αναφέρουν τα πεδία που μελετά η ΥΓ, που αφορούν στον προφορικό και γραπτό λόγο. Στο σχήμα 3.4 συνοψίζονται σε διαγραμματική μορφή:



Σχήμα 3. 4. Πεδία έρευνας ΥΓ

Αρχικά θα αναφερθούν κάποια βασικοί ορισμοί που χρησιμοποιούνται στην ΥΓ. Χρησιμοποιώντας τα συστατικά δομής ενός κειμένου από το μικρότερο προς το μεγαλύτερο, η μικρότερη μονάδα φυσικής γλώσσας ονομάζεται **μόρφημα** (morph). Το **επόμενο επίπεδο γλωσσικής μονάδας αποτελεί η λέξη**. Σύμφωνα με τους Bolshakov και Gelbukh (2004) ως **λέξη** θεωρείται κάθε **υπο-συμβολοσειρά σε ένα κείμενο από τον πρώτο οριοθέτη (delimiter) ως τον επόμενο οριοθέτη** (δηλαδή ένα κενό ή κάποιο σημείο στίξης).

Στη συνέχεια θα επεξηγηθούν οι διάφοροι χαρακτηρισμοί που αφορούν τον όρο **λέξη**. Με την έννοια **εμφάνισης λέξης** (wordoccurrence), υποδηλώνεται η επανάληψη λέξεων σε ένα κείμενο. Οι ομοιότητες των λέξεων, όπως το κοινό τους θέμα, μπορεί να αποτελέσει παράγοντα ομαδοποίησης τους μέσα στο συγκεκριμένο κείμενο, με βάση κάποιο κοινό νόημα, αν και μπορεί να έχουν διαφορετική μορφή. Συνεπώς το

φαινόμενο αυτό πρέπει να χαρακτηριστεί με κάποιο τρόπο. Έτσι, το σύνολο των συμβολοσειρών που έχουν το ίδιο νόημα αλλά εμφανίζονται σε διαφορετικές μορφές ονομάζεται *λέξημα* (lexeme), ενώ κάθε συμβολοσειρά του συνόλου αυτού ονομάζεται *μορφή λέξης* (wordform). Έτσι λοιπόν ως μορφή λέξης θεωρείται κάθε εμφάνιση λέξης αλλά ταυτόχρονα οι μορφές λέξεων μπορούν να επαναλαμβάνονται μέσα στο κείμενο.

Για να είναι εφικτή η επεξεργασία της φυσικής γλώσσας αποδίδονται σύμβολα για την ανάλυση των συστατικών (constituents) των κειμένων, μέσω των οποίων αυτά κατατάσσονται σε διάφορες γραμματικές κατηγορίες. Παρατίθεται ένα σύνολο τέτοιων συμβόλων από τους Bolshakov και Gelbukh (2004) στον πίνακα 1 καθώς και ορισμένοι κανόνες παραγωγής:

Πίνακας 3.1. Γραμματικά σύμβολα και κανόνες παραγωγής

Γραμματικά σύμβολα	Κανόνες παραγωγής
S – πρόταση	$S \rightarrow NP VP$
NP – ονοματική φράση (με κέντρο το ουσιαστικό)	$VP \rightarrow V NP$
VP – ρηματική φράση (με κέντρο το ρήμα)	$NP \rightarrow D N$
N – ουσιαστικό	$NP \rightarrow N$
V – ρήμα	
D – προσδιοριστές (a, an, the)	

Εξετάζοντας τους κανόνες παραγωγής ισχύει η εξής σχέση μεταξύ των συμβόλων: από τη μια πλευρά παρουσιάζονται οι κλάσεις μερών του λόγου με βάση τις οποίες κατηγοριοποιούνται οι λέξεις και από την άλλη παρουσιάζονται οι σχέσεις μεταξύ των κλάσεων, ως συστατικά του κειμένου. Πιο συγκεκριμένα, χρησιμοποιώντας σαν παράδειγμα τον πρώτο κανόνα παραγωγής ($S \rightarrow NPVP$), είναι κατανοητό ότι η ονοματική και ρηματική φράση αποτελούν συστατικά της πρότασης.

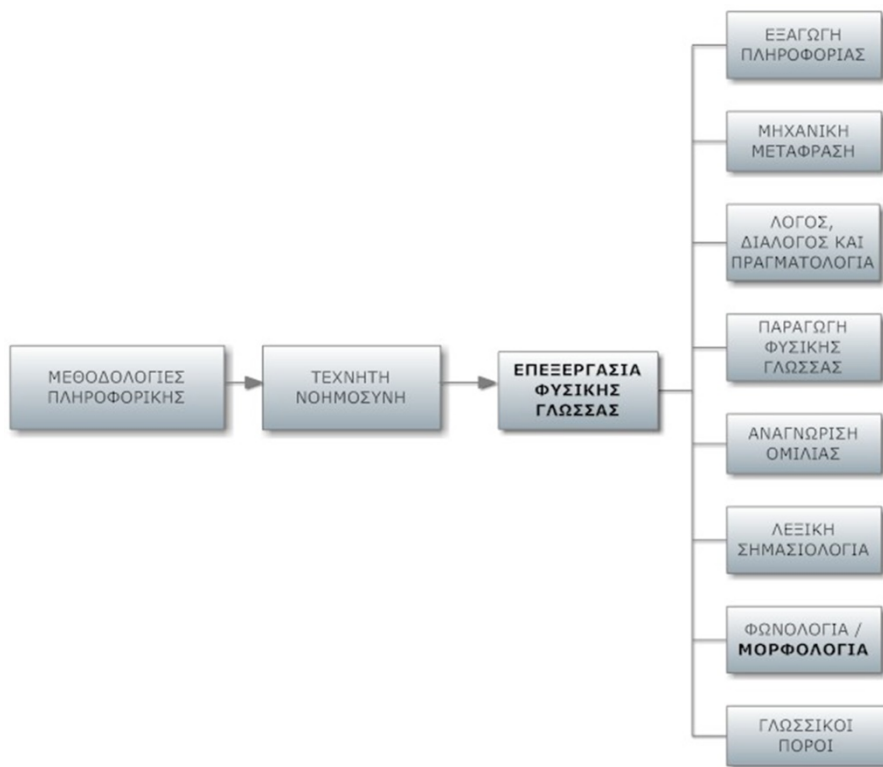
Στις παρακάτω ενότητες θα αναλυθούν τα πεδία της ΥΓ που αφορούν το γραπτό λόγο και τα κείμενα, δηλαδή η Μορφολογία, η Σύνταξη και η Σημασιολογία.

3.2.1 Μορφολογία

Σύμφωνα με Jurafsky και Martin (2000) η Μορφολογία (Morphology) αφορά τη **μελέτη των βασικών μονάδων που σχηματίζουν τις λέξεις, τα μορφήματα**, τα οποία αποτελούν τη θεμελιώδη νοηματική μονάδα στη γλώσσα. Αποτελεί από μόνη της έναν ξεχωριστό υπο-κλάδο της Γλωσσολογίας και αποτελεί αντικείμενο μελέτης και για την ΥΓ ειδικότερα (βλέπε σχήμα 3.4).

Ακόμη στον κλάδο της Επιστήμης των Υπολογιστών, με βάση το σύστημα ταξινόμησης της ACM (ανακτήθηκε 25/02/2015 από <http://www.acm.org/about/class/2012>), η Μορφολογία, χρησιμοποιείται και μελετάται και από την Τεχνητή Νοημοσύνη και πιο συγκεκριμένα αποτελεί υπο-κλάδο της Επεξεργασίας Φυσικής Γλώσσας όπως φαίνεται και στο σχήμα 3.5.

Στη Μορφολογία οι δύο βασικές ομάδες στις οποίες κατηγοριοποιούνται τα μορφήματα είναι τα θέματα (stems), τα οποία αποτελούν το κύριο μόρφημα κάθε λέξης και δίνουν το βασικό νόημα στη λέξη και τα προσφύματα (affixes), τα οποία δίνουν επιπλέον νοήματα σε μια λέξη. Τα προσφύματα με τη σειρά τους διακρίνονται και αυτά σε τέσσερις μεγάλες κατηγορίες: τα προθήματα (prefixes) που προηγούνται του θέματος, τα επιθήματα (suffixes) που έπονται του θέματος, τα ενθήματα (infixes) που εισάγονται μέσα στο θέμα και τα περιθήματα (circumfixes) τα οποία μπορεί να προηγούνται και να έπονται του θέματος. Μάλιστα μια λέξη μπορεί να έχει πάνω από ένα πρόσφυμα. Οι παραπάνω κατηγορίες χρησιμοποιούνται για τη συντακτική ανάλυση.



Σχήμα 3. 5. Επεξεργασία Φυσικής Γλώσσας και Μορφολογία

3.2.2 Σύνταξη

Ως συντακτική ανάλυση (parsing) οι Jurafsky και Martin (2000) ορίζουν την εισαγωγή δεδομένων (input) και ως αποτέλεσμα την εξαγωγή κάποιας δομής για τα δεδομένα αυτά. Έτσι στη συντακτική ανάλυση περιλαμβάνεται το θέμα της λέξης καθώς και άλλα μορφολογικά της χαρακτηριστικά, όπως το μέρος του λόγου που ανήκει. Προκειμένου να μπορεί να γίνεται μορφολογική ανάλυση αυτόματα, για να δημιουργηθεί δηλαδή ένας μορφολογικός συντακτικός αναλυτής (morphological parser) χρειάζονται τρία βασικά στοιχεία:

1. **Λεξικό (lexicon)**, με τη λίστα θεμάτων και προσφυμάτων και συνοδευτικές βασικές πληροφορίες για αυτά. Το λεξικό συνεργάζεται με τους μορφοτακτικούς περιορισμούς ώστε να καλύψει όλους τους πιθανούς συνδυασμούς σχηματισμού λέξεων, καθώς όλες οι λέξεις μιας γλώσσας είναι αδύνατο να περιληφθούν σε ένα λεξικό.
2. **Μορφοτακτικοί περιορισμοί (morphotactics)**, υποδεικνύουν την σειρά που ακολουθούν τα μορφήματα (χωρισμένα σε βασικές κλάσεις) προκειμένου να σχηματιστεί μια λέξη.
3. **Ορθογραφικοί κανόνες (orthographic rules)**.

Σύμφωνα με τους Jurafsky και Martin (2000) προκειμένου να αναγνωριστεί μια εισαγωγή δεδομένων αν αποτελεί λέξη και να πραγματοποιηθεί η περαιτέρω ανάλυση της, τότε χρησιμοποιούνται τα finitestate automata (FSAs), πρόκειται για λογισμικό, το οποίο χρησιμοποιεί το λεξικό σε συνδυασμό με τους μορφοτακτικούς κανόνες.

Η μορφολογική συντακτική ανάλυση μπορεί να πραγματοποιηθεί είτε χρησιμοποιώντας δύο είτε τρία επίπεδα, όταν περιλαμβάνει και τους κανόνες ορθογραφίας. Θα παρουσιαστεί η ανάλυση με τρία επίπεδα που θεωρείται και πιο πλήρης ώστε να γίνουν κατανοητά τα στάδια που περνάει ένα αυτόματος μορφολογικός συντακτικός αναλυτής για την ανάλυση μιας λέξης: μια λέξη μπορεί να διακριθεί σε τρία επίπεδα, το λεξικό (lexical), το οποίο αναπαριστά την λέξη ως απλή συνένωση μορφημάτων, το ενδιάμεσο (intermediate) το οποίο διορθώνει την ορθογραφία με βάση τους κανόνες και το επιφανειακό (surface), το οποίο δείχνει την πραγματική ορθογραφία της λέξης. Για να επιτευχθεί η μορφολογική συντακτική ανάλυση χρησιμοποιούνται κανόνες αντιστοίχισης (mapping rules), οι οποίοι αντιστοιχίζουν μια λέξη σε μια ακολουθία μορφημάτων συνοδευμένη με επιμέρους μορφολογικά χαρακτηριστικά. Έτσι ένας μετατροπέας αντιστοιχίζει μεταξύ του

finite-state transducer (FST), μια υπολογιστική συσκευή για τη μοντελοποίηση της Μορφολογίας, από ένα σύνολο συμβόλων σε ένα άλλο.

Σύμφωνα με τους Jurafsky και Martin (2000) υπάρχουν και μορφολογικοί συντακτικοί αναλυτές – μετατροπείς, οι οποίοι δε χρησιμοποιούν λεξικά και αυτοί χρησιμοποιούνται κυρίως στην ΑΠ.

3.2.3 Μέρη του λόγου και επισημειωτές

Σύμφωνα με τους Jurafsky και Martin (2000) ως **μέρη του λόγου** (partofspeech/POS) ορίζονται **οι κλάσεις στις οποίες ομαδοποιούνται οι λέξεις** και από τις οποίες αντλούνται αρκετές πληροφορίες σχετικά με τη λέξη αλλά και τις γειτονικές τις. Πολλές φορές στη βιβλιογραφία αναφέρονται και ως μορφολογικές κλάσεις ή λεξιλογικές ετικέτες (lexicaltags). Οι πληροφορίες αυτές που αντλούνται από το μέρος του λόγου της λέξης χρησιμοποιούνται σε πολλούς επιστημονικούς κλάδους όπως και στην ΑΠ, με πολλές διαφορετικές εφαρμογές. Ειδικότερα αναφέρουν οι Jurafsky και Martin (2000), την διαδικασία της περιστολής (stemming) ώστε να μπορούν να αναγνωριστούν τα είδη των προσφωμάτων/καταλήξεων που επιδέχεται μια λέξη, την ανάκτηση προκειμένου να πραγματοποιηθεί ανάκτηση συγκεκριμένων μερών του λόγου, τη δημιουργία αλγόριθμων αποσαφήνισης σημασίας λέξεων (automaticword-sensedisambiguatingalgorithms), τη συντακτική ανάλυση κειμένων και γενικότερες εφαρμογές εξαγωγής πληροφοριών.

Πιο συγκεκριμένα, σχετικά με τις κλάσεις των μερών του λόγου οι Jurafsky και Martin (2000) αναφέρουν πως **οι κλάσεις αυτές χωρίζονται σε δύο υπο-κλάσεις**: τις **κλειστές κλάσεις** (closedclass), όπου τα μέλη τους είναι σταθερά σε αντίθεση με τις ανοικτές κλάσεις (openclass), όπου υπάρχει πιθανότητα διαρκούς εμπλουτισμού των κλάσεων με νέες λέξεις που προκύπτουν είτε μέσω της επινόησης είτε δανεισμού από άλλες γλώσσες. Στις κλειστές κλάσεις ανήκουν: προθέσεις (prepositions), προσδιοριστές (determiners), αντωνυμίες (pronouns), σύνδεσμοι (conjunctions), βοηθητικά ρήματα (auxiliaries), μόρια (particles), αριθμοί (numerals). Στις ανοικτές ανήκουν: ουσιαστικά (nouns), ρήματα (verbs), επίθετα (adjectives), επιρρήματα (adverbs). Οι Jurafsky και Martin (2000) ορίζουν τη διαδικασία επισημείωσης μερών του λόγου (partofspeechtagging/tagging) ως τη διαδικασία με την οποία το λογισμικό αυτό αναγνωρίζει το μέρος του λόγου στο οποίο ανήκει μια λέξη. Η διαδικασία της επισημείωσης μπορεί να ταυτιστεί σε επίπεδο φυσικής γλώσσας με τη διαδικασία διαίρεσης σε σύμβολα. Οι επισημειωτές (taggers) είναι πολύ σημαντικοί στην ΑΠ αλλά το μεγαλύτερο τους πρόβλημα αποτελεί η ασάφεια των λέξεων και η σωστή ανάθεση μιας ετικέτας (tag), η οποία αντλείται από ένα σύνολο ετικετών (tagset).

Με βάση τους αλγόριθμους ανάθεσης ετικετών οι επισημειωτές χωρίζονται σε δύο κατηγορίες, τους επισημειωτές βασισμένους σε κανόνες (βάση δεδομένων με χειρόγραφους κανόνες αποσαφήνισης) και τους στοχαστικούς επισημειωτές (χρησιμοποιούν εκπαιδευμένο σώμα κειμένου για την αποσαφήνιση). Σύμφωνα με τους Jurafsky και Martin (2000) από το συνδυασμό των δύο παραπάνω κατηγοριών επισημειωτών έχουν προκύψει οι επισημειωτές Transformation-BasedTagging ή Brilltagging, οι οποίοι χρησιμοποιούν κανόνες προκειμένου να αναθέσουν ετικέτες σε λέξεις αλλά χρησιμοποιούν και την τεχνική εκμάθησης μηχανής, καθώς υποθέτουν ένα εκπαιδευμένο σώμα κειμένου (corpus) όπου ήδη έχουν ανατεθεί ετικέτες για αυτόματη πρόκληση των κανόνων. Ειδικότερα η λειτουργία τους είναι η εξής:

- Αρχικά πραγματοποιείται ανάθεση ετικετών στις λέξεις με βάση ένα σώμα κειμένου, που ήδη φέρει ετικέτες, όπως π.χ. το Brown corpus.
- Αφού ανατεθεί η πιθανότερη ετικέτα τότε εφαρμόζονται οι κανόνες μετατροπής και διορθώνονται οι αρχικές ετικέτες, αν χρειαστεί στην πορεία.
- Η εφαρμογή των κανόνων για ανάθεση ετικετών σε πρώτη φάση πραγματοποιείται σε μεγάλο μέρος του κειμένου με γενικούς κανόνες ενώ στη συνέχεια σταδιακά εφαρμόζονται κανόνες που εμπίπτουν σε όλο και μικρότερο μέρος κειμένου, φτάνοντας πια σε εντελώς εξειδικευμένους κανόνες για μεμονωμένες περιπτώσεις.

Όπως είναι φυσικό σε μια λέξη ανατίθενται αρκετές ετικέτες ώσπου να προκύψει η τελική απόφαση, η οποία διαμορφώνεται μέσα από την εφαρμογή των κανόνων. Απαραίτητο για την εφαρμογή όλων των προαναφερθέντων αλγορίθμων είναι ένα λεξικό με τα μέρη του λόγου για κάθε λέξη που υπάρχει (όσο αυτό είναι δυνατό), καθώς οι γλώσσες εμπλουτίζονται διαρκώς, συνεπώς είναι πρακτικά αδύνατο ένα λεξικό να τις περιέχει όλες.

Σύμφωνα με τους Jurafsky και Martin (2000), η σύνταξη αφορά το πως είναι οργανωμένες οι λέξεις και τα μέρη του λόγου μεταξύ τους, με σημαντικότερη την έννοια της συστατικότητας (constituency), η οποία αναφέρεται σε κλάσεις λέξεων/φράσεων οι οποίες συμπεριφέρονται ως μια μονάδα, δηλαδή ως ένα συστατικό. Όλες αυτές οι δομές μεταξύ των λέξεων και των συστατικών μοντελοποιούνται με μαθηματικά

συστήματα όπως το context-free grammar (CFG). Αυτά τα συστήματα αποτελούνται από ένα σύνολο από κανόνες βάση των οποίων ομαδοποιούνται σύμβολα της γλώσσας καθώς και από ένα λεξικό που περιέχει λέξεις και τα αντίστοιχα σύμβολα. Η συντακτική ανάλυση στην ΥΓ ταυτίζεται με την διαδικασία της περιστολής στην ΑΠ.

3.2.4 Σημασιολογία

Όπως φαίνεται και στην εικ. 17, η ΥΓ εκτός από την Μορφολογία και την Σύνταξη αφορά και την Σημασιολογία.

Όπως αναφέρει ο Kracht (2007) ως Σημασιολογία (Semantics) ορίζεται ο **τομέας της Γλωσσολογίας που ασχολείται με την σημασία των λέξεων**. Πιο συγκεκριμένα, ο Hayes (2010) συμπληρώνει πως η Σημασιολογία μελετά το **πώς η σημασία των λέξεων μεταφέρεται μέσω της γλώσσας**. Ως σημασία ορίζονται τα νοήματα τα οποία μεταφέρονται μέσα από τις λέξεις, τις προτάσεις. Ουσιαστικά η γλώσσα μπορεί να θεωρηθεί ως ένα από τα πιο περίπλοκα συστήματα συμβόλων. Έτσι και οι προτάσεις με τη σειρά τους αποτελούν σύμβολα τα οποία εκφράζουν τη σκέψη του ατόμου που τις δομεί. Συνεπώς μέσω της γλώσσας επιτυγχάνεται η μεταφορά των σκέψεων του. Όπως αναφέρει ο Kracht (2007), στην Σημασιολογία οι προτάσεις οι οποίες μπορούν να χαρακτηριστούν αληθής ή ψευδής ονομάζονται δηλώσεις (statements), οι οποίες εκφράζουν λογικές προτάσεις (propositions). Τέλος, ο Hayes (2010) αναφέρεται στον στόχο της Σημασιολογίας, ο οποίος είναι *«η μελέτη του πως η γλώσσα εμπεριέχει τη σκέψη, χωρίς ακόμα να υπάρχει μια καλώς ανεπτυγμένη θεωρία περί της σκέψης»*.

3.2.4.1 Προβλήματα στη Σημασιολογία

Το μεγάλο όμως εμπόδιο για την επικοινωνία και συνάμα το αδύνατο σημείο της γλώσσας αποτελεί, όπως αναφέρεται στη βιβλιογραφία (Karaman 2003), η αμφισημία (ambiguity), η οποία αν και αποτελεί μια αναπόσπαστη ιδιότητα της φυσικής γλώσσας, παρακωλύει τη διαδικασία της επικοινωνίας. Η έλλειψη αυτή της σαφήνειας του νοήματος αντιμετωπίζεται από τον άνθρωπο υποσυνείδητα καθώς εκτελούνται πραγματικές και σημασιολογικές διεργασίες, στις οποίες μεγάλο ρόλο διαδραματίζει το περιεχόμενο (βλέπε ενότητα 3.1.3), το οποίο περιβάλλει την λέξη που εμφανίζει την ασάφεια.

Σύμφωνα με τον Kennedy (2009) η αμφισημία (ambiguity) αφορά τη συσχέτιση μιας ακολουθίας χαρακτήρων με πολλές διαφορετικές σημασίες και χωρίζεται σε φωνολογική, λεξιλογική, δομική και πεδίου. Ειδικότερα όσον αφορά την σημασιολογική αμφισημία όπως αναφέρουν οι Jurafsky και Martin (2000) η αμφισημία είναι το φαινόμενο όπου το μέρος του λόγου στο οποίο ανήκει μια λέξη μέσα σε μια πρόταση δεν είναι σαφές. Έτσι μέσω της διαδικασίας της αποσαφήνισης (disambiguation) διευκρινίζεται ο ρόλος της λέξης.

Σύμφωνα με τους Jurafsky και Martin (2000) ως πολυσημία ορίζεται το φαινόμενο κατά το οποίο ένα λέξημα μπορεί να αντιστοιχεί σε πολλαπλές σχετικές έννοιες. Μάλιστα στη βιβλιογραφία (Konacs 2011) αναφέρεται πως το φαινόμενο της πολυσημίας έγκειται στο ότι δεν είναι εφικτή η διάκριση μεταξύ των διαφορετικών εννοιών μιας λέξης, καθώς και στο ότι δεν υπάρχει σαφής εικόνα ως προς το πόσες διαφορετικές έννοιες μπορεί να έχει μια λέξη. Ακόμη το νόημα μιας λέξης μπορεί να διακρίνεται σε κυριολεκτικό και μεταφορικό. Τέλος, ένα από τα μεγαλύτερα προβλήματα που αντιμετωπίζονται σε σχέση με την πολυσημία είναι η διάκριση της από την ομωνυμία (homonymy). Οι Jurafsky και Martin (2000) αναφέρουν την ομωνυμία ως τη σχέση μεταξύ λέξεων που έχουν την ίδια μορφή αλλά τα νοήματα τους δεν συσχετίζονται. Ακόμη στη βιβλιογραφία αναφέρεται (Konacs 2011) πως στην ομωνυμία οι λέξεις δε σχετίζονται ετυμολογικά ενώ προφέρονται με τον ίδιο τρόπο ή έχουν την ίδια ορθογραφία. Από την άλλη πλευρά στην πολυσημία η ετυμολογία είναι η ίδια, άρα υπάρχει σημασιολογική σχέση και τα διαφορετικά νοήματα πηγάζουν από τη μεταφορική χρήση της λέξης. Καθώς με την ομωνυμία και την πολυσημία η αμφισημία ενδυναμώνεται έγκειται στο περιεχόμενο η αποσαφήνιση του νοήματος. Τέλος, σύμφωνα με τους Jurafsky και Martin (2000) και η συνωνυμία αποτελεί ένα φαινόμενο της γλώσσας που παρακωλύει την κατανόηση της γλώσσας με αυτόματο τρόπο, καθώς πρόκειται για το φαινόμενο κατά το οποίο διαφορετικά λεξήματα αντιστοιχούν σε ένα κοινό νόημα.

Γενικότερα τόσο η πολυσημία όσο και η συνωνυμία αποτελούν σημαντικά ζητήματα και για την ΑΠ καθώς επηρεάζουν την ακρίβεια και την ανάκληση. Πιο συγκεκριμένα, όπως αναφέρουν οι Jurafsky και Martin (2000), χωρίς να είναι κανείς απόλυτος, η πολυσημία τείνει να μειώνει την ακρίβεια, καθώς επιστρέφει αποτελέσματα μη σχετικά ως προς τις πληροφοριακές ανάγκες του χρήστη ενώ η συνωνυμία τείνει

να μειώνει την ανάκληση, εφόσον έγγραφα που είναι σχετικά με τις πληροφοριακές ανάγκες του χρήστη παραλείπονται.

3.2.4.2 Μέτρα σημασιολογικής εγγύτητας (ομοιότητας)

Τα μέτρα σημασιολογικής εγγύτητας, όπως αναφέρουν οι Harispeetal. (2013), αποτελούν μέτρα, με την έννοια μαθηματικού εργαλείου, μέσω των οποίων είναι δυνατός ο υπολογισμός της σημασιολογικής συνάφειας στοιχείων, τα οποία μπορεί να είναι γλωσσικές μονάδες, έννοιες ή οντότητες, οι οποίες αντλούνται από κείμενα (αδόμητα ή ημι-δομημένα). Για τον υπολογισμό της εγγύτητας αναφέρουν πως υπάρχει πληθώρα μέτρων που υπολογίζουν την ομοιότητα ή διαφορά (γενικότερα την απόσταση) ανάμεσα σε συγκεκριμένες δομές δεδομένων (π.χ. διανύσματα) και σε τύπους δεδομένων (π.χ. αριθμητικά δεδομένα ή συμβολοσειρές). Η εφαρμογή των σημασιολογικών μέτρων έχει διεπιστημονικό χαρακτήρα και αφορά τις Γνωσιακές Επιστήμες, τη Γλωσσολογία, την Επεξεργασία Φυσικής Γλώσσας, το Σημασιολογικό Ιστό και άλλους επιστημονικούς κλάδους.

Όπως αναφέρεται στη βιβλιογραφία (Harispeetal. 2013) μέσω της σημασιολογικής εγγύτητας ένα στοιχείο προς σύγκριση προέρχεται από έναν σημασιολογικό χώρο, όπου π.χ. το στοιχείο θα μπορούσε να αντιστοιχεί σε μια πρόταση και ο σημασιολογικός χώρος σε ένα κείμενο. Μάλιστα η έννοια του σημασιολογικού χώρου μπορεί να αντιστοιχηθεί με έναν δειγματικό χώρο με σημασιολογική υπόσταση και τα αντίστοιχα στοιχεία που περιλαμβάνει δύναται να αναπαρασταθούν μέσω μιας συγκεκριμένης δομής δεδομένων όπως π.χ. ένα διάνυσμα. Έτσι και η αναπαράσταση αυτή παίρνει πλέον σημασιολογικές διαστάσεις.

Τα σημασιολογικά μέτρα ομοιότητας κατηγοριοποιούνται όπως αναφέρεται στη βιβλιογραφία (Harispeetal. 2013) σύμφωνα με το είδος των στοιχείων προς σύγκριση, τη σημασιολογική πηγή (semanticproxy) από όπου προέρχονται οι πληροφορίες για τα στοιχεία και τέλος την κανονική μορφή (canonicalform) που χρησιμοποιείται για την αναπαράσταση των στοιχείων. Για τη σχεδίαση των σημασιολογικών μέτρων ορίζεται μια συνάρτηση υπολογισμού της ομοιότητας.

Ειδικότερα για την περίπτωση αναπαράστασης λέξεων, όπως αναφέρεται στη βιβλιογραφία (Harispeetal. 2013) η κλασική κανονική μορφή υλοποιείται μέσω της αναπαράστασης με διανύσματα, η οποία για την ΑΠ αντιστοιχεί στο γνωστό μοντέλο VSM.

Έτσι σύμφωνα με τη βιβλιογραφία (Harispeetal. 2013) τα σημασιολογικά μέτρα ομοιότητας χωρίζονται σε τρεις μεγάλες κατηγορίες: τα κατανεμημένα (distributional), τα βασισμένα σε γνώση (knowledge-based) και το συνδυασμό αυτών. Πιο συγκεκριμένα τα κατανεμημένα αφορούν σύγκριση γλωσσικών μονάδων, οι οποίες προέρχονται από σημασιολογική πηγή που αποτελεί κείμενο, ενώ τα βασισμένα σε γνώση αφορούν μονάδες όπως έννοιες, ομάδες εννοιών, οι οποίες προέρχονται από σημασιολογική πηγή δομημένης γνώσης, όπως θησαυροί ή οντολογίες.

Ειδικότερη σημασία έχουν τα **κατανεμημένα** καθώς αφορούν σε κείμενα και χωρίζονται στις εξής προσεγγίσεις:

- **Γεωμετρικές** (geometric), τα στοιχεία που συγκρίνονται αναπαρίστανται ως διανύσματα και το μέτρο ομοιότητας που χρησιμοποιείται περισσότερο είναι η μέτρηση του συνημίτονου της γωνίας που σχηματίζουν μεταξύ τους τα διανύσματα (βλέπε ενότητα 2.2.1).
- **Βασισμένες σε ασαφή σύνολα** (fuzzy – setbased), συγκρίνεται συνήθως ο αριθμός κοινής εμφάνισης συγκρινόμενων στοιχείων και συχνά χρησιμοποιούνται σχήματα απόδοσης βάρους.
- **Πιθανοτικές** (probabilistic), αφορά τη δύναμη της σχέσης μεταξύ των συγκρινόμενων στοιχείων και την κοινή πληροφορία που πιθανώς φέρουν, η οποία μεταφράζεται στην πιθανότητα τα δύο στοιχεία να συνυπάρχουν στην ίδια συλλογή.

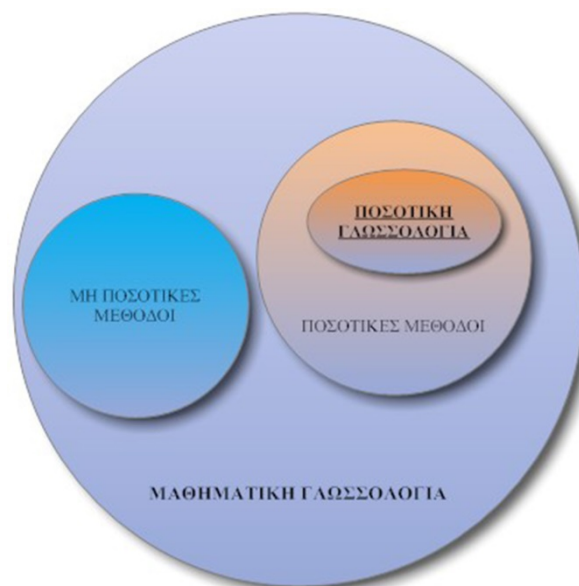
•

3.3 Ποσοτική Γλωσσολογία

Η συγκέντρωση δεδομένων, η επεξεργασία τους και η έκφραση τους με ποσοτικό τρόπο με σκοπό τη διεξαγωγή συμπερασμάτων από αυτά, αποτελεί μια σημαντική εργασία στο πλαίσιο πολλών επιστημονικών κλάδων. Σύμφωνα με τον Johnson (2008) για την επιστήμη της Γλωσσολογίας ειδικότερα αποτελεί πρακτική δεκαετιών για πολλούς υπο-κλάδους της. Σε αυτούς συγκαταλέγεται και η ΥΓ όπου τα παραπάνω αποτελούν βασικό ένα κομμάτι εκπαίδευσης της.

Εκτός από την ΥΓ που έχει αναλυθεί έως τώρα, όπως ήδη έχει αναφερθεί (βλέπε ενότητα 1.3), από τη στροφή της Γλωσσολογίας προς την ποσοτική μελέτη έχει δημιουργηθεί ένας ξεχωριστός υπο-κλάδος, αυτός

της ΠΓ. Μάλιστα οι Bolshakov και Gelbukh (2004) κατατάσσουν την ΠΓ (Quantitative Linguistics) ως υπο-κλάδο της Μαθηματικής Γλωσσολογίας (Mathematical Linguistics). Σύμφωνα με το Oxforddictionaries ως ΠΓ ορίζεται: «*Η συγκριτική μελέτη της συχνότητας και της κατανομής των λέξεων και των συντακτικών δομών σε διαφορετικά κείμενα*». Ακόμη στη βιβλιογραφία αναφέρεται πως (Liu και Huang 2012) η ΠΓ αφορά στα διάφορα γλωσσικά φαινόμενα, γλωσσικές δομές, δομικές ιδιότητες και τους μεταξύ τους συσχετισμούς στις δραστηριότητες επικοινωνίας στην πραγματική ζωή. Μέσω διαφόρων ποσοτικών τεχνικών η ΠΓ διεξάγει ακριβείς μετρήσεις, παρατηρήσεις, προσομοιώσεις, μοντελοποιήσεις και επεξήγηση αυτών των φαινομένων, με σκοπό την ανακάλυψη μαθηματικών νόμων που αφορούν τα γλωσσικά φαινόμενα. Ακόμη οι Bolshakov και Gelbukh (2004) συμπληρώνουν πως μέσω της ΠΓ παρέχονται οι μέθοδοι για τη λήψη αποφάσεων στην επεξεργασία κειμένου (textprocessing), με βάση τα στατιστικά δεδομένα. Τέλος, η Tesitelova (1992) συμπληρώνει πως η ΠΓ αφορά σε ποσοτικές μεθόδους που εφαρμόζονται στο πλαίσιο της Μαθηματικής Γλωσσολογίας. Σχηματικά η ΠΓ αναπαρίσταται στο σχήμα 3.6. Σύμφωνα με τη βιβλιογραφία (Johnson 2008) οι βασικοί στόχοι της ΠΓ είναι οι εξής: (α) μετατροπή δεδομένων (μείωση των ποσοτικών δεδομένων στα ουσιαστικά τους τμήματα ώστε να περιγραφούν τάσεις και κοινά στοιχεία), (β) διεξαγωγή συμπερασμάτων (μέσω ελέγχου υποθέσεων), (γ) έρευνα και περιγραφή σχέσεων μεταξύ των δεδομένων και τέλος, (δ) εξερεύνηση και περιγραφή διαδικασιών που έχουν ως βάση τη θεωρία Πιθανοτήτων, όπως η θεωρία της Πληροφορίας.



Σχήμα 3. 6. ΠΓ ως υπο-κλάδος της Μαθηματικής Γλωσσολογίας

3.3.1 Στατιστικοί νόμοι στην ΠΓ

Όπως αναφέρθηκε ήδη στην παραπάνω ενότητα, η ΠΓ περιγράφει τα διάφορα υπό μελέτη γλωσσικά φαινόμενα μέσω της μαθηματικής έκφρασης και πιο συγκεκριμένα μαθηματικών ή αλλιώς στατιστικών νόμων. Μάλιστα στην βιβλιογραφία (Hogan 2014) οι νόμοι αυτοί κατηγοριοποιούνται σε δύο ομάδες στην ΠΓ: στους νόμους κατανομής (distributional) που αφορούν στις κατανομές πιθανοτήτων με χαρακτηριστικό παράδειγμα το νόμο του Zipf και στους λειτουργικούς (functional) νόμους, οι οποίοι αφορούν στη σύνδεση μεταξύ γλωσσικών ιδιοτήτων, με χαρακτηριστικό παράδειγμα το νόμο Menzies – Altmann.

Σε αυτό το σημείο κρίνεται απαραίτητη η επεξήγηση του ορισμού της έννοιας του «νόμου». Όπως αναφέρεται στη βιβλιογραφία (Hogan 2014) ο επιστημονικός νόμος ουσιαστικά αποτελεί μια υπόθεση η οποία ισχύει πάντοτε για τα αντικείμενα που αφορά, η οποία από την μια πλευρά συνδέεται με άλλες υποθέσεις ενός επιστημονικού κλάδου και από την άλλη επιβεβαιώνεται μέσω εμπειρικών δεδομένων. Μάλιστα στη βιβλιογραφία (Hogan 2014) ο επιστημονικός νόμος αποτελεί τη βάση για την κατασκευή μιας θεωρίας, καθώς και αυτή με τη σειρά της αποτελείται από ένα σύνολο νόμων που συσχετίζονται μεταξύ τους, δηλαδή ένα σύστημα νόμων. Οι νόμοι χρησιμοποιούνται προκειμένου να υποστηρίξουν και να τεκμηριώσουν επιστημονικές εξηγήσεις. Τέλος, να επεξηγηθεί η διάκριση των νόμων και των θεωριών από τα εμπειρικά

δεδομένα. Όπως αναφέρει ο Hempel (1964) οι εμπειρικές επιστήμες αφορούν στην περιγραφή φαινομένων του κόσμου σύμφωνα με την εμπειρία και στην καθιέρωση γενικών αρχών ώστε τα φαινόμενα αυτά να μπορούν να επεξηγηθούν και να προβλεφθούν.

Στη Γλωσσολογία υπάρχει πληθώρα νόμων σε σχέση με στατιστικές παρατηρήσεις και φαινόμενα που αφορούν τη γλώσσα. Παρακάτω θα αναλυθούν οι νόμοι των Menzerath – Altmann (MA) και Zipf. Ακόμη θα αναλυθεί η μεταξύ τους σχέση.

3.3.2 Νόμος των Menzerath-Altmann

Ένας λοιπόν από τους σημαντικότερους νόμους της ΠΓ είναι ο νόμος MA. Σύμφωνα με τη βιβλιογραφία (Buk και Rovenchak 2008), ο παραπάνω νόμος αφορά στην περιγραφή της σχέσης μεταξύ δομήματος (construct) και συστατικού (constituent), όπου όσο μεγαλώνει το πρώτο, τόσο μικραίνει το δεύτερο. Το 1928 ο Paul Menzerath, πειραματικός ψυχολόγος και φωνολόγος, παρατήρησε ότι όσο ο μέσος όρος του μήκους μιας συλλαβής μειώνονταν τόσο ο αριθμός των συλλαβών που απαρτίζουν τη λέξη αυξανόταν. Σύμφωνα με τον Eroglu (2013) ο Menzerath, ο οποίος θεωρείται ένας από τους πρωτοπόρους ερευνητές στην ΠΓ πρώτος ανέδειξε την αρνητική συσχέτιση στη σχέση μεταξύ του μήκους ενός γλωσσικού δομήματος με τα συστατικά του. Σύμφωνα με τη βιβλιογραφία (Buk και Rovenchak 2008) αργότερα ο Gabriel Altmann εξέφρασε με μαθηματική μορφή την παραπάνω σχέση μεταξύ δομήματος και συστατικού του Menzerath έτσι ώστε σήμερα ο παραπάνω νόμος είναι γνωστός ως νόμος των MA. Παρότι ο Menzerath αρχικά αναφέρθηκε στη σχέση μεταξύ συλλαβής - λέξης είναι ξεκάθαρο ότι ο παραπάνω νόμος μπορεί να εφαρμοστεί σε διάφορα επίπεδα γλωσσικών μονάδων αρκεί να περιγράψει μια σχέση δομήματος - συστατικού, σύμφωνα με τη βιβλιογραφία (Mikros και Milicka 2014).

Σύμφωνα με τον Eroglu (2013), παρότι ο νόμος MA αποτελεί έναν από τους βασικότερους νόμους της ΠΓ, η χρήση του δεν αφορά μόνο στην επιστήμη της Γλωσσολογίας αλλά βρίσκει εφαρμογή σχεδόν σε οποιαδήποτε περιγραφή οργανωτικής δομής σε πληθώρα επιστημονικών πεδίων όπως για την περιγραφή οργανωτικής δομής μουσικών κειμένων. Μάλιστα οι Baixeries et al. (2013) αναφέρουν σε άρθρο τους τη χρήση του νόμου MA για την περιγραφή της σχέσης μεταξύ γονιδιωμάτων και χρωμοσωμάτων, όπου παρατήρησαν την ίδια αρνητική συσχέτιση, δηλαδή πως όσο μεγαλύτερο είναι το γονιδίωμα τόσο μικρότερα είναι τα χρωμοσώματα (σε ζεύγη βάσεων).

Όπως αναφέρει ο Altmann (1980), ένα βασικό ζήτημα στο νόμο MA αποτελεί το τι θεωρείται συστατικό. Δηλαδή ποιες μονάδες θεωρούνται συστατικά, οι αμέσως επόμενες στην ιεραρχία ή και άλλες, προερχόμενες από μετέπειτα επίπεδα ιεραρχίας. Με βάση αυτό το ζήτημα, προβλήματα στην εφαρμογή της μαθηματικής έκφρασης του νόμου των MA μπορούν να παρατηρηθούν για γλώσσες που περιέχουν μη συλλαβικές λέξεις.

Πέραν του θεωρητικού μέρους του νόμου MA που αναπτύχθηκε παραπάνω, θα παρατεθούν και οι μαθηματικές εκφράσεις του νόμου. Σύμφωνα με τον Altmann (1980), στην παρακάτω εξίσωση εκφράζεται ένας «σταθερός ρυθμός μείωσης του μήκους του συστατικού y» μέσω του τύπου (3.1):

$$\frac{y'}{y} = -c \quad (3.1)$$

Μέσω της μαθηματικής ολοκλήρωσης (integration), ο παραπάνω μαθηματικός τύπος παίρνει τη μορφή του (3.2):

$$y = Ae^{-cx} \quad (3.2)$$

Όπου το x αντιστοιχεί στο δόμημα και το c σε μια μεταβλητή. Έτσι μέσω της εξίσωσης (3.3), παρουσιάζεται, όπως αναφέρεται στη βιβλιογραφία (Mikros και Milick 2014), η σχέση του μήκους δομήματος x και μήκους συστατικού y, η οποία περιγράφεται από μια μονότονη φθίνουσα συνάρτηση (monotonically decreasing function).

Στην παρακάτω διαφορετική εξίσωση από τη βιβλιογραφία (Kulacka και Macutek 2007), ορίζεται ξανά η σχέση μεταξύ συστατικού και δομήματος. Όπως φαίνεται από το μαθηματικό τύπο (3.3) το συστατικό y' είναι ανάλογο του μέσου μήκους y και αντιστρόφως ανάλογο του μήκους x.

$$y' = \frac{cy}{x} \quad (3.3)$$

Σύμφωνα με τον Eroglu (2013), «ο νόμος του MA αποτελεί ένα συνεχόμενο μοντέλο κατανομής πιθανοτήτων (*probability distribution model*), το οποίο χρησιμοποιείται για να περιγράψει την πιθανοτική σχέση μεταξύ των διακριτών αποτελεσμάτων των ποσοτήτων» όπως φαίνεται στο μαθηματικό τύπο (3.4):

$$y(x | A, b, c) = Ax^b e^{-cx} \quad (3.4)$$

Ξανά, το y αντιστοιχεί σε συστατικό, το x σε δόμημα και οι A , b , c αναπαριστούν ελεύθερες παραμέτρους.

3.3.3 Νόμος του Zipf

Σύμφωνα με τον Wyllys (1981) ο νόμος του Zipf περιγράφει τη σχέση μεταξύ της συχνότητας εμφάνισης λέξεων σε ένα σώμα κειμένου και της κατάταξής τους, η οποία περιγράφηκε από τον George Kingsley Zipf, καθηγητή φιλολογίας και γλωσσολόγος στο Harvard University. Το 1935 στο βιβλίο του “*The Psychology of Language*”, ο Zipf περιγράφει την αλγεβρική μαθηματική έκφραση, γνωστή αργότερα ως νόμος του Zipf. Στην επιστήμη των μαθηματικών ο νόμος του Zipf κατατάσσεται στους νόμους των δυνάμεων (*power laws*). Με βάση το Oxford Dictionary (2015) ως *power law* ορίζεται «μια σχέση μεταξύ δύο ποσοτήτων έτσι ώστε η μια είναι ανάλογη σε μια σταθερή δύναμη της άλλης». Σύμφωνα με τη βιβλιογραφία (Powers 1998) ο νόμος του Zipf βρίσκει εφαρμογή σε πληθώρα επιστημονικών κλάδων που ασχολούνται με τη φυσική γλώσσα όπως αυτών της Γλωσσολογίας και ειδικότερα της ΠΓ και της Υπολογιστικής Ψυχολογίας. Παρατηρείται στη διεθνή βιβλιογραφία (Piantadosi 2014) ότι ο νόμος του Zipf βρίσκει εφαρμογή εκτός από τη φυσική γλώσσα και σε πληθώρα άλλων επιστημών όπως στη μουσική, στα υπολογιστικά συστήματα, στο διαδίκτυο και στα φυσικά και βιολογικά συστήματα.

Σύμφωνα με τους Manning και Schutze (1999) ο νόμος του Zipf μπορεί να χρησιμοποιηθεί ως μια γενικευμένη περιγραφή της κατανομής συχνότητας των λέξεων στις ανθρώπινες γλώσσες και η περιγραφή αυτή διαχωρίζει τη συχνότητα εμφάνισης των λέξεων σε χαμηλή, μεσαία και υψηλή συχνότητα εμφάνισης λέξεων. Οι Manning και Schutze (1999) αναφέρουν πως με το νόμο του Zipf μπορεί να ερευνηθεί η σχέση μεταξύ της συχνότητας εμφάνισης μιας λέξης f σε ένα σώμα κειμένου και της θέσης κατάταξης r της λέξης αυτής, αφού έχει ήδη μετρηθεί η συχνότητα εμφάνισης για όλες τις λέξεις του σώματος κειμένου και έχει συνταχθεί μια λίστα σε φθίνουσα σειρά ξεκινώντας με τη λέξη με τη συχνότερη εμφάνιση. Μάλιστα τονίζουν ότι χρησιμοποιώντας το νόμο του Zipf τα δεδομένα στη γραφική παράσταση σχετικά με τη χρήση των περισσότερων λέξεων θα είναι συνήθως εξαιρετικά αραιά (*sparse*).

Στην εργασία του Wyllys (1981) περιγράφεται ο νόμος του Zipf αναλυτικότερα. Για να επεξηγηθεί, υποτίθεται η ύπαρξη ενός σώματος κειμένου σε φυσική γλώσσα και όπως και παραπάνω μετρίεται η συχνότητα εμφάνισης των λέξεων σε αυτό ώστε να σχηματιστεί η λίστα κατάταξης των λέξεων σε φθίνουσα σειρά με την πιο συχνή λέξη να έχει την πρώτη θέση. Ως νόμος του Zipf ορίζεται η παρακάτω μαθηματική έκφραση (3.5):

$$r \cdot f = c \quad (3.5)$$

όπου το r αντιστοιχεί στην κατάταξη μιας λέξης, το f στη συχνότητα εμφάνισης της λέξης και c αντιστοιχεί σε μια σταθερά, η οποία εξαρτάται από το σώμα κειμένου.

Σύμφωνα με τη βιβλιογραφία (Sorell 2012), τοποθετώντας τις λέξεις ενός σώματος κειμένου σε φθίνουσα σειρά ξεκινώντας από αυτήν με τη μεγαλύτερη συχνότητα, τότε η δεύτερη συχνότερη λέξη θα εμφανίζεται περίπου τις μισές φορές από ότι η πρώτη και η τρίτη συχνότερη λέξη περίπου 1/3 φορές από ότι η πρώτη κ.ο.κ. Έτσι πολλαπλασιάζοντας την κατάταξη r με τη συχνότητα f της κάθε λέξης, η σταθερά c θα πρέπει να παραμένει περίπου ίδια για κάθε λέξη. Υποδεικνύει λοιπόν μια αντιστρόφως ανάλογη σχέση μεταξύ κατάταξης και συχνότητας των λέξεων ενός κειμένου.

Στην εργασία του Wyllys (1981) ο νόμος του Zipf εκτός από την παραπάνω αλγεβρική του έκφραση περιγράφεται επίσης ως ισοδύναμος με τη γραφική αναπαράσταση:

$$\log r \cdot \log f = \log c \quad (3.6)$$

Όπου στη σχεδίαση των ζευγών σημείων που προκύπτουν, ο λογάριθμος της κατάταξης r τοποθετείται στον οριζόντιο άξονα και ο λογάριθμος της συχνότητας f στον κάθετο άξονα και έτσι τα σημεία σχηματίζουν μια ελαφρά καμπύλη γραμμή, γνωστή ως καμπύλη Zipf (Zipf's curves).

Ο Wyllys (1981) αναφέρει ότι ο υπολογισμός με βάση το νόμο του Zipf έχει πιο έγκυρα αποτελέσματα κυρίως όσον αφορά την κατάταξη λέξεων με μεσαία τάξη εμφάνισης παρά για τις λέξεις με πολύ υψηλή ή χαμηλή συχνότητα εμφάνισης. Ακόμη αναφέρει ότι η εργασία του Zipf δείχνει πως το μέγεθος του δείγματος θα πρέπει να αποτελείται από τουλάχιστον 5000 λέξεις ώστε το $r \times f$ να είναι σταθερό, ακόμη και για τις μεσαίες κατατάξεις.

Σύμφωνα με τον Altmann (2002) ο νόμος του Zipf εστιάζει στις σχέσεις μεταξύ των διαφορετικών οντοτήτων της γλώσσας και αναφέρει πολλά διαφορετικά είδη σχέσεων που έχουν παρατηρηθεί μεταξύ των οντοτήτων αυτών. Στη βιβλιογραφία (Hřebíček 2002) οι οντότητες αυτές διακρίνονται για το νόμο του Zipf, σε λεξιλογικές μονάδες και σώματα κειμένου, τα οποία στο πλαίσιο μιας συγκεκριμένης δομής κειμένου έχουν αμοιβαίες σχέσεις μεταξύ τους. Οι σχέσεις αυτές που αναπτύσσονται μεταξύ των οντοτήτων έχουν διάφορες ιδιότητες εκ των οποίων οι δύο παρακάτω, όπως αναφέρονται στη βιβλιογραφία (Hřebíček 2002):

1. «Κάθε παρατηρούμενη συνεχής ακολουθία ήχου μιας γλώσσας λειτουργεί ως φορέας της μη συνεχούς ακολουθίας κωδικών συμβόλων διαφορετικών γλωσσικών επιπέδων».
2. «Οι μονάδες διαφορετικών επιπέδων μπορούν να περιγραφούν ως σύνολα που διακρίνονται από αυτοομοιότητα» (δηλαδή την ιδιότητα ενός σχήματος να είναι όμοιο με ένα ή περισσότερα τμήματά του).

Οι παραπάνω ιδιότητες προκύπτουν από την ιεραρχική σχέση μεταξύ των οντοτήτων, όπως αναφέρει και ο Altmann (2002), δηλαδή τις σχέσεις ανάμεσα σε διαφορετικά επίπεδα, οι οποίες εκφράζονται κυρίως μέσω του νόμου MA. Με βάση το νόμο MA λοιπόν μπορούμε να δούμε το κείμενο ως μια δομούμενη γλωσσική μονάδα, όπου η βασική δομική της μονάδα είναι ο κώδικας της γλώσσας. Συνεπώς διαφαίνεται ήδη η σχέση μεταξύ των δύο νόμων, MA και Zipf, καθώς οι δύο παραπάνω ιδιότητες προκύπτουν μέσω του νόμου των MA και αποτελούν βασικές ιδιότητες που διέπουν τις σχέσεις μεταξύ των γλωσσικών μονάδων, οι οποίες αφορούν κυρίως το νόμο του Zipf. Η σχέση μεταξύ των δύο νόμων αποδεικνύεται μαθηματικά μέσω μιας υποπερίπτωσης της εξίσωσης (16) (βλέπε ενότητα 3.3.1.1), όπου περιγράφεται ο νόμος MA. Ειδικότερα μέσω της συγκεκριμένης υπο-περίπτωσης του, όπως αναφέρεται σε Eroglu (2013), όταν $b \neq 0$ και $c=0$ τότε κανείς οδηγείται στην εξίσωση του νόμου Zipf, όπως φαίνεται παρακάτω στον τύπο (3.7):

$$y(x | A, b) = Ax^{-b} \quad (3.7)$$

Συνεπώς κάτω από συγκεκριμένες συνθήκες ο νόμος του Zipf, αποτελεί υπο-περίπτωση του νόμου MA.

3.3.4 Θεωρία της Πληροφορίας και ΠΓ

Όπως ήδη έχει αναφερθεί παραπάνω παρατηρείται η εστίαση της ενασχόλησης των μεθόδων ΠΓ και των τάσεων στην έρευνα της ΠΓ σχετικά με την εφαρμογή διαφόρων μαθηματικών και υπολογιστικών θεωριών, όπως η θεωρία Πληροφορίας και η θεωρία Πιθανοτήτων. Για το λόγο αυτό σε αυτή την ενότητα θα παρουσιαστεί η θεωρία της πληροφορίας (Shannon C. E. 1998), καθώς θα επιχειρηθεί η σύνδεσή της με το σχήμα απόδοσης βάρους (tf – idf) και τη θεωρία πιθανοτήτων.

Όπως αναφέρεται στη βιβλιογραφία (Nadel 2005) ο Shannon θεωρείται ο θεμελιωτής της θεωρίας της Πληροφορίας, η οποία μπορεί να χαρακτηριστεί ως μια μαθηματική θεωρία για την επικοινωνία. Όρισε στο πλαίσιο της διαδικασίας επικοινωνίας ένα σύστημα αποτελούμενο από έναν δέκτη, ένα κανάλι μεταφοράς και έναν αποδέκτη. Ο σκοπός του συστήματος αυτού είναι η μεταφορά κάποιας πληροφορίας σε μορφή μηνύματος. Πιο συγκεκριμένα κάθε μήνυμα περιέχει μια ποσότητα πληροφορίας, με την οποία ασχολήθηκε ο Shannon, ο οποίος εισήγαγε τη μέτρηση τόσο της ποσότητας της πληροφορίας αυτής, όσο και της χωρητικότητας του καναλιού μετάδοσης.

Όπως αναφέρεται στη βιβλιογραφία (Nadel 2005), ο ορισμός της πληροφορίας του Shannon είναι αρκετά τεχνικός και προσεγγίζει την ποσότητα της πληροφορίας συνδυάζοντας τη θεωρία Πιθανοτήτων. Πιο συγκεκριμένα, θεωρεί την πληροφορία αυτή ως ένα ενδεχόμενο δειγματικού χώρου για το οποίο υπάρχει αβεβαιότητα εάν θα συμβεί ή όχι και η αβεβαιότητα μειώνεται μονάχα με την παρατήρηση πραγματοποίησης του. Τότε μπορεί κανείς να είναι σίγουρος για το αποτέλεσμα της πιθανότητας του. Έτσι η ποσότητα κατά την οποία μειώνεται η αβεβαιότητα για το πιθανό αποτέλεσμα αποτελεί την περιεχόμενη πληροφορία για ένα

ενδεχόμενο. Αντιστοιχίσε δηλαδή την πραγματοποίηση ενός ενδεχομένου με μια πιθανότητα πραγματοποίησης του. Όπως π.χ. κατά την ρίψη ενός ζαριού, όπου θα αντιστοιχεί η πιθανότητα να ισχύει μια από έξι πιθανές τιμές και τότε η πιθανότητα που αντιστοιχεί στην πληροφορία είναι ίση με 1/6.

Στη βιβλιογραφία (Nadel 2005) αναφέρεται ακόμη πως η πληροφορία μεταφέρεται από σύμβολα, άρα στα σύμβολα αναλογεί η πιθανότητα εμφάνισης. Ο Shannon ασχολήθηκε αρχικά με την αντιστοίχιση ενός συμβόλου, η οποία όμως μπορεί να επεκταθεί και σε μια ροή συμβόλων, όπως π.χ. είναι μια λέξη. Υπολόγισε το μέσο όρο πληροφορίας που δύναται να μεταφερθεί ανά σύμβολο σε μια ροή συμβόλων, ποσότητα που ονόμασε ως εντροπία.

Τα δύο βασικά προβλήματα της θεωρίας Πληροφορίας, σύμφωνα με τη βιβλιογραφία (Nadel 2005) είναι η αποδοτικότητα μετάδοσης (συμπύεση δεδομένων) και η αξιοπιστία μετάδοσης όταν το κανάλι έχει θόρυβο, ο οποίος συνδέεται με τον υπολογισμό της εντροπίας.

Στη βιβλιογραφία (Robertson 2004) αναφέρεται ακόμη πως η θεωρία του Shannon μπορεί να συνδεθεί με το συστατικό *idf*, το οποίο χρησιμοποιείται κυρίως στο σχήμα απόδοσης βαρών *tf-idf*. Το *idf* συστατικό ορίζεται από τον παρακάτω μαθηματικό τύπο (3.8):

$$idf(t_i) = \log \frac{N}{n_i} \quad (3.8)$$

Όπου το *idf* υποδηλώνει τη σπανιότητα ενός όρου σε μια συλλογή. Παρατηρώντας το λόγο του λογαρίθμου μπορεί κανείς να παρατηρήσει το λόγο για τον υπολογισμό μιας πιθανότητας, αλλά αντεστραμμένο. Έτσι θα μπορούσε να υπολογιστεί η πιθανότητα για ένα τυχαίο έγγραφο να περιέχει έναν όρο t_i , από τον μαθηματικό τύπο (3.9):

$$P(t_i) = \frac{n_i}{N} \quad (3.9)$$

Έτσι μέσω των τύπων (20) και (21), είναι εφικτή η σύνδεση του *idf* συστατικού με τη θεωρία της Πληροφορίας του Shannon, η οποία είναι εμφανής στον παρακάτω μαθηματικό τύπο (3.10):

$$idf(t_i) = -\log P(t_i) \quad (3.10)$$

Να σημειωθεί πως σύμφωνα με τον Nadel (2005) για τη θεωρία της Πληροφορίας η ποσότητα που αντιστοιχεί στο $-\log P(t_i)$, αποτελεί την ποσότητα πληροφορίας στη μετάδοση μηνύματος, η οποία παίρνει μέρος στη συνάρτηση υπολογισμού της εντροπίας. Τα δε μηνύματα μετάδοσης πληροφορίας του Shannon μπορούν κάλλιστα να αντιστοιχηθούν σε μεταβλητές, που αντιστοιχούν σε ένα δειγματικό χώρο (με όλα τα πιθανά μηνύματα), σε μια συνάρτηση υπολογισμού για την πιθανότητα και σε κάποιο μέτρο πιθανότητας.

Βιβλιογραφία

- Ahlsen, E. (2006). *Introduction to Neurolinguistics*. Amsterdam/Philadelphia: John Benjamins Publishing
- Akobeng, A. K. (2007). Understanding diagnostic tests 1: sensitivity, specificity and predictive values. *Acta Paediatr.* 96 (3). P. 338-41.
- Altmann, G. (2002). Zipfian linguistics. *Glottometrics*. 3. P. 19-26.
- Baeza-Yates, R., Ribeiro-Neto, B., & others. (1999). Modern information retrieval (T. 463). ACM press New York. Ανακτήθηκε από http://mail.im.tku.edu.tw/seke/slide/baeza-yates/chap10_user_interfaces_and_visualization-modern_ir.pdf
- Baixerries, J., Hernandez-Fernandez, A., Forns, N., & Ferrer-i-Cancho, R. (2013). The parameters of the Menzerath-Altmann Law in genomes. *Journal of Quantitative Linguistics*, 20(2), 94–104.
- Bendersky, M., & Croft, W. B. (2012). Modeling higher-order term dependencies in information retrieval using query hypergraphs. Στο Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval (σσ 941–950). ACM. Ανακτήθηκε από <http://dl.acm.org/citation.cfm?id=2348408>
- Blair, D. C., & Kimbrough, S. O. (2002). Exemplary documents: a foundation for information retrieval design. *Information processing & management*, 38(3), 363–379.
- Bokos, M. P. G., Papavlasopoulos, N. K. S., & Avlonitis, M. (2007a). Specific selection of FFT amplitudes from audio sports and news broadcasting for classification purposes. Ανακτήθηκε από <http://www.emis.ams.org/journals/JGAA/accepted/2007/Poulos+2007.11.1.pdf>
- Bokos, M. P. G., Papavlasopoulos, N. K. S., & Avlonitis, M. (2007b). Specific selection of FFT amplitudes from audio sports and news broadcasting for classification purposes. Ανακτήθηκε από <http://www.emis.ams.org/journals/JGAA/accepted/2007/Poulos+2007.11.1.pdf>
- Bolshakov, I. A., & Gelbukh, A. (2004). Computational Linguistics Models, Resources, Applications.
- Brants, T. (2003). Natural Language Processing in Information Retrieval. Στο CLIN. Ανακτήθηκε από <http://www.clips.ua.ac.be/clin2003/proc/03Brants.pdf>
- Buk, S., & Rovenchak, A. (2008). Menzerath–Altmann Law for Syntactic Structures in Ukrainian. *Glottology*, 1(1), 10–17.
- Büttcher, S., Clarke, C. L., & Cormack, G. V. (2010). Information retrieval: Implementing and evaluating search engines. Mit Press. Ανακτήθηκε από <https://www.google.com/books>
- Carstens, W. (1999). Text-linguistics: Relevant linguistics. School of Languages and Arts, Potchefstroom University for CHE. Ανακτήθηκε από <http://www.pala.ac.uk/uploads/2/5/1/0/25105678/carstens.pdf>
- Cerulo, L., & Canfora, G. (2004). A taxonomy of information retrieval models and tools. *CIT. Journal of computing and information technology*, 12(3), 175–194.
- Crestani, F., & Wu, S. (2006). Testing the cluster hypothesis in distributed information retrieval. *Information Processing & Management*, 42(5), 1137–1150.
- Croft, W. B., Metzler, D., & Strohman, T. (2010). Search engines: Information retrieval in practice. Addison-Wesley Reading. Ανακτήθηκε από http://library.mpib-berlin.mpg.de/toc/z2009_2465.pdf
- Dale, E., & Chall, J. S. (1949). The concept of readability. *Elementary English*, 26(1), 19–26.
- De Beaugrande, R. (1981). Introduction to text linguistics. Ανακτήθηκε από <http://www.citeulike.org/group/236/article/202549>
- Diggle, P. J., & Chetwynd, A. G. (2011). Statistics and scientific method: an introduction for students and researchers. Oxford University Press.

- Dubay, W. H. (2004). *The Principles of Readability A brief introduction to readability research*. Impact Information, Costa Mesa, CA.
- Dubin, D. (2004). The most influential paper Gerard Salton never wrote. Ανακτήθηκε από <https://www.ideals.illinois.edu/handle/2142/1697>
- Eroglu, S. (2013). Menzerath–Altmann law for distinct word distribution analysis in a large text. *Physica A: Statistical Mechanics and its Applications*, 392(12), 2775–2780.
- Espunya i Prat, A. (1994). Computational linguistics: a brief introduction. Στο *Links & Letters* (σσ 009–23). Ανακτήθηκε από <http://ddd.uab.cat/pub/lal/11337397n1/11337397n1p9.pdf>
- Everitt, B. S., & Skrondal, A. (2002). *The Cambridge dictionary of statistics*. Cambridge: Cambridge. Ανακτήθηκε από <http://www.maa.org/publications/maa-reviews/the-cambridge-dictionary-of-statistics>
- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters*, 27(8), 861–874.
- Fisher, R. A. (1925). *Statistical methods for research workers*. Genesis Publishing Pvt Ltd.
- Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32(200), 675–701.
- Galvez, C., de Moya-Anegón, F., & Solana, V. H. (2005). Term conflation methods in information retrieval: non-linguistic and linguistic approaches. *Journal of Documentation*, 61(4), 520–547.
- Gries, S. T. (2013). *Statistics for linguistics with R: a practical introduction*. Walter de Gruyter. Ανακτήθηκε από https://books.google.com/books?hl=en&lr=&id=-SFmpJ5N-98C&oi=fnd&pg=PP5&dq=Statistics+for+linguistics+with+R.+&ots=EJtROz94e-&sig=_7bDzRBLUXoXRBjUDqH1vqHTXFU
- Grossman, D. A., & Frieder, O. (2012). *Information retrieval: Algorithms and heuristics* (T. 15). Springer Science & Business Media. Ανακτήθηκε από <https://books.google.com/books>
- Harispe, S., Ranwez, S., Janaqi, S., & Montmain, J. (2013). Semantic Measures for the Comparison of Units of Language, Concepts or Instances from Text and Knowledge Base Analysis. *arXiv preprint arXiv:1310.1285*. Ανακτήθηκε από <http://arxiv.org/abs/1310.1285>
- Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *science*, 298(5598), 1569–1579.
- Hayes, B., Curtiss, S., Szabolcsi, A., Stowell, T., Stabler, E., Sportiche, D., ... others. (2001). *Linguistics: An introduction to linguistic theory*. Wiley-Blackwell.
- Hempel, C. G. (1952). Fundamentals of concept formation in empirical science.(Int. Encyclopedia Unified Science, Vol. II. No. 7). Ανακτήθηκε από <http://psycnet.apa.org/psycinfo/1953-03119-000>
- Hoey, M. (2004). Lexical priming and the properties of text. *Corpora and discourse*, 385–412.
- Hogan, P. C. (2011). *The Cambridge encyclopedia of the language sciences*. Cambridge University Press Cambridge, NY. Ανακτήθηκε από http://www.langtoninfo.com/web_content
- Hřebíček, L. (2002). Zipf's Law and Text. *Glottometrics*, 27.
- Ingwersen, P. (1992). *Information Retrieval Interaction*. London: Taylor Graham Publishing
- Johnson, K. (2008). *Quantitative Methods in Linguistics*. Malden (USA): Blackwell.
- Joho, H. and Jose, J. M. (2006). A Comparative Study of the Effectiveness of Search Result Presentation on the Web. *Advances in Information Retrieval: Lecture Notes in Computer Science*. 3936. P. 302-313
- Jurafsky, D. and Martin, H. J. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall series in artificial intelligence. Upper Saddle River, NJ: Prentice Hall PTR

- Kamps, J. (2009). Presenting Structured Text Retrieval Results. *Encyclopedia of Database Systems*. P. 2130-2134
- Karaman, B. I. (2003). Polysemy in Natural Language: Case Studies on the Structural Description of Polysemous Lexemes in English, German and Turkish. PhD thesis. University of Surrey. Retrieved 22/09/2014 from <http://epubs.surrey.ac.uk/2816/1/412061.pdf>
- Kennedy, C. (2009). Ambiguity and Vagueness: an Overview. In: Maienborn, C., Heusinger, K. and Portner, P. (Eds.). *Semantics: an International Handbook of Natural Language Meaning*. Handbooks of Linguistics and Communication Science (HSK). 33 (1). Berlin: De Gruyter Mouton
- Kovacs, E. (2011). Polysemy in Traditional vs Cognitive Linguistics. *Eger Journal of English Studies*. XI. P. 3–19
- Kowalski, G. (2011). *Information Retrieval Architecture and Algorithms*. New York: Springer.
- Kracht, M. (2007). Introduction to linguistics. Los Angeles: Department of Linguistics, UCLA. Retrieved 08/12/2014 from: <http://www.whomes.uni-bielefeld.de/mkracht/html/ling-intro.pdf>
- Kulacka, A. and Macutek, J. (2007). A discrete formula for the Menzerath - Altmann law. *Journal of Quantitative Linguistics*. 14 (1). P. 23-32
- Kuroepka, D. (2004). Modelle zur Repräsentation natürlichsprachlicher Dokumente: Ontologie-basiertes, Information-Filtering und -Retrieval mit relationalen Datenbanken. *Advances in Information Systems and Management Science*. 10. Berlin: Logos.
- Lalkhen, A.G. and McCluskey, A. (2008). Clinical tests: sensitivity and specificity. *Continuing Education in Anaesthesia, Critical Care and Pain*. 8. P. 221–223.
- Lan, M., Tan, C. L. and Su, J. (2009). Supervised and Traditional Term Weighting Methods for Automatic Text Categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 31(4). P. 721-735
- Legendre, P. (2005). Species Associations: The Kendall Coefficient of Concordance Revisited. *Journal of Agricultural, Biological, and Environmental Statistics*. 10 (2). P. 226–245.
- Legendre, P. (2010). Coefficient of concordance. *Encyclopedia of Research Design*. 1. P. 164-169.
- Lewis, D. D. and Sparck Jones, K. (1996). Natural language processing for information retrieval. *Communications of the ACM*. 39 (1). P. 92-101.
- Manning, C., Raghavan, P. and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge: Cambridge University Press.
- Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.
- Metzler, D. and Croft, B. W. (2005). Modeling Query Term Dependencies in Information Retrieval with Markov Random Fields. *SIGIR '05 Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*. P. 472-479.
- Mikk, J. (2005). Text comprehensibility (Textverständlichkeit). In Köhler, R., Altmann, G. and Piotrowski, R. G. (eds.) *Quantitative Linguistik - Quantitative Linguistics. Ein Internationales Handbuch*. Berlin: Walter de Gruyter
- Mikros, G. and Milicka, J. (2014). Distribution of the Menzerath's law on the syllable level in Greek texts. In Altmann, G. et al. (eds.) *Empirical Approaches to Text and Language Analysis*. Lüdenschied: RAM-Verlag.
- Miller, G.A. (2003). The cognitive revolution: a historical perspective. *Trends in Cognitive Science*. 7 (3). P. 141–144
- Moens, M. F. (2006). *Information Extraction: Algorithms and Prospects in a Retrieval Context*. The Information Retrieval Series. Dordrecht: Springer.

- Nadel L. (2005). Encyclopedia of Cognitive Science. Wiley
- Neideen, T. and Brasel K. (2007). Understanding Statistical Tests. Journal of Surgical Education. 64 (2). P. 93-96
- Onwuchekwa, E. O. (2011). Information Retrieval Methods in Libraries and Information Centers. An International Multidisciplinary Journal, Ethiopia. 5(6), P. 108-120.
- Panik M. J. (2012). Statistical Inference: a short course. Hoboken, NJ: Wiley
- Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: a critical review and future directions. Psychonomic Bulletin & Review. 21 (5). P. 1112-1130
- Powers, D. M. W. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation. Journal of Machine Learning Technologies. 2 (1). P. 37-63
- Power law: definitions of power law in Oxford Dictionary (British and World English) (2015). Retrieved 08/01/2015 from: <http://www.oxforddictionaries.com/definition/english/power-law>
- Poulos, M. et al. (2007). Specific Selection of FFT Amplitudes from Audio Sports and News Broadcasting for Classification Purposes. Journal of Graph Algorithms and Applications. 11 (1). P. 277–307
- Poulimenou S. et al. (2014). Keywords Extraction from Articles' Title for Ontological Purposes. In Proceedings of the 2014 International Conference on Pure Mathematics, Applied Mathematics, Computational Methods (PMAMCM 2014). P. 120-125.
- Poulimenou, S., Stamou, S., Papavlasopoulos, S., & Poulos, M. (χ.χ.). Short Text Coherence Hypothesis. Ανακτήθηκε από http://www.researchgate.net/profile/Marios_Poulos/publication/271167995_Short_Text_Coherence_Hypothesis/links/54bf5d570cf28ce68e6b4c71.pdf
- Powers, D. M. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. Ανακτήθηκε από <http://dspace2.flinders.edu.au/xmlui/handle/2328/27165>
- Pulvermüller, F. (2002). The neuroscience of language: on brain circuits of words and serial order. Cambridge University Press. Ανακτήθηκε από <https://www.google.com/books>
- Raghavan, V. V., & Wong, S. M. (1986). A critical analysis of vector space model for information retrieval. Journal of the American Society for information Science, 37(5), 279–287.
- Richards, J. C., & Schmidt, R. W. (2013). Longman dictionary of language teaching and applied linguistics. Routledge. Ανακτήθηκε από <https://www.google.com/books>
- Robertson, S. (2004). Understanding inverse document frequency: on theoretical arguments for IDF. Journal of documentation, 60(5), 503–520.
- Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information processing & management, 24(5), 513–523.
- Singhal, A. (2001). Modern information retrieval: A brief overview. IEEE Data Eng. Bull., 24(4), 35–43.
- Spoerri, A. (1995). InfoCrystal, a visual tool for information retrieval. Massachusetts Institute of Technology. Ανακτήθηκε από <http://dspace.mit.edu/handle/1721.1/36946>
- Squire, L., Berg, D., Bloom, F. E., du Lac, S., Ghosh, A., & Spitzer, N. C. (2012). Fundamental neuroscience. Academic Press. Ανακτήθηκε από <https://www.google.com>
- Strzalkowski, T., Lin, F., Wang, J., & Perez-Carballo, J. (1999). Evaluating natural language processing techniques in information retrieval. Στο Natural language information retrieval (σσ 113–145). Springer. Ανακτήθηκε από http://link.springer.com/chapter/10.1007/978-94-017-2388-6_5
- Tanskanen, S.-K. (2006). Collaborating towards coherence: Lexical cohesion in English discourse (T. 146). John Benjamins Publishing. Ανακτήθηκε από <https://www.google.com/books>
- Tešitelova, M. (1983). On the state of Quantitative Linguistics in studies of Czech. Prague bulletin of mathematical linguistics, (40), 15–30.

- Turney, P. D., Pantel, P., & others. (2010). From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37(1), 141–188.
- Wei, L. H. H. (2012). Quantitative Linguistics: State of the Art, Theories and Methods. *Journal of Zhejiang University (Humanities and Social Sciences)*, 2, 017.
- Wong, H. B., & Lim, G. H. (2011). Measures of diagnostic accuracy: sensitivity, specificity, PPV and NPV. *Proceedings of Singapore Healthcare*▪ Volume, 20(4). Ανακτήθηκε από http://www.singhealthacademy.edu.sg/documents/publications/vol20no42011/11_stats_032-0911wonghb.pdf
- Woods, W. A., Bookman, L. A., Houston, A., Kuhns, R. J., Martin, P., & Green, S. (2000). Linguistic knowledge can improve information retrieval. Στο *Proceedings of the sixth conference on Applied natural language processing* (σσ 262–267). Association for Computational Linguistics. Ανακτήθηκε από <http://dl.acm.org/citation.cfm?id=974183>
- Wu, S., Bi, Y., & Zeng, X. (2010). Retrieval result presentation and evaluation. Στο *Knowledge Science, Engineering and Management* (σσ 125–136). Springer. Ανακτήθηκε από http://link.springer.com/chapter/10.1007/978-3-642-15280-1_14
- Wyllys, R. E. (1981). Empirical and theoretical bases of Zipf's law. *Library Trends*, 30(1), 53–64.
- Zamanian, M., & Heydari, P. (2012). Readability of texts: State of the art. *Theory and Practice in Language Studies*, 2(1), 43–53.
- Zamir, O., & Etzioni, O. (1999). Grouper: a dynamic clustering interface to Web search results. *Computer Networks*, 31(11), 1361–1374.
- Zar, J. H., & others. (1999). *Biostatistical analysis*. Pearson Education India.
- Zhu, W., Zeng, N., Wang, N., & others. (2010). Sensitivity, specificity, accuracy, associated confidence interval and ROC analysis with practical SAS® implementations. *NESUG proceedings: health care and life sciences*, Baltimore, Maryland, 1–9.

4. Οντολογικές Προσεγγίσεις

Περίληψη

Στο τέταρτο κεφάλαιο παρουσιάζονται με άλλα κριτήρια άλλες γλώσσες ή ακόμη και οι προαναφερθείσες που αποτελούν μοντέλα για οντολογική δομή και συλλογιστική. Επίσης γίνεται ανάπτυξη των γλωσσών προσδιορισμού οντολογίας με κριτήριο την κατηγοριοποίηση / ταξινόμησή τους ανάλογα με τα πρότυπα με τα οποία είναι συμβατές. Τέλος γίνεται προσπάθεια σύγκρισης των κυριοτέρων οντολογιών με έμφαση στις δυνατότητες και ταυτόχρονα στις αδυναμίες τους.

Προαπαιτούμενες γνώσεις

RDF, XML, Σχήματα

4.1 Οντολογίες

Στη φιλοσοφία, μια οντολογία αποτελεί μια θεωρία που αφορά στη φύση της ύπαρξης, στο τι είδους πράγματα υπάρχουν στον κόσμο. Οι ερευνητές της τεχνητής νοημοσύνης και του Ιστού συναίνεσαν σε αυτόν τον όρο για τη δική τους επαγγελματική διάλεκτο: με τον όρο οντολογία αναφέρονται σε ένα τεκμήριο ή ένα αρχείο που ορίζει τυπικά τις σχέσεις ανάμεσα στους διάφορους όρους. Το συνηθέστερο είδος οντολογίας που αφορά στον Ιστό αποτελείται από μια ταξινόμια και από ένα σύνολο κανόνων εξαγωγής συμπερασμάτων. Η ταξινόμια αυτή ορίζει τάξεις αντικειμένων και τις μεταξύ τους σχέσεις. Π.χ., μια *διεύθυνση* μπορεί να οριστεί ως είδος *τοποθεσίας*, ενώ οι ταχυδρομικοί τομείς να ορίζονται μόνο ως *τοποθεσίες*, κλπ. Οι κατηγορίες, οι υποκατηγορίες και οι σχέσεις μεταξύ των οντοτήτων συνιστούν ένα πολύ ισχυρό εργαλείο για τη χρήση του Ιστού. Μπορούμε να εκφράσουμε ένα μεγάλο αριθμό σχέσεων μεταξύ των οντοτήτων προσδίδοντας ιδιότητες στις κατηγορίες και επιτρέποντας στις υποκατηγορίες να κληρονομήσουν αυτές τις ιδιότητες. Εάν ο ταχυδρομικός τομέας πρέπει να είναι του τύπου *πόλη* και όλες οι πόλεις έχουν τις ιστοσελίδες τους, τότε μπορούμε να μιλάμε για ιστοσελίδες που σχετίζονται με κάποιον ταχυδρομικό τομέα ακόμα κι αν καμία βάση δεδομένων δεν έχει άμεση ζεύξη ενός ταχυδρομικού τομέα με μια ιστοσελίδα.

Τυπολογία της Οντολογίας (ταξινόμια οντολογιών σύμφωνα με τον Mizoguchi)

Για να αναλογιστούμε πώς χρησιμοποιείται η γνώση, βοηθήστε μας να αντιληφθούμε τι είναι μια οντολογία. Μέχρι σήμερα, λίγες συζητήσεις έχουν γίνει για το πώς χρησιμοποιούμε την οντολογία. Πολλοί ερευνητές λένε πως «χρησιμοποιούν» γνώση / οντολογία χωρίς να ορίζουν τι εννοούν με το «χρησιμοποιώ», δηλαδή, ποιος τη χρησιμοποιεί και με ποιους τρόπους. Εδώ πραγματευόμαστε το ζήτημα αυτό. Μια οντολογία υποδιαιρείται επιπλέον σε υποκατηγορίες από την άποψη της επαναχρησιμοποίησης της γνώσης ως εξής:

Οντολογία

Οντολογία του χώρου εργασίας

Αυτή είναι μία οντολογία για το χώρο εργασίας, που επηρεάζει τα χαρακτηριστικά των έργων, προδιαγράφοντας αρκετές οριακές προϋποθέσεις που χαρακτηρίζουν και δικαιολογούν τη συμπεριφορά επίλυσης προβλημάτων στο χώρο εργασίας. Οι οντολογίες χώρου εργασίας και έργων προδιαγράφουν το περιβάλλον στο οποίο η γνώση ενός τομέα ερμηνεύεται και χρησιμοποιείται κατά την επίλυση του προβλήματος.

Παραδείγματα (Αντιμετώπιση προβλημάτων για κυκλώματα): Πιστότητα/Αποδοτικότητα/ Υψηλή Αξιοπιστία/ κτλ.

Οντολογία έργων

Η οντολογία έργων είναι ένα σύστημα λεξιλογίου για την περιγραφή της δομής της επίλυσης προβλημάτων όλων των υπάρχοντων έργων, ανεξαρτήτως τομέα. Δεν καλύπτει τη δομή ελέγχου, αλλά τα στοιχεία και τις αρχές των συμπερασμάτων των μονάδων που λαμβάνουν χώρα κατά την εκτέλεση των έργων. Η γνώση έργων με τη σειρά της προδιαγράφει τη γνώση του τομέα δίνοντας ρόλους σε κάθε αντικείμενο και σχέσεις μεταξύ τους.

Παραδείγματα (Έργα χρονοδιαγράμματος): Αποδέκτης χρονοδιαγράμματος/ πόρος χρονοδιαγράμματος/ στόχος/ περιορισμός/ διαθεσιμότητα/ φορτίο/επιλογή/ ανάθεση/ ταξινόμηση/ αφαίρεση/ χαλάρωση/ πρόσθεση/ κτλ.

Οντολογία τομέα

Εξαρτώμενη από το έργο οντολογία

Μια δομή έργου δεν απαιτεί όλη τη γνώση του τομέα, αλλά κάποια συγκεκριμένη γνώση του τομέα σε κάποιο συγκεκριμένο οργανισμό. Ονομάζουμε αυτό τον ειδικό τύπο γνώσης τομέα οντολογία Ε-τομέα, επειδή εξαρτάται από το έργο.

Οντολογία Ε-Τομέα

Παραδείγματα (Χρονοδιάγραμμα δουλειάς-καταστήματος): Δουλειά/ σειρά/ γραμμή/ ημερομηνία παράδοσης/ διαθεσιμότητα μηχανημάτων/ καθυστέρηση/ φορτίο/ κόστος/ κτλ.

Ανεξάρτητη από το έργο οντολογία

Επειδή οι οντολογίες αντικειμένου και δραστηριότητας σχετίζονται με δραστηριότητες, τις ονομάζουμε οντολογία συναφής με τη δραστηριότητα και οντολογία πεδίου, ανεξάρτητη από τη δραστηριότητα οντολογία.

Οντολογία συναφής με τη δραστηριότητα

Η οντολογία αυτή σχετίζεται με δραστηριότητες που λαμβάνουν χώρα στον τομέα και σχεδιάζεται έχοντας στο νου την προσομοίωση της δραστηριότητας του τομέα, όπως η επιχειρηματική οντολογία. Υπάρχουν δύο κύριες δραστηριότητες σε ένα τομέα. Η μία είναι η συμπεριφορά ενός αντικειμένου και η άλλη είναι η οργανωτική ή οι ανθρώπινες δραστηριότητες. Τα ρήματα παίζουν σημαντικό ρόλο στην οντολογία αυτή, ωστόσο, είναι διαφορετικά απ' αυτά της οντολογίας έργων. Τα υποκείμενα των πρώτων ρημάτων είναι αντικείμενα, συστατικά ή παράγοντες που εμπλέκονται στις δραστηριότητες του ενδιαφέροντος, ενώ αυτά των τελευταίων είναι ειδικοί τομέα.

Οντολογία αντικειμένου

Η οντολογία αυτή καλύπτει τη δομή, συμπεριφορά και λειτουργία του αντικειμένου.

Παραδείγματα (Πίνακες κυκλωμάτων): συστατικό/ σύνδεση/ γραμμή/ τσιπ/ βελόνα/ πύλη/ δίαυλος/ κατάσταση/ ρόλος/ κτλ.

Οντολογία δραστηριότητας (Επιχειρηματική οντολογία)

Παραδείγματα: χρήση/ κατανάλωση/ παραγωγή/ κυκλοφορία/ κατάσταση/ πόρος/ δέσμευση/ εφικτό/ πλήρες/ ανέφικτο/ κτλ.

Ανεξάρτητη από τη δραστηριότητα οντολογία

Οντολογία πεδίου

Η οντολογία αυτή σχετίζεται με θεωρίες και αρχές που διέπουν τον τομέα. Περιέχει αρχικές έννοιες που εμφανίζονται στις θεωρίες και τις σχέσεις, τις φόρμουλες, τις μονάδες που συνιστούν τις θεωρίες και τις αρχές.

Μονάδες

Παραδείγματα: mole/ χιλιόγραμμα/ μέτρο/ αμπέρ/ ακτίνιο/ κτλ.

Μαθηματικά μηχανολογίας

Παραδείγματα: Φυσικό ποσότητα/ φυσική διάσταση/ μονάδα μέτρησης/ μονόμετρη ποσότητα/ γραμμική άλγεβρα/ φυσικό στοιχείο/ κτλ.

Γενική / Κοινή οντολογία

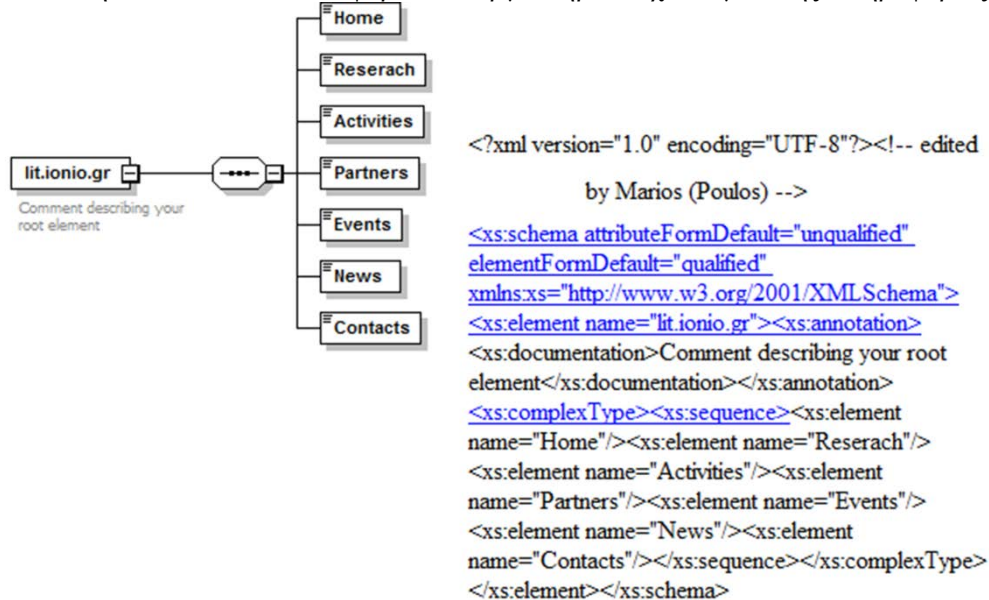
Παραδείγματα: Πράγματα/ Γεγονότα/ Χρόνος/ Χώρος/ Αιτιότητα

Συμπεριφορά/ Λειτουργία, κτλ.

Αυτό δείχνει πως υπάρχουν πολλές κατηγορίες οντολογιών. Αλλά αυτό δεν σημαίνει πως η έρευνα για την οντολογία ξεστρατίζει, αλλά σημαίνει πως ο πραγματικός κόσμος είναι πλούσιος. Η αναγνώριση της ποικιλίας της γνώσης, και επομένως αυτής της ίδιας της οντολογίας, μπορεί να είναι ένα από τα θέματα έρευνας που βαθαίνει την κατανόηση της γνώσης για μας. Η πιο πάνω ταξινόμηση της οντολογίας διακρίνει την εξαρτώμενη από το έργο οντολογία και την ανεξάρτητη από το έργο. Η δεύτερη έχει συχνά συζητηθεί στη βιβλιογραφία. Η πρώτη, της οποίας τη σημασία οι συγγραφείς έχουν τονίσει, είναι η αρχική έννοια των συγγραφέων. Όσον αφορά στις οντολογίες, οι κανόνες εξαγωγής συμπερασμάτων παρέχουν περισσότερες δυνατότητες. Μια οντολογία μπορεί να εκφράζει τον κανόνα 'Εάν ο ταχυδρομικός τομέας μιας πόλης σχετίζεται με τον ταχυδρομικό τομέα μιας πολιτείας και στη διεύθυνση χρησιμοποιείται αυτός ο ταχυδρομικός τομέας, τότε σε αυτή τη διεύθυνση περιλαμβάνεται ο σχετιζόμενος ταχυδρομικός κώδικας της πολιτείας.' Συνεπώς, ένα πρόγραμμα θα μπορούσε, π.χ., να εξάγει το συμπέρασμα ότι η διεύθυνση του Ιονίου Πανεπιστημίου στην πόλη Κέρκυρα πρέπει να βρίσκεται στο νησί της Κέρκυρας στα Επτάνησα, τα οποία βρίσκονται στην Ελλάδα, και άρα να ακολουθεί το μορφότυπο διευθύνσεων κατά τα ελληνικά πρότυπα. Ο Η/Υ αν και δεν καταλαβαίνει καμία από αυτές τις πληροφορίες, μπορεί τώρα να χειρίζεται τους όρους πολύ πιο αποτελεσματικά ώστε να βγαίνει νόημα και να χρησιμεύει καλύτερα στον ανθρώπινο χρήστη. Με τη βοήθεια των οντολογικών σελίδων στον Ιστό, πολλά προβλήματα ορολογίας και όχι μόνο αρχίζουν να βρίσκουν λύση. Η σημασία των όρων ή των κωδικών XML που χρησιμοποιούνται σε μια ιστοσελίδα μπορεί

να οριστεί μέσω δεικτών (indices) που υπάρχουν στη σελίδα και οδηγούν σε μια οντολογία. Οι οντολογίες (ή άλλες υπηρεσίες που παρέχονται στον Ιστό), βέβαια, θα πρέπει να παρέχουν σχέσεις ισοδυναμίας ώστε μία ή δύο οντολογίες να περιέχουν την πληροφορία ότι, π.χ., ο ταχυδρομικός μου τομέας ισοδυναμεί με τον ταχυδρομικό κώδικα κάποιου άλλου.

Οι οντολογίες μπορούν να επαυξηθούν τη λειτουργικότητα του Ιστού με πολλούς τρόπους (βλ σχήμα 4.1). Μπορούν να χρησιμοποιηθούν απλά, και να βελτιώσουν την ακρίβεια των αναζητήσεων στον Ιστό, το πρόγραμμα αναζήτησης μπορεί να ψάξει μόνο εκείνες τις σελίδες που αναφέρονται σε μία ακριβή έννοια και όχι σε εκείνες που χρησιμοποιούν ασαφείς ή διφορούμενες λέξεις κλειδιά. Σε πιο εξελιγμένες εφαρμογές, οι οντολογίες θα χρησιμοποιηθούν για τον συσχετισμό πληροφοριών σε μια σελίδα με συναφείς δομές γνώσης και κανόνες εξαγωγής συμπερασμάτων. Παράδειγμα μιας σημασμένης σελίδας, στην εικόνα που ακολουθεί σε αναπαράσταση σχήματος. Εάν στείλετε το δικό σας φυλλομετρητή Ιστού σε αυτή τη σελίδα, θα δείτε την κανονική ιστοσελίδα να τιτλοφορείται «Εργαστήριο Τεχνολογιών της Πληροφορίας (ETΠ)»



Σχήμα 4. 1. Η αναπαράσταση της ιστοσελίδας του Εργαστηρίου Τεχνολογιών της Πληροφορίας (ETΠ) σε XML schema

Ένας χρήστης θα μπορούσε εύκολα να βρει διάφορες πληροφορίες στις δραστηριότητες του εργαστηρίου μέσω της σήμανσης που παρέχεται από το σχήμα. Αντίθετα, η ανεύρεση παρόμοιων πληροφοριών θα ήταν εξαιρετικά περίπλοκη για ένα ηλεκτρονικό πρόγραμμα εφόσον θα έπρεπε όχι μόνο να μαντέψει ότι οι πληροφορίες περιέχονται στο βιογραφικό, αλλά και να καταλαβαίνει την αγγλική γλώσσα που χρησιμοποιείται εκεί.

Για τους H/Y, η σελίδα έχει ζεύξη με μια οντολογική σελίδα που ορίζει τις πληροφορίες για τα τμήματα πληροφορικής επιστήμης.

Επιπλέον, αυτή η σήμανση διευκολύνει τη δημιουργία προγραμμάτων που θα μπορούν να χειρίζονται πιο περίπλοκα ερωτήματα, οι απαντήσεις των οποίων δεν βρίσκονται σε μια και μόνη σελίδα. Π. χ., εάν ψάχνω να βρω την κυρία Παπά που συνάντησα πέρσι σε ένα συνέδριο και δε θυμάμαι το όνομα της αλλά θυμάμαι ότι δούλεψε για έναν από τους πελάτες μου κι ότι ο υιός της φοιτά στην ιδιαίτερη μου πατρίδα, τότε ένα ευφυές πρόγραμμα αναζήτησης μπορεί να παρακάμψει όλες τις σελίδες που περιέχουν το όνομα αυτό (ή λέξεις που αφορούν τον παπά, τον Πάπα, τα πρώτα συνθετικά με παπά-Γιώργο, κτλ.) και να βρει εκείνες που αναφέρουν ότι δουλεύει σε μια εταιρία που βρίσκεται στις λίστες των πελατών μου και να βρει μέσω ζεύξης τις ιστοσελίδες των παιδιών της κτλ.

Σε κάθε οντολογία καθορίζονται οι όροι που απαιτούνται για την αναπαράσταση και την περιγραφή μιας περιοχής γνώσης. Οι οντολογίες χρησιμοποιούνται από ανθρώπους, βάσεις δεδομένων και εφαρμογές που ανταλλάσσουν πληροφορίες σχετικές με το αντικείμενο (περιοχή) π.χ. ιατρική, κατασκευή εργαλείων, ακίνητη περιουσία, επισκευή αυτοκινήτων, οικονομική διαχείριση και άλλα. Οι οντολογίες περιλαμβάνουν ορισμούς βασικών εννοιών χρησιμοποιήσιμους από υπολογιστές στην εκάστοτε περιοχή και τις μεταξύ τους σχέσεις. Κωδικοποιούν τη γνώση σε μια περιοχή καθώς και γνώση που διαπερνά περιοχές. Με αυτό τον τρόπο κάνουν τη γνώση επαναχρησιμοποιήσιμη.

4.2 Διαμεσολαβητές (Agents)

Διαμεσολαβητές του Ιστού βασισμένοι σε οντολογία

Φανταστείτε πως ψάχνετε στον παγκόσμιο Ιστό την προσωπική ιστοσελίδα κάποιων Κου και Κας Σολωμού, τους οποίους γνωρίσατε σε ένα συνέδριο πέρσι. Δεν θυμάστε τα βαφτιστικά ονόματά τους, αλλά θυμάστε πως και οι δύο δουλεύουν για ένα εργοδότη που σχετίζεται με την πρωτοβουλία χρηματοδότησης 123-4567 του Υπουργείου Εθνικής Άμυνας. Αυτές θα είναι σίγουρα αρκετές πληροφορίες για να βρείτε τα άτομα αυτά, δεδομένης μιας λογικά δομημένης βάσης γνώσεων που περιέχει όλα τα συναφή γεγονότα. Αρχικά, αυτές φαίνονται επίσης αρκετές πληροφορίες για να βρείτε τις προσωπικές ιστοσελίδες τους ψάχνοντας τον παγκόσμιο Ιστό, αλλά σύντομα ανακαλύπτετε διαφορετικά πράγματα.

Χρησιμοποιώντας ένα υπάρχοντα κατάλογο του Ιστού φτιαγμένο από ανθρώπους, μπορείτε να βρείτε τον ιστότοπο του Υπουργείου Εθνικής Άμυνας, αλλά μαθαίνετε πως εκατοντάδες υπεργολάβοι και ερευνητικές ομάδες δουλεύουν για την πρωτοβουλία 123-4567. Η έρευνα των υφιστάμενων ευρετηρίων του ιστού για τον όρο «Σολωμού» παράγει χιλιάδες αποτελέσματα για την ιχθυολογία, και η αναζήτηση για το «ΥΠΕΘΑ» και το «123-4567» μας παρέχει εκατοντάδες και εκατοντάδες επιτυχίες (hits) σχετικά με την πρωτοβουλία. Δυστυχώς, η αναζήτηση όλων αυτών μαζί δεν παράγει τίποτα, επειδή κανένα πρόσωπο δεν περιλαμβάνει την πρωτοβουλία στη σελίδα του. Η άσκοπη φυλλομέτρηση στον Ιστό είναι άκαρπη. Τι μπορείτε να κάνετε;

Αυτό το σενάριο είναι σύνηθες για πολλούς ανθρώπους στον παγκόσμιο Ιστό. Ένα μείζον πρόβλημα με την αναζήτηση στον Ιστό σήμερα είναι πως τα διαθέσιμα δεδομένα έχουν ελάχιστη σημασιολογική οργάνωση πέραν της απλής δομικής διαρρύθμισης του κειμένου, των δηλωμένων λέξεων-κλειδιά, των τίτλων και των περιλήψεων. Καθώς ο Ιστός επεκτείνεται εκθετικά σε μέγεθος, η έλλειψη αυτή της οργάνωσης κάνει πολλή δύσκολη την αποδοτική σταχυολόγηση γνώσης από τον Ιστό, ακόμη και με προηγμένης τεχνολογίας τεχνικές επεξεργασίας φυσικής γλώσσας, μηχανισμούς ευρετηρίασης ή τη βοήθεια του στρατού των ειδικών σε θέματα εισαγωγής δεδομένων που συγκεντρώνουν χειροποίητους καταλόγους του Ιστού. Εν ολίγοις, δεν υπάρχει αποδοτικός τρόπος χρήσης του παγκόσμιου Ιστού για να απαντηθεί μια ερωτήρηση όπως:

Να βρεθούν οι ιστοσελίδες για όλα τα x , y και z , έτσι ώστε το x να είναι πρόσωπο, το y να είναι πρόσωπο και το z να είναι ένας οργανισμός, όπου `lastName( $x$ , "Σολωμού")`, `lastName( $y$ , "Σολωμού")`, `employee( $z$ ,  $x$ )`, `employee( $z$ ,  $y$ )`, `marriedTo( $x$ ,  $y$ )`, `andinvolvedIn( $z$ , "ΥΠΕΘΑ 123-4567")`.

Η αναζήτηση στον Ιστό όπως γινόταν μέχρι σήμερα

Η πρωτεύουσα επιδίωξη της HTML και του HTTP είναι η βοήθεια της παρουσίας και της πλοήγησης του Διαδικτύου στο επίπεδο του χρήστη· η αυτοματοποιημένη αναζήτηση ή η εκλεπτυσμένη συλλογή γνώσης πάντοτε είχαν λιγότερη προτεραιότητα. Δεδομένης αυτής της έμφασης, σχετικά λίγοι μηχανισμοί έχουν συσταθεί για να επισημαίνουν τα έγγραφα με χρήσιμες σημασιολογικές πληροφορίες, πέραν των προσανατολισμένων σε έγγραφα πληροφοριών όπως «περίληψη» ή «πίνακας περιεχομένων». Ως αποτέλεσμα, οι κοινότεροι μηχανισμοί ευρετηρίασης και τα ρομπότ διαμεσολάβησης στον παγκόσμιο Ιστό εμπίπτουν γενικά σε μια από τις τρεις κατηγορίες:

- Μηχανές δοκιμαστικής ευρετηρίασης.
- Επίμοχθοι κατάλογοι περασμένοι με το χέρι.
- Ιδιωτικά ρομπότ που χρησιμοποιούν ad-hoc μεθόδους για να συλλέξουν περιορισμένες σημασιολογικές πληροφορίες για τις σελίδες (όπως «Όλοι όσοι συνδέονται με εμένα» ή «Όλοι οι ασύνδετοι σύνδεσμοι της σελίδας»).

Κάθε προσέγγιση έχει μειονεκτήματα. Τα ευρετήρια κειμένου νοσούν επειδή συσχετίζουν το σημασιολογικό νόημα των ιστοσελίδων με πραγματικό λεξιλογικό ή συντακτικό περιεχόμενο. Από το προηγούμενό μας παράδειγμα, η αναζήτηση για ένα ευρετήριο λέξης-κλειδί υπό το «Σολωμού» παρήγαγε τεράστιους αριθμούς ιστοσελίδων, από τις οποίες σχεδόν καμιά δεν αφορά ζωντανούς ανθρώπους με το όνομα Σολωμού. Το «Σολωμός» έχει πολλές χρήσεις εκτός από το να είναι επίθετο. Αν και τα ευρετήρια κειμένου βελτιώνονται, η συνολική ποσότητα των πληροφοριών στον Ιστό αυξάνεται επίσης ραγδαία.

Ένα μείζον μειονέκτημα των καταλόγων που είναι περασμένοι με το χέρι είναι οι ανθρωποώρες που απαιτούνται για να φτιαχτούν. Δεδομένου του μεγέθους του παγκόσμιου Ιστού, και το ρυθμό με τον οποίο μεγαλώνει, η καταλογογράφηση ακόμη και ενός ταπεινού ποσοστού ιστοσελίδων αποτελεί ηράκλειο άθλο. Επιπρόσθετα, τα κριτήρια που χρησιμοποιούνται για την κατασκευή οποιουδήποτε καταλόγου μπορούν να αποδειχθούν ορθογώνια σε σχέση με το ενδιαφέρον του χρήστη.

Τα εξειδικευμένα ρομπότ που επιχειρούν να συλλέξουν σημασιολογικές πληροφορίες από τον ιστό συνήθως συλλέγουν μόνο τις περιορισμένες σημασιολογικές πληροφορίες που είναι τεκμαρτές από τις υπάρχουσες ετικέτες της HTML. Οι υφιστάμενη κατάσταση της τεχνολογίας επεξεργασίας φυσικής γλώσσας κάνει δύσκολο να τεκμαίρουμε πολύ σημασιολογικό νόημα από το ίδιο το σώμα του κειμένου κατά μια λογική αναλογία (αν υπάρχει). Από την εμπειρία μας, ακόμη και αν περιορίσουμε την κατανόηση της φυσικής γλώσσας από ένα ρομπότ του ιστού σε ένα μικρό θέμα, όπως τις ιστοσελίδες για την Επιστήμη των Υπολογιστών, αποδεικνύεται ακόμη απροσδόκητα δύσκολο να εφαρμοστεί, και όπως πολλές ad-hoc μέθοδοι, τέτοιοι αλγόριθμοι είναι εξαιρετικά εύθραυστοι. Περαιτέρω, καμιά από αυτές τις προσεγγίσεις (εκτός ίσως την τελευταία, για συγκεκριμένους τομείς) δεν επιτρέπει τα συμπεράσματα για σχέσεις *ανάμεσα* στις ιστοσελίδες, πέραν από απλά γεγονότα σχετικά με τη σύνδεση. Εκλεπτυσμένες επερωτήσεις όπως στο αρχικό παράδειγμα μας «Σολωμού» είναι επομένως σαφώς ανέφικτες.

Λύση αποτελεί η προσθήκη σημασιολογίας στην HTML

Αντί να προσπαθούμε να σταχυολογήσουμε γνώση από την υπάρχουσα HTML, μία άλλη προσέγγιση είναι να δώσουμε στους δημιουργούς τη δυνατότητα να ενσωματώσουν τη γνώση απευθείας μέσα στις σελίδες της HTML, απλουστεύοντας την ανεύρεση και φύλαξη των γνώσεων αυτών από διαμεσολαβητές χρήστη και ρομπότ. Ο ξεκάθαρος τρόπος για να γίνει αυτό είναι η παροχή στους συγγραφείς ενός καθαρού υπερσυνόλου της HTML που προσθέτει μια σύνταξη σήμανσης γνώσεως· δηλαδή, να τους δώσουμε τη δυνατότητα να χαρακτηρίσουν απευθείας της ιστοσελίδες τους και να εκθέσουν λεπτομερώς τις σχέσεις και τα γνωρίσματα των ιστοσελίδων τους σε μορφή που διαβάζεται από τις μηχανές χρησιμοποιώντας την HTML.

Χρησιμοποιώντας μια τέτοια γλώσσα, ένα έγγραφο θα αξίωνε πως είναι η προσωπική ιστοσελίδα ενός μεταπτυχιακού φοιτητή. Ένα σύνδεσμος από τη σελίδα αυτή προς μια ερευνητική ομάδα θα μπορούσε να δηλώσει πως ο μεταπτυχιακός φοιτητής εργάζεται στην ομάδα αυτή ως βοηθός έρευνας. Και η ιστοσελίδα θα μπορούσε να υποστηρίξει πως το «Σολωμού» είναι το επώνυμο του μεταπτυχιακού φοιτητή. Οι αξιώσει αυτές *δεν* είναι απλές λέξεις-κλειδιά· μάλλον είναι σημασιολογικές ετικέτες, καθορισμένες σε κάποιο «επίσημο» σύνολο γνωρισμάτων και σχέσεων (μια *οντολογία*). Στο παράδειγμα αυτό, η οντολογία θα περιλάμβανε γνωρίσματα όπως «lastName», ταξινομήσεις όπως «Person» και σχέσεις όπως «employee». Τα συστήματα που συλλέγουν αξιώσεις σχετικά με τα γνωρίσματα και τις σχέσεις αυτές θα μπορούσαν να χρησιμοποιήσουν το αποτέλεσμα της συλλεγμένης γνώσης για να παράσχουν απαντήσεις σε εκλεπτυσμένες επερωτήσεις που βασίζονται στη γνώση.

Επιπλέον, οι διαμεσολαβητές χρήστη ή τα ρομπότ θα μπορούσαν να χρησιμοποιήσουν τις συλλεγμένες σημασιολογικές πληροφορίες για να τελειοποιήσουν τη διεργασία έρπησης του Ιστού (web-crawling). Για παράδειγμα, σκεφτείτε έναν ευφυή διαμεσολαβητή που η αποστολή του είναι να συλλέξει ιστοσελίδες σχετικά με την ιχθυολογία. Αν ο διαμεσολαβητής αυτός χρησιμοποιούσε ένα μηχανισμό που ανέτρεχε σε ένα θησαυρό ή έψαχνε λέξεις κλειδιά, θα μπορούσε τυχαία να αποφασίσει πως η ιστοσελίδα της Έλενας Σολωμού, και οι ιστοσελίδες που συνδέονται από αυτήν, είναι καλοί υποψήφιοι έρευνας για το θέμα αυτό. Αυτό θα μπορούσε να είναι φυσικά ένα μεγάλο λάθος, όχι μόνο για τους προφανείς λόγους, αλλά επίσης επειδή οι σύνδεσμοι της Έλενας Σολωμού γίνονται προς το υπόλοιπο του Πανεπιστημίου Αθηνών (όπου εργάζεται). Ο εξυπηρετητής Ιστού του Πανεπιστημίου Αθηνών είναι πολύ, πολύ μεγάλος και το ρομπότ μπορεί να σπαταλήσει πάρα πολύ χρόνο σε άσκοπη αναζήτηση. Ωστόσο, αν ο διαμεσολαβητής συνέλεγε σημασιολογικές ετικέτες από την ιστοσελίδα της Έλενας Σολωμού που αποτελούσαν ένδειξη πως το Σολωμού ήταν το επώνυμό της, τότε ο διαμεσολαβητής θα είχε τη φρονιμάδα να μην αναζητήσει αυτή την ιστοσελίδα και του συνδέσμούς της.

Η πεπερασμένη αποτελεσματικότητα της γλώσσας σήμανσης HTML

Η HTML 3.0 περιλαμβάνει αρκετούς αδύναμους μηχανισμούς για σημασιολογική σήμανση (οι υπό-ετικέτες REL, REV και CLASS και η ετικέτα META). Η HTML 5.0 προχωρεί τους μηχανισμούς αυτούς. Δυστυχώς, τα στοιχεία σημασιολογικής σήμανσης στην HTML έχουν μέχρι τώρα χρησιμοποιηθεί κυρίως για μετά-πληροφορίες εγγράφων (όπως δηλωμένες λέξεις κλειδιά) ή για σχέσεις που προσανατολίζονται στο υπερκείμενο (hypertext) (όπως «περίληψη» ή «πίνακας περιεχομένων»). Περαιτέρω, οι σχέσεις μπορούν να εδραιωθούν μόνο κατά μήκος των συνδέσμων του υπερκειμένου (χρησιμοποιώντας το <LINK> ή το <A>). Φαίνεται πως η επιδίωξη του υφισταμένου συνόλου σημασιολογικών ετικετών σήμανσης της HTML είναι μόνο η παροχή σημασιολογίας για εφαρμογές υπερκειμένου, ή άλλες λειτουργίες προσανατολισμένες στο έγγραφο.

Για να επιληφθούν μερικών από τα προβλήματα, ειδικοί ερευνητές επιχείρησαν να συνταιριάσουν την HTML με το μοντέλο Οντότητας-Σχέσης (ER - Entity-Relationship) βάσεων δεδομένων. Αυτό γίνεται

προσθέτοντας στην HTML ένα απλό σύνολο ετικετών που καθορίζουν «οντότητες» εντός των εγγράφων, προσθέτοντας ταμπέλες σε τμήματα του κυρίου κειμένου ως «γνωρίσματα» των οντοτήτων αυτών και καθορίζοντας σχέσεις από μια οντότητα προς εξωτερικές οντότητες. Οι Dobson και Burill συσχετίζουν ένα μοναδικό κλειδί με κάθε οντότητα, και δημιουργούν σχέσεις όχι ανάμεσα σε συνδέσμους URL, αλλά ανάμεσα σε κλειδιά.

Αν και το σχέδιο ER των Dobson και Burill είναι μια σημαντική βελτίωση στον υφιστάμενο μηχανισμό της HTML, δεν προνοεί για *οντολογικές* δηλώσεις. Για παράδειγμα, το σχέδιό τους δεν δίνει κάποιο σαφή μηχανισμό για ταξινόμηση μέσω μιας «isa» ιεραρχίας κλάσεων. Ωστόσο, μια από τις πιο σημαντικές χρήσεις της σημασιολογίας στα έγγραφα είναι η κατηγοριοποίησή τους σύμφωνα με κάποιο ταξινομικό σύστημα ή μια ταξινομία. Για παράδειγμα, τα έντυπα έγγραφα συχνά ταξινομούνται χρησιμοποιώντας ιεραρχικά σχήματα, όπως οι θεματικές επικεφαλίδες της Εθνικής Βιβλιοθήκης ή της Βιβλιοθήκης του Κογκρέσου, το Δεκαδικό σύστημα Dewey ή η Παγκόσμια Δεκαδική Ταξινόμηση (UDC). Παρόμοια, ένας καλός σημασιολογικός μηχανισμός για τα έγγραφα του παγκόσμιου Ιστού χρειάζεται τη δυνατότητα να διενεργεί ευέλικτη, ιεραρχική ταξινόμηση. Η δυνατότητα καθιέρωσης σχέσεων ανάμεσα στις οντότητες του Ιστού είναι σημαντική, αλλά δευτερεύουσα της δυνατότητας της ταξινόμησης αυτών των οντοτήτων.

Επιπλέον, το μοντέλο ER δεν επιτρέπει την προδιαγραφή συμπερασμάτων που μπορούν να προκύψουν από σχέσεις που δίνονται στις ιστοσελίδες. Ακόμη και απλές προδιαγραφές όπως συμπεράσματα μεταβατικής λήξης μπορεί να είναι χρήσιμα: αν η προσωπική ιστοσελίδα της Έλενας Σολωμού αξιώνει πως δουλεύει για την ερευνητική ομάδα ΧΨ, και αυτή η ερευνητική ομάδα είναι μέρος του Τμήματος της Επιστήμης των Υπολογιστών, τμήμα του Τμήματος της Πληροφορικής και Θετικών Επιστημών, Σχολής του Πανεπιστημίου Αθηνών στην Πανεπιστημιούπολη Ζωγράφου, τμήμα του Νομού της Αττικής, δεν θα πρέπει να δηλώσει πως δουλεύει για *όλες* αυτές τις οντότητες· ένα τέτοιο γεγονός θα πρέπει να είναι τεκμαρτό. Οι αντίστροφες σχέσεις είναι επίσης χρήσιμες: αν είναι γνωστό πως ο Γεώργιος Σολωμού είναι νυμφευμένος με την Έλενα Σολωμού, το αντίθετο θα πρέπει να είναι αυτόματα τεκμαρτό, χωρίς να πρέπει ο Γεώργιος ή η Έλενα να πρέπει να το πουν. Μέσω της πρόσθεσης ισχυρότερων δυνατοτήτων συμπερασματικού κανόνα, θα μπορούσε να παρασχεθεί πλήρης Γνωσιακή Βάση σημασιολογία.

Η χρησιμοποίηση της οντολογικής γλώσσας SHOE

Αναπτύξαμε ένα μικρό υπερσύνολο της HTML που παρέχει πολλούς από τους μηχανισμούς αυτούς. Το σχέδιο αυτό ονομάζεται SHOE: SimpleHTMLOntologyExpressions (Απλές Εκφράσεις Οντολογίας της HTML). Το SHOE σκοπεύει να παράσχει στους διαμεσολαβητές του χρήστη εύκολη πρόσβαση σε σημασιολογική γνώση που μπορεί να διαβαστεί από μηχανές στο Web. Το κάνει αυτό προσθέτοντας στην HTML μια απλή γλώσσα αναπαράστασης της γνώσης. Το SHOE έχει τη δυνατότητα να:

- Καθορίζει οντολογίες χρησιμοποιώντας την HTML, που τακτοποιούν τους κανόνες ταξινόμησης και σχέσεων των οντοτήτων.
- Δημιουργεί νέες οντολογίες που επεκτείνουν υφιστάμενες οντολογίες.
- Δηλώνει οντότητες, τόσο για ολόκληρα κείμενα όσο και για υπο-τμήματα κειμένων.
- Δηλώνει γνωρίσματα οντοτήτων.
- Ταξινομεί τις οντότητες υπό ένα «isa» ταξινομικό σύστημα.

Η γλώσσα SHOE εσκεμμένα δεν έχει τη σημασιολογική πληρότητα ισχυρότερων σχημάτων αναπαράστασης της γνώσης· τουναντίον, επιχειρεί να επιτύχει ένα ισοζύγιο ανάμεσα στη σημασιολογία αναπαράστασης της πλήρους γνώσης και τους μοναδικούς διαμεσολαβητές του παγκοσμίου Ιστού, που πρέπει ραγδαία να συλλέξουν και να επερωτήσουν πολύ μεγάλα σώματα πληροφοριών σε ένα κατανεμημένο, και συχνά χαοτικό τομέα. Η τυπική προδιαγραφή παρέχει μια «isa» ταξινόμηση και σχέσεις απλής βάσης.

Μια προτεινόμενη προδιαγραφή της SHOE προσθέτει συμπερασματικούς κανόνες υπό τη μορφή προτάσεων χωρίς άρνηση-προσθέτουμε σημασιολογία συντηρητικά. Η πιο τρέχουσα προδιαγραφή, συμπεριλαμβανομένων των συμπερασματικών κανόνων, μπορεί να βρεθεί στο <http://www.cs.umd.edu/projects/plus/SHOE/spec.html>.

Η μεσολάβηση ενός ευφυή διαμεσολαβητή επανομαζόμενος Exposé σε συνεργασία με την SHOE

Για να επιδείξουμε τη χρήση της SHOE, αναπτύσσουμε επίσης ένα διαμεσολαβητή που έρπει τον Ιστό, το Exposé, που αναλύει γραμματικά έγγραφα της HTML που έχουν δυνατότητες της SHOE και προσθέτει γνώση SHOE στην εσωτερική του Γνωσιακή Βάση. Το Exposé τρέχει σε MacintoshCommonLisp ή C, χρησιμοποιώντας το PARKA (Evet, Anderson και Hendler, 1993), το μαζικά παράλληλο σύστημα σημασιολογικού δικτύου του Πανεπιστημίου του Μαϊντλαντ, για την αναπαράσταση γνώσης του.

Το Exposé είναι ουσιαστικά δύο διαμεσολαβητές. Ο πρώτος διαμεσολαβητής αναζητεί και φέρνει οντολογίες από τον ιστό όπως έχει εντολές, φορτώνοντάς τις στη βάση γνώσεων PARKA που έχει. Όταν αναζητείται και φέρνεται μια οντολογία SHOE, που οι γονικές οντολογίες είναι άγνωστες στο Exposé, ο διαμεσολαβητής θα αναζητήσει και θα φέρει πρώτα τους γονείς, με σκοπό να περιορίσει κατάλληλα την οντολογία στη βάση γνώσεων του Exposé.

προσθέτοντας γνώση της SHOE από τα έγγραφα αυτά στη βάση γνώσεων του Exposé. Όταν ανακαλύπτει ένα έγγραφο SHOE, το Exposé χρησιμοποιεί τη βάση γνώσεων του για να ερμηνεύσει τα δεδομένα SHOE στο πλαίσιο γνωστών οντολογιών, να καθορίσει σφάλματα και πλεονάζουσες αξιώσεις, και να βρει νέες ιστοσελίδες για να επισκεφθεί. Αν το Exposé αποφασίσει πως τα δεδομένα SHOE αυτά είναι χρήσιμα, τότε προσθέτει τα δεδομένα στη βάση γνώσεων του. Αν ο δεύτερος διαμεσολαβητής δεν έχει τις απαιτούμενες οντολογίες για να ερμηνεύσει ένα έγγραφο SHOE, μπορεί να ζητήσει από τον πρώτο διαμεσολαβητή να φορτώσει και να περιορίσει τις οντολογίες αυτές πριν να συνεχίσει.

Το Exposé χρειάζεται μια βάση γνώσεων για να φυλάζει όχι μόνο νέο-ανακαλυφθείσα γνώση SHOE, αλλά επίσης και ένα σύνολο οντολογιών, με τις οποίες ερμηνεύει τη γνώση SHOE καθώς αυτή αναλύεται γραμματικά. Αυτή είναι ένα πολύ μεγάλο ποσό πληροφορίας: αισιόδοξα, ο Ιστός θα μπορούσε να παράγει εκατοντάδες κοινώς χρησιμοποιημένες οντολογίες και εκατοντάδες χιλιάδες έγγραφα με δυνατότητα SHOE. Το Exposé πρέπει να επερωτήσει τη γνώση αυτή αρκετές φορές για κάθε έγγραφο. Το PARKA παρέχει την ιπποδύναμη που βοηθά το Exposé να το κάνει αυτό· το PARKA έχει βελτιστοποιηθεί για να παρέχει επερωτήσεις πραγματικού χρόνου σε βάσεις γνώσεων με εκατομμύρια οντότητες δεδομένων και ισχυρισμούς.

Αφού το Exposé έχει συγκεντρώσει τη γνώση από τον Ιστό, μπορούμε στη συνέχεια να χρησιμοποιήσουμε τη γνώση αυτή για να απαντήσουμε σε εκλεπτυσμένες επερωτήσεις σχετικά με τις οντότητες αυτές και τις σχέσεις τους. Για παράδειγμα, μετά τη συγκέντρωση αξιωμάτων από την ιστοσελίδα της Έλενας Σολωμού από το Exposé, μπορούμε να επερωτήσουμε το PARKA για να βρει αυτήν και το σύζυγό της. Αυτό είναι ισοδύναμο με την επερώτηση του PARKA με:

```
(query! '(:and
  (!instanceOf ?X #!Person) (!instanceOf ?Y #!Person)
  (!instanceOf ?Z #!Organization)
  (!lastName ?X "Σολωμού") (!lastName ?Y "Σολωμού")
  (!employee ?Z ?X) (!employee ?Z ?Y)
  (!marriedTo ?X ?Y) (!involvedIn ?Z "ΥΠΕΘΑ 123-4567")))
```

Σε συνδυασμό με το Exposé, σχεδιάζουμε εφαρμογές Java (γραφικούς σχολιαστές σελίδων, μηχανισμούς επερώτησης κτλ), για να βοηθήσουν τους χρήστες να σχολιάζουν ιστοσελίδες με σημασιολογική γνώση και να επερωτούν τους εξυπηρετητές ρομπότ με την SHOE.

Η πραγματική δύναμη του Σημασιολογικού Ιστού θα αποτελέσει πραγματικότητα όταν δημιουργηθούν πολλά προγράμματα που να συλλέγουν περιεχόμενο από τον Ιστό (Webcontent) από διάφορες πηγές, να επεξεργάζονται τις πληροφορίες και να ανταλλάσσουν τα αποτελέσματα τους με άλλα προγράμματα. Η αποτελεσματικότητα αυτών των διαμεσολαβητών λογισμικού θα αυξάνεται σταδιακά καθώς θα αρχίσει να διατίθεται ολοένα και περισσότερο μηχαναγνώσιμο περιεχόμενο από τον Ιστό και αυτοματοποιημένες υπηρεσίες (συμπεριλαμβανομένων και άλλων διαμεσολαβητών). Ο Σημασιολογικός Ιστός προάγει αυτή τη συνεργεία: ακόμη και διαμεσολαβητές που δεν σχεδιάστηκαν να συνεργάζονται μεταξύ τους μπορούν να μεταφέρουν δεδομένα το ένα στο άλλο όταν τα δεδομένα αυτά συνοδεύονται από σημασιολογικό περιεχόμενο.

Μια σημαντική πτυχή της λειτουργικότητας των πρακτόρων θα είναι η ανταλλαγή 'αποδείξεων' που θα είναι γραμμένες στην γλώσσα ενοποίησης του Σημασιολογικού Ιστού (δηλ. στη γλώσσα που εκφράζει την εξαγωγή λογικών συμπερασμάτων σύμφωνα με συγκεκριμένους κανόνες και πληροφορίες όπως ορίζονται από τις οντολογίες). Εάν, π.χ., η κύρια Παπά εντοπιστεί από μια online υπηρεσία στην Κύπρο και θέλετε να επιβεβαιώσετε την πληροφορία, τότε ο H/Y σάς ζητάει από την υπηρεσία αυτή απόδειξη της απάντησης που σας έστειλε. Η απόδειξη αυτή παρέχεται άμεσα αφού μεταφραστεί ο εσωτερικός συλλογισμός της υπηρεσίας στη γλώσσα ενοποίησης του Σημασιολογικού Ιστού. Μια μηχανή εξαγωγής συμπερασμάτων στον H/Y σας επαληθεύει άμεσα ότι πράγματι η συγκεκριμένη κυρία ταιριάζει με αυτή που ψάχνεται. Μπορεί, αν το επιθυμείτε, να σας παρουσιάσει τις σχετικές ιστοσελίδες. Αν και απέχουμε πολύ από την σε βάθος εκμετάλλευση των δυνατοτήτων του Σημασιολογικού Ιστού, ορισμένα προγράμματα μπορούν ήδη να ανταλλάσσουν αποδείξεις με αυτό τον τρόπο, χρησιμοποιώντας τις πρώτες διαθέσιμες εκδόσεις της γλώσσας ενοποίησης.

Ήδη υπάρχουν πολλές ιστοπαγείς υπηρεσίες (Web-based services) που δε χρησιμοποιούν σημασιολογία. Ωστόσο, προγράμματα όπως οι διαμεσολαβητές δεν μπορούν να εντοπίσουν με άλλο τρόπο μια υπηρεσία που θα εκτελέσει μια συγκεκριμένη λειτουργία. Αυτή η διαδικασία, γνωστή ως ανακάλυψη υπηρεσίας (servicediscovery) μπορεί να λάβει χώρα μόνο όταν υπάρχει μια κοινή γλώσσα που να περιγράφει μια υπηρεσία και κατά τέτοιο τρόπο που να μπορεί να δώσει σε άλλους πράκτορες να 'καταλάβουν' τόσο τη λειτουργία που προσφέρεται όσο και πώς να την εκμεταλλευτεί. Οι υπηρεσίες και οι πράκτορες μπορούν να διαφημίζουν τη λειτουργία τους εάν, π.χ., εναποθέσουν τέτοιες περιγραφές σε καταλόγους ανάλογους προς τις κίτρινες σελίδες του Χρυσού Οδηγού.

Σήμερα, υπάρχουν διαθέσιμα ορισμένα σχήματα ανακάλυψης υπηρεσίας χαμηλού επιπέδου (low-level service-discovery schemes), όπως τα: Universal Plug and Play της Microsoft (εστιάζει στη σύνδεση διαφόρων ειδών συσκευών) και το Jini της Sun Microsystems (στοχεύει στη σύνδεση υπηρεσιών). Αυτές οι πρωτοβουλίες, ωστόσο, χτυπούν το πρόβλημα σε δομικό ή συντακτικό επίπεδο και βασίζονται κυρίως στην καθιέρωση ενός προκαθορισμένου συνόλου περιγραφών λειτουργικότητας.

Αντίθετα, ο Σημασιολογικός Ιστός είναι πιο ευέλικτος. Οι διαμεσολαβητές των καταναλωτών και των παραγωγών μπορούν να οδηγηθούν σε ένα κοινό επίπεδο κατανόησης ανταλλάσσοντας οντολογίες, οι οποίες παρέχουν το απαιτούμενο λεξιλόγιο για συζήτηση. Οι διαμεσολαβητές μπορούν ακόμα και να 'εξασφαλίζουν' νέες ικανότητες συλλογισμού όταν ανακαλύπτουν νέες οντολογίες. Επίσης, η σημασιολογία καθιστά πιο εύκολη την εκμετάλλευση μιας υπηρεσίας που ταιριάζει μόνο μερικώς σε μια αίτηση μας.

Μια τυπική διαδικασία θα αφορά στη δημιουργία μιας 'αλυσίδας αξιών' σύμφωνα με την οποία υποσυστήματα πληροφοριών θα περνούν από τον ένα πράκτορα στον άλλο, η καθεμία από τις οποίες θα 'προσθέτει αξία', προκειμένου να κατασκευαστεί το τελικό προϊόν που ζητήθηκε από τον τελικό χρήστη. Είναι βέβαιο ότι για να δημιουργηθούν περίπλοκες αλυσίδες αξιών αυτόματα και κατά βούληση, ορισμένοι διαμεσολαβητές θα εκμεταλλευτούν και τις τεχνολογίες τεχνητής νοημοσύνης εκτός από το Σημασιολογικό Ιστό. Ο Σημασιολογικός Ιστός, όμως, θα δώσει τις βάσεις και το πλαίσιο ώστε να γίνουν πιο εφικτές αυτές οι τεχνολογίες.

Αργότερα, ο Σημασιολογικός Ιστός θα σπάσει τα δεσμά του εικονικού κόσμου (virtual realm) και θα επεκταθεί στον φυσικό μας κόσμο. Οι URI μπορούν να καταδεικνύουν οτιδήποτε, ακόμα και φυσικές οντότητες, πράγμα που σημαίνει ότι μπορούμε να χρησιμοποιούμε το RDF για να περιγράψουμε συσκευές όπως κινητά τηλέφωνα και τηλεοράσεις. Τέτοιες συσκευές μπορούν να διαφημίζουν τη λειτουργικότητα τους (δηλ. τι μπορούν να κάνουν και πώς μπορούμε να τις ελέγξουμε) όπως γίνεται περίπου και με τους διαμεσολαβητές λογισμικού. Αυτή η σημασιολογική προσέγγιση, όντας πολύ πιο ευέλικτη από τα σχήματα χαμηλού επιπέδου τύπου Universal Plug and Play, ανοίγει νέους και τρομερά ενδιαφέροντες ορίζοντες.

Π.χ., εάν ο απλός συντονισμός των οικιακών συσκευών που γνωρίζουμε σήμερα αντικατασταθεί με σημασιολογικές περιγραφές των δυνατοτήτων και της λειτουργικότητας τους, τότε θα επιτύχουμε τον αυτοματισμό με ελάχιστη ανθρώπινη παρέμβαση. Αντί, δηλ., να προγραμματίζουμε την κάθε ηλεκτρική συσκευή χωριστά, θα μπορούμε να προγραμματίσουμε μια συγκεκριμένη λειτουργία άπαξ δια παντός: π.χ., κάθε *τοπική* συσκευή που σύμφωνα με το διαφημιστικό φυλλάδιο έχει *ρυθμιστή ακουστικής έντασης* θα χαμηλώνει αμέσως μόλις χτυπήσει το τηλέφωνο.

Τα πρώτα σταθερά βήματα προς αυτή την κατεύθυνση έχουν ήδη γίνει: υπάρχουν άνθρωποι που εργάζονται για την εξέλιξη ενός προτύπου που θα περιγράφει τις λειτουργικές δυνατότητες των συσκευών (π.χ., το μέγεθος μιας οθόνης) και τις προτιμήσεις των χρηστών. Αυτό το πρότυπο βασίζεται σε έναν RDF και ονομάζεται Composite Capability/ Preference Profile (CC/ PP). Αρχικά, θα επιτρέψει στα κινητά τηλέφωνα και σε μη σταθερούς πελάτες του Ιστού να περιγράψουν τα χαρακτηριστικά τους έτσι το περιεχόμενο του Ιστού να προσαρμοστεί πάνω σε αυτούς. Αργότερα, όταν θα προστεθεί η πολυχρησιμότητα των γλωσσών για το χειρισμό των οντολογιών και της λογικής σε όλη της την έκταση, οι συσκευές θα μπορούν αυτόματα να ψάχνουν και να χρησιμοποιούν υπηρεσίες και άλλες συσκευές για περαιτέρω πληροφορίες ή λειτουργικότητα. Δεν είναι δύσκολο να φανταστούμε τον φούρνο μικροκυμάτων που έχει τη δυνατότητα ενεργοποίησής του μέσω Ιστού (Web-enabled) να συμβουλευεται την ιστοσελίδα της εταιρίας κατεψυγμένων για τις ιδανικές συνθήκες μαγειρέματος.

4.3 Οντολογικές Γλώσσες

Οι οντολογίες έχουν αποδειχθεί ιδιαιτέρως χρήσιμες σε πολλές εφαρμογές, όπως τα συστήματα διαμεσολαβητών, συστήματα διαχείρισης γνώσης, και πλατφόρμες ηλεκτρονικού εμπορίου. Μπορούν, επίσης, να παράγουν φυσική γλώσσα, να ενσωματώνουν ευφρείς πληροφορίες, να παρέχουν πρόσβαση στο Διαδίκτυο

βάσει σημασιολογίας, και να εξάγουν πληροφορίες από διάφορα κείμενα παράλληλα με τη χρήση τους σε άλλες εφαρμογές προκειμένου να δηλώσουν ρητά τη γνώση που περικλείεται σε αυτές.

Ωστόσο, εκτός από τη χρήση τους σε εφαρμογές όπου η γνώση διαδραματίζει κεντρικό ρόλο, μπορούν να πυροδοτήσουν σημαντικές αλλαγές στο περιεχόμενο του Ιστού με την τρέχουσα μορφή. Οδηγούμαστε, δηλ., στην τρίτη γενιά του Ιστού, τον γνωστό ως Σημασιολογικό Ιστό, ο οποίος έχει οριστεί ως 'η εννοιολογική δόμηση του Ιστού με ρητό / κατηγορηματικό μηχαναγνώσιμο τρόπο. Ο ορισμός αυτός δε διαφέρει και πολύ από εκείνον της οντολογίας: 'Η οντολογία είναι μια τυπική, κατηγορηματική προδιαγραφή κοινής εννοιολογικής αναπαράστασης'.

Ο νέος αυτός Ιστός όπως και οι γλώσσες που καλούνται να αναπαραστήσουν αυτές τις σημασιολογικές πληροφορίες έχουν ένα κοινό παρανομαστή: την επίλυση του προβλήματος ανταλλαγής των ετερογενών δεδομένων σε αυτό το ετερογενές περιβάλλον. Στόχος των ερευνητών είναι η δημιουργία της καταλληλότερης γλώσσας για την αναπαράσταση των νέων αναγκών.

Ορισμένες από τις οντολογικές γλώσσες βασίζονται στη σύνταξη XML, όπως οι *OntologyExchangeLanguage* (XOL), η *SHOE* που παλιά βασίζονταν στην *HTML*, και η *OntologyMarkupLanguage* (OML). Αντίθετα, οι *ResourceDescriptionFramework* (RDF/ Πλαίσιο Περιγραφής Πόρων) και *RDF Schema* είναι γλώσσες που δημιουργήθηκαν από τις ομάδες εργασίας της Κοινοπραξίας του Παγκοσμίου Ιστού (*WorldWideWebConsortium* (W3C)). Τέλος, υπάρχουν ακόμα 2 γλώσσες που χτίζονται επάνω στις *RDF(S)*- δηλαδή η συνένωση των *RDF* και *RDF Schema*- προκειμένου να βελτιώσουν τα χαρακτηριστικά τους: οι *OntologyInferenceLayer* (OIL) και *DAML+OIL*.

Ακολουθεί αναλυτική παρουσίαση των επικρατέστερων οντολογικών γλωσσών όπως αυτές προαναφέρθηκαν.

XML-based Ontology Exchange Language

Η XML σχεδιάστηκε από την κοινότητα βιοπληροφορικής των ΗΠΑ (*USbioinformaticscommunity*) για την ανταλλαγή οντολογικών ορισμών ανάμεσα σε ετερογενή σύνολα συστημάτων λογισμικού του πεδίου τους, κι αφού πρώτα ερευνητές μελέτησαν ανάγκες αναπαράστασης των γνώσεων των συγκεκριμένων ειδικών. Για τη δημιουργία της (XOL) βασίστηκαν στην *Ontolingua* και στην *OML*, συγχωνεύοντας έτσι την υψηλή εκφραστικότητα της *OKBC-Lite*, ενός υποσυνόλου του πρωτοκόλλου *OpenKnowledgeBasedConnectivity* (*OpenKnowledgeBasedConnectivityprotocol*) και της σύνταξης *OML*, που βασίζεται στην XML. Η δημιουργία οντολογιών με XML είναι αδύνατη εφόσον δεν υπάρχουν τα κατάλληλα εργαλεία. Ωστόσο, αφού τα αρχεία XOL χρησιμοποιούν τη σύνταξη XML, μπορούμε να χρησιμοποιούμε έναν επεξεργαστή XML για τη συγγραφή XOL αρχείων.

Επέκταση Απλής HTML Οντολογίας SHOE (Simple HTML Ontology Extension)

Η *SHOE*, που εξελίχθηκε στο Πανεπιστήμιο του Maryland και χρησιμοποιήθηκε για την ανάπτυξη της *OML*, δημιουργήθηκε ως επέκταση της *HTML*, ενσωματώνοντας μηχαναγνώσιμη σημασιολογική γνώση σε έγγραφα *HTML* ή άλλα κείμενα του Ιστού. Πρόσφατα, το Πανεπιστήμιο του Maryland προσάρμοσε τη σύνταξη *SHOE* στην XML. Με τη *SHOE*, οι πράκτορες μπορούν να συγκεντρώσουν σημαντικές πληροφορίες για τις σελίδες και τα έγγραφα του Ιστού, βελτιώνοντας τους μηχανισμούς αναζήτησης, και τη συλλογή γνώσης. Αυτή η διαδικασία συντελείται σε 3 φάσεις: ορισμός μιας οντολογίας, σχολιασμός των σελίδων *HTML* με οντολογικές πληροφορίες για αυτές και άλλες σελίδες, και χρήση ενός διαμεσολαβητή για τη σημασιολογική ανάκτηση πληροφοριών, ο οποίος θα ψάχνει όλες τις υπάρχουσες σελίδες και θα κρατάει ενήμερες τις πληροφορίες.

Ο Σχολιαστής Γνώσης (*KnowledgeAnnotator*) σχολιάζει τις οντολογικές πληροφορίες στις σελίδες *HTML*.

OML (Ontology Markup Language)

HOML εξελίχθηκε στο Πανεπιστήμιο της Ουάσινγκτον και βασίζεται εν μέρει στην *SHOE*. Αρχικώς, θεωρούνταν μια XML αλληλουχία της *SHOE*, γι' αυτό και οι πολλές ομοιότητες μεταξύ των *OML* και *SHOE*. Υπάρχουν τέσσερα επίπεδα της *OML*: η *OML Core* (*Πυρηνική OML*) σχετίζεται με τις λογικές πτυχές της γλώσσας και περιλαμβάνεται σε όλα τα υπόλοιπα επίπεδα· η *SimpleOML* (*Απλή OML*) χαρτογραφείται άμεσα στην *RDF(S)*· η *AbbreviatedOML* (*Συντετμημένη OML*) περιλαμβάνει εννοιολογικά γραφικά χαρακτηριστικά (*conceptualgraphsfeatures*)· και η *StandardOML* (*Καθιερωμένη OML*) αποτελεί την πιο εκφραστική έκδοση της *OML*. Επιλέξαμε την *SimpleOML* γιατί τα ανώτερα επίπεδα της δεν παρέχουν περισσότερα συστατικά μέρη από εκείνα που έχουμε ορίσει στο πλαίσιο μας. Αυτά τα ανώτερα επίπεδα σχετίζονται στενά με την αναπαράσταση των εννοιολογικών γραφημάτων (*conceptualgraphs*). Δεν υπάρχουν άλλα εργαλεία για την συγγραφή οντολογιών *OML* εκτός από τα υπάρχοντα γενικής χρήσης εργαλεία έκδοσης της XML.

RDF (Resource Description Framework) και RDF Schema

Η RDF εξελίχθηκε από το W3C για την περιγραφή πόρων του Ιστού, επιτρέποντας τον προσδιορισμό της σημασιολογίας των δεδομένων που βασίζονταν στην XML με έναν τυποποιημένο, διαλειτουργικό τρόπο. Διαθέτει, επίσης, μηχανισμούς για να αναπαριστά ρητώς υπηρεσίες, διαδικασίες, και επιχειρηματικά πρότυπα, ενώ επιτρέπει την αναγνώριση μη ρητά διατυπωμένων πληροφοριών. Το πρότυπο δεδομένων RDF ισοδυναμεί με το σημασιολογικό φορμαλισμό των δικτύων. Αποτελείται από 3 τύπους αντικειμένων: (α) οι πόροι περιγράφονται από εκφράσεις RDF και καθορίζονται πάντα από τους URIs (UniversalResourceIdentifiers), (β) οι ιδιότητες ορίζουν συγκεκριμένες πτυχές, χαρακτηριστικά, χαρακτηριστικά γνωρίσματα (attributes), ή σχέσεις που χρησιμοποιούνται για την περιγραφή ενός πόρου· και (γ) δηλώσεις (statements), οι οποίες προσδίδουν μια αξία/ τιμή (value) σε κάθε ιδιότητα (property) ενός συγκεκριμένου πόρου (αυτή η αξία μπορεί να είναι μια άλλη δήλωση RDF). Στο πρότυπο δεδομένων RDF δεν παρέχονται μηχανισμοί για τον ορισμό των σχέσεων ιδιοτήτων (properties) και πόρων· αυτό είναι δουλειά της RDFS. Η RDFS προσφέρει αρχέγονα για τον ορισμό προτύπων γνώσεων που πλησιάζουν περισσότερο στις προσεγγίσεις που είναι βασισμένες σε πλαίσια. Η RDFS χρησιμοποιείται ευρέως ως μορφότυπο αναπαράστασης σε πολλά εργαλεία και ερευνητικά προγράμματα, όπως τα Amaya, Protégé, Mozilla, SilRI, κ.ο.κ.

OIL (Ontology Interchange Language)

Η OIL, η οποία εξελίχθηκε στο ερευνητικό πρόγραμμα OntoKnowledge (www.ontoknowledge.org/OIL), επιτρέπει τη σημασιολογική διαλειτουργικότητα μεταξύ των πόρων του Ιστού. Η σύνταξη και η σημασιολογία της βασίζεται στις γλώσσες OKBC, XOL, και RDFS, παρέχοντας σχεδιαστικά πρότυπα (modelingprimitives) που χρησιμοποιούνται σε πλαισιοπαγείς προσεγγίσεις στην οντολογική μηχανίκευση (ontologicalengineering) (έννοιες, ταξινομίες εννοιών, σχέσεις, κ.ο.κ.). Επίσης, παρέχει μορφική σημασιολογία (formalsemantics) και συλλογιστική στήριξη (reasoningsupport) που εμφανίζεται σε προσεγγίσεις περιγραφικής λογικής (descriptionlogicapproaches) (ένα υποσύνολο λογικής πρώτης τάξης (firstorderlogic) που διατηρεί υψηλή εκφραστική δύναμη, εκτός από αποφασιστικότητα και έναν επαρκή μηχανισμό συμπερασματολογίας (inferencemechanism)).

Η OIL χτίστηκε επάνω στην RDF(S) και αποτελείται από τα εξής επίπεδα: το *Core OIL* ομαδοποιεί τα αρχέγονα της OIL που χαρτογραφούνται άμεσα στα αρχέγονα της RDF(S)· το *StandardOIL* είναι το πλήρες πρότυπο OIL (OILmodel) και χρησιμοποιεί περισσότερα αρχέγονα από εκείνα που ορίζονται στην RDF(S)· το *InstanceOIL* προσθέτει υποστάσεις εννοιών και ρόλους στο προηγούμενο πρότυπο· και το *HeavyOIL* αποτελεί το επίπεδο στο οποίο θα βασιστούν μελλοντικές επεκτάσεις της OIL. Τα ερευνητικά προγράμματα OIEd, Protégé 2011, και WebODE μπορούν να χρησιμοποιηθούν για τη συγγραφή οντολογιών OIL. Η σύνταξη της OIL εκφράζεται στην XML αλλά μπορεί να αναπαριστάται και στην ASCII. Εμείς θα χρησιμοποιήσουμε την ASCII για τα παραδείγματα μας.

DAML (DAPRA Agent Markup Language) + OIL

Η DAML+OIL εξελίχθηκε από μια μικτή επιτροπή των ΗΠΑ και της Ευρωπαϊκής Ένωσης (IST) στα πλαίσια της DAML, ένα ερευνητικό πρόγραμμα που επιτρέπει τη σημασιολογική διαλειτουργικότητα στην XML. Γι' αυτό και η DAML+OIL έχει κοινό στόχο με την OIL. Η DAML+OIL χτίστηκε στην RDF(S). Το όνομα της υποδηλώνει την ύπαρξη στενής σχέσης με την OIL. Αντικαθιστά τον αρχικό προσδιορισμό, ο οποίος ονομαζόταν DAML+ONT, και βασιζόταν εξίσου στην γλώσσα OIL. Τα OIEd, OntoEdit, Protégé 2000, και WebODE είναι εργαλεία που μπορούν να συγγράψουν οντολογίες DAML+OIL.

OWL (WebOntologyLanguage)

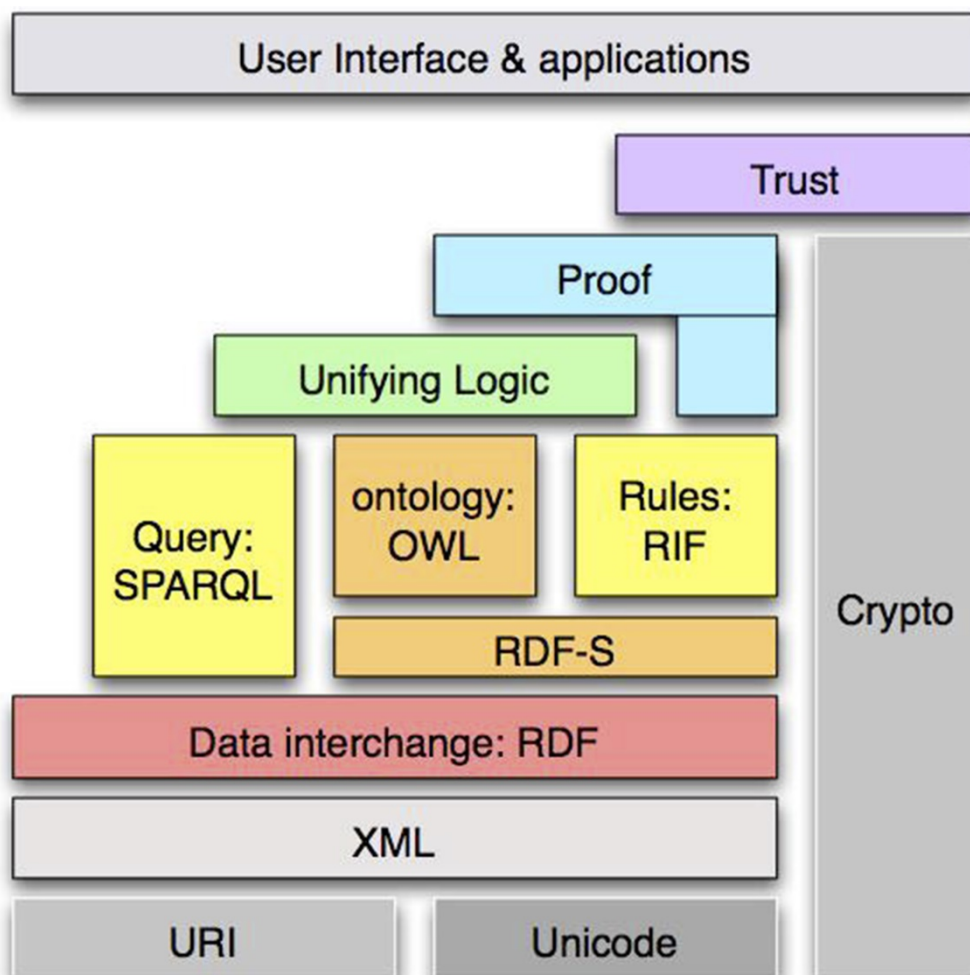
Η OWL σήμερα αποτελεί Οντολογική Γλώσσα του Παγκόσμιου Ιστού και αναδεικνύεται ως πρότυπο οντολογίας από την W3C. Η OWL είναι γλώσσα για τον ορισμό δομημένων οντολογιών που δίνουν τη δυνατότητα για ενσωμάτωση και διαλειτουργικότητα δεδομένων, σε εύρος εφαρμογών. Η OWL δίνει τη δυνατότητα πραγματοποίησης εύρους περιγραφικών εφαρμογών, συμπεριλαμβανομένων και της διαχείρισης θεματικών πυλών (portals) του Παγκόσμιου Ιστού, αναζητήσεων που βασίζονται στο περιεχόμενο, διαχείρισης συλλογών, ενδυναμώνοντας τους ευφυείς διαμεσολαβητές (intelligentagents), τις υπηρεσίες του Παγκόσμιου Ιστού και άλλες υπολογιστικές εφαρμογές. Η γλώσσα OWL ορίζεται μέσω 6 εγγραφών: την επισκόπηση (OWLOverview), τη σημασιολογία και το αφηρημένο συντακτικό (OWLSemanticsandAbstractSyntax), τις απαιτήσεις και σενάρια χρήσης (OWLUseandRequirements), τις περιπτώσεις δοκιμών (OWLTestCases), τον οδηγό (OWLGuide) και την αναφορά (OWLReference). Ενώ οι προηγούμενες γλώσσες (που αναφέραμε) χρησίμευαν στην ανάπτυξη εργαλείων και οντολογιών για συγκεκριμένες κοινότητες χρηστών π.χ. εφαρμογές ηλεκτρονικού εμπορίου εταιρειών, δεν ορίζονταν να είναι συμβατές με την αρχιτεκτονική του Παγκόσμιου Ιστού γενικά αλλά και ούτε με τον Σημασιολογικό Ιστό ειδικότερα. Η OWL είναι Οντολογική Γλώσσα και επεμβαίνει αποτελεσματικά χρησιμοποιώντας τα URIs για

ονοματοδοσία και διασυνδέσεις μέσω RDF για την προσθήκη δυνατοτήτων όπως: (α) επεκτασιμότητα σύμφωνα με τις ανάγκες στον Παγκόσμιο Ιστό (β) δυνατότητα κατανομής σε εύρος συστημάτων (γ) διατηρούν το προνόμιο να είναι οι οντολογίες ανοιχτές και προσβάσιμες και το κυριότερο (δ) συμβατότητα με τα πρότυπα του Παγκόσμιου Ιστού για προσβασιμότητα και διεθνοποίηση.

Η OWL στηρίζεται πάνω στο μοντέλο και σχήμα RDF, προσθέτοντας πλουσιότερο λεξιλόγιο για την περιγραφή ιδιοτήτων και τάξεων όπως σχέσεις μεταξύ τάξεων π.χ. μη συσχετιση, αριθμητικές σχέσεις π.χ. “ακριβώς ένα”, ισότητα, πλουσιότερη πληκτρολόγηση ιδιοτήτων, χαρακτηριστικά ιδιοτήτων π.χ. συμμετρία και τέλος αριθμημένες τάξεις.

Υπάρχουν πολλές εφαρμογές της Γλώσσας OWL, ενώ οι πρώτοι που υιοθέτησαν αυτά τα πρότυπα περιλαμβάνουν κοινότητες βιο-πληροφορικής και ιατρικής, ιδιωτικές επιχειρήσεις και κυβερνήσεις.

Σήμερα η Κοινοπραξία του Παγκόσμιου Ιστού (W3C) εξέδωσε την Οντολογική Γλώσσα του Παγκόσμιου Ιστού (OWL) ως υποψήφια σύσταση του W3C. Ο TimBerners-Lee, Διευθυντής του W3C είπε ότι η OWL είναι ένα σημαντικό βήμα για να γίνουν τα δεδομένα του Παγκόσμιου Ιστού πιο επεξεργάσιμα από τις μηχανές και επαναχρησιμοποιήσιμα σε εύρος εφαρμογών, τονίζοντας ότι ενθαρρυνόμαστε να δούμε την OWL να χρησιμοποιείται ήδη ως ανοιχτό πρότυπο για την ανάπτυξη οντολογιών μεγάλης κλίμακας στον Παγκόσμιο Ιστό (Βλέπε Σχήμα 4.2).



Σχήμα 4. 2. Επεξήγηση των επιπέδων αρχιτεκτονικής διαστρωμάτωσης του Σημασιολογικού Ιστού, όπως διατυπώθηκε από τον TimBerners-Lee στο Συνέδριο του 2000 για την XML.

Επίπεδο 1: URI και UNICODE

- URI σημαίνει Ενιαίος Αναγνωρίσιμος Πόρος
- Υπάρχουν πολλές υποκατηγορίες URIs τα URN και URL

- URN σημαίνει Παγκόσμιοι Πόροι Ονομάτων επιτρέποντας τα πράγματα να είναι μοναδικά αναγνωρίσιμα.
- URL σημαίνει Παγκόσμιος Εντοπιστής Πόρου επιτρέποντας τους πόρους του Διαδικτύου να ανακτώνται όπως για παράδειγμα <http://www.ionio.gr>
- Το Unicode είναι ο αντικαταστάτης του ASCII και επιτυγχάνει να αντιμετωπίζει της πολλαπλές ομιλούμενες γλώσσες.

Επίπεδο 2: XML

Η XML είναι ένα μορφότυπο γλώσσα πρότυπο για την αλληλουχία των δεδομένων χρησιμοποιώντας ετικέτες (tags)

Προέρχεται από την εξέλιξη του SGML προτύπου και είναι παρόμοια με εκείνο του HTML για παράδειγμα (βλ σχήμα 4.3):

```
<WV-CSP-Message xmlns="http://www.wireless-village.org/CSP1.0">
  <TransactionContent xmlns="http://www.wireless-village.org/TRC1.0">
    <CapabilityList>
      <ClientType>MOBILE_PHONE</ClientType>
      <InitialDeliveryMethod>P</InitialDeliveryMethod>
      <AcceptedContentType>text/plain</AcceptedContentType>
      <AcceptedTransferEncoding>BASE64</AcceptedTransferEncoding>
      <AcceptedContentLength>32767</AcceptedContentLength>
    </CapabilityList>
  </TransactionContent>
</WV-CSP-Message>
```

Σχήμα 4.3. Παράδειγμα Εξέλιξης του SGML προτύπου

Υπάρχουν πολλά εργαλεία προσιτά για τη γλώσσα XML όπως π.χ.

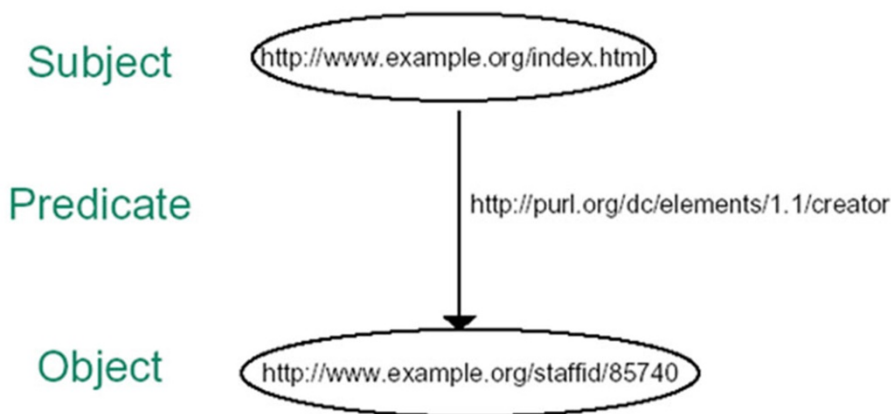
- XSLT για τον μετασχηματισμό
- DOM (Document Object Model) και SAX κατατμητές (parsers)
- Σχήμα Γλώσσες όπως XMLS για επαλήθευση

Επίπεδο 2: NS (Namespaces) Χώροι ονομάτων

- Τα XMLNamespaces είναι μια επέκταση της XML. Είναι συλλογές ονομάτων της XML και χρησιμοποιούνται σαν ονόματα στοιχείων και γνωρισμάτων. Συνοδεύονται από το URI ή το URL από το οποίο προέρχονται.
- Σε μια Namespace ένα URI είναι υπόδειγμα με το όνομα qNames π.χ.
- Το qName αποτελείται από
- έναν χαρακτηριστική (qualifier) που υποδεικνύει το λεξιλόγιο
- ένα τμήμα (fragment) που υποδεικνύει τα στοιχεία στο λεξιλόγιο
- Ομαδοποιούμε κοινά στοιχεία σε ένα λεξιλόγιο π.χ. μοιράζονται τον ίδιο χαρακτηριστή
- Στον Σημασιολογικό Ιστό υποτίθεται ότι θα υπάρχουν πολλά και διάφορα επικαλυπτόμενα λεξιλόγια. Τα Namespaces παρέχουν ένα μέσο μοναδικής αναγνώρισης κάθε επιμέρους στοιχείου σε κάθε λεξιλόγιο

Επίπεδο 3: RDF

Η RDF είναι μια γραφική παράσταση που αποτελείται από τρία μέρη όπου κάθε τριπλό μέρος αντιπροσωπεύει μια δήλωση για έναν πόρο που αποτελείται από : υποκείμενο – κατηγορημα – αντικείμενο π.χ (βλ. Σχήμα 4.4).



Σχήμα 4. 4. Παράδειγμα RDF γραφικής παράστασης

Η πρότυπη θεωρία *ModelTheoryRDF* είναι μια λογική θεωρία που καθορίζει πώς λογικά ερμηνεύεται ένα πρότυπο *RDF*

- Παρέχει την τυπική σημασιολογία δηλ. πώς να ερμηνεύσει λογικά το πρότυπο παρά τη σημασιολογία δηλ. την έννοια και την ερμηνεία του προτύπου *RDF*
- Βελτιστοποιείται για την αποδοτικότητα: συμβατικά κατά την προσθήκη νέων δηλώσεων στις βάσεις γνώσεων που πρέπει να εξετάσουμε εάν υπάρχει αντίφαση
- Η *ModelTheory* αποφεύγει αυτό το πρόβλημα με το να επιτρέψει ενδεχομένως αντιφατικές δηλώσεις που βεβαιώνονται ταυτόχρονα, με την εξαίρεση των αλληλοσυγκρουόμενων τύπων δεδομένων
- Δεν παρέχει οποιαδήποτε πρόσθετη σημασιολογία για τα περιεχόμενα τις συλλογές και τους περιορισμούς

Επίπεδο 4: RDF Schema

- Η RDF Σχήμα είναι μια γλώσσα για την περιγραφή των λεξιλογίων RDF.
- Μπορεί να περιγράψει ιεραρχίες κατηγορίας

```

rdfs:Resource
  example:LivingThing
    example:Dog
    example:Human
  example:EmailAddress
  
```

- Και ιεραρχίες κτήσης

```

rdf:Property
  example:relative
    example:parent
    example:father
    example:mother
  example:email
  
```

- Επιτρέπει τους τομείς και τη σειρά των ιδιοτήτων για να είναι

```

example:email      Domain=example:Human, Range=example:EmailAddress
  
```

- Πρόσφατες αναθεωρήσεις σε RDFS υποστηρίζουν την δακτυλογράφηση δεδομένων
- Η RDFS δεν μπορεί να δείξει την ισοδυναμία έτσι δεν υποστηρίζει λειτουργικότητα μεταξύ των διαφορετικών λεξιλογίων ή άλλων εκδοχών λεξιλογίων

Επίπεδο 5: Οντολογικές Γλώσσες

- Η OWL και η DAML+OIL είναι πιο σύνθετες γλώσσες για να περιγράψουν RDF λεξιλόγια

- Μπορούν να περιγράψουν τους περιορισμούς μεγέθους (cardinality) αριθμού στοιχείων συνόλου στις ιδιότητες π.χ. ότι ένα πρόσωπο έχει ακριβώς ένα βιολογικό πατέρα
- Μεταβατικότητα για παράδειγμα

Εάν ο Α έχει πρόγονο τον Β και ο Β έχει πρόγονο τον Γ τότε ο Α έχει πρόγονο τον Γ

- Ένα δεδομένο κτήμα (property) είναι ένα μοναδικό προσδιοριστικό
- Δυο διαφορετικές κατηγορίες (classes) αντιπροσωπεύουν την ίδια έννοια
- Δυο διαφορετικές περιπτώσεις αντιπροσωπεύουν το ίδιο άτομο
- Νέες κατηγορίες που δημιουργούνται από την άποψη δυο άλλων κατηγοριών π.χ. ενώσεις και διατομές
- Δυο κατηγορίες χωρίζουν για παράδειγμα κανένας πόρος δεν είναι περίπτωση και των δυο κατηγοριών

Ωστόσο υπάρχουν κι άλλες Οντολογικές Γλώσσες όπως η SHOE, η OIL, η Ontobroker. Ο συνδυασμός των καλύτερων χαρακτηριστικών γνωρισμάτων των RDFS/OIL και SHOE έχουν οδηγήσει στη δημιουργία της DAML+OIL γλώσσας που επίσης βασίζεται στην RDF. Η συγκεκριμένη DAML+OIL γλώσσα είναι η βάση της γλώσσας OWL που αποτελεί παρούσα σύσταση της W3C.

Υπέρνω του οντολογικού επιπέδου υπάρχει ένα ακόμη επίπεδο, το επίπεδο πλαίσιο λογικής (logicframework) αυτό μπορεί ενεργοποιήσει τις σημασιολογικές ιστοπαγείς μηχανές αναζήτησης να συγκεντρώσει συμπερασματικά νέα γνώση από την ήδη υπάρχουσα. Αυτό θα επιτρέψει στα λογισμικά προγράμματα των διαμεσολαβητών να φέρουν απόδειξη (proof) με το ακολουθούν τους σημασιολογικούς συνδέσμους και επιτηρώντας τα λογικά βήματα της ακολουθίας. Κατά συνέπεια μια εκτίμηση των πηγών και των διαδικασιών θα είναι απευθείας (online) προσβάσιμη για την πιστοποίηση των πόρων. Το ερώτημα κατά πόσο εμπιστευόμαστε (trust) αυτή την εκτίμηση παραμένει ανοιχτό και αναπάντητο.

Ψηφιακή υπογραφή (digitalsignature)

Ένα άλλο σημείο ζωτικής σημασίας θα είναι οι ψηφιακές υπογραφές: πρόκειται για κρυπτογραφημένες ομάδες δεδομένων που μπορούν να χρησιμοποιούν οι Η/Υ και οι διαμεσολαβητές ώστε να επαληθεύουν το ότι η επισυναπτόμενη πληροφορία παρέχεται από μια εμπιστευμένη πηγή. Η ψηφιακή υπογραφή εμφανίζεται εδώ για να διευκολύνει την κατάσταση με τον προσδιορισμό του συντάκτη των τεκμηρίων και των πηγών. Όταν όλα αυτά που αναφέρθηκαν γίνουν πραγματικότητα τότε οι χρήστες που χρησιμοποιούν τον Ιστό θα μπορούν και να τον εμπιστευτούν στο επίπεδο των πληροφοριών που βλέπουν, τα πειστήρια γεγονότα που εισπράττουν, οι αξιώσεις που προβάλουν οδηγούν στην ως εκ τούτου δημιουργία ενός Ιστού εμπιστοσύνης (weboftrust).

4.4 Οντολογικός Μοντελισμός

Η σημασιολογική διεγχειρησιμότητα για την διασύνδεση των πληροφοριακών κέντρων διαφορετικών κομμάτων, η συγκροτημένη ανάπτυξη των συστημάτων πληροφόρησης, η διαχείριση περιβαλλόντων περιοχών υπολογισμών που βρίσκονται παντού και συστημάτων αναγνώρισης περιεχομένου, και τα ταχύρρυθμα συστήματα πληροφόρησης είναι μερικά μόνο παραδείγματα περιοχών που στηρίζονται κατά πολύ στην πρόοδο του οντολογικού μοντελισμού. Ιδιαίτερα εντατικές είναι οι τρέχουσες εξελίξεις για το Σημασιολογικό Ιστό που σχεδιάζεται να παρέχει πιο πολλές αυτοματοποιημένες υπηρεσίες βασισμένες σε στοιχεία σημασιολογίας που εξελίσσονται μέσω μηχανών. Οι οντολογίες ως ενιαίες και κοινές θεωρίες χώρου αποτελούν ένα κεφάλαιο κλειδί στην πρόσβαση που βασίζεται στο περιεχόμενο, τη διεγχειρησιμότητα, και την επικοινωνία μέσα στο Δίκτυο, δίδοντάς του ένα ποιοτικά καινούργιο επίπεδο υπηρεσιών. Η σύγκλιση Ηλεκτρονικών Βιβλιοθηκών και τεχνολογιών Σημασιολογικού Ιστού αναμένονται να δώσουν περισσότερη σημασιολογία στην πρόσβαση συλλογών ετερογενών πληροφοριών. Οι οντολογίες έγιναν μια ευρύς και εντατική περιοχή έρευνας και ανάπτυξης που αξίζει ειδική ανάλυση. Στη συνέχεια γίνεται μια σύντομη παρουσίαση με μορφή έρευνας για τις τεχνολογίες των οντολογικών γλωσσών και όσων δεν αναφέρθηκαν προηγουμένως, ενώ ταυτόχρονα ομαδοποιούνται σύμφωνα με τα οντολογικά μοντέλα. Κατά κύριο λόγο αυτό το κεφάλαιο εστιάζεται σε προσεγγίσεις για οντολογικό μοντελισμό που διακρίνονται σε λεκτικά μοντέλα, μοντέλα με βάση τη λογική, μοντέλα που βασίζονται στη δομή (αντικείμενο) και στα νόθα μοντέλα. Η έρευνα παρέχει μια ανάλυση των σχετικών ευκολιών του μοντελισμού. Γίνεται επίσης σύντομη αναφορά σε σχέδια που εστιάζουν στη δομή των οντολογιών ως πρακτικές εφαρμογές. Εξαιτίας της σημασιολογικής και προγραμματικής ετερογένειας των ευκολιών του οντολογισμού μοντελισμού, η οντολογική ενοποίηση γίνεται

ένα δύσκολο θέμα που τίθεται υπό συζήτηση. Τέλος, η παρουσίαση συνεισφέρει στην ανάλυση της διαφοράς μεταξύ εννοιολογικών στοιχείων μοντελισμού και της οντολογικής μηχανικής.

Κατηγορίες μοντέλων για οντολογική δομή και συλλογισμό

Ο οντολογικός μοντελισμός στις Τεχνολογίες της Πληροφορίας έχει υποστεί μια σημαντική εξέλιξη. Τα μοντέλα και οι γλώσσες που χρησιμοποιούνται για την οντολογική δομή και αιτιολόγηση μπορούν να κατηγοριοποιηθούν ως εξής:

Λεκτικά μοντέλα (verbal models)

Ανεπίσημα γλωσσολογικά μοντέλα χρησιμοποιούνται συχνά για τον προσδιορισμό των οντολογιών. Σε τέτοια μοντέλα οι οντολογικές έννοιες ορίζονται από λεκτικούς ορισμούς όπως σε ένα επεξηγηματικό λεξικό. Κάποια είδη βασικών σχέσεων μπορούν να εδραιωθούν ανάμεσα στις έννοιες. Ένα γλωσσάρι όρων σε κάποιον υποκειμενικό χώρο, ένας θησαυρός με τους όρους του και τις σχέσεις του που να ορίζει όρους της απλής γλώσσας μπορούν να θεωρηθούν οντολογίες. Μέθοδοι ανάκτησης πληροφοριών εφαρμόζονται για την εδραίωση σχέσεων ανάμεσα σε έννοιες που ορίζονται λεκτικά καθώς και για την έρευνα εννοιών σχετικών με κάποιο ερώτημα.

Μοντέλα που βασίζονται στη λογική (logic-based models)

Σε αντίθεση με τα λεκτικά μοντέλα, οι οντολογίες που βασίζονται στη λογική ορίζονται τυπικά και έχουν την ικανότητα επίσημου συλλογισμού. Ένα από τα πρώτα μοντέλα βασισμένα στη λογική ήταν το Ontolingua. Κατηγορηματικές εκφράσεις στην Ontolingua παρουσιάζονται στη γλώσσα KIF (Knowledge Interchange Format), που βασίζονται στη λογική της πρώτης σειράς. Η KIF είναι πολύ εκφραστική και εξαιτίας αυτού δε βολεύει στην αυτόματη εξαγωγή συμπερασμάτων. Η εξέλιξη οντολογιών σ' αυτό το μοντέλο ρυθμίζεται από ειδική τεχνική για την προσθήκη ξεκάθαρων προσδιορισμών. Η Ontolingua έχει ορισθεί ως η γλώσσα που θα λειτουργεί ως μεσάζοντας στις ανταλλαγές ετερογενών οντολογιών. Το οντολογικό μοντέλο OKBC (Open Knowledge Base Connectivity) χρησιμοποιεί ένα μοντέλο πλαισίου ως βάση. Η Βάση Συνδετικότητας Ανοιχτής Γνώσης (Open Knowledge Base Connectivity OKBC) API θεωρείται ως ο παράγοντας κλειδί που κατέστησε ικανή την αρχιτεκτονική ενός θησαυρού διασκορπισμένων οντολογιών. Το μοντέλο OKBC λειτουργεί ως μια γλώσσα επικοινωνίας για την οντολογία η οποία παρουσιάζεται μέσω του OKBC. Η XOL έχει αναπτυχθεί με σκοπό να ικανοποιήσει την ανάγκη για μια γλώσσα με τη σημασιολογία των συστημάτων αναπαράστασης γνώσης με αντικειμενική κατεύθυνση, αλλά με τη σύνταξη της XML. Οι ορισμοί οντολογίας που η XOL έχει σχεδιαστεί να κωδικοποιήσει συμπεριλαμβάνουν και το σχέδιο πληροφοριών (μετα-στοιχεία), όπως ορισμοί τάξεων από βάσεις δεδομένων αντικειμένων καθώς και πληροφορίες χωρίς σχέδιο (γεγονότα εδάφους), όπως ορισμοί αντικειμένων από βάσεις δεδομένων αντικειμένων. Η XOL μοιάζει με παλιότερες γλώσσες ανταλλαγής οντολογιών. Συγκεκριμένα, χρησιμοποιεί το μοντέλο της OKBC. Άλλη κατηγορία σύγχρονων επίσημων οντολογικών μοντέλων βασίζεται σε διάφορα είδη περιγραφικής λογικής. Αυτή είναι μια οικογένεια φορμαλισμών αναπαράστασης γνώσεων που βασίζονται στη λογική κατάλληλων για την αναπαράσταση και για τον συλλογισμό σχετικά με τις τερμινολογικές γνώσεις και οντολογίες. Χαρακτηρίζονται κυρίως από ένα σύνολο κατασκευαστών που επιτρέπουν τη δημιουργία σύνθετων εννοιών και ρόλων από ατομικές έννοιες και ρόλους. Αυτά τα μοντέλα αρχίζουν να βολεύουν σε εργασίες όπως ο έλεγχος της ένταξης ανάμεσα σε έννοιες και στη δυνατότητα ύπαρξης ικανοποιητικών ορισμών εννοιών καθώς και ομάδων διασυνδεδεμένων εννοιών. Παραδείγματα συστημάτων συλλογισμού που χρησιμοποιούνται συχνά για τον οντολογικό μοντελισμό είναι τα KL-ONE, LOOM, CLASSIC, FaCT. Οι τυπικές υπηρεσίες συλλογισμού που παρέχονται είναι η κατηγοριοποίηση, η συνέπεια και ο στιγμιαίος έλεγχος. Η έννοια της ερμηνείας εισάγεται για να θεωρηθούν οι έννοιες περιγραφικής λογικής ως ομάδες αντικειμένων. Ο έλεγχος της δυνατότητας ικανοποίησης αποτελείται από μια δήλωση απόδειξης ότι υπάρχει τουλάχιστον μία ερμηνεία σύμφωνα με την οποία η έννοια ανταποκρίνεται σε μια γεμάτη ομάδα. Η ένταξη ανάμεσα στις έννοιες σημαίνει ότι για κάθε ερμηνεία μία ομάδα που ερμηνεύει μία έννοια είναι υποομάδα μιας ομάδας που ερμηνεύει την άλλη. Η απόδειξη της ένταξης περιορίζεται σ' αυτήν της δυνατότητας ικανοποίησης. Η OIL είναι η πρώτη γλώσσα οντολογικής έκφρασης που είναι κατάλληλα εδραιωμένη σύμφωνα με τα μέτρα του W3C. Πρόκειται για εξέλιξη ήδη υπάρχοντων προτάσεων όπως οι OKBC, XOL, RDF. Είναι επίσης η πρώτη γλώσσα που βασίζεται στην έκφραση μέσω δικτύου με σκοπό οντολογικούς ορισμούς με τυπική σημασιολογία και υπηρεσίες συλλογισμού παρεχόμενες από μια περιγραφική λογική. Η πιο σημαντική ανάπτυξη των οντολογικών μοντέλων που βασίζονται στη λογική λαμβάνει χώρα στο πλαίσιο της ανάπτυξης του Σημασιολογικού Ιστού.

Η DAML +OIL είναι ένα πολύ γνωστό μοντέλο το οποίο χρησιμοποιεί την περιγραφική λογική SHIQ που έχει λογικά καλή εκφραστική δύναμη. Είναι το αποτέλεσμα της συγχώνευσης των γλωσσών DAML και OIL. Βασίζεται επίσης σε παλιότερα μέτρα W3C όπως την RDF και την RDF Σχήμα και τα επεκτείνει με πιο πλούσια αρχέτυπα μοντελισμού. Αυτό το μοντέλο σχεδιάζεται ως η βάση για την Οντολογική Γλώσσα Δικτύου (OWL Web Ontology Language). Η OWL επεκτείνει την RDF και τη RDF Σχήμα με την παροχή επιπλέον λεξιλογίου μαζί με μια τυπική σημασιολογία. Έχει τα θεμέλια της στην DAML +OIL και παρέχει επίσης υποστήριξη στη δημιουργία προγραμμάτων οντολογίας. Η OWL έχει τρεις διαλέκτους: την OWLLite, την OWL DL (DescriptionLogics) και την OWL Full. Αυτά τα είδη οντολογικής έκφρασης διαφέρουν ως προς την περιπλοκότητά τους και έχουν διαφορετικές εφαρμογές ανάλογα με την απαιτούμενη απλότητα εξαγωγής συμπεράσματος ή την επισημότητα των περιγραφών. Μια σύγκριση των γλωσσών οντολογικής έκφρασης που βασίζονται στη λογική δίδεται στον πίνακα που ακολουθεί.

Δομικά (αντικειμενικά) μοντέλα (structural-object models)

Διάφορες προσεγγίσεις είναι γνωστές για την εφαρμογή μοντέλων δομικών (αντικειμένων) στοιχείων για να ορίσουν τις οντολογίες. Μια προσέγγιση σχετικά με ένα εξελιγμένο οντολογικό μοντέλο κατασκευής σε ένα περιβάλλον μεσολάβησης προορισμένο για συγχώνευση ετερογενών πηγών πληροφοριών σε διάφορους τομείς έχει παρουσιαστεί από τον (Kalin). Μία μεσολαβητική οντολογική γλώσσα (MOL) μπορεί να εξαρτάται από μια υποκειμενική περιοχή και να ορίζεται στη φάση της μεσολαβητικής σταθεροποίησης. Από την άλλη, για διαφορετικές πηγές πληροφοριών μπορεί να χρησιμοποιηθούν διαφορετικά οντολογικά μοντέλα (γλώσσες) για να ορίσουν τις οντολογίες τους. Χρειάζεται ανατρεπόμενος σχεδιασμός της πηγής οντολογικών μοντέλων σε MOL για την καταγραφή πηγών πληροφόρησης στον μεσολαβητή. Μια προσέγγιση για τέτοιου είδους ανατρεπόμενου σχεδιασμού έχει αναπτυχθεί για μια τάξη των πηγών πληροφοριών του Δικτύου θεωρώντας ότι τέτοιες πηγές εφαρμόζουν το οντολογικό μοντέλο DAML+OIL. Μια υποομάδα ενός μοντέλου που είναι νόθο, προσανατολίζεται προς το αντικείμενο και έχει ημι-δομημένα κανονικά στοιχεία μεσολάβησης χρησιμοποιείται ως βάση για την MOL. Έχει παρουσιαστεί μια κατασκευή ενός ανατρεπόμενου σχεδιασμού της DAML+OIL σε προέκταση της βάσης της MOL. Αυτός ο σχεδιασμός είναι μια απαραίτητη προϋπόθεση για να μπορέσουν οι πηγές πληροφόρησης να ενταχθούν στο πλαίσιο και να καταγραφούν στον μεσολαβητή. Ο σχεδιασμός δείχνει πώς μπορεί να κατασκευαστεί η εξελιγμένη MOL. Η προτεινόμενη προσέγγιση προσανατολίζεται σε ψηφιακές βιβλιοθήκες όπου η ανάκτηση εστιάζεται στο περιεχόμενο των πληροφοριών παρά στις οντότητες των πληροφοριών. Άλλο παράδειγμα δομικού μοντέλου είναι η πρωτοβουλία της OMG να εκφράσει Σημασιολογικές Οντολογίες Ιστού σε UML. Η Αίτηση για Προτάσεις της OMG (ontol) ζητά κανονιστικούς προσδιορισμούς για:

Η ODM αναμένεται να είναι μια γλώσσα τόσο πλούσια όσο και η UML και μολονότι υπάρχει μια σημαντική σημασιολογική μερική επικάλυψη ανάμεσα στις γλώσσες καμιά δεν είναι υποσύνολο της άλλης. Η RFP ζητά έναν περιορισμένο σχεδιασμό δύο κατευθύνσεων ανάμεσα σε μια περίπτωση του μεταμοντέλου ODM και μιας περίπτωσης του μεταμοντέλου UML2. 'Περιορισμένος' σημαίνει ότι οι επαναλαμβανόμενες εφαρμογές δεν οδηγούν σε μοντέλα συνεχώς περισσότερων χρηστών. Αυτό γίνεται ώστε η ανάπτυξη των οντολογιών να μπορεί να υποστηριχθεί και να ανταλλαχθεί ανάμεσα στους χώρους της UML και της οντολογικής γλώσσας. Προτάσεις αναμένονται να προσφέρουν ένα γλωσσικό σχεδιασμό για την DAML +OIL, ή για άλλες οντολογικές γλώσσες, με τα ίδια χαρακτηριστικά όπως αυτά που περιγράφονται για την OWL. Προτάσεις αναμένονται να ξεκινήσουν από τον Σεπτέμβριο του 2015.

Νόθα μοντέλα

Για τον εμπλουτισμό μιας εκφραστικής δύναμης του οντολογικού μοντέλου μπορεί να υπάρξει η ανάγκη να χρησιμοποιηθούν ευκολίες από τα λεκτικά μοντέλα, από τα μοντέλα που βασίζονται στη λογική ή τα δομικά μοντέλα στην ίδια και την αυτήν οντολογία. Για παράδειγμα, οντολογίες στο πρόγραμμα SYNTHESIS χρησιμοποιούν ένα νόθο μοντέλο. Το πρόγραμμα SYNTHESIS αποσκοπεί στην ανάπτυξη των συστημάτων πληροφόρησης που στηρίζεται σε επί μέρους συστατικά. Για τους συστατικούς προσδιορισμούς χρησιμοποιεί ένα ενοποιημένο μοντέλο που μπορεί να αναπτυχθεί και που προσανατολίζεται προς το αντικείμενο. Οι οντολογίες επίσης χρησιμοποιούν αυτό το μοντέλο. Το οντολογικό μοντέλο έχει λεκτικές ευκολίες που επιτρέπουν τον ορισμό οντολογικών ή γλωσσικών εννοιών, κατηγορίες ταξινόμησης, πολύγλωσσους ορισμούς. Την ίδια στιγμή οι οντολογικές έννοιες ορίζονται ως τύποι αντικειμένου. Η επιλογή ειδών οντολογικού μοντελισμού επιτρέπει τη χρήση διαφορετικών σεναρίων συλλογισμού, ακολουθώντας την αρχή ότι η περιπλοκότητα δεν είναι κτήμα ενός μοντέλου αλλά κτήμα μιας άσκησης. Οι οντολογικές έννοιες στο SYNTHESIS μπορούν να συγκριθούν σε δύο επίπεδα: οι γλωσσικές μέθοδοι εγκαθιστούν αδύναμη την

ομοιότητα εννοιών, οι επίσημες μέθοδοι θεωρούν τις έννοιες ως τύπους, ελέγχουν την συνέπεια και εγκαθιστούν τη στενή τους ομοιότητα. Το πρόγραμμα SymOntos (formi) [και το διαδοχικό του SymOntosX (miss)] ασχολείται με θέματα επιχειρησιακών οντολογικών συστημάτων διοίκησης. Βασίζεται στο μοντέλο OPAL (Αντικείμενο, Διαδικασία, Γλώσσα Ηθοποιών-Object, Process, ActorLanguage) το οποίο επιτρέπει τον ορισμό εννοιών σύμφωνα με την προσέγγιση του πλαισίου. Ο ηθοποιός, το αντικείμενο και η διαδικασία είναι κατηγορίες εννοιών του μοντέλου για τον ορισμό περιπτώσεων ενεργητικών και παθητικών εννοιών. Το Μοντέλο OPAL συνδυάζει χαρακτηριστικά του λεκτικού και του δομικού οντολογικού μοντέλου.

Δυνάμεις και αδυναμίες σύγχρονων οντολογικών γλωσσών

Το τμήμα αυτό ολοκληρώνει την ανάλυση των γλωσσών του Σημασιολογικού Ιστού. Συνοψίζουμε τις δυνάμεις και αδυναμίες των γλωσσών που παρουσιάστηκαν και υποδεικνύουμε στις υποκείμενες αρχές σχεδιασμού των Γλωσσών του Σημασιολογικού Ιστού.

RDF(S). Η εκφραστική της δύναμη είναι αρκετά περιορισμένη και οι δυνατότητες αιτιολόγησης δεν είναι και οι δυνατότερες ανάμεσα στις διαφορετικές γλώσσες, παρέχοντας ένα περιορισμένο μηχανισμό αιτιολόγησης που είναι κατάλληλος μόνο για τον έλεγχο περιορισμού. Λαμβάνεται υπόψη με τμηματικές διευκολύνσεις διαλειτουργικότητας, όπου μπορούν να οριστούν κανόνες αποτύπωσης. Έχει σύνταξη που βασίζεται στην XML. Υπάρχουν πολλά εργαλεία και παραδείγματα που θα μπορούσαν είτε να χρησιμοποιηθούν είτε να ακολουθηθούν για την εκμάθηση της γλώσσας, που την κάνει ευρύτατα διαδεδομένη. Αναφορικά με τη διεθνοποίηση, υποστηρίζει διαφορετικές φυσικές γλώσσες και είναι συμβατή με την HTML, της οποίας θεωρείται πως είναι ένα υπερσύνολο. Η κοινότητα αναπτύσσει και βελτιώνει ενεργά τη γλώσσα αυτή.

OIL. Η OIL λαμβάνεται υπόψη με μια πλουσιότερη εκφραστική δύναμη από την RDFS για τον ορισμό των οντολογιών. Οι δυνατότητες αιτιολόγησης της OIL παρέχουν έλεγχο ατομικής συνοχής και επιτρέπει την αλληλοσύνδεση ενδό-οντολογικών σχέσεων και τον έλεγχο για σχέσεις που υποδηλώνονται. Σχετικά με τη διαλειτουργικότητα, η OIL επιτρέπει μερικό ορισμό των κανόνων αποτύπωσης. Ενσωματώνει διευκολύνσεις διεθνοποίησης που υποστηρίζουν διαφορετικές φυσικές γλώσσες. Η OIL είναι εύκολη στη χρήση· υπάρχει πολλή τεκμηρίωση και παραδείγματα γι' αυτήν, καθώς και εργαλεία και υποστήριξη γι' αυτά. Εφόσον αποτελεί έγνοια η συμβατότητα, ο σχεδιασμός της βασίζεται στα πρότυπα της Περιγραφικής Λογικής, της Λογικής F και του Web (RDFs και XML). Η κεντρική δομή της OIL συμπίπτει με την παράσταση δομής της RDF, εκτός από τα χαρακτηριστικά reification της RDFS. Η OIL δεν είναι πλέον υπό ανάπτυξη.

DAML+OIL. Οι δυνατότητες αιτιολόγησης είναι χρήσιμες για την κοινή χρήση της οντολογίας. Σχετικά με τη διαλειτουργικότητα, επιτρέπει το μερικό ορισμό των κανόνων αποτύπωσης. Η αιτιολογία στην DAML+OIL είναι ειδικά ταιριαστή για τη συντήρηση αιτιολογίας υποστήριξης σχεδιασμού της DL και την ανάπτυξη των οντολογιών. Η εκφραστική της δύναμη είναι πολύ πλουσιότερη απ' αυτήν των προκατόχων της· υποστηρίζει διαφορετικές φυσικές γλώσσες· είναι αρκετά εύκολη στη χρήση, και όσον αφορά τη συμβατότητα, είναι σημαντικό να σημειώσουμε πως υποστηρίζει την πλήρη έκταση των τύπων δεδομένων της παράστασης δομής της XML, καθώς βασίζεται στα υπάρχοντα πρότυπα του Web, XML και RDF. Τέλος, λαμβάνεται υπόψη με μερικές διευκολύνσεις διαλειτουργικότητας, όπου κανόνες αποτύπωσης μπορούν να οριστούν.

OWL. Οι λειτουργικότητες αιτιολόγησης της OWL θα μπορούσαν να χρησιμοποιηθούν, όπως στην περίπτωση της DAML+OIL, για να παράσχουν δυνατότητες κοινής χρήσης. Αντίθετα από τις γλώσσες που παρουσιάστηκαν μέχρι τώρα, η OWL παρέχει ενσωματωμένες λειτουργικότητες έκδοσης. Ο μηχανισμός αιτιολόγησης της είναι ο ίδιος με της DAML+OIL, και βασίζεται στη θεώρηση του ανοιχτού κόσμου (OWA). Είναι εξοπλισμένη μια πλούσια εκφραστική δύναμη και λαμβάνεται υπόψη με μια διαστρωμένη αρχιτεκτονική για προσαρμοστικότητα. Η ευκολία της χρήσης είναι ένα κοινό χαρακτηριστικό όλων των γλωσσών που παρουσιάστηκαν μέχρι τώρα. Υποστηρίζει διαφορετικές φυσικές γλώσσες, όπως οι άλλες γλώσσες και, σχετικά με τη συμβατότητα, θα πρέπει να σκιαγραφηθεί πως η OWL βασίζεται στην OIL και DAML+OIL, που την κάνει συμβατή και με τις δύο. Η OWL χωρίζεται σε τρεις υπό-γλώσσες, κατάλληλες για διαφορετικούς σκοπούς (OWL-Lite, OWL-DL, OWL-Full). Το κύριο μειονέκτημα της OWL είναι το γεγονός πως είναι ακόμη υπό ανάπτυξη. Ως τα υποκείμενα κριτήρια σχεδιασμού των γλωσσών του Σημασιολογικού Ιστού, έχουν υπολογιστεί οι ακόλουθες διαστάσεις:

Συμβατότητα. Όλες οι γλώσσες βασίζονται στη σύνταξη της XML ή της RDF. Αυτό επιτρέπει συμβατότητα με τα πρότυπα του web. Τα υπάρχοντα εργαλεία που υποστηρίζουν την XML ή την RDF

μπορούν εύκολα να ξαναχρησιμοποιηθούν για την OIL, τη DAML+OIL και την OWL. Έτσι, γίνεται δυνατός ο πύργος διαστρωμένης γλώσσας όπως επιδιώκεται από το W3C.

Σημασιολογία. Η προσθήκη της σημασιολογίας στις υπάρχουσες πληροφορίες είναι ο κύριος στόχος των γλωσσών του σημασιολογικού ιστού. Αλλά το πώς θα προστεθεί και αναπαρασταθεί σαφής, συγκεκριμένη και κατανοητή από τη μηχανή σημασιολογία δεν είναι ένα ασήμαντο έργο. Είναι μια σαφής κατεύθυνση για το σχεδιασμό των γλωσσών. Διαφορετική σημασιολογία χρειάζεται διαφορετική εκφραστική δύναμη· επομένως, η διαστρωματωμένη δομή είναι ουσιώδης.

Διαστρωματωμένη δομή. Δεν θα υπάρχει μία μόνο γλώσσα που πληροί όλες τις ανάγκες των διαφόρων χρηστών του web. Σε ένα διαστρωματωμένο σχεδιασμό, ένας απλός πυρήνας μπορεί να στεγάσει απλές ταξινομίες και σχέσεις, ενώ επιπρόσθετα στρώματα εκφραστικότητας, λειτουργικότητας και πολυπλοκότητας μπορούν να προστεθούν για ομάδες που απαιτούν περισσότερη εκφραστική δύναμη. Επίσης, το βάρος της προσαρμοστικότητας και της συντήρησης μπορεί να κατανεμηθεί στα διαφορετικά στρώματα της γλώσσας, και επομένως να μετριαστεί αναλόγως. Οι γλώσσες που παρουσιάζουν μια τέτοια σαφή διαστρωματωμένη δομή είναι η OIL και η OWL.

Βιβλιογραφία

- Arroyo, S., Lara, R., Ding, Y., Stollberg, M., & Fensel, D. (2004). Semantic web languages. strengths and weakness. Στο *International Conference in Applied Computing*. Citeseer. Ανακτήθηκε από <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.88.5910&rep=rep1&type=pdf>
- Bedford, D. A. (2014). Understanding and managing taxonomies as economic goods and services. *Bulletin of the American Society for Information Science and Technology*, 40(4), 15–22.
- Berners-Lee, T. (1998a). What the Semantic Web can represent. Ανακτήθηκε από <http://www.citeulike.org/group/17638/article/11988245>
- Berners-Lee, T. (1998b). *Why RDF model is different from the XML model*.
- Berners-Lee, T., & Hendler, J. (2001). *Scientific Publishing on the “Semantic Web”*. Nature Webdebates.
- Berners-Lee, T., Hendler, J., Lassila, O., & others. (2001). The semantic web. *Scientific american*, 284(5), 28–37.
- Berners-Lee, T., & others. (1998). *Semantic web road map*. Ανακτήθηκε από http://www.researchgate.net/profile/Tim_Berners-Lee/publication/200034151_Semantic_Web_Road_Map/links/00b495273611f87093000000.pdf
- Butler, M. H. (2003). Is the semantic web hype. *Retrieved May, 31, 2006*.
- Conklin, J. (1987). Hypertext: A survey and introduction. *IEEE Computer*, 20(9), 17–41.
- Corcho, O., & Gómez-Pérez, A. (2000). A roadmap to ontology specification languages. Στο *Knowledge Engineering and Knowledge Management Methods, Models, and Tools* (σσ 80–96). Springer. Ανακτήθηκε από http://link.springer.com/chapter/10.1007/3-540-39967-4_7
- Daconta, M. C., Obrst, L. J., & Smith, K. T. (2003). *The Semantic Web: a guide to the future of XML, Web services, and knowledge management*. John Wiley & Sons. Ανακτήθηκε από <https://www.google.com/books>
- Davies, J., Fensel, D., & Van Harmelen, F. (2003). *Towards the semantic web: ontology-driven knowledge management*. John Wiley & Sons. Ανακτήθηκε από <https://www.google.com/books>
- Dobson, S. A., & Burrill, V. A. (1995). Lightweight databases. *Computer Networks and ISDN Systems*, 27(6), 1009–1015.
- Gómez-Pérez, A., & Corcho, O. (2002). Ontology languages for the semantic web. *Intelligent Systems, IEEE*, 17(1), 54–60.
- Hearst, M., Elliott, A., English, J., Sinha, R., Swearingen, K., & Yee, K.-P. (2002). Finding the flow in web site search. *Communications of the ACM*, 45(9), 42–49.
- Hossain, M. A., & El Saddik, A. (2004). Creating Knowledge and Semantic Web Applications on the Web. Στο *Proc. 2nd Intl. Symposium on Innovation in Information and Communication Technology (ISIICT)*, Amman, Jordan, April.
- Hyvönen, E., Mäkelä, E., Salminen, M., Valo, A., Viljanen, K., Saarela, S., ... Kettula, S. (2005). MuseumFinland—Finnish museums on the semantic web. *Web Semantics: Science, Services and Agents on the World Wide Web*, 3(2), 224–241.
- Hyvönen, E., Salminen, M., Kettula, S., & Junnila, M. (2004a). A content creation process for the semantic web. Στο *Proceeding of OntoLex* (σσ 1–7). Ανακτήθηκε από https://128.214.112.31/files/admin/sab/SAB04/material/SeCo/13_contentCreation.pdf
- Hyvönen, E., Salminen, M., Kettula, S., & Junnila, M. (2004b). A content creation process for the semantic web. Στο *Proceeding of OntoLex* (σσ 1–7). Ανακτήθηκε από https://128.214.112.31/files/admin/sab/SAB04/material/SeCo/13_contentCreation.pdf

- Hyvönen, E., Valo, A., Viljanen, K., & Holi, M. (2003). Publishing semantic web content as semantically linked HTML pages. Στο *Proceedings of XML Finland 2003, Kuopio, Finland*. Ανακτήθηκε από <https://www.seco.tkk.fi/publications/2003/hyvonen-valo-et-al-publishing-semantic-web-2003.pdf>
- Kalinichenko, L., Missikoff, M., Schiappelli, F., & Skvortsov, N. (2003). Ontological modeling. Στο *Proc. of the 5th Russian Conference on Digital Libraries RCDL2003, St.-Petersburg, Russia* (σσ 7–13). Citeseer. Ανακτήθηκε από <http://citeseerx.ist.psu.edu>
- Kanal, L. N., Kumar, V., Kitano, H., & Suttner, C. B. (2014). Providing computationally effective knowledge representation via massive parallelism. *Parallel Processing for Artificial Intelligence* 1, 115.
- Lee, B. T., & Fischetti, M. (1999). *Web Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by Its Inventor*. Ανακτήθηκε από <http://www.citeulike.org/group>
- Luke, S., Spector, L., Rager, D., & Hendler, J. (1997). Ontology-based web agents. Στο *Proceedings of the first international conference on Autonomous agents* (σσ 59–66). ACM. Ανακτήθηκε από <http://dl.acm.org/citation.cfm?id=267668>
- Lu, S., Dong, M., & Fotouhi, F. (2002). The Semantic Web: opportunities and challenges for next-generation Web applications. Στο *International Journal of Information Research*. Citeseer. Ανακτήθηκε από <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.20.2117>
- Manola, F., Miller, E., & McBride, B. (2004). *Rdf primer w3c recommendation 10 february 2004*. Ανακτήθηκε από http://publishing-portal.net/download/dot/ontopage/sub_MaMi2004a.pdf
- Mizoguchi, R., & Ikeda, M. (1998). Towards ontology engineering. *Journal-Japanese Society for Artificial Intelligence*, 13, 9–10.
- Pollitt, A. S. (1998). *The key role of classification and indexing in view-based searching*. Citeseer. Ανακτήθηκε από <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.109.626&rep=rep1&type=pdf>
- Su, X., & Ilebrikke, L. (2002). A comparative study of ontology languages and tools. Στο *Advanced Information Systems Engineering* (σσ 761–765). Springer. Ανακτήθηκε από http://link.springer.com/chapter/10.1007/3-540-47961-9_62

5. Στατιστική Επεξεργασία της Πληροφορίας

Περίληψη

Στο κεφάλαιο αυτό αναπτύσσεται η μεθοδολογία συλλογής και επεξεργασίας δεδομένων από άμεση παρατήρηση μέσω πειραμάτων και δειγματοληψίας, ώστε να γίνει μια εμπέδωση γνώσεων στο τμήμα της προ-επεξεργασίας και συλλογής των δεδομένων. Στην συνέχεια θα γίνει αναφορά και ανάλυση στις μεθοδολογίες εξαγωγής χαρακτηριστικών από δεδομένα, χρησιμοποιώντας τις προαναφερόμενες διαδικασίες επεξεργασίας των δεδομένων. Τέλος, θα παρουσιασθούν οι θεμελιώδεις διαδικασίες ταξινόμησης και κατηγοριοποίησης βασισμένες σε στατιστικά σημασιολογικά δίκτυα όπως είναι τα νευρωνικά δίκτυα και τα δένδρα αποφάσεων.

Προαπαιτούμενες γνώσεις

Στατιστική Ανάλυση, Νευρωνικά Δίκτυα

5.1 Στοιχεία Στατιστικής

Στο προηγούμενο κεφάλαιο εξετάστηκε πως τα δεδομένα μπορούν να περιγραφούν με συντακτική διαδικασία ώστε να δημιουργηθούν συνθήκες κυρίως μέσω των οντολογικών προσεγγίσεων ώστε αυτά να διασυνδεθούν σημασιολογικά. Η διαδικασία αυτή προς το παρόν δεν αποπνέει αισιοδοξία, γιατί αυτή απαιτεί υπερβολικά μεγάλη μη-αυτοματοποιημένη επίβλεψη και παρέμβαση. Η λύση μίας περισσότερο αυτοματοποιημένης διαδικασίας μπορεί να προέλθει μέσω μιας στατιστικής επεξεργασίας των δεδομένων. Η προσέγγιση αυτή μπορεί να επιτευχθεί αν προσδιορισθεί με μια πιθανοτική μεθοδολογία ότι η ιδιότητα αυτή μπορεί να ποσοτικοποιηθεί και να είναι ικανή και αναγκαία ώστε τα εξεταζόμενα δεδομένα να διαχωριστούν σε ομάδες. Η διαδικασία αυτή επιτυγχάνεται μέσω της στατιστικής η οποία έχει ως αντικείμενο τις εφαρμογές σε πραγματικά προβλήματα κυρίως προβλέψεων οι οποίες απαιτούν εξαγωγή συμπερασμάτων. Η στατιστική διαδικασία στηρίζεται στις παρατηρήσεις και δανείζεται εργαλεία από τους επιστημονικό κλάδο κυρίως της θεωρίας των πιθανοτήτων τυχαίων μεταβλητών.

Η συνήθης πρακτική όπου ακολουθείται στις παρατηρήσεις είναι η προσέγγιση της πιθανότητας εμφάνισης της τιμής της ιδιότητας του δείγματος. Η μεθοδολογία ποσοτικοποίησης θεωρείται ιδιαίτερως σημαντική και εξαρτάται από το είδος και τη φύση των δεδομένων. Στο έκτο (6) κεφαλαίο αναπτύσσονται παραδείγματα εφαρμογής αυτής της ποσοτικοποίησης.

Τα δεδομένα αποτελούν ένα στατιστικό όρος για τις πληροφορίες που έχουν σχέση με την υπό εξέταση μεταβλητή. Είναι δηλαδή αριθμοί και γεγονότα, που συλλέγονται, παρουσιάζονται, αναλύονται και ερμηνεύονται. Τα μη κατεργασμένα δεδομένα συλλέγονται στην αρχική τους μορφή ως πρώτο βήμα. Είναι δυνατόν να υποστηριχθεί ότι τα δεδομένα αποτελούν την πρωτόλεια πληροφορία για τη Στατιστική. Διακρίνονται τέσσερις (4) κλίμακες μετρήσεων, οι οποίες προσδιορίζουν το σύνολο των πληροφοριών που περιλαμβάνονται εντός των δεδομένων:

α) *Ονομαστική κλίμακα*: δεδομένα αυτής της κλίμακας συνιστούν έναν αριθμό παρατηρήσεων, οι οποίες έχουν σχέση με ξεχωριστές κατηγορίες οι οποίες είναι

- μη ποσοτικοποιημένες, όπως επάγγελμα, οικογενειακή κατάσταση, φύλο, κ.λπ.
- ποσοτικοποιημένες, όπως ο αριθμός ISSN ενός περιοδικού δημοσιεύματος κ.λπ.). Η ονομαστική κλίμακα επεξεργάζεται αριθμούς που ταυτοποιούν κατηγορίες (ταυτοποίηση). Για παράδειγμα, αν διαθέταμε (2) δύο κατηγορίες, τότε θα τις δώναμε τιμές από το 1 έως και 2. Οι αριθμητικές τιμές της ονομαστικής κλίμακας μέτρησης (1 και 2) δεν δείχνουν κάποια ποσότητα ή σύνολο, αλλά δείχνουν τα ονόματα των κατηγοριών.

β) *Τακτική κλίμακα*: αποτελούν δεδομένα που απορρέουν από βαθμολογίες κατηγοριών παρατηρήσεων, τα οποία κατατάσσονται ή διατάσσονται όπως έχουν οριστεί από μια επιβλέπουσα αρχή. Παραδείγματα τακτικής κλίμακας αποτελεί η κατάταξη των πόλεων σύμφωνα με τον αριθμό των υπαρχόντων μουσείων τους ή με το επίπεδο της εξυπηρέτησης που προσφέρεται στις συγκοινωνίες τους (άριστο, καλό, μέτριο, ασθενές, κλπ).

γ) *Διαστημική κλίμακα*: Σε αυτή την περίπτωση τα δεδομένα αποτελούνται από τις παρατηρήσεις, οι οποίες μετρούνται: σε κλίμακα οποία είναι δομημένη με τέτοιον τρόπο ώστε ίσες αποστάσεις μεταξύ των αριθμών να αντιστοιχούν σε ίσες αποστάσεις μεταξύ των τιμών του χαρακτηριστικού που μετράται, ενώ το 0 επιλέγεται ως σταθερό μέτρο σύγκρισης. Για παράδειγμα, βαθμοί θερμοκρασίας Κελσίου: ή η βαθμολογία του τεστ της

γνώσης της αγγλικής γλώσσας FirstCertificate (450, 500, 700 κ.λπ.). Τα προαναφερόμενα δεδομένα λαμβάνουν πάντοτε αριθμητικές τιμές. Χαρακτηριστικό παράδειγμα κλιμάκων μέτρησης διαστημάτων αποτελούν οι κλίμακες θερμοκρασίας Φαρενάιτ και Κελσίου. Παρότι για την κλίμακα Φαρενάιτ οι 32 βαθμοί ορίζονται ως το σημείο πήξης του νερού ενώ για την κλίμακα Κελσίου το αντίστοιχο σημείο ορίζεται το 0, και για τις δύο κλίμακες θερμοκρασίας μπορούμε να αναφέρουμε ότι οι 25 βαθμοί είναι κατά 5 βαθμούς θερμότεροι από τους 20 και οι 35 βαθμοί είναι κατά 10 βαθμούς θερμότεροι από τους 25. Όμως, το σύνολο της θερμοκρασίας που συναπτά μια μεταβολή σε καταγεγραμμένη θερμοκρασία είναι διαφορετικό και για τις δύο κλίμακες.

δ) *Αναλογική κλίμακα*: Εδώ διακρίνουμε παρατηρήσεις, οι οποίες καθορίζονται από κλίμακα, της οποίας η δομή είναι τέτοια, ώστε ίσες αποστάσεις μεταξύ αριθμών δύναται να αντιστοιχισθούν από ίσες αποστάσεις μεταξύ των τιμών της μετρήσιμης μεταβλητής, ενώ ο αριθμός (0) αντιπροσωπεύει την απουσία του χαρακτηριστικού που μελετάται. Συμπερασματικά, η τιμή μίας τυχαίας μεταβλητής μπορεί να συγκριθεί με την αναλογική κλίμακα, διότι λαμβάνει ένα πραγματικά μηδενικό σημείο. Για παράδειγμα, οι ηλικίες των ατόμων είναι δεδομένα αριθμοδεικτών και το (0) μηδέν αντιπροσωπεύει την ημερομηνία της γέννησης. Ένας επιπρόσθετος διαχωρισμός για τα δύο προαναφερόμενα είδη δεδομένων είναι τα συνεχή ή τα διακριτά. Μια μεταβλητή θεωρείται διακριτή όταν δέχεται συγκεκριμένες αριθμητικές τιμές. Ως συνεχή δεδομένα θεωρούνται οι παρατηρήσεις οι οποίες δύναται να περιλαμβάνουν οποιαδήποτε τιμή ενός πλήθους αριθμού τιμών. Μια συνεχής μεταβλητή δύναται να λαμβάνει οποιαδήποτε αριθμητική τιμή π.χ., το σωματικό βάρος ή το ύψος ενός ατόμου και εν γένει κάθε μη καθορισμένη ή αόριστη τιμή.

Τα δεδομένα συνήθως συλλέγονται στις ακόλουθες μορφές:

1. *Δεδομένα από Χρονολογικές Σειρές*: τα οποία χρησιμοποιούνται για την εξαγωγή συμπερασμάτων σχετικά με τη μεταβολή μιας μεταβλητής κατά τη διάρκεια χρονικών περιόδων. Δηλαδή, απεικονίζουν μετρήσεις σε διακριτά σημεία μέσα στο χρόνο (έτη, τρίμηνα ή μήνες, ημέρες).
2. *Διαστρωματικά στοιχεία*: είναι τα αντλούμενα δεδομένα από έναν πληθυσμό σε δεδομένη χρονική περίοδο. Για παράδειγμα, οι έρευνες για την δημοτική δαπάνη αποτελούνται από έναν αριθμό επιμέρους δημοτικών δαπανών, των οποίων οι δαπάνες παρουσιάζονται για ένα συγκεκριμένο έτος.
3. *Δυναμικά διαστρωματικά στοιχεία*: είναι ένας συνδυασμός των παραπάνω δύο κατηγοριών, αφού αναφέρονται σε δεδομένα που αποτελούνται από μια μικρή χρονολογική σειρά για κάποιες διαστρωματικές μονάδες που μελετώνται. Μπορούμε να ισχυριστούμε ότι έχουμε επαναλαμβανόμενα διαστρωματικά δεδομένα, τα οποία συλλέγονται από τις ίδιες πηγές για κάποιο συγκεκριμένο χρονικό διάστημα. Το συγκεκριμένο είδος δεδομένων μπορεί να έχει σε σχέση με έρευνα σχετικά με χώρες, περιοχές, επιχειρήσεις, νοικοκυριά κ.λπ. Μπορεί να παραχθεί λαμβάνοντας παρατηρήσεις συγκεκριμένων χρονολογικών σειρών για μια ποικιλία διαστρωματικών οντοτήτων.

Στο κεφάλαιο αυτό γίνεται μία αναφορά στη μελέτη της Επαγωγική Στατιστικής η οποία αποτελεί το τμήμα της Στατιστικής που επικεντρώνεται με τις μεθόδους που καθιστούν δυνατή τη γενίκευση των δειγματοληπτικών συμπερασμάτων. Οι μέθοδοι της Επαγωγικής Στατιστικής που επικεντρώνονται με την εκτίμηση των παραμέτρων ενός πληθυσμού καλούνται παραμετρικές μέθοδοι, ενώ αυτές που αποκλίνουν από τέτοιες εκτιμήσεις παραμέτρων ονομάζονται: - μη παραμετρικές (ελεύθερες κατανομών). Για παράδειγμα, η ανάλυση της διασποράς (ANOVA) αποτελεί μια παραμετρική διαδικασία καθώς συγκρίνει τους μέσους κάποιων πληθυσμού ενώ η διαδικασία X^2 που θα εξεταστεί στο κεφάλαιο αυτό ως τεστ ανεξαρτησίας, ομογένειας και καλής προσαρμογής είναι μια μη παραμετρική μέθοδο: καθώς εξετάζεται για το αν συσχετίζονται στατιστικά δύο χαρακτηριστικές ιδιότητες ενός πληθυσμού και δεν σχετίζεται με παραμέτρους. Όπως έχει αναφερθεί, τα δεδομένα ανήκουν σε μία από τις τέσσερις κλίμακες που προαναφέρθηκαν: ονομαστική, τακτική, διαστημική και αναλογική.

Αν τα δεδομένα έχουν πραγματοποιηθεί με μετρήσεις τακτικής κλίμακας, τότε υπολογισμοί αριθμητικών μέσων και διακυμάνσεων δεν έχουν νόημα, που αυτό σημαίνει ότι οι παραμετρικές μέθοδοι δεν μπορούν να χρησιμοποιηθούν. Στη πλειονότητα οι παραμετρικές στατιστικές μέθοδοι απαιτούν τη χρήση διαστημικών ή αναλογικών δεδομένων. Σε τέτοιου είδους δεδομένα ο υπολογισμός περιγραφικών στατιστικών έχει νόημα καθώς η ερμηνεία των στατιστικών αυτών έχει κάποια σημασία. Οι στατιστικές μέθοδοι που παρουσιάστηκαν προηγούμενα είναι γνωστές ως παραμετρικές μέθοδοι και εφαρμόζονται στα μη-κατηγορικά, δομένα δειγματοληπτικών από κανονικά κατανεμημένους πληθυσμούς των οποίων οι διακυμάνσεις είναι ίσες. Οι μη-παραμετρικές διαδικασίες δοκιμάζονται για ανάλυση κατηγορικών και μη δεδομένων, των οποίων οι διακυμάνσεις είναι άνισες και των οποίων οι πληθυσμοί προέλευσης δεν ακολουθούν κανονική κατανομή. Διαπιστώνεται ότι οι μη παραμετρικές διαδικασίες απαιτούν λιγότερο περιοριστικές υποθέσεις για το επίπεδο μέτρησης των δεδομένων και το είδος της κατανομής του πληθυσμού από τον οποίο προέρχεται το δείγμα.

5.2 Σχεδιασμός Πειραμάτων

Η Στατιστική αποσκοπεί στο σχεδιασμό κάποιου πειράματος που διεξάγεται από έναν ερευνητή. Χρησιμοποιείται η **θεωρία Συνόλων** ώστε να καθοριστεί το πεδίο εφαρμογής του πειράματος. Επιπρόσθετα η θεωρία Πιθανοτήτων μπορεί να αποτελέσει λύση μέσω της οποίας μπορούν να ερμηνευθούν τα στοχαστικά πειράματα όπως αναφέρεται από τους Δαμιανού, Παπαδάτος και Χαραλαμπίδης (2003).

Σύμφωνα με Μπούτσικα (2003) τα πειράματα χωρίζονται σε δύο κατηγορίες:

- **Αιτιοκρατικά** (deterministic), όπου οι γνωστές μεταβλητές (η έννοια επεξηγείται στην περιγραφική στατιστική) επαρκούν για την πρόβλεψη των αποτελεσμάτων του πειράματος.
- **Στοχαστικά** (stochastic / probabilistic), όπου οι γνωστές μεταβλητές δεν επαρκούν για την πρόβλεψη των αποτελεσμάτων του πειράματος.

5.2.1 Περιγραφική Στατιστική

Βασικοί ορισμοί – δεδομένα - μεταβλητές

Προτού παρουσιαστούν οι επιμέρους ενότητες που αφορούν την Περιγραφική Στατιστική θα γίνει μια σύντομη εισαγωγή σε βασικούς ορισμούς: σύμφωνα με τους Τσίμπος και Γεωργιακώδης (1999) ο υπό μελέτη **πληθυσμός** αναφέρεται στο σύνολο των δεδομένων μιας μελέτης ενώ το **δείγμα** αφορά μέρος του πληθυσμού, ώστε να μη χρειαστεί η εξονυχιστική ανάλυση ολόκληρου του πληθυσμού.

Τα διαφορετικά είδη δεδομένων υπό μελέτη είναι πολύ σημαντικό ζήτημα ώστε να επιλεγεί η κατάλληλη στατιστική προσέγγιση. Έτσι, τα είδη των δεδομένων με βάση τη βιβλιογραφία (Τσίμπος και Γεωργιακώδης 1999, Panik 2012) χωρίζονται σε δύο κατηγορίες ανάλογα με την τιμή που μπορούν να πάρουν:

- **Αριθμητικά/Ποσοτικά**: Όταν η πληροφορία που προκύπτει από τα δεδομένα έχει αριθμητική τιμή. Αυτά διακρίνονται στις εξής υπο-κατηγορίες:
- **Συνεχή**: οι αριθμητικές τιμές εκφράζουν κάποιο διάστημα τιμών, όπως π.χ. έναν δεκαδικό αριθμό.
- **Ασυνεχή/Διακριτά**: οι αριθμητικές τιμές αφορούν ακέραιους αριθμούς.
- **Ονομαστικά/Ποιοτικά**: όταν η πληροφορία που προκύπτει από τα δεδομένα έχει ως τιμή κάποια λέξη που την περιγράφει. Όπως είναι κατανοητό στα ποιοτικά δεδομένα δεν μπορούν να εφαρμοστούν μαθηματικές πράξεις, αλλά μπορούν οι ποιοτικές τους τιμές αυθαίρετα να αντικατασταθούν από αριθμητικές για τη διευκόλυνση της διεξαγωγής της στατιστικής μελέτης.

Γενικότερα σύμφωνα με τη βιβλιογραφία (Τσίμπος και Γεωργιακώδης 1999, Panik 2012) οι τιμές των δεδομένων και των μεταβλητών (που αναλύονται στη συνέχεια) μπορούν να ενταχθούν σε τέσσερις κλίμακες μέτρησης ξεκινώντας από τη χαμηλότερη προς την υψηλότερη βαθμίδα. Η χαμηλότερη είναι η (α) ονομαστική (nominal), έπειτα ακολουθεί η (β) τακτική (ordinal), στη συνέχεια η (γ) διαστημική (interval) και τέλος, η (δ) αναλογική (ratio). Ειδικότερα ο Panik (2012) αναλύει τις παραπάνω κλίμακες με κάθε κλίμακα να φέρει ένα νέο χαρακτηριστικό το οποίο προστίθεται στην αμέσως επόμενη κλίμακα μαζί με το νέο. Έτσι μέσω της (α) ονομαστικής κλίμακας είναι εφικτός μονάχα ο προσδιορισμός κατηγοριών. Μέσω της (β) τακτικής αναφέρεται η δυνατότητα διάταξης και κατάταξης των στοιχείων. Μέσω της (γ) διαστημικής κλίμακας εντάσσεται η ιδιότητα καθορισμού της απόστασης μεταξύ των μετρήσεων και τέλος μέσω της (δ) αναλογικής προστίθεται μια μηδενική κατάσταση, μέσω της οποίας είναι εφικτές οι συγκρίσεις μεταξύ των μετρήσεων. Ανάλογα λοιπόν το είδος των δεδομένων και των μεταβλητών καθορίζονται και οι στατιστικές μέθοδοι που ακολουθούνται.

Έως τώρα έχει γίνει αναφορά στα δεδομένα ενός πληθυσμού, τα οποία συλλέγονται και επεξεργάζονται με κατάλληλο τρόπο. Η εξέταση όμως αυτή πραγματοποιείται ως προς κάποιο χαρακτηριστικό ή κάποια ιδιότητα για το δείγμα. Σύμφωνα με τη βιβλιογραφία (Τσίμπος και Γεωργιακώδης 1999, Panik 2012) το σύνολο των χαρακτηριστικών αυτών ονομάζονται πληθυσμιακά χαρακτηριστικά. Οι δε τιμές που παίρνουν τα πληθυσμιακά χαρακτηριστικά αποτελούν τιμές των δεδομένων ενώ το κάθε πληθυσμιακό χαρακτηριστικό προς εξέταση ονομάζεται μεταβλητή, η οποία με βάση τις μετρήσεις που πραγματοποιούνται έχει και αυτή τιμές, που μπορούν πάλι να υπάγονται σε μια από τις **κατηγορίες των ειδών δεδομένων** που αναπτύχθηκαν παραπάνω.

Σύμφωνα με τη βιβλιογραφία (Τσίμπος και Γεωργιακώδης 1999, Panik 2012), μια μεταβλητή γράφεται με κεφαλαίους λατινικούς χαρακτήρες όπως π.χ. *X*. Αυτή μετράται για κάθε στοιχείο *i*, το οποίο

ξεκινά από το πρώτο στοιχείο μέχρι το τελευταίο του πληθυσμού, με μέγεθος πληθυσμού $N = (i = 1, 2, \dots, N)$ ενώ όταν αφορά εξέταση δείγματος πληθυσμού (δηλαδή εξέταση τμήματος του πληθυσμού) αντιστοιχεί σε μέγεθος δείγματος $n = (i = 1, 2, \dots, n)$. Έτσι η εξεταζόμενη μεταβλητή για κάποιο στοιχείο i του πληθυσμού παίρνει τη μορφή x_i . Σύμφωνα με τον Panik (2012), στον παρακάτω τύπο (5.1) εκφράζεται το άθροισμα όλων των τιμών της μεταβλητής για το δείγμα του πληθυσμού:

$$X_1 + X_2 + \dots + X_n = \sum_{i=1}^n X_i \quad (5.1)$$

Ακόμη όπως αναφέρεται στη βιβλιογραφία (Τσίμπος και Γεωργιακώδης 1999, Panik 2012), οι μεταβλητές διακρίνονται σε:

- **Εξαρτημένες και ανεξάρτητες** ανάλογα με το αν εξαρτώνται ή όχι από τις τιμές άλλων μεταβλητών.
- **Τυχαίες/Στοχαστικές και μη τυχαίες/προσδιορίσιμες** ανάλογα με το αν μπορούν να προσδιοριστούν εκ των προτέρων. Έτσι οι τιμές των πρώτων διέπονται από την θεωρία των Πιθανοτήτων προκειμένου να προσδιοριστούν, ενώ οι δεύτερες προσδιορίζονται εκ των προτέρων.

5.2.2 Κατανομή συχνότητας και παρουσίαση δεδομένων

Η μέτρηση των συχνοτήτων εμφάνισης των μεταβλητών είναι πολύ σημαντική για την ερμηνεία πειραμάτων και φαινομένων. Έτσι όπως αναφέρουν και οι Τσίμπος και Γεωργιακώδης (1999), οι συχνότητες (frequencies) είναι αριθμητικές τιμές που εκφράζουν την καταμέτρηση της συχνότητας εμφάνισης τιμών που λαμβάνουν οι μεταβλητές και συμβολίζονται ως f_i , όπου το $i = 1, 2, \dots, k$ και αποτελεί κάποια κατηγορία της μεταβλητής υπό εξέταση. Οι συχνότητες μπορούν να εκφραστούν σε ποσοστιαία ή αναλογική κατανομή και τότε ονομάζονται σχετικές. Μάλιστα σύμφωνα με τους Τσίμπος και Γεωργιακώδης (1999) «επειδή οι σχετικές συχνότητες δηλώνουν την κατ' αναλογία συχνότητα εμφάνισης κάποιου πληθυσμιακού χαρακτηριστικού στο σύνολο των εκ παρατηρήσεως μετρήσεων, μπορούμε να θεωρήσουμε ότι επέχουν θέση πιθανοτήτων». Όπως συμπληρώνουν οι Αγγελής και Δημάκη (2011) η κατανομή πιθανότητας (probability distribution) μιας τυχαίας μεταβλητής δίνει πληροφορίες για την πιθανότητα εμφάνισης των δυνατών τιμών της. Ειδικότερα σύμφωνα με τη βιβλιογραφία (Everitt 2006) η κατανομή πιθανότητας μιας διακριτής τυχαίας μεταβλητής είναι ουσιαστικά ένας μαθηματικός τύπος που έχει ως αποτέλεσμα την πιθανότητα για κάθε τιμή της μεταβλητής. Ειδικότερα, το μέτρο πιθανότητας όπως αναφέρεται τοποθετείται μεταξύ του διαστήματος των τιμών 0 έως 1, με την τιμή 1 να αντιστοιχεί σε ένα βέβαιο να συμβεί ενδεχόμενο. Αντίστοιχα όταν η τυχαία μεταβλητή είναι συνεχής ο μαθηματικός τύπος καθορίζει ένα διάστημα τιμών και όσον αφορά τη γραφική απεικόνιση του αποτελεί μια καμπύλη. Ένα παράδειγμα κατανομής πιθανότητας αποτελεί η κανονική κατανομή. Άλλος όρος που μπορεί να χρησιμοποιηθεί είναι η πυκνότητα πιθανότητας (probability density). Για την περιγραφή της τυχαίας μεταβλητής απαιτείται η χρήση μέτρων κεντρικής τάσης τα οποία αναφέρονται παρακάτω.

Ο όρος αθροιστική συχνότητα είναι επίσης πολύ σημαντικός στη Στατιστική, για την ανάλυση του όμως πρώτα θα χρειαστεί η επεξήγηση του όρου της κλάσης. Τα ποσοτικά δεδομένα, σύμφωνα με τη βιβλιογραφία (Τσίμπος και Γεωργιακώδης 1999, Panik 2012), εντάσσονται είτε σε διαστημική είτε σε αναλογική κλίμακα και η κατάταξη τους πραγματοποιείται μέσω δημιουργίας κλάσεων. Για να οριστούν οι κλάσεις, πρέπει να υπολογιστεί το εύρος (range) της μεταβλητής, το οποίο αποτελεί τη διαφορά μεταξύ της μέγιστης και ελάχιστης τιμής που παίρνει η μεταβλητή. Στη συνέχεια από τον παρακάτω τύπο (24), μπορεί να υπολογιστεί το πλάτος των διαστημάτων κλάσεων δ όπως ονομάζεται, όπου R αντιστοιχεί στο εύρος της μεταβλητής και k στον επιθυμητό αριθμό κλάσεων:

$$\delta = \frac{R}{k} \quad (5.2)$$

Η παραπάνω αναφορά στον αριθμό κλάσεων, το εύρος και το πλάτος των διαστημάτων κλάσεων κρίθηκε απαραίτητη ώστε να αναλυθεί παρακάτω έννοια των αθροιστικών συχνοτήτων. Οι συχνότητες έχουν ήδη αναλυθεί. Σύμφωνα λοιπόν με τους Τσίμπος και Γεωργιακώδης (1999) με τον όρο αθροιστικές συχνότητες F_i (cumulative frequencies) εννοείται το άθροισμα των συχνοτήτων, το οποίο για όλες τις κλάσεις πρέπει να ισούται με όλα τα στοιχεία του πληθυσμού. Έτσι εκφράζεται από τον παρακάτω τύπο (5.3):

$$\sum_{i=1}^k f_i = n \quad (5.3)$$

Ειδικότερα σύμφωνα με τον Everitt (2006) η συνάρτηση αθροιστικής κατανομής (cumulative distribution function) είναι μια κατανομή που εμφανίζει το σύνολο των τιμών μιας τυχαίας μεταβλητής, οι οποίες είναι λιγότερες/περισσότερες από τις δεδομένες τιμές. Η κατανομή αθροιστικών συχνοτήτων (cumulative frequency distribution) κατατάσσει τα παραπάνω σε πίνακα.

Στην περιγραφική στατιστική για την **παρουσίαση των δεδομένων** χρησιμοποιούνται οι **πίνακες** και οι **γραφικές παραστάσεις**. Στους μεν πίνακες περιέχονται τα συγκεντρωμένα στοιχεία, στις δε γραφικές παραστάσεις οι ίδιες πληροφορίες αναπαρίστανται με γραφικό τρόπο.

Πρέπει να σημειωθεί ότι για τις γραφικές παραστάσεις που αντιπροσωπεύουν ποιοτικά δεδομένα τότε χρησιμοποιούνται τα κυκλικά διαγράμματα ή ραβδογράμματα.

Σύμφωνα με τους Τσίμπος και Γεωργιακώδης (1999), στα ποσοτικά δεδομένα αντιστοιχούν τα ιστογράμματα συχνοτήτων (αφορούν τον βαθμό συκέντρωσης των δεδομένων σε σχέση με τις κλάσεις), τα πολύγωνα και οι καμπύλες συχνοτήτων (για σύγκριση πολλών κατανομών ταυτόχρονα όταν αφορούν συνεχή δεδομένα, ώστε να εξαχθούν συναρτήσεις που περιγράφουν την σχέση τιμών μεταβλητών και συχνότητας εμφάνισης τους στον πληθυσμό), οι αψίδες ή πολύγωνα αθροιστικών συχνοτήτων και τα φυλλογραφήματα.

Ειδικότερα, σύμφωνα τους με Τσίμπος και Γεωργιακώδης (1999) ανάλογα το σχήμα των καμπυλών των γραφικών παραστάσεων, υποδεικνύονται τέσσερις κατηγορίες κατανομών: μονοκόρυφες, σχήματος U, σχήματος J, σύνθετες κατανομές. Τα άκρα των καμπυλών ονομάζονται ουρές (tails). Η μονοκόρυφη κατανομή όταν έχει το χαρακτηριστικό σχήμα καμπάνας τότε είναι συμμετρική και το εμβαδόν του σχήματος της είναι ίσο με το σύνολο των συχνοτήτων. Η χρήση αυτής της κατανομής είναι πολύ συνηθισμένη στην Επαγωγική Στατιστική και όταν είναι συμμετρική ονομάζεται κανονική ή γκαουσιανή κατανομή (normal/gaussian distribution) ενώ σε άλλη περίπτωση χαρακτηρίζεται ασύμμετρη θετικά ή αρνητικά, το οποίο αφορά τις ουρές της κατανομής.

5.2.3 Βασικά στατιστικά μέτρα

Για την επιτυχημένη μελέτη των κατανομών προκειμένου να περιγραφούν τα δεδομένα είναι απαραίτητη η χρήση των παρακάτω στατιστικών περιγραφικών μέτρων τα οποία με βάση τη βιβλιογραφία (Τσίμπος και Γεωργιακώδης 1999, Panik 2012) χωρίζονται σε πέντε κατηγορίες:

1. Μέτρα κεντρικής τάσης (measures of central tendency): όπου υπολογίζονται ποικίλοι μέσοι όροι. Π.χ. αριθμητικός μέσος ή μέση τιμή.
2. Μέτρα θέσης (measures of location): τα οποία αφορούν τη θέση στην κατανομή για τις διάφορες τιμές της μεταβλητής. Π.χ. Διάμεσος
3. Μέτρα διασποράς ή διασκόρπισης (measures of variability, dispersion): τα οποία αφορούν τη συκέντρωση ή μη γύρω από την κεντρική τιμή της μεταβλητής και τη μεταβλητότητα των δεδομένων. Π.χ. εύρος, διακύμανση, τυπική απόκλιση και συντελεστής μεταβλητότητας.
4. Μέτρα ασυμμετρίας (measures of skewness): υπολογίζεται ο βαθμός ασυμμετρίας.
5. Μέτρα κύρτωσης (kurtosis measures): αφορά τον υπολογισμό αιχμηρότητας της καμπύλης.

Τα παραπάνω ονομάζονται στατιστικές όταν αφορούν την περιγραφή του δείγματος. Στην περίπτωση που αφορούν το σύνολο του πληθυσμού τότε ονομάζονται παράμετροι. Από τα παραπάνω μέτρα θα παρουσιαστούν ο αριθμητικός μέσος (κεντρικής τάσης), η διακύμανση και τυπική απόκλιση (κύμανσης), τα οποία αποτελούν κάποια από τα βασικότερα και ευρέως χρησιμοποιούμενα μέτρα.

Όπως αναφέρεται στη βιβλιογραφία (Τσίμπος και Γεωργιακώδης 1999) ο αριθμητικός μέσος ή και απλά μέσος υπολογίζεται από το λόγο του αθροίσματος των τιμών των δεδομένων προς είτε το μέγεθος του πληθυσμού (N) είτε του δείγματος (n) ανάλογα. Ο μέσος πληθυσμού συμβολίζεται με το γράμμα μ και υπολογίζεται από τον παρακάτω μαθηματικό τύπο (26):

$$\mu = \frac{\sum_{i=1}^N X_i}{N} = \frac{X_1 + X_2 + \dots + X_N}{N} \quad (5.4)$$

Αν αντικατασταθεί το N με n υπολογίζεται ο μέσος δείγματος, ο οποίος αντί για το γράμμα μ συμβολίζεται με $\bar{x}$.

Όπως αναφέρεται στη βιβλιογραφία (Τσίμπος και Γεωργιακώδης 1999), η διακύμανση (variance) και η τυπική απόκλιση (standard deviation) είναι μέτρα που υπολογίζουν τη διασπορά για τις τιμές των δεδομένων γύρω από το μέσο τους. Ειδικότερα η διακύμανση συμβολίζεται σ^2 ή $V(X)$, όταν αντιστοιχεί στη διακύμανση πληθυσμού, ενώ όταν αντιστοιχεί στο δείγμα συμβολίζεται με s^2 . Στην τυπική απόκλιση αντιστοιχεί το σύμβολο σ όταν αφορά τον πληθυσμό και s όταν αφορά δείγμα. Η διακύμανση και η τυπική απόκλιση ορίζονται από τους παρακάτω μαθηματικούς τύπους (5.5) και (5.6) αντίστοιχα:

$$\sigma^2 = \frac{\sum_{i=1}^N (X_i - \mu)^2}{N} \quad (5.5)$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^N (X_i - \mu)^2}{N}} \quad (5.6)$$

Μάλιστα μεγάλες τιμές για τη διακύμανση και τυπική απόκλιση υποδηλώνουν μεγάλη μεταβλητότητα, δηλαδή μεγάλη απόσταση μεταξύ μιας τιμής δεδομένων από το μέσο της. Τέλος, θα πρέπει τόσο η διακύμανση όσο και η τυπική απόκλιση να είναι μεγαλύτερες ή ίσες με το μηδέν. Όσον αφορά την διακύμανση και τυπική απόκλιση δείγματος στη βιβλιογραφία (Τσίμπος και Γεωργιακώδης 1999) αναφέρεται ο παρακάτω μαθηματικός τύπος (5.7):

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = \frac{1}{n-1} \left(\sum_{i=1}^n nx^2 \right) \quad (5.7)$$

5.2.4 Κανονική κατανομή και τυποποιημένη τιμή Z

Όταν ο πληθυσμός υπό μελέτη είναι κανονικής κατανομής, όπως αναφέρει ο Zar (2010), μπορεί να οριστεί ένας μέσος, σύμφωνα με τον οποίο ο μισός πληθυσμός είναι μικρότερος από αυτόν και ο άλλος μισός μεγαλύτερος λόγω της συμμετρίας που διέπει την κανονική κατανομή.

Για τον υπολογισμό ποσοστού του πληθυσμού που είναι μεγαλύτερο, για κάποια ποσότητα μεγαλύτερη ή μικρότερη του μέσου, τότε είναι απαραίτητη η κανονική απόκλιση σ του πληθυσμού. Μέσω της τυποποιημένης τιμής Z (Zar 2010) φαίνεται κατά πόσες κανονικές αποκλίσεις απέχει από τον μέσο (σε ποια θέση βρίσκεται σε σχέση με το μέσο δηλαδή) μια τιμή x_i , του πληθυσμού και υπολογίζεται από τον παρακάτω μαθηματικό τύπο (5.8):

$$Z = \frac{x_i - \mu}{\sigma} \quad (5.8)$$

Ο παραπάνω υπολογισμός αναφέρεται (Zar 2010) ως κανονικοποίηση (normalizing) ή τυποποίηση (standardizing) για κάποια τιμή x_i , και η τιμή Z αναφέρεται ως κανονική απόκλιση (normaldeviate) ή τυποποιημένη τιμή (standardscore). Η μέση τιμή ενός συνόλου τυποποιημένων τιμών είναι ίση με 0 και η διακύμανση είναι ίση με 1.

5.2.5 Μη κανονική κατανομή

Σύμφωνα με τον Zar (2010) όταν ο υπό μελέτη πληθυσμός προέρχεται από μη κανονική κατανομή, η κατανομή θα τείνει προς την κανονικότητα, όσο περισσότερο αυξάνεται το δείγμα n σύμφωνα με το **θεώρημα κεντρικού ορίου** (centrallimittheorem). Ήδη στην παραπάνω ενότητα έχει παρουσιαστεί η παράμετρος

διακύμανσης σ^2 . Έτσι για τη **διακύμανση μέσου** $\sigma_{\bar{x}}^2$ (δηλαδή η διακύμανση για την μέση τιμή, όπου στην συγκεκριμένη περίπτωση αφορά μόνο το δείγμα n και όχι τον πληθυσμό) παρατηρείται η μείωση της όσο μειώνεται το n όπως φαίνεται στον παρακάτω μαθηματικό τύπο (5.9):

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} \quad (5.9)$$

Όπως ήδη έχει αναφερθεί σε παραπάνω ενότητα η τυπική απόκλιση συμβολίζεται με το σύμβολο σ . Έτσι η τυπική απόκλιση μέσου $\sigma_{\bar{x}}$ αναφέρεται στη βιβλιογραφία (Zar 2010) και ως τυπικό σφάλμα (μπορεί να είναι σφάλμα τύπου Ι ή σφάλμα τύπου ΙΙ) και ορίζεται από την παρακάτω μαθηματική έκφραση (5.10):

$$\sigma_{\bar{x}} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}} \quad (5.10)$$

Σύμφωνα με το μαθηματικό τύπου υπολογισμού της τυποποιημένης τιμής Z (βλέπε εξίσωση (5.10), η τυποποιημένη τιμή κατανομής μέσων που αφορά την x_i εκφράζεται με τον παρακάτω μαθηματικό τύπο (5.11) (Zar 2010):

$$Z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} \quad (5.11)$$

Για τον υπολογισμό της τυποποιημένης τιμής πρέπει να είναι γνωστή η διακύμανσης σ^2 για ολόκληρο τον πληθυσμό, κάτι το οποίο είναι δύσκολο να υπολογιστεί. Συνεπώς αντ' αυτού υπολογίζεται η διακύμανση μέσου για τυχαίο δείγμα.

5.2.6 Επαγωγική Στατιστική και θεωρία πιθανοτήτων

Μέχρι τώρα αναλύθηκαν οι βασικοί ορισμοί της Στατιστικής και οι έννοιες και τα μέτρα που αφορούν ειδικότερα την Περιγραφική Στατιστική. Όσον αφορά την Επαγωγική Στατιστική, όπως ήδη έχει αναφερθεί παραπάνω, αποτελεί το τμήμα της Στατιστικής που ασχολείται με τη **διεξαγωγή συμπερασμάτων** για το σύνολο του πληθυσμού, μέσω της λήψης τυχαίου δείγματος. Μάλιστα όπως αναφέρεται στη βιβλιογραφία (Panik 2012) η Επαγωγική Στατιστική κατηγοριοποιείται σε δύο υπό-κλάδους: την Εκτιμητική και τον Έλεγχο Στατιστικών Υποθέσεων, στον οποίο και χρησιμοποιείται η θεωρία των Πιθανοτήτων.

Έτσι στο σημείο αυτό μπορεί να πραγματοποιηθεί μια πιο εκτεταμένη αναφορά στις θεωρίες Συνόλων και Πιθανοτήτων, καθώς κάποιοι βασικοί ορισμοί που τις αφορούν έχουν επεξηγηθεί στις παραπάνω ενότητες.

Σύμφωνα με τον Μπούτσικα (2003) το σύνολο των δυνατών αποτελεσμάτων ενός πειράματος ονομάζεται **δειγματικός χώρος** και τα μεμονωμένα αυτά δυνατά αποτελέσματα ονομάζονται **ενδεχόμενα ή γεγονότα**. Μάλιστα, όπως αναφέρουν οι Δαμιανού, Παπαδάτος και Χαραλαμπίδης (2003) ένας δειγματικός χώρος μπορεί να είναι αριθμήσιμος (πεπερασμένος ή άπειρος αριθμήσιμος) ή μη αριθμήσιμος. Ο αριθμήσιμος δειγματικός χώρος μπορεί να είναι διακριτός ή συνεχής. Κάποιες βασικές έννοιες της θεωρίας Συνόλων και Πιθανοτήτων, οι οποίες έχουν αντληθεί από τη βιβλιογραφία (Δαμιανού, Παπαδάτος και Χαραλαμπίδη 2003) είναι οι ακόλουθες: ο όρος **σύνολο** αναφέρεται σε μια συλλογή, η οποία περιέχει κάποια στοιχεία. Τα σύνολα συμβολίζονται με κεφαλαία γράμματα ενώ τα στοιχεία τους με πεζά. Η σχέση μεταξύ συνόλου - υποσυνόλου αφορά ένα σύνολο το οποίο περιλαμβάνει μέσα του ένα άλλο σύνολο. Στη θεωρία Συνόλων και Πιθανοτήτων χρησιμοποιούνται απεικονίσεις για την αναπαράσταση των συνόλων και των σχέσεων τους με άλλα σύνολα, τα γνωστά διαγράμματα Venn. Ο όρος **πιθανότητα** χρησιμοποιείται ώστε να υπολογιστεί το πόσο πιθανή είναι η πραγματοποίηση ενός ενδεχομένου του δειγματικού χώρου. Τα ενδεχόμενα (αναφέρονται επίσης και ως δειγματικά σημεία) και μπορεί να είναι ποσοτικής ή ποιοτικής φύσεως (π.χ. αριθμοί ή περιγραφές αντίστοιχα).

Ειδικότερα σε σχέση με όσα έχουν αναφερθεί στην προηγούμενη ενότητα σχετικά με τις τυχαίες μεταβλητές, όμοια για την εξέταση των ποιοτικών **ενδεχομένων** είναι δυνατή η αντιστοίχιση κάθε ενδεχομένου με έναν πραγματικό αριθμό. Ειδικότερα, σύμφωνα με Αγγελή και Δημάκη (2011), η

αντιστοίχιση αυτή πραγματοποιείται μέσω της συνάρτησης που εκφράζει την τυχαία μεταβλητή. Σχετικά με τις τυχαίες μεταβλητές έχει γίνει ήδη αναφορά. Όπως ήδη έχει αναφερθεί η συνάρτηση αυτή διαφέρει ανάλογα το είδος της τυχαίας μεταβλητής και η αντιστοίχιση μιας πιθανότητας για τα ενδεχόμενα της τυχαίας μεταβλητής ονομάζεται συνάρτηση πιθανότητας ή αλλιώς κατανομή πιθανότητας. Τέλος να τονιστεί ξανά πως η παρουσίαση των κατανομών πιθανοτήτων γίνεται κλασικά μέσω πινάκων ή γραφικών παραστάσεων (διάγραμμα πιθανότητας).

5.2.7 Έλεγχος Στατιστικών Υποθέσεων

Στην παρούσα διατριβή έχει χρησιμοποιηθεί ο υπο-κλάδος της Επαγωγικής Στατιστικής που αφορά τον **Έλεγχο Στατιστικών Υποθέσεων** για το προτεινόμενο μοντέλο. Έτσι, θα παρουσιαστούν τα βασικά σημεία του υπο-κλάδου αυτού, για την καλύτερη κατανόηση των διεργασιών που πραγματοποιήθηκαν για την ανάπτυξη του προτεινόμενου μοντέλου της διατριβής.

Σύμφωνα με τους Κατσάνος και Αβούρης (2008), ο Έλεγχος Στατιστικών Υποθέσεων (Statistical Hypothesis Testing) αφορά στην επαλήθευση **πρόβλεψης** που αφορά την μεταβλητή υπό μελέτη, η οποία ονομάζεται **υπόθεση (hypothesis)**. Έτσι, η υπόθεση μπορεί να θεωρηθεί ως ένας ισχυρισμός ή μια μη αποδεδειγμένη θεωρία όπως αναφέρει ο Panik (2012). Οι Κατσάνος και Αβούρης (2008) συνοψίζουν τη διαδικασία ελέγχου στατιστικής υπόθεσης στα εξής βήματα:

1. **Διατύπωση υποθέσεων:** στο στάδιο αυτό καθορίζεται ένα **ζευγάρι υποθέσεων, η μηδενική υπόθεση H_0 (null hypothesis) και η εναλλακτική H_1 ή H_A (alternative hypothesis)**. Σύμφωνα με τον Zar (2010) όταν πρόκειται για διεξαγωγή συμπερασμάτων σε σχέση με το μέσο όρο κάποιου πληθυσμού η **μηδενική υπόθεση συνήθως αποτελεί μια συνοπτική δήλωση που αφορά το μέσο του πληθυσμού (π.χ. $\mu=0$)**. Η μηδενική υπόθεση δηλώνεται έχοντας υπόψη ότι μπορεί να είναι ψευδής. Για το λόγο αυτό δηλώνεται και η αντίστοιχη εναλλακτική υπόθεση, η οποία μπορεί να είναι αληθής. Μάλιστα, για την απόρριψη της μηδενικής υπόθεσης θα πρέπει να ισχύει κάποιος βαθμός βεβαιότητας (που συνδέεται παρακάτω με το επίπεδο σημαντικότητας), όπως συμπληρώνουν οι Κατσάνος και Αβούρης (2008).
2. **Διαμόρφωση κριτηρίων:** όπως αναφέρουν οι Κατσάνος και Αβούρης (2008), στον Έλεγχο Στατιστικών Υποθέσεων το βασικό ζήτημα προς επίλυση είναι να διευκρινιστεί εάν οι παρατηρήσεις στις αλλαγές των τιμών αφορούν κάποιο σφάλμα δειγματοληψίας (δηλαδή απόκλιση των πραγματικών τιμών του πληθυσμού από αυτούς που εκτιμώνται βάση ενός αντιπροσωπευτικού δείγματος) ή στην επίδραση της μεταβλητής υπό-μελέτη. Για το λόγο αυτό τίθενται κριτήρια που ορίζουν την ακριβή διαφορά μεταξύ των παραπάνω περιπτώσεων, ώστε να υπάρχει σαφές πεδίο τιμών που οδηγεί στην απόρριψη της μηδενικής υπόθεσης. Πιο συγκεκριμένα, σύμφωνα με Panik (2012), ορίζεται ένα εύρος τιμών για τη μηδενική υπόθεση, το οποίο θα διασπαστεί σε δύο ξεχωριστά υπο-σύνολα έστω υπο-σύνολο R (αντιστοιχεί ουσιαστικά στην εναλλακτική υπόθεση εφόσον η μηδενική απορρίπτεται) και το συμπλήρωμα του $\bar{R}$. Τα υπο-σύνολα αυτά μπορούν να χαρακτηριστούν ως περιοχές απόρριψης και μη-απόρριψης αντίστοιχα. Έτσι, θα πρέπει να καθοριστεί από τη μια η περιοχή απόρριψης και από την άλλη ένα μέγεθος α , το οποίο ονομάζεται επίπεδο σημαντικότητας (level of significance). Σύμφωνα με Κατσάνος και Αβούρης (2008) αυτό *«καθορίζει τη μέγιστη πιθανότητα το αποτέλεσμα μιας στατιστικής ανάλυσης να οφείλεται σε σφάλματα ή τυχαίους παράγοντες»*. Σύμφωνα με τον Zar (2010) το κριτήριο απόρριψης αφορά στο πόσο μικρή πιθανότητα αποδέχεται ο ερευνητής ώστε να απορρίψει τη μηδενική υπόθεση. Συνηθισμένη τιμή πιθανότητας για την απόρριψη της μηδενικής υπόθεσης αποτελεί οποιαδήποτε τιμή αγγίζει το 5% και κάτω. Ακόμη το επίπεδο σημαντικότητας α συνδέεται και με την τιμή τυποποίησης Z η οποία όσο μεγαλύτερη είναι, τόσο μικρότερη είναι και η πιθανότητα η μηδενική υπόθεση να είναι αληθής
3. **Συλλογή δεδομένων:** αφορά τη συλλογή αντιπροσωπευτικού δείγματος και εφαρμογή στατιστικών μέτρων (υπολογισμοί και μετρήσεις).
4. **Αξιολόγηση μηδενικής υπόθεσης:** το τελικό αυτό στάδιο αφορά τον έλεγχο του κατά πόσο η μηδενική υπόθεση ισχύει ή όχι και αυτό είναι εφικτό μέσω διάφορων τύπων στατιστικών δεικτών ελέγχου όπως ονομάζονται. Μόλις υπολογιστούν οι τιμές των στατιστικών μέτρων σε συνδυασμό με τα κριτήρια που διαμορφώθηκαν πραγματοποιείται η λήψη απόφασης. Σύμφωνα με Panik (2012) τα είδη ελέγχου χωρίζονται σε μονόπλευρους ελέγχους (one-tail test) ή όπως απαντάται στη βιβλιογραφία κατευθυνόμενους και σε αμφίπλευρους (two-tailed test) ή μη-κατευθυνόμενους. Οι δε μίας πλευράς

χωρίζονται σε δεξιόπλευρους και αριστερόπλευρους. Όπως αναφέρουν οι Κατσάνος και Αβούρης (2008) μέσω των αμφίπλευρων ελέγχων μπορεί να παρατηρηθεί «αν η αλλαγή της ελεγχόμενης μεταβλητής έχει οποιαδήποτε επίδραση στην παρατηρούμενη μεταβλητή», ενώ μέσω των μονόπλευρων μπορεί να διερευνηθεί και ο τρόπος επίδρασης. Ακόμη, σύμφωνα με τον Zar (2010), μέσω των μονόπλευρων ελέγχων ο ερευνητής μπορεί να εστιάσει στην διαφορά που αφορά σε μια συγκεκριμένη κατεύθυνση προς μια παράμετρο πληθυσμού για μια υποτιθέμενη τιμή.

Καθότι ο καθορισμός κριτηρίου για την αποδοχή ή μη της μηδενικής της υπόθεσης (βήμα 2) αποτελεί κρίσιμο στάδιο για τον Στατιστικό Έλεγχο Υποθέσεων θα πραγματοποιηθεί περαιτέρω ανάλυση όσον αφορά το επίπεδο σημαντικότητας α .

Πιο συγκεκριμένα, το γεγονός ότι η μηδενική υπόθεση δύναται να είναι ψευδής οδηγεί στο συμπέρασμα διεξαγωγής κάποιου τυπικού σφάλματος, όπως αναφέρθηκε και παραπάνω, για το δείγμα προς εξέταση. Το τυπικό σφάλμα μπορεί να ανήκει σε δύο κατηγορίες όπως αναφέρει ο Zar (2010):

1. **Σφάλμα τύπου I:** απόρριψη μηδενικής υπόθεσης όταν όντως είναι αληθής με βαθμό σημαντικότητας α .
2. **Σφάλμα τύπου II:** μη απόρριψης της μηδενικής υπόθεσης, όταν όντως είναι ψευδής. Η πιθανότητα αυτού του σφάλματος αντιστοιχεί στο σύμβολο β . Ως δύναμη (power) στατιστικής υπόθεσης ορίζεται η τιμή $1-\beta$, η οποία εκφράζει την πιθανότητα απόρριψης της μηδενικής υπόθεσης όταν είναι πραγματικά ψευδής.

Όπως αναφέρει ο Zar (2010), οι δύο τύποι σφάλματος α και β συνδέονται μεταξύ τους αντιστρόφως ανάλογα, χαμηλές πιθανότητες του α , οδηγούν σε υψηλές πιθανότητες του β , αλλά και τα δύο μπορούν να μειωθούν ταυτόχρονα όταν αυξάνεται το μέγεθος n του δείγματος. Στον Έλεγχο Στατιστικών Υποθέσεων έχει ιδιαίτερη σημασία ο επιθυμητός συνδυασμός μεταξύ α και β . Γενικά όταν $\alpha=0.05$, τότε υπάρχει μικρή πιθανότητα σφάλματος τύπου I και μεγάλη πιθανότητα σφάλματος τύπου II.

Το επίπεδο σημαντικότητας, ο τύπος ελέγχου στατιστικών υποθέσεων που χρησιμοποιείται και η αντίστοιχη πιθανότητα, θα πρέπει να δηλώνονται στα αποτελέσματα. Ειδικότερα ο Zar (2010) αναφέρει πως όταν πραγματοποιείται απόρριψη της μηδενικής υπόθεσης με 5% επίπεδο σημαντικότητας μπορεί κανείς να αναφερθεί σε **σημαντική διαφορά** (significant difference), ενώ με 1% σε **πολύ σημαντική διαφορά** (highly significant difference).

Το επίπεδο σημαντικότητας α , αντικατοπτρίζεται μέσω της τιμής p (p -value) της πιθανότητας. Ειδικότερα λοιπόν το επίπεδο σημαντικότητας α πρέπει να επιλεγεί από τον ερευνητή. Εναλλακτικά όμως υπάρχει η δυνατότητα καθορισμού του επιπέδου αυτού μέσω των ίδιων των δεδομένων του δείγματος, τα οποία δίνουν το παρατηρούμενο επίπεδο σημαντικότητας, το οποίο αναφέρεται στη βιβλιογραφία (Panik, 2012) ως τιμή p (από το αγγλικό probability). Πιο συγκεκριμένα σύμφωνα με τον Panik (2012), μέσω της τιμής p , υποδηλώνεται η πιθανότητα να «λάβουμε μια υπολογισμένη τιμή από τον στατιστικό έλεγχο τουλάχιστον τόσο μεγάλη όσο αυτή που έχει παρατηρηθεί για τη μηδενική υπόθεση αν αυτή είναι αληθής».

Σύμφωνα με τον Panik (2012) με την τιμή p , μπορούν να διεξαχθούν συμπεράσματα για το αν η μηδενική υπόθεση είναι αληθής ή όχι. Η τιμή p υπολογίζει την περίπτωση που η μηδενική υπόθεση ισχύει. Πιο συγκεκριμένα, όσο πιο μεγάλη είναι η τιμή p τόσο πιο αληθής θεωρείται η μηδενική υπόθεση. Όσο πιο μικρή είναι η τιμή, δηλαδή όσο πιο σπάνιο θα θεωρούνταν ένα αποτέλεσμα δείγματος για τη μηδενική υπόθεση, οδηγεί στο συμπέρασμα ότι είναι ψευδής. Η τιμή p αφορά την παραπάνω απόδειξη και επιστρέφει πιθανοτική τιμή αν και κατά πόσο η μηδενική υπόθεση ικανοποιείται ή όχι. Η μικρή τιμή p υποδηλώνει το βαθμό ψεύδους της μηδενικής υπόθεσης.

Όπως αναφέρουν οι Κατσάνος και Αβούρης (2008), η Επαγωγική Στατιστική περιλαμβάνει μια μεγάλη ποικιλία στατιστικών μεθόδων για τον έλεγχο υποθέσεων και η επιλογή της κατάλληλης μεθόδου έγκειται στα εξής:

1. Στόχος διεξαγωγής πειράματος.
2. Πλήθος, είδος και τιμές μεταβλητών (ανεξάρτητες-εξαρτημένες).
3. Πλήθος δειγμάτων.
4. Είδος δεδομένων.
5. Είδος στατιστικού ελέγχου με βάση τις προϋποθέσεις χρήσης του.

Όπως είναι κατανοητό για τόσο μεγάλο εύρος στατιστικών μεθόδων η ταξινόμηση των ειδών ελέγχων είναι αρκετά δύσκολη. Ορισμένες κατηγοριοποιήσεις θα αναφερθούν στη συνέχεια.

Σύμφωνα με τη βιβλιογραφία (Griest appear in *International Encyclopedia of the Social and Behavioral Sciences*), σχεδόν **όλα τα είδη στατιστικών ελέγχων χρησιμοποιούν τα παρακάτω για τη διεξαγωγή ελέγχου υποθέσεων**: κατανομές, συχνότητες, μέσους όρους (π.χ. μέση τιμή), διασπορά και συσχετίσεις (π.χ. δείκτης συσχέτισης w του Kendall).

Μια πρώτη κατηγοριοποίηση των ελέγχων στατιστικών υποθέσεων με την οποία οι στατιστικοί έλεγχοι διακρίνονται **σε καλής προσαρμογής**, ασχολούνται με την σύγκριση κατανομών και απόκλιση από το αναμενόμενο **και σε ελέγχους ανεξαρτησίας**. Ειδικότερα για τους ελέγχους ανεξαρτησίας σύμφωνα με τη βιβλιογραφία (Griest appear in *International Encyclopedia of the Social and Behavioral Sciences*) διακρίνονται σε:

1. Μονοπαραγοντικούς (monofactorial), περιέχουν μόνο μια ανεξάρτητη μεταβλητή.
2. Πολυπαραγοντικούς (multifactorial), περιέχουν περισσότερες από μια μεταβλητές:
3. Μεταβλητές που σχετίζονται αλλά αλληλεπιδρούν μεταξύ τους.
4. Μεταβλητές που σχετίζονται αλλά δεν αλληλεπιδρούν μεταξύ τους.

Σύμφωνα με SAS Institute (1999) οι έλεγχοι στατιστικών υποθέσεων χωρίζονται σε **παραμετρικούς** και **μη-παραμετρικούς**. Οι παραμετρικοί έλεγχοι εξαρτώνται από τις προδιαγραφές που φέρει μια κατανομή πιθανότητας και από υποθέσεις σχετικές με την κατανομή ενώ οι μη-παραμετρικοί έλεγχοι όχι. Καθώς οι παραμετρικοί έλεγχοι συχνά υποθέτουν μια κανονική κατανομή πληθυσμού, αν αυτή η υπόθεση δεν ισχύει για το πείραμα τότε συνίσταται η χρήση μη παραμετρικών ελέγχων. Όπως αναφέρεται στη βιβλιογραφία (Neideen και Brasel 2007) ορισμένοι συνήθεις παραμετρικοί έλεγχοι είναι οι studentt-Test, z-Test, ANOVA και ορισμένοι μη παραμετρικοί είναι οι χ^2 , SpearmanRankCoefficient, Mann-WhitneyUTest, Kruskal-WallisTest.

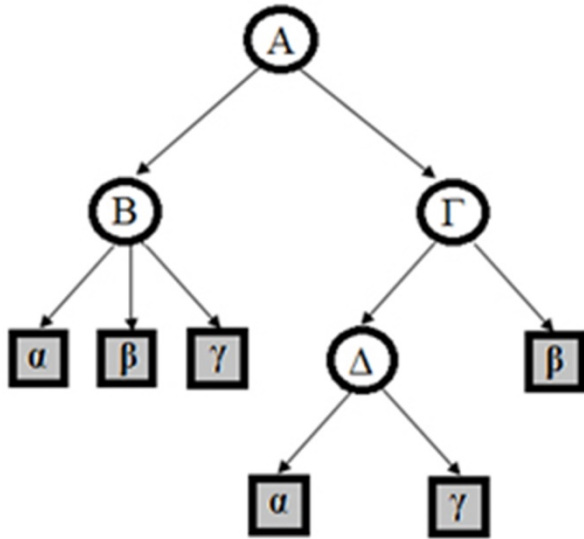
5.3 Συστήματα Ομαδοποίησης & Κατηγοριοποίησης

5.3.1 Δέντρα αποφάσεων SVM - BinaryTrees (BTS)

Η εξέλιξη της τεχνητής νοημοσύνης καθώς και οι προοδευτικά αυξανόμενες και σημαντικές εφαρμογές της, αιτιάζονται σε υψηλό βαθμό στις εξελίξεις στον κλάδο της – κατηγοριοποίησης δεδομένων (classification) ομαδοποίησης. Πρόκειται για κατηγοριοποιήσεις οιαδήποτε δεδομένων, για πολύ ευρύ φάσμα εφαρμογών τους και για απλουστευμένη και γρήγορη διαδικασία που όμως προσφέρεται ως σημαντικό εργαλείο στη λήψη αποφάσεων σχετικά με πολύπλευρα προβλήματα πρόβλεψης. Η πρόβλεψη μπορεί να έχει εφαρμογή σε πολυάριθμα ερευνητικά πεδία όπως στην Ιατρική, τη Βιολογία, στη Σεισμολογία, την Οικονομία και γενικά στην Επιστήμη της Πληροφορίας. Η μεθοδολογία λύσης ενός προβλήματος που προκύπτει από την ανάλυσή του, οδηγεί στη σχεδίαση ενός αλγορίθμου που συνιστά την ακολουθία βημάτων, που πρέπει να ακολουθηθούν για να επιλυθεί το πρόβλημα. Όπως αναφέρθηκε, υπάρχει περίπτωση να παρουσιασθούν περισσότερες από μία διαδικασίες για την επίλυση ενός προβλήματος. Συνεπώς, για να βρεθεί η βέλτιστη λύση απαιτείται να πραγματοποιηθούν κάποιες επιμέρους παραδοχές κατά τη διαδικασία σχεδίασης του αλγορίθμου. Υπάρχουν συγκεκριμένες διαδικασίες που συχνά χρησιμοποιούνται σε πληθώρα προβλημάτων. Έτσι, οι διαδικασίες αυτές θεωρούνται τετριμμένες εξαιτίας των κοινών χαρακτηριστικών τους κατά την επίλυση ενός προβλήματος. Η πρακτική αυτή σε ένα βαθμό διευκολύνει στην ένταξη κάποιου προβλήματος στην αντίστοιχη κατηγορία επίλυσής του (όσο αυτό είναι εφικτό). Γενικότερα, κάθε διαδικασία απαιτεί να υποστηρίξει τα εξής: να αντιμετωπίζει με συγκεκριμένη μεθοδολογία τα δεδομένα, να έχει συγκεκριμένη ακολουθία εντολών και να έχει τη δική της αποδοτικότητα. Επομένως κάθε διαδικασία έχει τα δικά της χαρακτηριστικά και τα δικά της χαρακτηριστικά. Ο αλγόριθμος αποτελεί πρωτίστως τη βάση όλων των μεθοδολογιών κατηγοριοποίησης.

Οι δυαδικοί αλγόριθμοι “SUPPORT VECTOR MACHINES-SVM” αποτελούν τη μεθοδολογία της μάθησης (learning methodology), ακριβώς επειδή σκοπός είναι να «εκπαιδευτεί» το σύστημα να σκέφτεται, δηλαδή να ξεχωρίζει, να αναγνωρίζει και να δημιουργεί ομαδοποιήσεις και κατηγοριοποιήσεις των δεδομένων [Meyer, D., Wien, F.T., 2014]. Είναι πολύ σημαντικό ότι είναι αλγόριθμοι με γενικότητα που να μπορεί να εφαρμοστούν σε πληθώρα προβλημάτων και έτσι δεν χρειάζεται να αναζητούνται νέοι αλγόριθμοι για κάθε νέο πρόβλημα. Εξασφαλίζεται στο σύστημα ευκαμψία και ελαστικότητα στις διάφορες απαιτήσεις που καλείται να αντιμετωπίσει. Επιπροσθέτως ένας γενικότερος αλγόριθμος παρουσιάζει πολλά πλεονεκτήματα. Εξασφαλίζει στατιστική σταθερότητα και υπολογιστική αποδοτικότητα. Δύναται δηλαδή να διαχειρίζεται πολυδιάστατα δεδομένα ακόμα και όταν συνυπάρχει θόρυβος ή μη ακριβή δεδομένα λόγω

λάθους. Επιπλέον οδηγεί πάντα, σχεδόν στα ίδια αποτελέσματα ακόμα και όταν τα δείγματα εισόδου κάθε φορά είναι διαφορετικά και τυχαία, φθάνει να προέρχονται από την ίδια πηγή. Τα δέντρα αποφάσεων (classification/decision trees) [Quinlan, J.R., 1986] είναι μια πολύ αποτελεσματική μέθοδος ανάλυσης της πληροφορίας [Rokach, L., 2007]. Με τη χρήση απλών συνθηκών ελέγχου μπορούν να εξαχθούν χρήσιμα συμπεράσματα χωρίς παράλληλα να είναι απαραίτητη η γραμμική σχέση της εξαρτημένης μεταβλητής (δηλαδή της μεταβλητής η οποία εξετάζεται) με τους λογικούς ελέγχους που εκτελούνται. Η δομή ενός δέντρου αποφάσεων είναι πολύ απλή. Κάθε κόμβος του δέντρου αναπαριστά ένα λογικό έλεγχο και αποτελεί μια ανεξάρτητη μεταβλητή (predictor). Κάθε κλαδί του δέντρου αναπαριστά το αποτέλεσμα ενός λογικού ελέγχου. Τέλος κάθε φύλλο (leaf) του δέντρου, δηλαδή κάθε κόμβος πέρα από τον οποίο δεν υπάρχουν άλλες διακλαδώσεις, αποτελεί μια τιμή (κλάση) της εξαρτημένης (predicted) μεταβλητής. Στο Σχήμα 5.1 παρουσιάζεται ένα παραδειγμα ενός δένδρου διαγράμματος αποφάσεων.



Σχήμα 5. 1 Παράδειγμα ενός δένδρου διαγράμματος αποφάσεων

Για τη δημιουργία ενός δένδρου αποφάσεων, αρχικά επιλέγεται μια ανεξάρτητη μεταβλητή που αποτελεί το κόμβο ρίζας. Στη συνέχεια, επιλέγονται η επόμενες ανεξάρτητες μεταβλητές (λογικοί έλεγχοι) έτσι ώστε τα υποσύνολα που δημιουργούνται να χωρίζονται με το βέλτιστο δυνατό τρόπο. Διάφορες μέθοδοι χρησιμοποιούνται για να προσδιορίσει ο βέλτιστος διαχωρισμός [Ben-Gal, I., Trister, C., 2014]. Μια από τις μεθόδους είναι ο υπολογισμός του κέρδους πληροφορίας (Information gain) [16, 26] που βασίζεται στην έννοια της εντροπίας όπως αναπτύχθηκε στη θεωρία της πληροφορίας. Συγκεκριμένα, αν ο διαχωρισμός οδηγεί σε νέους κόμβους με ίσο αριθμό κλάσεων τότε η αβεβαιότητα (εντροπία) είναι μέγιστη και η εξαγόμενη πληροφορία ελαχιστοποιείται. Αντίθετα, αν ο διαχωρισμός οδηγεί σε λίγες τιμές της εξαρτημένης μεταβλητής τότε ο διαχωρισμός αυτός είναι αποτελεσματικός και χρήσιμος για την εξαγωγή πληροφορίας. Όπως περιγράφηκε παραπάνω, η αμοιβαία πληροφορία $I(X,Y)$ εκφράζει την μείωση της αβεβαιότητας της μεταβλητής X λόγω της γνώσης της μεταβλητής Y . Κατά συνέπεια, η ποσότητα αυτή υποδεικνύει το αναμενόμενο κέρδος πληροφορίας. Αν θεωρήσουμε ότι X είναι η εξαρτημένη μεταβλητή και Y εκφράζει τις ανεξάρτητες μεταβλητές, επιλέγονται οι μεταβλητές Y οι οποίες συντελούν στο μεγαλύτερο κέρδος πληροφορίας. Το κέρδος πληροφορίας υπολογίζεται για τις μετρήσεις που βρίσκονται μόνο στο τελευταίο κόμβο. Έτσι αν θεωρήσουμε ότι έχουν επιλεγεί οι συνθήκες ελέγχου $Y_1=y_1$ και $Y_2=y_2$, η επόμενη συνθήκη ελέγχου Y_3 θα πρέπει να μεγιστοποιεί την ποσότητα:

$$I(X,Y/Y_1 = y_1, Y_2 = y_2) = H(X/Y_1 = y_1, Y_2 = y_2) - H(X/Y_1 = y_1, Y_2 = y_2, Y) \quad (5.12)$$

Στην οικογένεια των κατηγοριοποιητών διακρίνουμε διάφορες υποδιαίρεσεις με αντίστοιχα κριτήρια (γραμμικοί / μη - γραμμικοί, προσαρμοστικοί / μη - προσαρμοστικοί, με επίβλεψη / χωρίς επίβλεψη κ.τ.λ.).

Στη συνέχεια εξετάζονται εν συντομία οι κυριότερες λύσεις προκειμένου να επιλεγεί η καταλληλότερη για το πρόβλημά μας.

A. Γραμμική μέθοδος. Τα προ-αναφερόμενα διανύσματα χαρακτηριστικών τα οποία χρησιμοποιούνται ως διανύσματα εισόδου στον κατηγοριοποιητή συσχετίζονται με τα διανύσματα εξόδου, τα οποία προκύπτουν μετά τη διαδικασία εκμάθησης των εισαγομένων διανυσμάτων από ένα επιλεγμένο σύστημα κατηγοριοποίησης, με γραμμικό τρόπο. Ειδικότερα, αν θεωρηθούν x_i τα διανύσματα εισόδου και y_j τα διανύσματα εξόδου και μ_{ij} είναι τα διανύσματα διαχωρισμού ανά κατηγορία των διανυσμάτων εισόδου (βάρη), τότε η γραμμική συσχέτιση των διανυσμάτων x_i, y_i δίνεται από την κάτωθι σχέση:

$$y_j = \sum_{i=1}^k \mu_{ij} x_i \quad (5.13)$$

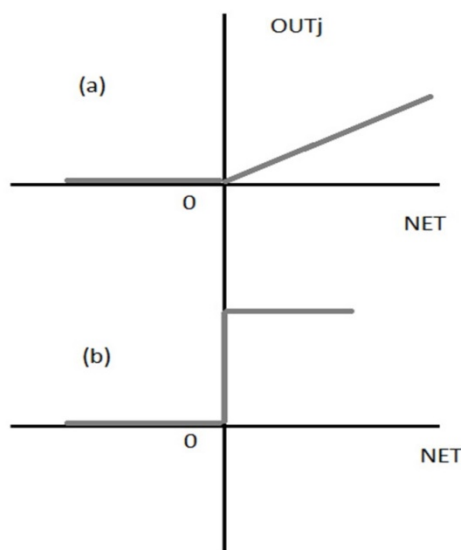
Για καλύτερη κατανόηση της μεθόδου θεωρείται ένα πολύ απλό σύστημα κατηγοριοποιητή που η βασική του ιδέα είναι η θεώρηση του αμφιβληστροειδής ως δισδιάστατης διάταξης φωτοαισθητήρων [Ρίζος, 1996]. Τα φωτοαισθητήρια συνδέονται πλήρως ή τυχαία με αμετάβλητα βάρη w_{ij} με ανιχνευτές χαρακτηριστικών (Τεχνητοί Νευρώνες, (TN)) που έχουν την ικανότητα να αναγνωρίζουν κάποια συγκεκριμένα πρότυπα. Για να δραστηριοποιηθεί κάποιος TN πρέπει το σταθμισμένο άθροισμα εισόδου να είναι ως εξής (μετασχηματισμός της εξίσωσης 5.14)

$$NET_j = \sum_{i=1}^k w_{ij} x_i \geq 0 \quad (5.14)$$

όπου i, j : είναι ο αριθμός των εισόδων και των εξόδων αντίστοιχα.

Ακόμη υπάρχει το επίπεδο εξόδου που αποτελείται από TN που λειτουργούν ως "ανιχνευτές προτύπων" και συνδέονται με μεταβλητά βάρη και τυχαίο τρόπο με κάποιους από τους ανιχνευτές χαρακτηριστικών. Ως συνάρτηση μεταφοράς χρησιμοποιείται η γραμμική συνάρτηση κατωφλίου (**Σχήμα 5.2**). Αυτά τα χαρακτηριστικά του κατηγοριοποιητή προέρχονται από φυσιολογικές παρατηρήσεις. Έτσι οι νευρώνες ενεργοποιούνται μόνον όταν η δραστηριότητα εισόδου υπερβεί κάποιο όριο.

Στη συνέχεια ο ρυθμός πυροδότησης αυξάνει ανάλογα με την αύξηση της εισόδου. Εναλλακτικά μπορούμε να χρησιμοποιήσουμε μια βηματική συνάρτηση μεταφοράς, η οποία κβαντίζει την έξοδο του TN στην τιμή 0, αν το σταθμισμένο άθροισμα εισόδου $NET_j \leq 0$ ή στην τιμή 1, αν το σταθμισμένο άθροισμα εισόδου $NET_j \geq 0$ (**σχήμα 5.2**).



Σχήμα 5. 2. Γραμμική συνάρτηση κατωφλίου που χρησιμοποιείται ως συνάρτηση μεταφοράς για τους ανιχνευτές χαρακτηριστικών και τους ανιχνευτές προτύπων. Η αύξηση της δραστηριοποίησης από κάποιο σημείο και μετά είναι γραμμική (A). Μία εναλλακτική συνάρτηση

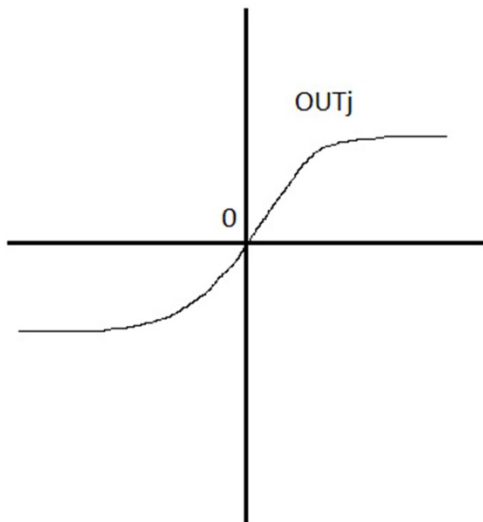
B. Μη γραμμική μέθοδος. Τα διανύσματα εξόδου τα οποία προέκυψαν από μία γραμμική διεργασία, μεταφέρουν συνήθως κάποιες ανεπιθύμητες συνιστώσες. Η απόρριψη των συνιστωσών αυτών γίνεται με την ενσωμάτωση στη γραμμική εξίσωση ενός μη γραμμικού τμήματος $\gamma_i(y_j)$ που είναι παρόμοιο με ένα βαθμωτό πεδίο. Κατόπιν τούτου η εξίσωση τροποποιείται δια της μη γραμμικής ολοκλήρωσής της, ως ακολούθως:

$$\frac{dy_i}{dt} = y_i - \gamma_i(y_j) \quad (5.15)$$

Για καλύτερη κατανόηση της μη γραμμικής μεθόδου αναφέρεται η περίπτωση του σιγμοειδούς μοντέλου (εικόνα 11). Το σιγμοειδές μοντέλο είναι αυτό που χρησιμοποιείται περισσότερο στην πράξη. Ονομάστηκε έτσι από τη μορφή της μη γραμμικότητάς του. Μία δημοφιλής επιλογή είναι η λογιστική συνάρτηση:

$$f(u) = \frac{1}{1 + e^{-au}} \quad (5.16)$$

Καθώς το $a \rightarrow +\infty$, το σιγμοειδές μοντέλο μετατρέπεται σε γραμμικό μοντέλο (βλ σχήμα 5.3).



Σχήμα 5.3. Μη γραμμική συνάρτηση καταωφλίου που χρησιμοποιείται ως συνάρτηση μεταφοράς για τους ανιχνευτές χαρακτηριστικών και τους ανιχνευτές προτύπων (σιγμοειδής συνάρτηση).

5.3.2 Μέθοδοι με επίβλεψη και χωρίς επίβλεψη

Αν θεωρηθεί ότι τα διανύσματα εισόδου (εδώ διανύσματα επεξεργασίας ΗΕΓ) αποτελούνται από d μεταβλητές, δηλαδή ο υπερχώρος των διανυσμάτων εισόδου έχει διάσταση d και κάθε μεταβλητή λαμβάνει N διακριτές τιμές, τότε ο υπόχωρος των διανυσμάτων αποτελείται από N^d διανύσματα εισόδου που λαμβάνουν N διακριτές τιμές. Μπορούμε να φανταστούμε ότι με την εκπαίδευση το σύστημα κατηγοριοποίησης σχηματίζει υποθέσεις για το ποίον της υποκείμενης συνάρτησης εισόδου / εξόδου. Τα εκπαιδευτικά διανύσματα εισόδου είναι τα "στοιχεία" που βοηθούν στην ανακάλυψη της σωστής συνάρτησης. Βέβαια όσο πιο μεγάλο είναι το μέγεθος του συστήματος κατηγοριοποίησης τόσο περισσότερες συναρτήσεις μπορεί να υλοποιήσει και επομένως αυξάνει η πιθανότητα η επιθυμητή συνάρτηση να βρίσκεται σε αυτό το σύνολο συναρτήσεων.

Αν λάβουμε υπόψη ότι το μέγεθος του συστήματος κατηγοριοποίησης των ΗΕΓραφικών χαρακτηριστικών διανυσμάτων είναι αρκετά μεγάλο ώστε να συμπεριλαμβάνει την επιθυμητή συνάρτηση, τότε μπορούμε να θεωρήσουμε ότι τα εκπαιδευτικά διανύσματα εισόδου μας δίνουν τη δυνατότητα απόρριψης των ανεπιθύμητων συναρτήσεων. Όσο περισσότερα εκπαιδευτικά διανύσματα διαθέτουμε τόσο

περισσότερες συναρτήσεις μπορούμε να απορρίψουμε. Αν όμως το μέγεθος του συστήματος κατηγοριοποίησης δεν είναι αρκετά μεγάλο τότε το σύστημα κατηγοριοποίησης αναζητεί τη συνάρτηση που ταιριάζει καλύτερα στην επιθυμητή.

Υπάρχει μια πληθώρα αλγορίθμων οι οποίοι χρησιμοποιούνται για την "εκπαίδευση" των συστημάτων κατηγοριοποίησης. Στην πράξη η εκπαίδευση κάποιου συστήματος κατηγοριοποίησης ισοδυναμεί με την ενδυνάμωση ή αποδυνάμωση των τιμών των βαρών (διανύσματα διαχωρισμού των διανυσμάτων εισόδου ανά κατηγορία προέλευσης), σύμφωνα με τους κανόνες που ορίζει ο αλγόριθμος εκπαίδευσης, μέχρι που τα βάρη να συγκλίνουν στις καταλληλότερες τιμές. Οι μεθοδολογίες εκπαίδευσης που χρησιμοποιούνται είναι:

- Με εποπτεία από εξωτερική πηγή ή με επίβλεψη (supervised).
- Χωρίς εποπτεία ή χωρίς επίβλεψη (unsupervised).

Κατά την *εκπαίδευση με επίβλεψη* τα ζευγάρια διανυσμάτων εισόδου - εξόδου δημιουργούν σχέση όσο το δυνατόν πιο ικανοποιητική σύμφωνα με την αργiori κατάταξη τους από τον επιβλέποντα κατηγοριοποιητή. Πιο συγκεκριμένα, κάποιος αλγόριθμος καθοδήγησης από εξωτερική πηγή ακολουθεί μια επαναληπτική διαδικασία βήμα προς βήμα σύμφωνα με την οποία τροποποιούνται οι τιμές των βαρών. Η επαναληπτική διαδικασία σταματάει όταν ελαχιστοποιηθεί κάποια συνάρτηση σφάλματος, η οποία δίνει το μέτρο της διαφοράς της πραγματικής εξόδου από την επιθυμητή έξοδο του συστήματος. Το βασικό μειονέκτημα της μεθόδου είναι η πιθανότητα παγίδευσης του συστήματος σε κάποιο τοπικό ελάχιστο. Κλασικό παράδειγμα συστήματος εκπαίδευσης με επίβλεψη εκτελεί το νευρωνικό δίκτυο Learning Vector Quantization (LVQ).

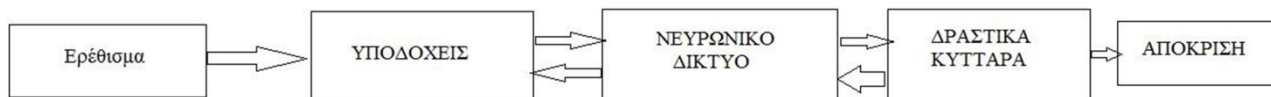
Κατά την *εκπαίδευση χωρίς επίβλεψη* παρουσιάζονται στο σύστημα κατηγοριοποίησης μόνο τα διανύσματα εισόδου τα οποία ταξινομούνται σε κατηγορίες ανάλογα με τη σχετική ομοιότητα τους (αυτο-οργάνωση του συστήματος κατηγοριοποίησης). Πιο συγκεκριμένα κάποιος αλγόριθμος χωρίς επίβλεψη πραγματοποιεί ψευδοτυχαίες μεταβολές των τιμών των βαρών. Από τις μεταβολές διατηρούνται μόνο αυτές που βελτιώνουν (ελαχιστοποιούν) τη συνάρτηση σφάλματος. Το σημαντικότερο μειονέκτημα αυτής της μεθόδου είναι ο μεγάλος χρόνος που απαιτείται για την εύρεση του ολικού ελαχίστου.

Συμπερασματικά, μία από τις σημαντικές διαφορές των δύο μεθόδων είναι ότι ενώ η μέθοδος με επίβλεψη χρειάζεται εκ των προτέρων την πληροφορία της οργάνωσης των διανυσμάτων σε ομάδες ομοιότητας, η μέθοδος χωρίς επίβλεψη χρησιμοποιεί, κατά το στάδιο εκπαίδευσης των διανυσμάτων εισόδου, κατηγορίες χωρίς να είναι γνωστή η πληροφορία προέλευσής τους. Χαρακτηριστικά παραδείγματα της κατηγορίας εκπαίδευσης χωρίς επίβλεψη είναι η μέθοδος εκπαίδευσης του νευρωνικού δικτύου Self Organising Map (SOM), ως επίσης και η ανάλυση τύπου clustering.

5.4 Τεχνητά Νευρωνικά Δίκτυα

5.4.1 Ο Ανθρώπινος Εγκέφαλος

Το ανθρώπινο νευρικό σύστημα μπορεί να αντιμετωπιστεί ως ένα σύστημα τριών σταδίων, όπως απεικονίζεται στο σχηματικό διάγραμμα του Σχ. 5.4 (Arbib, 1987). Το κέντρο του συστήματος είναι ο εγκέφαλος, ο οποίος αναπαριστάται από το νευρωνικό (νευρικό) δίκτυο αυτό προσλαμβάνει διαρκώς πληροφορίες, τις



Σχήμα 5.4. Σχηματική αναπαράσταση του νευρικού συστήματος.

επεξεργάζεται και λαμβάνει τις κατάλληλες αποφάσεις. Στο σχήμα παρουσιάζονται δύο ομάδες βελών. Αυτά με κατεύθυνση από αριστερά προς τα δεξιά υποδεικνύουν την μετάδοση των σημάτων πληροφορίας προς τα εμπρός (πρόσθια τροφοδότηση του συστήματος). Αυτά με κατεύθυνση από δεξιά προς τα αριστερά σηματοδοτούν την παρουσία ανάδρασης (feedback) στο σύστημα. Οι υποδοχείς μετατρέπουν τα ερεθίσματα που προέρχονται από το ανθρώπινο σώμα ή το εξωτερικό περιβάλλον σε ηλεκτρικά σήματα (ώσεις) που μεταφέρουν πληροφορία στο νευρικό δίκτυο (τον εγκέφαλο). Τα αποτελεσματικά τα ενεργής δραστηριότητας κύτταρα μετατρέπουν τα ηλεκτρικά σήματα που δημιουργούνται από το νευρικό δίκτυο σε αισθητές

αποκρίσεις οι οποίοι είναι οι έξοδοι του συστήματος. Ο αγώνας για την κατανόηση του εγκεφάλου έχει γίνει ευκολότερος χάρη στο πρωτοποριακό έργο του Ramon y Cajal (1911), ο οποίος εισήγαγε την ιδέα των νευρώνων ως δομικά συστατικά του εγκεφάλου. Εκτιμάται ότι οι νευρώνες είναι πέντε έως έξι τάξεις μεγέθους αργότεροι από τις λογικές πύλες που υλοποιούνται με τεχνολογία πυριτίου. Σ' ένα ολοκληρωμένο κύκλωμα τα συμβάντα λαμβάνουν χώρα στην κλίμακα των νανοδευτερολέπτων, ενώ στο ανθρώπινο νευρικό δίκτυο πραγματοποιούνται στην κλίμακα των χιλιοστών του δευτερολέπτου. Ωστόσο, ο εγκεφάλος υποκαθιστά το σχετικά αργό ρυθμό λειτουργίας ενός νευρώνα με την δημιουργία ενός τεράστιου πλήθους νευρώνων (νευρικών κυττάρων), με εξίσου μεγάλο πλήθος διασυνδέσεων μεταξύ τους. Έχει υπολογισθεί ότι υπάρχουν περίπου 10 δισεκατομμύρια νευρώνες στον εγκεφαλικό φλοιό του ανθρώπου και 60 τρισεκατομμύρια συνάψεις ή συνδέσεις (Shepherd και Koch, 1990). Το τελικό αποτέλεσμα είναι ότι ο εγκεφάλος αποτελεί μια εκπληκτικά αποτελεσματική και αποδοτική δομή. Συγκεκριμένα, η ενεργειακή αποδοτικότητα του εγκεφάλου είναι περίπου 106 joules (J) ανά λειτουργία ανά δευτερόλεπτο, ενώ η αντίστοιχη τιμή για τους καλύτερους υπολογιστές είναι τάξεις μεγέθους μεγαλύτερη. Οι συνάψεις, ή νευρικές απολήξεις, είναι οι στοιχειώδεις δομικές και λειτουργικές μονάδες που παίζουν διαμεσολαβητικό ρόλο κατά τις αλληλεπιδράσεις μεταξύ των νευρώνων. Το πιο κοινό είδος σύναψης είναι η χημική σύναψη, η οποία λειτουργεί ως εξής: Μια προσυναπτική διεργασία απελευθερώνει μια χημική ουσία (διαβιβαστής) η οποία διαχέεται στη συνοπτική ένωση μεταξύ των νευρώνων και κατόπιν επενεργεί σε μια μετασυναπτική διεργασία. Έτσι, μια σύναψη μετατρέπει ένα προσυναπτικό ηλεκτρικό σήμα σε ένα χημικό σήμα και κατόπιν μετατρέπει ξανά το χημικό σήμα σε μετασυναπτικό ηλεκτρικό σήμα (Shepherd και Koch, 1990). Με όρους της ηλεκτρονικής, ένα τέτοιο στοιχείο λέγεται ότι είναι ένα δίθυρο στοιχείο. Στις «παραδοσιακές» περιγραφές της οργάνωσης του νευρικού συστήματος, θεωρείται ότι μία σύναψη είναι μια απλή σύνδεση που μπορεί να προκαλέσει διέγερση ή αναστολή, αλλά όχι και τα δύο στο νευρώνα-αποδέκτη (Haykin, S.S., 2009).

Παραπάνω αναφέραμε ότι η πλαστικότητα επιτρέπει στο εξελισσόμενο νευρικό σύστημα να προσαρμόζεται στο περιβάλλον του (Eggermont, 1990, Churchland και Sejnowski, 1992). Στον εγκεφαλο ενός ενήλικα, η πλαστικότητα μπορεί να αποδοθεί σε δύο μηχανισμούς: τη δημιουργία νέων συναπτικών συνδέσεων μεταξύ νευρώνων και την τροποποίηση των υφιστάμενων συνάψεων. Οι άξονες (οι γραμμές μετάδοσης) και οι δενδρίτες (οι ζώνες αποδεκτών) είναι δύο είδη μορφολογικά διαφορετικών νηματίων των κυττάρων ένας άξονας έχει πιο ομαλή επιφάνεια, λιγότερους κλάδους και μεγαλύτερο μήκος, ενώ ένας δενδρίτης έχει ακανόνιστη επιφάνεια και περισσότερους κλάδους (Freeman, 1975). Οι νευρώνες που υπάρχουν σε διάφορα μέρη του εγκεφάλου ποικίλουν σε μέγεθος και σχήμα. Όμοια με πολλούς άλλους τύπους νευρώνων, λαμβάνει τις περισσότερες εισόδους του από τις δενδριτικές άκανθες. Το πυραμοειδές κύτταρο μπορεί να δέχεται 10,000 ή περισσότερες συναπτικές επαφές και μπορεί να τις προβάλλει σε χιλιάδες κύτταρα-αποδέκτες (Haykin, S.S., 2009)..

Στην πλειονότητά τους, οι νευρώνες κωδικοποιούν τις εξόδους τους σαν μια σειρά παλμών τάσης σύντομης διάρκειας. Αυτοί οι παλμοί, ευρέως γνωστοί ως δυναμικά δράσης (action potentials, spikes) 3, ξεκινούν από (ή κοντά στο) κυτταρικό σώμα των νευρώνων και κατόπιν διαδίδονται, διαμέσου μεμονωμένων νευρώνων, με σταθερή ταχύτητα και πλάτος σήματος. Οι λόγοι χρήσης των δυναμικών δράσης για την επικοινωνία μεταξύ των νευρώνων βασίζονται στη φυσική των αξόνων. Ο άξονας ενός νευρώνα είναι πολύ μακρύς και λεπτός και χαρακτηρίζεται από υψηλή ηλεκτρική αντίσταση και πολύ μεγάλη χωρητικότητα. Αυτά τα δύο στοιχεία κατανέμονται σε όλη την έκταση του άξονα. Άρα, ο άξονας μπορεί να μοντελοποιηθεί ως μια γραμμή μετάδοσης (υλοποιημένη με μια αντίσταση και μια χωρητικότητα) και σ' αυτό οφείλεται η κοινή ορολογία που χρησιμοποιείται για την περιγραφή της μετάδοσης σήματος και στις δύο περιπτώσεις. Η ανάλυση αυτού του μηχανισμού μετάδοσης αποκαλύπτει ότι όταν εφαρμόζεται μια τάση (δυναμικό) στο ένα άκρο του άξονα, η τιμή της εξασθενεί εκθετικά με την απόσταση, φτάνοντας σε αμελητέα στάθμη στο άλλο άκρο. Τα δυναμικά δράσης παρέχουν έναν τρόπο παράκαμψης αυτού του προβλήματος κατά τη μετάδοση των σημάτων (Anderson, 1995). Στον εγκεφαλο, υπάρχουν ανατομικές οργανώσεις τόσο σε μικρή όσο και σε μεγάλη κλίμακα- διαφορετικές λειτουργίες λαμβάνουν χώρα στα χαμηλότερα και υψηλότερα επίπεδα αυτών των οργανώσεων. Το Σχ. 5.5 παρουσιάζει μια ιεραρχία των επιπέδων οργάνωσης του εγκεφάλου, όπως προκύπτει από την λεπτομερή ανάλυση των τοπικών περιοχών του εγκεφάλου (Shepherd και Koch, 1990- Churchland και Sejnowski, 1992). Οι συνάψεις αντιπροσωπεύουν το πλέον θεμελιώδες επίπεδο και βασίζονται σε συγκεκριμένα μόρια και ιόντα για τη δράση τους. Στα επόμενα επίπεδα της ιεραρχίας έχουμε νευρωνικά «μικροκύκλωματα», δενδριτικά δέντρα και κατόπιν νευρώνες. Ένα νευρωνικό μικροκύκλωμα αναφέρεται σε μια ομάδα συνάψεων οργανωμένων σε κατάλληλα μοτίβα διασυνδέσεων για την παραγωγή μιας ορισμένης λειτουργίας. Είναι δυνατόν να προσομοιωθεί ένα νευρωνικό μικροκύκλωμα με ένα ολοκληρωμένο κύκλωμα τεχνολογίας πυριτίου, αποτελούμενο από ένα σύνολο διασυνδεδεμένων τρανζίστορ.

Το μέγεθος των μικροκυκλωμάτων μετρίεται στην κλίμακα των μικρόμετρων και η ταχύτητα λειτουργίας τους στην κλίμακα των χιλιοστών του δευτερολέπτου. Τα νευρωνικά μικροκυκλώματα ομαδοποιούνται και σχηματίζουν υπομονάδες δενδριτών μέσα στα δενδριτικά δέντρα μεμονωμένων νευρώνων. Ένας ολόκληρος νευρώνας, με μέγεθος περίπου 100 μμ, περιέχει αρκετές υπομονάδες δενδριτών.

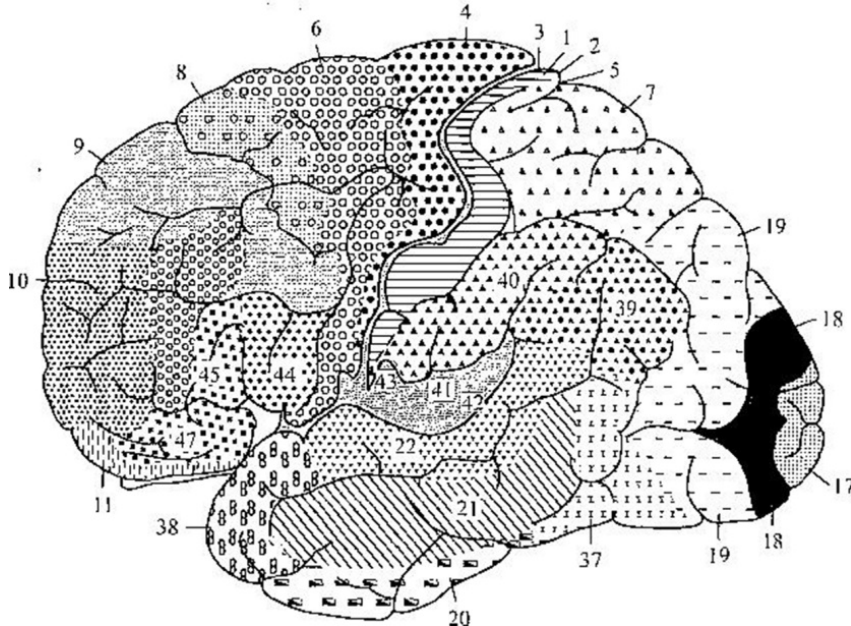
Στο επόμενο επίπεδο πολυπλοκότητας έχουμε *τοπικά κυκλώματα* (με μέγεθος περίπου 1 mm), αποτελούμενα από νευρώνες με όμοιες ή διαφορετικές ιδιότητες· αυτές οι ομάδες νευρώνων εκτελούν λειτουργίες που είναι χαρακτηριστικές για μια συγκεκριμένη περιοχή του εγκεφάλου. Ακολουθούνται από διαπεριφερειακά κυκλώματα

που εκτείνονται σε περισσότερες της μιας περιοχές και αποτελούνται από «διαδρομές», στήλες και τοπογραφικούς χάρτες που εμπλέκουν πολλαπλές περιοχές από διαφορετικά μέρη του εγκεφάλου.

Οι *τοπογραφικοί χάρτες* είναι οργανωμένοι με τρόπο ώστε να αντιδρούν στην εισερχόμενη αισθητηριακή πληροφορία. Συνήθως είναι διευθετημένοι σε στρώσεις, όπως στο βραχίονα του άνω διδυμίου, όπου οι χάρτες του οπτικού, ακουστικού και σωματοαισθητικού φλοιού του εγκεφάλου είναι τοποθετημένοι σε διαδοχικές στρώσεις με τρόπο ώστε τα ερεθίσματα που προέρχονται από αντίστοιχα σημεία του χώρου να βρίσκονται το ένα πάνω ή κάτω από το άλλο.

Το Σχήμα 5.5 παρουσιάζει ένα χάρτη της κυτταρικής αρχιτεκτονικής του εγκεφαλικού φλοιού, όπως αποτυπώθηκε από τον Brodmann (Brodal, 1981). Το σχήμα εμφανίζεται ότι διαφορετικές αισθητήριες εισοδοί χαρτογραφούνται σε αντίστοιχες περιοχές του εγκεφαλικού φλοιού με καθορισμένο τρόπο. Στο τελικό επίπεδο πολυπλοκότητας, οι τοπογραφικοί χάρτες και άλλα διαπεριφερειακά κυκλώματα μεταβιβάζουν συγκεκριμένους τύπους συμπεριφοράς στο *κεντρικό νευρικό σύστημα*.

Είναι σημαντικό να αναγνωρίσουμε ότι τα δομικά επίπεδα οργάνωσης που περιγράφονται εδώ είναι ένα μοναδικό χαρακτηριστικό του εγκεφάλου. Δεν υφίστανται στους ψηφιακούς υπολογιστές και δεν μπορούν ούτε καν να προσεγγιστούν με τεχνητά νευρωνικά δίκτυα. Ωστόσο, προχωράμε, αν και αργά, προς μια ιεραρχία υπολογιστικών επιπέδων όμοια με αυτή που περιγράφεται στο Σχ. 5.4. Οι τεχνητοί νευρώνες που χρησιμοποιούμε για την κατασκευή νευρωνικών δικτύων είναι κυριολεκτικά πρωτόγονοι συγκριτικά με αυτούς του εγκεφάλου μας. Τα νευρωνικά δίκτυα που είμαστε επί του παρόντος σε θέση να σχεδιάσουμε είναι εξίσου πρωτόγονα σε σχέση με τα τοπικά και διαπεριφερειακά κυκλώματα του εγκεφάλου. Ωστόσο, έχουμε κάνει σημαντική πρόοδο σε πολλά μέτωπα. Με πηγή έμπνευσης την αναλογία με τη νευροφυσιολογία του εγκεφάλου και τη μεγάλη γκάμα θεωρητικών και υπολογιστικών εργαλείων που έχουμε στη διάθεσή μας σήμερα, είναι σίγουρο ότι οι γνώσεις μας πάνω στα τεχνητά νευρωνικά δίκτυα και τις εφαρμογές τους θα συνεχίζουν να αυξάνονται, τόσο σε βάθος όσο και σε εύρος (Haykin, S.S., 2009).



Σχήμα 5. 5. Η αρχιτεκτονική του εγκεφαλικού φλοιού σε κυτταρικό επίπεδο. Οι διάφορες περιοχές ταυτοποιούνται από πάχος των στρώσεων τους και τα είδη των κυττάρων που περιέχουν. Ορισμένες από τις βασικές αισθητηριακές περιοχές είναι οι εξής: Κινητικός φλοιό

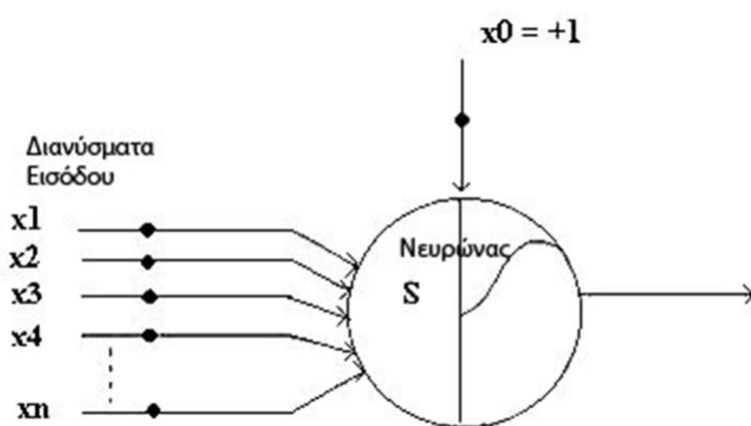
5.4.2 Μοντέλα Νευρώνων

Ένας νευρώνας είναι μια μονάδα επεξεργασίας πληροφορίας, η οποία είναι θεμελιώδης για τη λειτουργία ενός νευρωνικού δικτύου. Το σχηματικό διάγραμμα του Σχ. 5.6 παρουσιάζει το μοντέλο ενός νευρώνα που αποτελεί τη βάση για τη σχεδίαση μιας μεγάλης οικογένειας νευρωνικών δικτύων, τα οποία μελετώνται σε επόμενα κεφάλαια. Εδώ, θα επισημάνουμε τρία βασικά στοιχεία αυτού του μοντέλου:

1. Ένα σύνολο συνάψεων (ή διασυνδέσεων), κάθε μία εκ των οποίων χαρακτηρίζεται από το δικό της βάρος, ή δύναμη. Συγκεκριμένα, ένα σήμα x_j στην είσοδο της σύναψης j που συνδέεται με το νευρώνα S πολλαπλασιάζεται επί το συνοπτικό βάρος W/g . Εδώ είναι σημαντικό να εξηγήσουμε τον τρόπο αναγραφής των δεικτών στο συμβολισμό του συναπτικού βάρους, w_{kj} . Ο πρώτος δείκτης στο w_{kj} αναφέρεται στον εν λόγω νευρώνα και ο δεύτερος δείκτης αναφέρεται στο άκρο εισόδου της σύναψης στην οποία αναφέρεται το βάρος. Ανόμοια με το βάρος μιας σύναψης στον ανθρώπινο εγκέφαλο, το συναπτικό βάρος ενός τεχνητού νευρώνα μπορεί να λαμβάνει και αρνητικές και θετικές τιμές.
2. Έναν συσσωρευτή (adder) για την πρόσθεση των σημάτων εισόδου, τα οποία σταθμίζονται από τα αντίστοιχα συναπτικά βάρη του νευρώνα - οι λειτουργίες που περιγράφονται εδώ συνιστούν έναν γραμμικό συνδυαστή (linear combiner).
3. Μια *συνάρτηση ενεργοποίησης (activation function)* για τον περιορισμό του πλάτους του σήματος εξόδου ενός νευρώνα. Η ενεργοποίησης υπό μορφή συνάρτησης γνωστή ως *συνάρτηση περιορισμού (squashing function)*, μειώνει το επιτρεπτό εύρος πλάτους του σήματος εξόδου σε κάποια συγκεκριμένη τιμή. Το κανονικοποιημένο πλήθος τιμών πλάτους της εξόδου ενός νευρώνα αποδίδεται ως μοναδιαίο κλειστό διάστημα, με τη μορφή $[0,1]$, ή $[-1,1]$.

Το Τεχνητό Νευρωνικό Δίκτυο (ΤΝΔ) είναι ένα σύστημα υπολογισμού εκτέλεσης προσομοιωμένων πράξεων αναλόγων με τα Πραγματικά Νευρωνικά Δίκτυα (ΠΝΔ). Μερικά από τα πλεονεκτήματα των ΤΝΔ είναι:

- Επεξεργασία με παράλληλο τρόπο.
- Ικανότητα Μνήμης.
- Δυνατότητα εκμάθησης τυχαίων συναρτήσεων εισόδου / εξόδου.
- Δυνατότητα προσαρμογής.
- Απλές μονάδες υπολογισμού (τεχνητοί νευρώνες, ΤΝ, βλ. σχήμα 5.6).



Σχήμα 5. 6. Σχηματική αναπαράσταση ενός Τεχνητού Νευρώνα ΤΝ με μεταβλητά βάρη (μαύρες βούλες) και μία σταθερή είσοδο (κατώφλι).

Ένα ΤΝΔ αποτελείται από ΤΝ που απεικονίζουν την υπολογιστική μηχανή του (Σχήμα 5.6). Οι ΤΝ είναι στρωματοποιημένοι σε επίπεδα, ή στιβάδες (layers). Συνήθως συναντάται ένα επίπεδο εισόδου (με ή χωρίς υπολογιστικές ικανότητες) όπου εισέρχεται το διάνυσμα εκπαίδευσης εισόδου σε ένα ή περισσότερα κρυφά επίπεδα όπου εκτελείται η (μη) γραμμική επεξεργασία των πληροφοριών και τέλος δημιουργείται ένα

επίπεδο εξόδου που διαθέτει υπολογιστική δυνατότητα και αποδίδει τα αποτελέσματα. Ο αριθμός των TN μπορεί να αποτελείται από μερικούς TN έως μερικές χιλιάδες TN.

Οι TN υλοποιούν την ίδια συνάρτηση μεταφοράς, χωρίς αυτό να αποτελεί κανόνα. Η σύνδεση των TN επιτυγχάνεται με «βάρη στάθμισης» για την κάθε έξοδο TN. Η αρχιτεκτονική καθώς και η συνάρτηση μεταφοράς που πραγματοποιεί κάθε TN μπορεί να διαφέρει. Επιπλέον εφαρμόζεται ένας αλγόριθμος εκπαίδευσης ο οποίος διαφοροποιεί τις τιμές των βαρών, έτσι ώστε το TND να πραγματοποιήσει την προσδοκώμενη απεικόνιση των διανυσμάτων εισόδου στα διανύσματα εξόδου. Όλα αυτά τα χαρακτηριστικά συνιστούν ένα μοντέλο TND. Μερικά γνωστά μοντέλα TND είναι τα ακόλουθα:

- Multilayer Perceptron (Back Error Propagation, BP)
- Counter Propagation (CP)
- Recurrent Back Propagation Network (RBP)
- Hopfield Network
- Radial Basis Function (RBF)
- Self Organizing Map (SOM)
- Learning Vector Quantizer (LVQ)
- Hamming Network (Ha)
- Probabilistic Neural Network (PNN)

Στην παρούσα ενότητα δοκιμάστηκαν δύο (2) από τα πιο δημοφιλή νευρωνικά δίκτυα, το LVQ και το RBF.

5.4.1 Learning Vector Quantizer (LVQ)

Το νευρωνικό δίκτυο τύπου LVQ προτάθηκε από τον Kohonen το 1990. Τυπικά το LVQ αποτελείται από το επίπεδο εισόδου, από το κρυφό επίπεδο που χωρίζεται σε ομάδες TN ανάλογα με τον αριθμό των κατηγοριών ταξινόμησης και από το επίπεδο εξόδου που διαθέτει τόσους TN όσες και οι κατηγορίες ταξινόμησης. Το επίπεδο εισόδου συνδέεται πλήρως και με μεταβλητά βάρη με το κρυφό επίπεδο. Οι TN κάποιας ομάδας του κρυφού επιπέδου συνδέονται μόνο με έναν TN του επιπέδου εξόδου, δηλαδή με τον TN που αντιπροσωπεύει την κατηγορία τους και μάλιστα με σταθερά αμετάβλητα βάρη (Σχήμα 5.7). Πριν από την εκπαίδευση όλα τα μεταβλητά βάρη έχουν μικρές τυχαίες τιμές. Ο αλγόριθμος εκπαίδευσης του LVQ διαθέτει:

- Στοιχεία εκπαίδευσης με επίβλεψη επειδή επιλέγουμε αυθαίρετα κάποιον TN του επιπέδου εξόδου για να εκπροσωπήσει μια κατηγορία.
- Ανταγωνιστικό τρόπο εκπαίδευσης επειδή οι TN του κρυφού επιπέδου που ανήκουν στην ίδια κατηγορία, ανταγωνίζονται μεταξύ τους για το ποιος θα αντιπροσωπεύσει το διάνυσμα εισόδου.

Στην παρούσα εργασία θεωρείται ότι το ΗΕΓφικό προ-επεξεργασμένο σήμα είναι το κάτωθι εκπαιδευτικό διάνυσμα εισόδου:

$$x = \{x_1, x_2 \dots x_N\} \quad (5.17)$$

Κατά την εκπαίδευση του LVQ υπολογίζουμε την Ευκλείδεια απόσταση d_i του διανύσματος εισόδου x από το διάνυσμα βαρών w_i του i κρυφού επιπέδου ως εξής:

$$d_i = \|x - w\| = \left[(x_1 - w_1)^2 + (x_2 - w_2)^2 + \dots + (x_N - w_N)^2 \right]^{1/2} \quad (5.18)$$

Με αυτό τον τρόπο επιλέγεται ο Τεχνητός Νευρώνας (TN) που έχει τη μικρότερη απόσταση από το διάνυσμα εισόδου. Αν ο "νικητής" ανήκει στη σωστή κατηγορία ταξινόμησης, τότε κάθε εισερχόμενο σε αυτόν βάρος ενδυναμώνεται σύμφωνα με τον εξής κανόνα:

$$w_{ij} = w_{ij} + a * (x_j - w_{ij}) \quad (5.19)$$

Το a είναι ο ρυθμός εκπαίδευσης που μπορεί να είναι μια συνάρτηση του χρόνου, συνήθως φθίνουσα. Αντίθετα αν ο TN που επιλέχθηκε ανήκει σε λανθασμένη κατηγορία ταξινόμησης, τότε κάθε εισερχόμενο σε αυτόν βάρος αποδυναμώνεται σύμφωνα με τον εξής κανόνα:

$$w_{ij} = w_{ij} - a * (x_j - w_{ij}) \quad (5.20)$$

Η διαδικασία που περιγράφεται από την εξίσωση 5.20 ονομάζεται "απόκρουση" (repulsion). Στο σημείο αυτό μπορεί να παρατηρήσουμε ότι αν το διάνυσμα των βαρών κάποιου TN τύχει να βρεθεί, στην αρχή της διαδικασίας εκπαίδευσης, κοντά στο διάνυσμα εισόδου, τότε αυτός ο TN επιλέγεται πιο συχνά από τους άλλους TN, επειδή τα διανύσματα των βαρών των άλλων TN βρίσκονται πιο μακριά από τα διανύσματα εισόδου και κατά συνέπεια αδρανούν.

Για καλύτερη κατανόηση της αρχιτεκτονικής του LVQ TNA περιγράφεται ένα μοντέλο (2 x 2). Συγκεκριμένα, το προαναφερόμενο LVQ αποτελείται από τα διανύσματα εισόδου [p(1), p(2)] διαστάσεως (2 x 2) και δύο κατηγορίες ταξινόμησης των διανυσμάτων εξόδου [a2(1), a2(2)]. Η βασική αρχιτεκτονική του συγκεκριμένου LVQ αποτελείται από δύο επίπεδα νευρώνων, ένα ανταγωνιστικό επίπεδο, και ένα γραμμικό επίπεδο. Οι ανταγωνιστικοί νευρώνες κατηγοριοποιούν τα διανύσματα εισόδου σε υποκατηγορίες, ενώ το γραμμικό επίπεδο ομαδοποιεί τις υποκατηγορίες στις αντίστοιχες κλάσεις εξόδου.

Λαμβάνοντας υπόψη τις προαναφερόμενες ιδιότητες ο επιλεγμένος τύπος TNA-LVQ προσφέρει τα κάτωθι πλεονεκτήματα εκμετάλλευσης των ιδιοτήτων που παρουσιάζει η ταξινόμηση των διανυσμάτων:

- Έχει την ικανότητα να κατηγοριοποιεί επιτυχημένα διανύσματα εισόδου μεγάλων διαστάσεων. Στη συγκεκριμένη περίπτωση, αντιμετωπίζει με επιτυχία τα διανύσματα που παράγει η μέθοδος επεξεργασίας, μέσω μετασχηματισμού Fourier, διότι είναι γνωστό ότι τα ΗΕΓφικά διανύσματα που προέρχονται από αυτή τη μέθοδο επεξεργασίας έχουν μεγάλη διάσταση, ώστε να υπάρχει ικανοποιητική ανάλυση στο πεδίο της συχνότητας.
- Έχει την ικανότητα να κατηγοριοποιεί τα διανύσματα εισόδου σε υποκατηγορίες που χωρίζονται μεταξύ τους με μη γραμμικό τρόπο. Αυτό είναι πολύ σημαντικό γιατί όπως προαναφέρθηκε και προηγούμενα αφαιρούνται με αυτό τον τρόπο οι ανεπιθύμητες συνιστώσες των διανυσμάτων εξόδου.

5.4.2 Νευρωνικό δίκτυο Τύπου Radial Basis Function (RBF)

Ένα άλλο δημοφιλές νευρωνικό δίκτυο που έχει επιλεγεί στην παρούσα διατριβή είναι το νευρωνικό δίκτυο RBF [Broomhead, 1988]. Η αρχιτεκτονική του νευρωνικού δικτύου RBF στηρίζεται σε τοπολογική διασταύρωση δύο μεθόδων εκπαίδευσης με και χωρίς επίβλεψη, (μη) εποπτείας. Η κατασκευή του αλγορίθμου στηρίζεται στη βασική μορφή τριών επιπέδων. Το επίπεδο εισόδου είναι κατασκευασμένο από νευρώνες εισόδου (αισθητήρες). Το δεύτερο επίπεδο είναι κρυφό και δύναται να δεχθεί αρκετά υψηλής διαστάσεως διανύσματα, κατάλληλα τροποποιημένα από το προηγούμενο επίπεδο. Τέλος το τρίτο επίπεδο αποτελεί το επίπεδο εξόδου και είναι κατασκευασμένο για να ανταποκρίνεται κατάλληλα στη δραστηριοποίηση του επιπέδου εισόδου. Ο μετασχηματισμός από το πρώτο επίπεδο εισόδου στο δεύτερο (κρυφό) επίπεδο είναι μη γραμμικός, ενώ ο μετασχηματισμός από το δεύτερο (κρυφό) επίπεδο στο επίπεδο εξόδου γίνεται με γραμμικό τρόπο. Η μαθηματική δικαιολογία της φύσης των προαναφερομένων μετασχηματισμών ανά επίπεδο παρέχεται από το θεώρημα του [Cover, 1965]. Συγκεκριμένα, θεωρείται μία οικογένεια από επιφάνειες, η κάθε μία από τις οποίες χωρίζεται σε δύο περιοχές (regions). Κάθε προ-

επεξεργασμένο διάνυσμα ΗΕΓραφήματος $x = \{x_1, x_2, \dots, x_N\}$ μπορεί να εκχωρηθεί σε μία από τις δύο κατηγορίες X^+ και X^- . Η διχοτόμος (binary partition) η οποία εκχωρεί το διάνυσμα σε μια από τις προαναφερόμενες επιφάνειες καθορίζεται από ένα σύνολο συναρτήσεων $\{\phi_i x \mid i = 1, 2, \dots, M\}$ οι οποίες περιγράφονται ως κάτωθι:

$$\phi(x) = [\phi_1(x), \phi_2(x), \dots, \phi_M(x)]^T \quad (5.21)$$

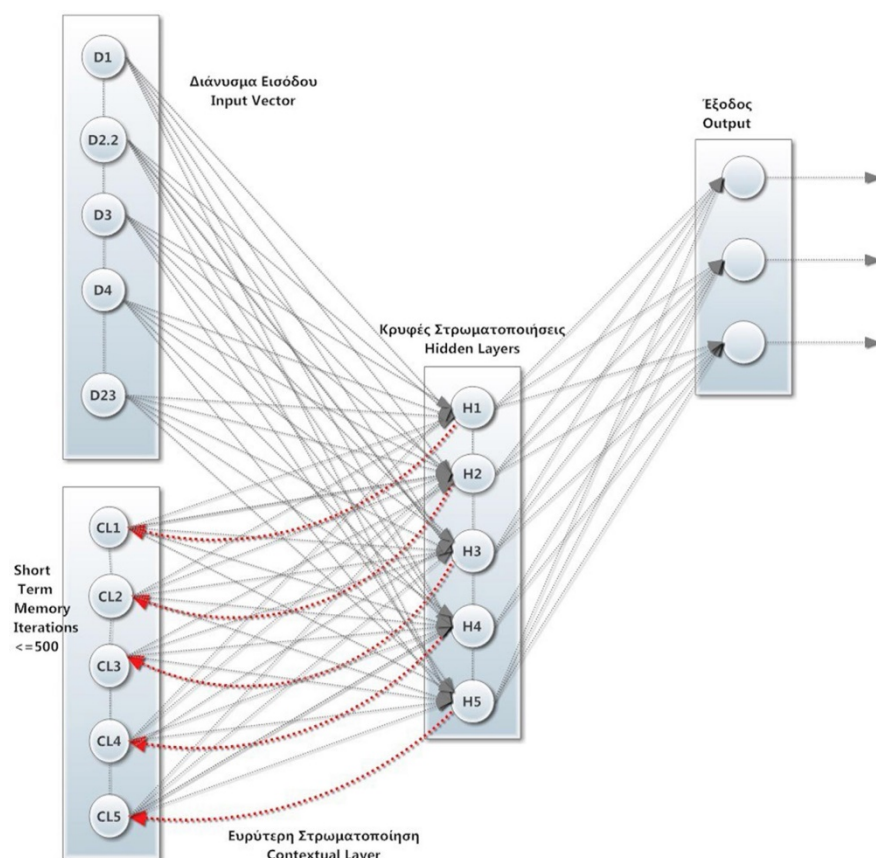
Στη συγκεκριμένη περίπτωση το σύνολο αυτών των συναρτήσεων είναι υπεύθυνο για το μη γραμμικό μετασχηματισμό μεταξύ των δύο πρώτων επιπέδων. Αν θεωρηθεί ένα διάνυσμα $\mathbf{w}$ διάστασης $(1 \times m)$ το οποίο προκύπτει από τον προηγούμενο μη γραμμικό μετασχηματισμό, τότε σύμφωνα με το θεώρημα Cover, η εκχώρηση του $\mathbf{w}$ διανύσματος σε μία από τις προαναφερόμενες κατηγορίες X^+ και X^- παρέχεται από τα κάτωθι κριτήρια:

$$\mathbf{w}^T \ddot{\phi}(\mathbf{x}) \geq 0, \mathbf{x} \in X^+ \quad (5.22)$$

ή

$$\mathbf{w}^T \ddot{\phi}(\mathbf{x}) \geq 0, \mathbf{x} \in X^+ \quad (5.23)$$

Σύμφωνα με τα ανωτέρω, η διαδικασία εκπαίδευσης ξεκινά με την εκπαίδευση των εισερχομένων διανυσμάτων (προ-επεξεργασμένα ΗΕΓραφήματα) στο πρώτο επίπεδο που έχει μη εποπτικές ιδιότητες. Αποτέλεσμα είναι ο υπολογισμός των κέντρων και των πλατών ορισμένων κανονικών (Gaussian) κατανομών. Αυτά τα κέντρα κωδικοποιούνται εντός των βαρών του χωρίς επίβλεψη επιπέδου χρησιμοποιώντας ανταγωνιστικό τρόπο εκμάθησης, κατά τη διάρκεια της οποίας ένας αλγόριθμος εκμάθησης χωρίς επίβλεψη υπολογίζει τα πλάτη χρησιμοποιώντας τα κέντρα των γειτονικών κόμβων. Στη συνέχεια το δεύτερο κρυφό επίπεδο συμπεριφέρεται ως εποπτευόν δεχόμενο την ταξινόμηση που έχει γίνει από το μη εποπτευόν πρώτο επίπεδο ως εξωτερική πηγή. Η διαδικασία συνεχίζεται με τη γραμμική επεξεργασία στο τρίτο επίπεδο με επίβλεψη όπου παράγεται η τελική πληροφορία ταξινόμησης στην έξοδο (Βλ. Σχήμα 5.7).



Σχήμα 5.7. Υβριδική μορφή ενός τυπικού LVQ νευρωνικού δικτύου.

Σύμφωνα με τα προηγούμενα, το RBF νευρωνικό δίκτυο έχει τις παρακάτω ιδιότητες:

- Το νευρωνικό δίκτυο RBF αποτελεί μία αποδοτική λύση δοκιμής ενός συστήματος με ιδιότητες εκμάθησης χωρίς επίβλεψη. Έτσι ο συνδυασμός του με το νευρωνικό δίκτυο LVQ προσφέρουν εναλλακτικές μέθοδοι εκπαίδευσης με ισοδύναμες πιθανότητες επιτυχίας.

- Το νευρωνικό δίκτυο RBF έχει τη δυνατότητα όπως και το νευρωνικό δίκτυο LVQ να κατηγοριοποιεί τα εισερχόμενα διανύσματα με μη γραμμικό τρόπο.

Από την άλλη πλευρά το νευρωνικό δίκτυο RBF παρουσιάζει προβλήματα όταν δέχεται διανύσματα εισόδου μεγάλης διάστασης, επειδή όταν αυτά έχουν διάσταση πολύ μεγαλύτερη από το βαθμό ελευθερίας του συστήματος εισόδου, τότε οι συναρτήσεις του νευρωνικού δικτύου RBF απομονώνονται μέχρι ο αριθμός τους να γίνει ίσος με τον αριθμό της διάστασης του διανύσματος εισόδου. Αυτό έχει ως αποτέλεσμα την περιορισμένη δυνατότητα του συστήματος (overdetermined problem). Κατά συνέπεια φαίνεται εκ πρώτης όψεως ανεδαφική η χρησιμοποίηση του νευρωνικού δικτύου RBF για την κατηγοριοποίηση διανυσμάτων προερχόμενων από μετασχηματισμό Fourier (μη παραμετρική επεξεργασία), διότι τότε παράγονται διανύσματα διαστάσεως πολύ μεγαλύτερης του βαθμού ελευθερίας του συστήματος.

Βιβλιογραφία

- Αγγελής, Β. και Δημάκη, Κ. (2011). *Στατιστική*. Τόμος Α: Στατιστική, Πιθανότητες, Στατιστική Συμπερασματολογία. Θεσσαλονίκη: Εκδόσεις Σοφία.
- Αδαμόπουλος, Λ., Δαμιανός, Χ. και Σβέρκος Α. (2014). *Μαθηματικά και στοιχεία στατιστικής*. Αθήνα: Οργανισμός Εκδόσεως διδακτικών βιβλίων.
- Anderson, J.E., Holcomb, P.J., 1995. Auditory and visual semantic priming using different stimulus onset asynchronies: An event-related brain potential study. *Psychophysiology* 32, 177–190.
- Ben-Gal, I., Trister, C., 2014. Parallel construction of decision trees with consistently non-increasing expected number of tests. *Applied Stochastic Models in Business and Industry*.
- Brodal, A., 1981. *Neurological anatomy in relation to clinical medicine*. Oxford University Press, USA.
- Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC press.
- Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1(1), 81–106.
- Cajal, S. R. (1911). Histogénèse du cervelet. In *Histologie du Système Nerveux de l'Homme et des Vertébrés* (Vol. 2, pp. 80–106). Instituto Ramón y Cajal del CSIC Madrid.
- Churchland, P.S., Sejnowski, T.J., 1992. *The computational brain*. The MIT press.
- Δαμιανού Χ., Παπαδάτος Ν. και Χαραλαμπίδης Χ. Α. (2003). *Εισαγωγή στις πιθανότητες και τη στατιστική: διδακτικές σημειώσεις*. Τμήμα Μαθηματικών Πανεπιστημίου Αθηνών, Αθήνα.
- Ding, Y. (2008). *Treating missing data in classification trees*. ProQuest. Retrieved from <https://www.google.com/books>
- Eggermont, J.J., 1990. *The correlative brain*. Springer.
- Εμβαλωτής, Α., Κατσης, Α. και Σιδερίδης Γ. (2006). *Στατιστική μεθοδολογία εκπαιδευτικής έρευνας*. Ιωάννινα: Πανεπιστήμιο Ιωαννίνων.
- Haykin, S.S., 2009. *Neural networks and learning machines*. Pearson Education Upper Saddle River.
- Hyafil, L., & Rivest, R. L. (1976). Constructing optimal binary decision trees is NP-complete. *Information Processing Letters*, 5(1), 15–17.
- Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, 78(9), 1464–1480.
- Lee, J.-Y. (2008). *A new approach of top-down induction of decision trees for knowledge discovery*. ProQuest. Retrieved from <https://www.google.com/books>
- Maimon, O. (2007). *Data Mining With Decision Trees: Theory and Applications*. Series in Machine Perception & Artificial Intelligence. World Scientific.
- Meyer, D., & Wien, F. T. (2014). Support vector machines. *The Interface to Libsvm in Package e1071*. Retrieved from <ftp://ftp.cse.yzu.edu.tw/CRAN/web/packages/e1071/vignettes/svmdoc.pdf>
- Rokach, L., 2007. *Data mining with decision trees: theory and applications*. World scientific.
- Murthy, S. K. (1998). Automatic construction of decision trees from data: A multi-disciplinary survey. *Data Mining and Knowledge Discovery*, 2(4), 345–389.
- Μπούτσικας Ν. Β. (2003). *Σημειώσεις στατιστικής II*. Τμήμα Οικονομικής Επιστήμης. Πανεπιστήμιο Πειραιώς. Retrieved 21/04/2015 from
- Quinlan, J.R., 1986. Induction of decision trees. *Machine learning* 1, 81–106
- Quinlan, J. R. (1987). Simplifying decision trees. *International Journal of Man-Machine Studies*, 27(3), 221–234.

- G. Runger, and E. Tuv, "Bias of importance measures for multi-valued attributes and solutions," presented at the Proceedings of the 21st international conference on Artificial neural networks - Volume Part II, Espoo, Finland, 2011.
- Shepherd, G.M., Koch, C., 1990. Dendritic electrotonus and synaptic integration. *The Synaptic Organization of the Brain* 439–473.
- Τσίμπος, Κ. και Γεωργιακώδης, Φ. (1999). *Περιγραφική και διερευνητική στατιστική ανάλυση δεδομένων*. 1. Αθήνα: Εκδόσεις Αθην. Σταμούλης.
- Yao, Y. Y. (2003). Information-theoretic measures for knowledge discovery and data mining. In *Entropy Measures, Maximum Entropy Principle and Emerging Applications* (pp. 115–136). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-3-540-36212-8_6
- Yuan, Y., & Shaw, M. J. (1995). Induction of fuzzy decision trees. *Fuzzy Sets and Systems*, 69(2), 125–139.
- Zar, J. H. (2010). *Biostatistical Analysis*. 4th ed. Upper Saddle River, NJ: Prentice Hall

6. Παραδείγματα εξαγωγής Γνώσης

Περίληψη

Στόχος κεφαλαίου: να παρουσιαστούν παραδείγματα εφαρμογής εξαγωγής γνώσης χρησιμοποιώντας και τη στατιστική και την συντακτική προσέγγιση που αναπτύχθηκε στα προηγούμενα κεφάλαια.

Προαπαιτούμενες γνώσεις

Υπολογιστική Γεωμετρία, Σχήματα, Στατιστική Ανάλυση, Matlab

6.1 Εμπειρικές Μέθοδοι Κατηγοριοποίησης/Ταυτοποίησης

Πέρα από τις προαναφερόμενες μεθόδους κατηγοριοποίησης έχουν κατά καιρούς προταθεί πλήθος μεθόδων που αντιμετωπίζουν με αυξημένη επιτυχία συγκεκριμένα προβλήματα χωρίς να στηρίζονται σε μία γενικευμένη μεθοδολογία. Οι μέθοδοι αυτές αναφέρονται ως εμπειρικές, σε αντιδιαστολή με τις προηγούμενες. Για το σκοπό της παρούσας εργασίας αναπτύχθηκε και εφαρμόστηκε μία τέτοια νέα μέθοδος κατηγοριοποίησης βασισμένη σε Αλγορίθμους Υπολογιστικής Γεωμετρίας (Computational Geometry Algorithms, CGA). Ο σκοπός της ανάπτυξης της μεθόδου αυτής είναι να εκμεταλλευτεί τις ιδιαιτερότητες των ιδιοτήτων ενός συνόλου δεδομένων ώστε να γίνει με μεγαλύτερη επιτυχία η κατηγοριοποίησή του.

6.1.1 Η θεμελιώδης αρχή της Μεθόδου

Η μέθοδος CGA δύναται να θεωρηθεί ως εμπειρική μέθοδος και στηρίζεται στην εμπειρική φιλοσοφική θεώρηση της απλούστευσης της αρχικής πληροφορίας που αναφέρεται στον Αριστοτέλη. Η θεμελιώδης αυτή φιλοσοφική θεώρηση υποστηρίζει την ελάττωση της αρχικής πληροφορίας σε διακριτά τμήματα, ώστε να διευκολυνθεί η μελέτη της.

Σύμφωνα με αυτή τη φιλοσοφική θεώρηση, στην παρούσα εργασία επιχειρείται να απομονωθεί ένα σύνολο τιμών που γεωμετρικά αποτελεί ένα γνήσιο υποσύνολο όλων των τιμών ενός συνόλου ΗΕΓραφημάτων του ιδίου ατόμου. Προς αυτή την κατεύθυνση θεωρείται ένα σύνολο n σημείων στο καρτεσιανό επίπεδο, έστω $A = \{a_1, a_2, a_3, \dots, a_n\}$.

Σε οποιανδήποτε από τις δύο περιπτώσεις τα χαρακτηριστικά διανύσματα εξάγονται με τη μεθοδολογία που αναπτύχθηκε στα κεφάλαια 4 & 5.

Στη συνέχεια ορίζεται ένα υποσύνολο σημείων X , επί του αρχικού συνόλου σημείων A , ($X \subseteq A$), το οποίο καθορίζει τις κορυφές ενός κυρτού πολυγώνου P στο καρτεσιανό επίπεδο, έτσι ώστε η επιφάνεια του πολυγώνου P να περικλείει όλα τα n σημεία (τιμές) του διανύσματος ή διαφορετικά το κυρτό σύνολο των τιμών X οριοθετεί μία κυρτή επιφάνεια που μέσα σε αυτή κινούνται οι τιμές των χαρακτηριστικών ΗΕΓραφικών διανυσμάτων (σύνολο A). Η οριοθέτηση αυτή είναι σημαντική διότι το πολύγωνο καθορίζει μία κλειστή περιοχή στην οποία περιλαμβάνονται όλα τα χαρακτηριστικά σημεία του κάθε διανύσματος.

Αν την ιδιότητα αυτή τη μεταφέρουμε στη σφαίρα της ασαφούς λογικής (fuzzy logic) τότε δύναται να θεωρηθεί ότι το σύνολο των κανονικοποιημένων σημείων A , τα οποία κινούνται εντός του διαστήματος $[0,1]$, απαρτίζει ένα κυρτό ασαφές σύνολο (convex fuzzy set), [Zadeh, 1996]. Επίσης θεωρείται ένα καθολικό σύνολο σημείων X . Τότε, η συνάρτηση μ_A από την οποία καθορίζεται το κυρτό ασαφές σύνολο A προσδιορίζεται από την κάτωθι μαθηματική σχέση:

$$\mu_A : X \rightarrow [0,1] \quad (6.1)$$

όπου $[0, 1]$ είναι το εύρος των τιμών που δέχεται το καθολικό σύνολο X .

Με τον ίδιο τρόπο, θεωρείται ένα δεύτερο σύνολο σημείων B , το οποίο ανάλογα του A , εάν κανονικοποιηθεί, μπορεί να θεωρηθεί ως ένα κυρτό ασαφές σύνολο, του οποίου η συνάρτηση καθορίζεται ως εξής :

$$\mu_B : X \rightarrow [0,1] \quad (6.2)$$

Η ετερογένεια μεταξύ των προαναφερομένων κυρτών συνόλων A, B σε σχέση με τη θέση τους στο Ευκλείδειο επίπεδο μπορεί να υπολογιστεί χρησιμοποιώντας το κλασικό θεώρημα διαχωρισμού των κυρτών συνόλων. Ο βαθμός διαχωρισμού D [Zadeh, 1996] για τα κυρτά σύνολα είναι:

$$D=1-M \quad (6.3)$$

όπου M είναι ο μέγιστος βαθμός τομής των δύο κυρτών συνόλων A και B :

$$M = \sup_X \min [\mu_A(X), \mu_B(X)] \quad (6.4)$$

Επίσης πρέπει να θεωρηθεί ότι η αναμενόμενη τιμή D είναι χαμηλότερη για κυρτά σύνολα (ΗΕΓραφημάτων) τα οποία ανήκουν στο ίδιο άτομο (περίπτωση 1) ενώ θεωρείται υψηλότερη για κυρτά σύνολα (ΗΕΓραφημάτων) τα οποία ανήκουν σε διαφορετικό άτομο (περίπτωση 2). Αυτή η θεώρηση ευσταθεί εφόσον αληθεύει η υπόθεση ότι τα κυρτά σύνολα ενός ατόμου έχουν κοινή τομή ("γενετική ιδιότητα").

Λαμβάνοντας υπόψη ότι τα προηγούμενα όρια των τιμών των βαθμών διαχωρισιμότητας (degrees of separability) (D) για τα A και B σύνολα είναι (0) και (1) αντίστοιχα, όπου το όριο μηδέν (0) αναπαριστά τις ειδικές περιπτώσεις τομής των συνόλων A, B όταν $A \cap B = A$ ή $A \cap B = B$, και το όριο ένα (1) αναπαριστά την ειδική περίπτωση τομής των συνόλων A, B όταν $A \cap B = \emptyset$. Τότε για n επαναλαμβανόμενες τομές, $A_1 \cap A_2, \dots, \cap A_n$ των $A_1, A_2, \dots, A_n$ κυρτών συνόλων διατυπώνεται η ακόλουθη πρόταση:

Πρόταση: Έστω $\{x_n\}$ η ακολουθία από $n-1$ τομές των $A_1, A_2, \dots, A_n$ κυρτών συνόλων, όπου $\{x_n\}$ είναι

$$\{x_n\} = \{A_1 \cap A_2, (A_1 \cap A_2) \cap A_3, \dots, (A_1 \cap A_2 \cap A_3 \cap \dots \cap A_{n-1}) \cap A_n\} \quad (6.5)$$

ή

$$\{x_n\} = \{m_1 \cap A_2, m_2 \cap A_3, \dots, m_{n-1} \cap A_n\},$$

$$\text{για } n \in \mathbb{N} \text{ και έστω } \{d_n\} = \{D_1, D_2, \dots, D_{n-1}\}$$

είναι η ακολουθία βαθμών διαχωρισιμότητας των αντίστοιχων ζευγών κυρτών συνόλων, m_i και A_j , για $i=1, \dots, n-1$, και $j=2, \dots, n$. Τότε η ακολουθία $\{d_n\}$ συγκλίνει στο 0 στην περίπτωση όπου τα ασαφή κυρτά σύνολα $[A_1, A_2, \dots, A_n]$ ανήκουν στο ίδιο άτομο. Στην περίπτωση που τα παραπάνω σύνολα ανήκουν σε μη συσχετισμένα άτομα, τότε η ακολουθία $\{d_n\}$ συγκλίνει στο 1.

Τέλος η ικανοποίηση της παραπάνω πρόταση οδηγεί στα εξής συμπεράσματα:

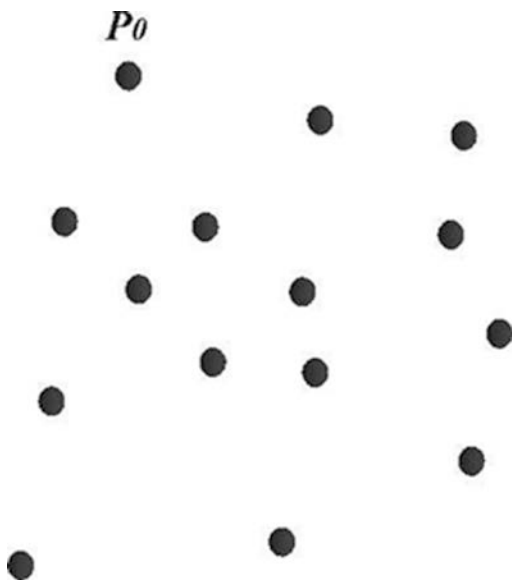
Έστω ότι όλες οι περιπτώσεις του βαθμού διαχωρισιμότητας μεταξύ των n τομών των συναρτήσεων από συσχετισμένα κυρτά ασαφή σύνολα δεν συγκλίνουν στο 0. Τότε η υπόθεση ότι όλα αυτά τα σύνολα μεταφέρουν κάποια κοινά, «γενετικά» χαρακτηριστικά απορρίπτεται και οδηγούμαστε στο συμπέρασμα ότι οι μέθοδοι (προ) επεξεργασίας δεν είναι αποτελεσματικές. Διαφορετικά η υπόθεση της πρότασης δεν απορρίπτεται.

6.1.2 Ανάπτυξη αλγορίθμου κατασκευής κυρτού πολυγώνου

Στην παρούσα παράγραφο περιγράφονται τα βήματα που πρέπει να ακολουθηθούν ώστε να υλοποιηθεί η πρόταση της προηγούμενης παραγράφου (3.4.4.1). Συγκεκριμένα, τα βήματα είναι τα ακόλουθα:

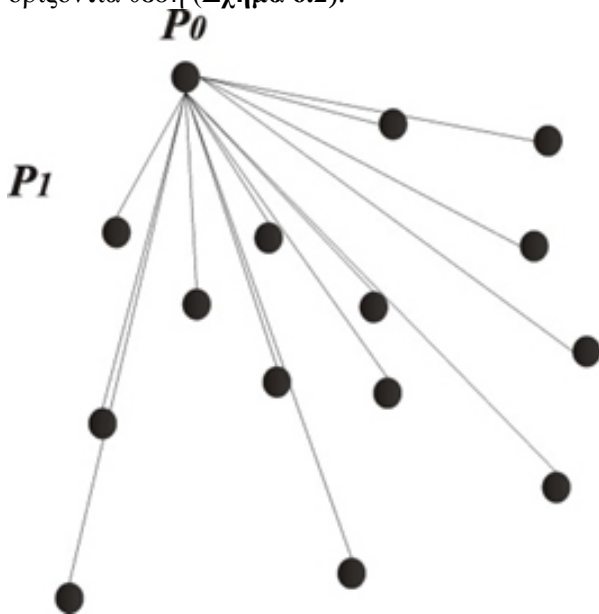
Κατασκευή κυρτού πολυγώνου από n σημεία στο καρτεσιανό επίπεδο. Το 1972 ο Graham, 1972 πρότεινε έναν αλγόριθμο πολυπλοκότητας $O(n \log n)$ ο οποίος υπολόγιζε το κυρτό κέλυφος από " n " δεδομένα σημεία στο Καρτεσιανό επίπεδο. Ο αλγόριθμος αυτός υλοποιήθηκε στα ακόλουθα τρία στάδια:

Στάδιο 1. Ευρίσκεται το σημείο το οποίο αντιπροσωπεύει την υψηλότερη τιμή των συντεταγμένων του άξονα y . Το σημείο αυτό (P_0 , **Σχήμα 6.1**) στη συνέχεια χρησιμοποιείται ως άξονας περιστροφής



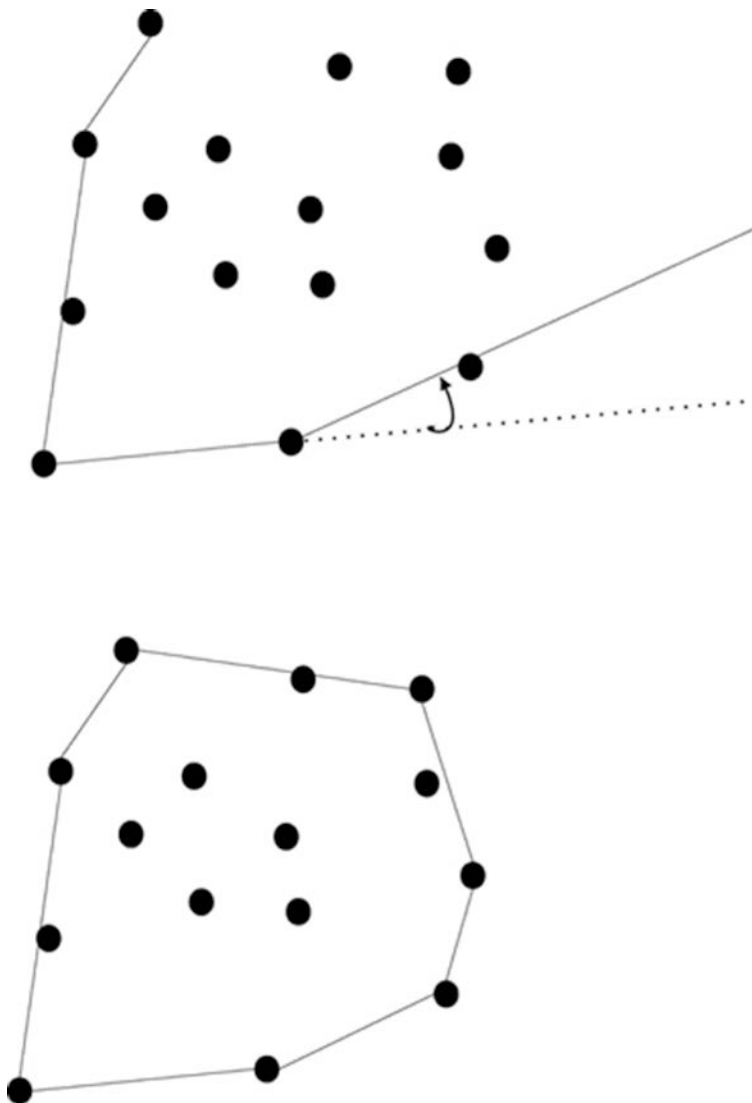
Σχήμα 6.1. Διάταξη των σημείων ως προς την υψηλότερη τιμή.

Στάδιο 2. Τα σημεία αριθμούνται σύμφωνα με αύξουσα γωνία σάρωσης του άξονα περιστροφής. Συγκεκριμένα, μία οριζόντια ακτίνα που έχει κέντρο το P_0 αρχίζει να στρέφεται αντίθετα με τη φορά των δεικτών του ρολογιού. Το πρώτο σημείο που θα συναντήσει το αριθμεί ως P_1 . Με τον ίδιο τρόπο, τα υπολειπόμενα $n-2$ σημεία αριθμούνται, μέχρι η περιστρεφόμενη ακτίνα να επιστρέψει στην αρχική της οριζόντια θέση (Σχήμα 6.2).



Σχήμα 6.2. Ταξινόμηση σημείων σύμφωνα με τη γωνία σάρωσης.

Στάδιο 3. Η οριζόντια ακτίνα αρχίζει ξανά να περιστρέφεται με την ίδια φορά γύρω από το P_0 και το πρώτο σημείο που θα συναντήσει η ακτίνα, καθίσταται νέο σημείο-κέντρο περιστροφής. Η διαδικασία αυτή συνεχίζεται μέχρι να ξαναγίνει το P_0 σημείο-κέντρο περιστροφής (Σχήμα 6.3). Τότε έχει σχηματισθεί το κυρτό κέλυφος.



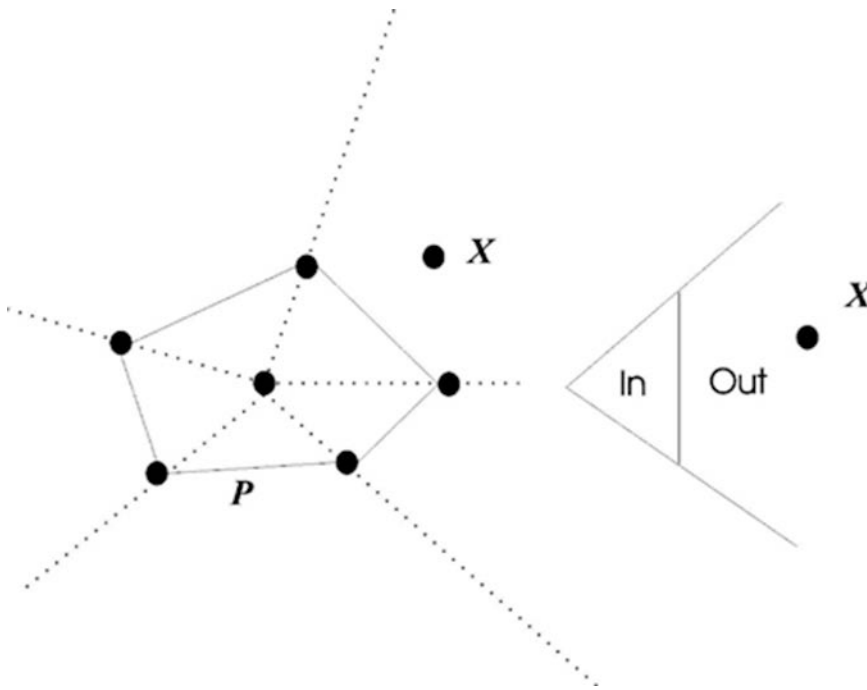
Σχήμα 6.3. Τελική φάση δημιουργίας του κυρτού κελύφους.

6.1.2.1 Τομή δύο κυρτών πολυγώνων

Η επίλυση του προβλήματος υπολογισμού της τομής δύο ομοεπιπέδων κυρτών πολυγώνων επετεύχθη με τον αλγόριθμο του Shamos, 1975, ο οποίος σε γραμμικό χρόνο είχε υπολογιστική πολυπλοκότητα ($O(n)$). Η επιτυχία αυτού του αλγορίθμου εξαρτάται από ένα προπαρασκευαστικό στάδιο το οποίο καθορίζει κατά πόσον ένα δεδομένο σημείο βρίσκεται μέσα ή έξω από ένα γνωστό, κυρτό, επίπεδο πολύγωνο. Ο αλγόριθμος που υλοποιεί τα παραπάνω περιγράφεται λεπτομερώς στα ακόλουθα δύο στάδια:

Στάδιο 1. Διαδικασία προσδιορισμού θέσης σημείου (εσωτερική, εξωτερική) ως προς ένα ομοεπίπεδο πολύγωνο.

Εξετάζεται ο προσδιορισμός της θέσης ενός σημείου x ως προς ένα κυρτό πολύγωνο P (εσωτερική, εξωτερική). Στη συνέχεια έστω A το διάνυσμα που ορίζεται από το σημείο x και από μία τυχαία κορυφή του κυρτού πολυγώνου P . Ομοίως έστω ως B ένα διάνυσμα που ορίζεται από το σημείο x και από μία διαφορετική κορυφή του κυρτού πολυγώνου P . Τότε η θέση του x (εσωτερική, εξωτερική) καθορίζεται προϋποθέτοντας ότι ο υπολογισμός της γωνίας των δύο διανυσμάτων γίνεται με δεξιόστροφη φορά (**Σχήμα 6.4**).



Σχήμα 6.4. Προσδιορισμός θέσης ενός σημείου x ως προς ένα επίπεδο κυρτό πολύγωνο P (εσωτερική, εξωτερική).

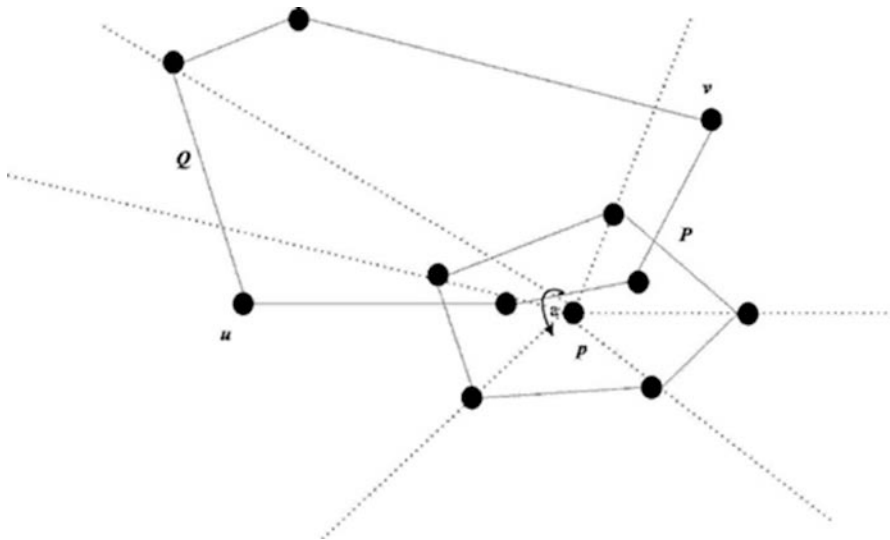
$$\begin{aligned} \text{εάν } \mathbf{A} \times \mathbf{B} < 0 \text{ τότε } x \notin H(A), \\ \text{ή} \\ \text{εάν } \mathbf{A} \times \mathbf{B} > 0 \text{ τότε } x \in H(A) \end{aligned} \quad (6.6)$$

όπου $H(A)$ θεωρείται το κλειστό ημιεπίπεδο που περιεχίει το σύνολο των σημείων A .

Στάδιο 2. Ανάπτυξη του αλγορίθμου τομής δύο κυρτών πολυγώνων.

Υποτίθεται ότι δύο ομοεπίπεδα πολύγωνα P και Q , έχουν w_1 και w_2 κορυφές αντίστοιχα. Τα βήματα για τον υπολογισμό της τομής των πολυγώνων αυτών είναι τα ακόλουθα (Σχήμα 6.5):

1. Εξετάζονται διαδοχικά οι κορυφές του P πολυγώνου ως προς το αν είναι εσωτερικά σημεία του Q πολυγώνου.
2. Τα τυχόν εξωτερικά σημεία $\{a_k\}$, όπου $k=0,1,\dots,n$ και n είναι το πλήθος των κορυφών του πολυγώνου P , αποτελούν το μέρος ή το σύνολο των νέων κορυφών του κυρτού πολυγώνου PQ , που αποτελεί την τομή των δύο κυρτών πολυγώνων P και Q .
3. Επίσης προσδιορίζεται ένα νέο σύνολο σημείων $\{\beta_\mu\}$ το οποίο προσδιορίζεται από τις τομές των πλευρών των πολυγώνων P και Q (όπου $\mu=0,1,\dots,w$, και $w=w_1+w_2$ είναι ο αριθμός του συνόλου των κορυφών των πολυγώνων P και Q).
4. Τότε, οι κορυφές του ζητούμενου κυρτού πολυγώνου PQ ορίζονται από ένα νέο σύνολο σημείων που προσδιορίζεται από την ένωση των συνόλων $\{a_k\}$ και $\{\beta_\mu\}$.

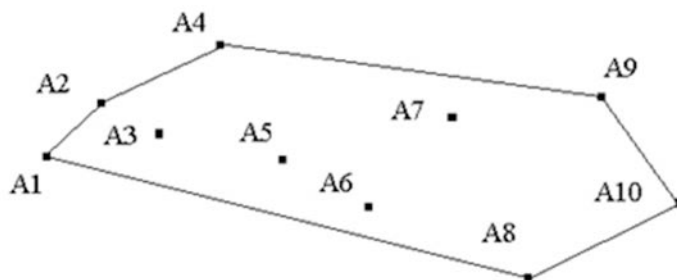


Σχήμα 6.5. Τελική φάση προσδιορισμού των συντεταγμένων των κορυφών της τελικής τομής των δύο πολυγώνων.

6.1.3 Υλοποίηση αλγορίθμων υπολογιστικής γεωμετρίας

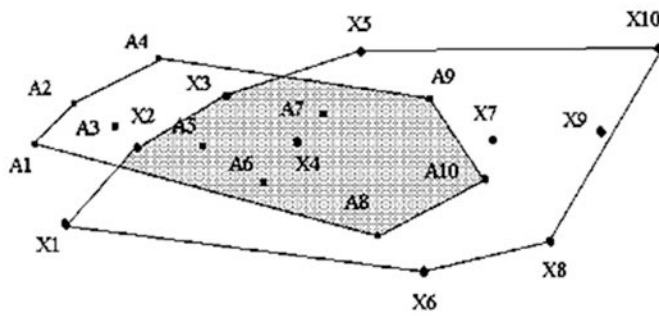
Για απλούστευση της παρουσίασης της μεθόδου, θεωρείται ότι τα δεδομένα ανήκουν στο δισδιάστατο χώρο και αποτελούνται από 10 τιμές (στοιχεία) έκαστο. Οι τιμές του κάθε διανύσματος θεωρούνται ότι προήλθαν από μία από τις προαναφερόμενες μεθόδους επεξεργασίας (κεφάλαια 4, 5). Η συνδυασμένη υλοποίηση των προαναφερόμενων αλγορίθμων υπολογιστικής γεωμετρίας περιγράφονται στα παρακάτω βήματα:

1. Οι τιμές του κάθε διανύσματος τοποθετούνται στο καρτεσιανό επίπεδο σύμφωνα με τη σειρά εξαγωγής από την επιλεγμένη μέθοδο επεξεργασίας. Στον οριζόντιο άξονα τοποθετείται ο αύξων δείκτης τους $i=1, 2, \dots, 10$. Στη συνέχεια από τις τιμές αυτές κατασκευάζεται ένα κυρτό πολύγωνο σύμφωνα με τον αλγόριθμο που αναπτύχθηκε στην παράγραφο. Με τον ίδιο τρόπο κατασκευάζονται και τα υπόλοιπα κυρτά πολύγωνα από δεδομένα διάστασης 1×10 (**Σχήμα 6.6**).



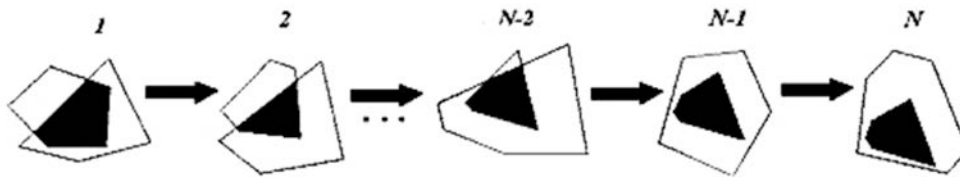
Σχήμα 6.6. Κατασκευή ενός κυρτού πολυγώνου, δοθέντος διανύσματος-παραδείγματος 10 σημείων.

2. Ακολούθως εφαρμόζεται ο αλγόριθμος υπολογισμού της τομής δύο κυρτών πολυγώνων όπως περιγράφηκε στην παράγραφο (**Σχήμα 6.7**). Ο αλγόριθμος εφαρμόζεται επαναληπτικά μεταξύ διαδοχικών ζευγών πολυγώνων που προέρχονται από δεδομένα του ίδιας ομάδας ιδιοτήτων.



Σχήμα 6.7. Τομή δύο κυρτών πολυγώνων στο επίπεδο.

Το 1^ο ΗΕΓ τέμνεται με το 2^ο, η τομή τους με το 3^ο, κ.ο.κ. Για τον τερματισμό της επανάληψης τίθεται σε εφαρμογή το κριτήριο σύγκλισης των επαναλαμβανόμενων τομών που θεμελιώθηκε στην παράγραφο 3.4.4.1. Συγκεκριμένα, η τομή επαναλαμβάνεται έστω n φορές, μέχρις ότου η τελευταία προκύπτουσα τομή παραμείνει αμετάβλητη σε σχέση με την προ-τελευταία. Η τελική τομή τότε ονομάζεται «χαρακτηριστικό πολύγωνο» (Σχήμα 6.8) και χαρακτηρίζει το άτομο.



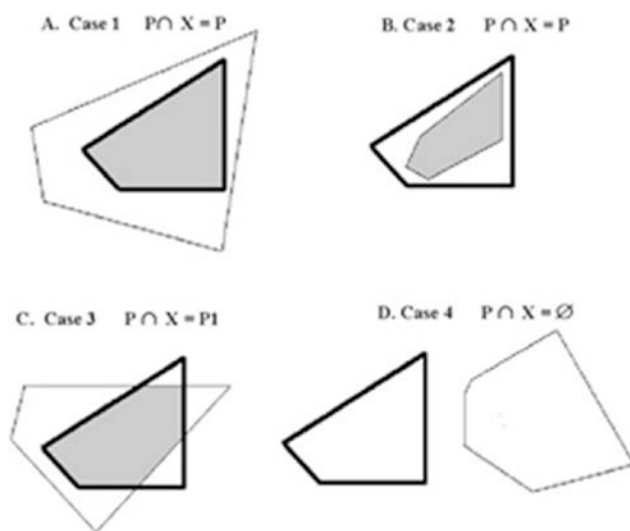
Σχήμα 6.8. Εξαγωγή χαρακτηριστικού πολυγώνου μετά από διαδοχικές τομές κυρτών πολυγώνων.

3. Το εξαγόμενο χαρακτηριστικό πολύγωνο υποβάλλεται σε έλεγχο τομής σύμφωνα με το βήμα 2. Συγκεκριμένα, αν συμβολιστεί το χαρακτηριστικό πολύγωνο ως P και ένα υποψήφιο προς προσδιορισμό ταυτότητας πολύγωνο ως X , τότε οι περιπτώσεις που είναι δυνατόν να προκύψουν είναι οι παρακάτω τέσσερις:

Περίπτωση 1:	$P \cap X = P$
Περίπτωση 2:	$P \cap X = X$
Περίπτωση 3:	$P \cap X = P_1$
Περίπτωση 4:	$P \cap X = \emptyset$

Στο σχήμα 6.9 παρατηρείται το χαρακτηριστικό πολύγωνο P με έντονα μαύρο περίγραμμα, και το υποψήφιο προς ταυτοποίηση πολύγωνο X με λεπτό περίγραμμα. Επίσης, P_1 είναι το κυρτό πολύγωνο που προκύπτει από την τομή των P και X πολυγώνων, περίπτωση 3, δηλαδή $P_1 \subset X$ και $P_1 \subset P$.

Σύμφωνα λοιπόν με τα προηγούμενα, η θετική αναγνώριση, δηλαδή η υιοθέτηση της απόφασης ότι το πολύγωνο X ανήκει στην ίδια οικογένεια δεδομένων που παρήγαγε το πολύγωνο P , παράγεται μόνο στην περίπτωση 1, ενώ για τις υπόλοιπες περιπτώσεις χρησιμοποιείται ο όρος «αρνητική αναγνώριση» δηλαδή το σύστημα αποφασίζει ότι το πολύγωνο X δεν ανήκει στην ίδια ομάδα δεδομένων που παρήγαγε το πολύγωνο P .



Σχήμα 6.9. Περιπτώσεις κατηγοριοποίησης του αγνώστου πολυγώνου X στην ίδια κατηγορία (άτομο) με το χαρακτηριστικό πολύγωνο P , σύμφωνα με το μέγεθος και τη θέση των χαρακτηριστικών πολυγώνων.

Παρακάτω δίδεται παράδειγμα εφαρμογής κειμένου σε υπολογιστική γεωμετρία. Συγκεκριμένα αναπαρίσταται ένα κείμενο 100 λέξεων περίπου σε μορφή κρεμμυδιού. Η σημασιολογία αυτής της εφαρμογής είναι να εξαχθεί το μικρότερο εσωτερικό πολύγωνο που αναπαριστά την κεντρική ιδέα του κειμένου ([βλ. video1](#))

6.2 Ανάλυση Κειμένων με Στατιστικά Εργαλεία

Το παρακάτω κείμενο που ακολουθεί αποτελεί ελεύθερη απόδοση του άρθρου [1]. Στην θεωρία Ποσοτικής Γλωσσολογίας, η λεξιλογική συνεκτικότητα των κειμένων ασχέτως της κατανομής λέξεων θεωρείται ένα πολύ σημαντικό επιστημονικό πεδίο. Σύμφωνα με Carstens (2001) η γλωσσική συνεκτικότητα κειμένων ορίζεται ως «οι τρόποι με τους οποίους τα συστατικά των προτάσεων ενός κειμένου δηλαδή, οι λέξεις που ακούμε και χρησιμοποιούμε είναι αμοιβαία συνδεδεμένες μεταξύ τους (γραμματικά και λεξιλογικά)». Οι Haliday και Hasan (1976) υπογραμμίζουν την συνοχή ως «την σημασιολογική σχέση μεταξύ ενός στοιχείου και ενός άλλου μέσα στο κείμενο και κάποιου άλλου που θεωρείται κρίσιμο για την ερμηνεία τους». Επιπλέον όπως αναφέρεται από Fahnestock (1983), η συνεκτικότητα, η οποία πηγάζει από σωστή σύνθεση κειμένου, είναι ένα κρίσιμο στοιχείο ώστε η ροή των ιδεών να πραγματοποιείται ομαλά σε όλο το κείμενο και ώστε η αναγνωσιμότητα να διατηρεί υψηλά επίπεδα για την κατανόηση του νοήματος του κειμένου. Με βάση τους Richardsetal. (1992), ως αναγνωσιμότητα ορίζεται: «πόσο εύκολα το γραπτό υλικό μπορεί να αναγνωσθεί και να κατανοηθεί. Αυτό εξαρτάται από πολλούς παράγοντες συμπεριλαμβανομένου και του μέσου μήκους μιας πρότασης, τον αριθμό καινούριων λέξεων που περιλαμβάνονται σε αυτά και την γραμματική πολυπλοκότητα της γλώσσας που χρησιμοποιείται στο απόσπασμα κειμένου». Ακόμη η έννοια της αναγνωσιμότητας συνδέεται με αυτήν της κατανόησης. Ο Sparks (2012) αναφέρει ότι η κατανόηση λόγου εμπεριέχει την κατασκευή νοήματος από εκτεταμένα τμήματα της γλώσσας. Επιπλέον ότι η επιτυχής κατανόηση μεγαλύτερων μονάδων κειμένου απαιτεί την λήψη συμπερασμάτων ώστε να συνδεθούν οι ιδέες τόσο στο κείμενο τοπικά όσο και σε παγκόσμιο πλαίσιο λόγου.

Όπως αναφέρεται από Eroglu (2013) η γλωσσολογική οργάνωση στα κείμενα μπορεί να επιτευχθεί όταν ισχύει ο νόμος των Menzerath-Altmann (νόμος MA) (Altmann, G., 1980). Ο νόμος MA θεωρείται βασικός νόμος στην ΠΓ, όπου σύμφωνα με Baixeriesetal. (2013) παρατηρούνται οι σχέσεις ανάμεσα στο μέγεθος του συνόλου και στο μέγεθος των επιμέρους μερών της γλώσσας. Έτσι σύμφωνα με ErogluS. (2013), για τον νόμο MA, όσο μεγαλύτερο είναι το γλωσσικό δόμημα, τόσο μικρότερα θα είναι τα συστατικά που το αποτελούν, όπου δόμημα θεωρείται το σύνολο και συστατικό μέρος του συνόλου. Ο νόμος MA είναι βασικός και σημαντικός νόμος στην ΠΓ και εστιάζει κυρίως στην στατιστική ανάλυση μεγάλων κειμένων. Ειδικότερα, σύμφωνα με Hrebicek (2002), ο νόμος MA δεν ισχύει σε εξαιρετικά μικρά κείμενα, όταν μικρά κείμενα θεωρούνται απλές ή περίπλοκες προτάσεις. Όμως τα μικρά κείμενα θεωρούνται πολύ σημαντικά σε πολλούς

επιστημονικούς τομείς, όπως στην διαδικτυακή επικοινωνία και το ηλεκτρονικό εμπόριο αλλά και στην γρήγορη αναζήτηση στο διαδίκτυο. Σύμφωνα με GeSongetal. (2014) θεωρείται μεγάλη πρόκληση η ταξινόμηση μικρών κειμένων καθώς ο περιορισμένος αριθμός λέξεων τους δεν μπορεί να αναπαραστήσει ούτε τον χώρο χαρακτηριστικών ούτε τις πραγματικές σχέσεις μεταξύ λέξεων και εγγράφων. Καθώς τα μικρά κείμενα έχουν μικρό μήκος σύμφωνα με XiaojunQuan (2009) τα κλασικά μέτρα ομοιότητας και οι αντίστοιχες αναπαραστάσεις των αντικειμένων τους δεν μπορούν να εφαρμοστούν με επιτυχία.

Είναι γνωστό πως οι άνθρωποι χρησιμοποιούν προτάσεις για επικοινωνούν μεταξύ τους επιτυχημένα και ορισμένες παράμετροι πρέπει να ληφθούν υπόψη, καθώς η σωστή δόμηση προτάσεων αποτελεί ένα πολύ σημαντικό ζήτημα. Προκειμένου η επικοινωνία να είναι επιτυχής πρέπει να υπάρχει ένα μέσο μήκος πρότασης ώστε αυτή να μην προκαλεί σύγχυση ή να θεωρείται περίπλοκη. Η λεξιλογική συνεκτικότητα της πρότασης έχει οριστεί εμπειρικά από Kornai (2008), στον δημοσιογραφικό λόγο με μέσο μήκος πρότασης να είναι πάνω από 15 λέξεις. Ακόμη Tesitelova (1992) αναφέρει ότι κατά την δεκαετία του '80 υπήρξε ενδιαφέρον στην επιστημονική έρευνα της πρότασης ως συντακτικού φαινομένου. Επιπλέον, η πολυπλοκότητα της σημασίας της πρότασης, εξαρτάται από το μήκος της πρότασης. Σύμφωνα με Cutts (2009) το μέσο μήκος πρότασης πρέπει να βρίσκεται ανάμεσα σε 15-20 λέξεις, ώστε να διατηρεί την αναγνωσιμότητα αν και το μέσο μήκος δεν είναι δυνατόν να επιτυγχάνεται πάντα. Οι Taskaretal. (2004) στην εργασία τους σχετικά με περιγραφική ανάλυση και τεχνικές δυναμικών προσεγγίσεων, επιτελούν τα πειράματα τους θέτοντας τον περιορισμό του μήκους πρότασης να είναι ίσο ή λιγότερο από 15 λέξεις. Άλλα επιστημονικά πεδία που εκφράζουν ενδιαφέρον σχετικά με το μήκος της πρότασης αφορούν την έρευνα που αφορά την ανθρώπινη μνήμη, όπως η γνωστική ψυχολογία και οι νευροεπιστήμες. Σύμφωνα με Baddeley (2003), στο μοντέλο του για την μνήμη εργασίας που αποτελείται από τρία τμήματα, αντιμετώπισε προβλήματα σχετικά με την αλληλεπίδραση με την μακροπρόθεσμη μνήμη, όπου ο περιορισμός των 15 λέξεων αναφέρεται ξανά. Στα πειράματα τους οι Daveman και Carpenter (1980) υποστηρίζουν ότι χρειάζονται 5 δευτερόλεπτα για έναν άνθρωπο να διαβάσει μια πρόταση. Επιπλέον οι Andersonetal. (2001) στην μελέτη τους σχετικά με την μνήμη προτάσεων, περιγράφουν διάφορα μοντέλα μνήμης και παρατήρησαν ότι ο ανθρώπινος εγκέφαλος χρειάζεται μερικές εκατοντάδες χιλιοστά του δευτερολέπτου για να επεξεργαστεί μια λέξη. Στην ίδια έρευνα αναφέρεται η διεξαγωγή του πειράματος του Zimny (1987), όπου μια διαδικασία παρουσίασης λέξης προς λέξη ολοκληρώθηκε σε 300 χιλιοστά του δευτερολέπτου ανά λέξη. Συνεπώς από τις παραπάνω έρευνες μπορεί κανείς να συμπεράνει ότι μια τυπική πρόταση θα πρέπει να περιέχει 16-17 λέξεις. Λαμβάνοντας υπόψη όλα τα παραπάνω η ανάλυση της συνεκτικότητας στα μικρά κείμενα και ειδικότερα στις προτάσεις μπορεί να θεωρηθεί ένα πολύ σημαντικό επιστημονικό πεδίο για διερεύνηση.

Ο σκοπός αυτού του άρθρου είναι η καθιέρωση μιας στατιστικής υπόθεσης που επιβεβαιώνει τις εμπειρικές παρατηρήσεις σχετικά με την συνεκτικότητα των μικρών κειμένων ή μιας πρότασης. Έτσι μπορεί να καλυφθεί το κενό που αφήνει ο νόμος MA σχετικά με τα μικρά κείμενα και η μελέτη αυτή μπορεί να αποτελέσει τον θεμέλιο λίθο για την ανάλυση μικρών κειμένων.

Για την εφαρμογή των παραπάνω, ανιχνεύσαμε μεταβλητές που θεωρούνται κρίσιμες για την συνεκτικότητα κειμένου. Η συνεκτικότητα κειμένου εξετάστηκε σε σχέση με το αντίκτυπο κάθε συστατικού σε σχέση με το δόμημα. Η συσχέτιση αυτή πραγματοποιήθηκε με την χρήση τριών μεταβλητών. Η απόδειξη της παραπάνω υπόθεσης πραγματοποιήθηκε μέσω του δείκτη συμφωνίας Kendall. Τα συμπεράσματα που προέκυψαν από αυτήν την μεθοδολογία, εισάγουν μια καινοτομία για το πεδίο της υπολογιστικής γλωσσολογίας διότι οι παραπάνω πειραματικές γλωσσολογικές παρατηρήσεις σχετικά με την σταθερότητα των μικρών κειμένων στα περίπου 15 συστατικά επιβεβαιώνονται, και γεγονός που καταδεικνύει την στατιστική συσχέτιση των μικρών κειμένων μέσω αυτών των τριών μεταβλητών.

Το άρθρο αυτό χωρίζεται στις παρακάτω ενότητες:

- Μεθοδολογία, όπου παρουσιάζεται εκτενώς ο αλγόριθμος και η στατιστική αξιολόγηση του.
- Πειραματικό μέρος, όπου ο αλγόριθμος εφαρμόζεται σε ευρύ δείγμα μικρών κειμένων.

6.2.1 Μεθοδολογία

Σύμφωνα με Salton, WongandYang (1975) τα έγγραφα μπορούν να αναπαρασταθούν ως διανύσματα, ώστε να ευρετηριαστούν και να υπολογιστεί ένας βαθμός ομοιότητας μεταξύ τους. Όπως αναφέρουν οι Raghavan και Wong (1986) τα διανύσματα είναι ιδιαίτερος χρήσιμα καθώς υπακούν σε βασικά αξιώματα και κανόνες της άλγεβρας. Το VSM χρησιμοποιείται σε διάφορα επιστημονικά πεδία όπως συστήματα φιλτραρίσματος πληροφορίας και αλγόριθμοι κατάταξης με βάση την συνάφεια. Οι Turney και Patel (2010)

αναφέρουν πως με την επέκταση της χρήσης του VSM σε σημασιολογικό πλαίσιο στην επεξεργασία της γλώσσας, μπορεί να επιφέρει λαμπρά αποτελέσματα. Το VSM είναι ένα αλγεβρικό μοντέλο το οποίο καθιστά δυνατή την αναπαράσταση οποιουδήποτε αντικειμένου εγγράφου, όπως ένα κείμενο, μια πρόταση, μια φράση, μια λέξη ή ένα μόρφημα. Η αναπαράσταση μέσω VSM αναλύεται σε τρία βήματα.

Στο πρώτο βήμα, οι όροι που περιέχονται σε ένα κείμενο (τυπικά λέξεις ή μικρές φράσεις) εξάγονται δημιουργώντας ένα ευρετήριο εγγράφου. Η ευρετηρίαση αυτή εκτελείται μέσω δύο εναλλακτικών μεθόδων. Η γλωσσολογική μέθοδος βασίζεται στην συγκέντρωση λειτουργικών λέξεων, οι οποίες έχουν υψηλή συχνότητα και χαμηλή συχνότητα, που αντικατοπτρίζονται σημασιολογικά στο κείμενο. Από την άλλη ή μη γλωσσολογική μέθοδος βασίζεται σε διαφορετικές διαδικασίες ευρετηρίασης όπως πιθανοτική και αυτόματη ευρετηρίαση.

Στο δεύτερο βήμα, τα βάρη των ευρετηριαζόμενων όρων καθορίζονται σύμφωνα με την συνάφεια τους με το ερώτημα του χρήστη για μια πιθανή διαδικασία ανάκτησης πληροφορίας. Τα βάρη όρων εφαρμόστηκαν ελέγχοντας την ευαισθησία – ειδικότητα της αναζήτησης, όπου η ειδικότητα σχετίζεται με την ακρίβεια και η ευαισθησία με την ανάκληση. Το πιο ευρέως χρησιμοποιούμενο σχήμα απόδοσης βάρους στην ΑΠ είναι το $tf - idf$, το οποίο περιέχει 3 κυρίαρχες μεταβλητές για τον καθορισμό του βάρους, που συνδέονται με την συχνότητα όρου στο έγγραφο, την συχνότητα όρου στην συλλογή εγγράφων και την κανονικοποίηση μεγέθους. Αυτές οι τρεις μεταβλητές συνδυάζονται μέσω της πράξης του πολλαπλασιασμού, ώστε να υπολογίσουν το αποτέλεσμα για το βάρος του όρου.

Ως παράδειγμα, ο ορισμός της αλγεβρικής έκφρασης ενός εγγράφου μέσω VSM πραγματοποιείται χρησιμοποιώντας την ακόλουθη εξίσωση:

$$\vec{V}_j = [w_{1,j}, w_{2,j}, \dots, w_{t,j}] \quad (6.7)$$

Όπου το j αντιστοιχεί στον αριθμό συστατικών ενός εγγράφου και t αναπαριστά τον αριθμό των μεταβλητών που καθορίζουν το βάρος στο προτεινόμενο μοντέλο.

Τέλος, στο τρίτο βήμα το κείμενο κατατάσσεται σύμφωνα με το μέτρο ομοιότητας ως προς ένα αντίστοιχο ερώτημα. Η ομοιότητα στο VSM μοντέλο υπολογίζεται μέσω συνδυαστικών μεταβλητών και βασίζεται στο κανονικοποιημένο εσωτερικό γινόμενο ανάμεσα στο διάνυσμα εγγράφου και διάνυσμα ερωτήματος, όπου η επικάλυψη μεταξύ διανυσμάτων υποδεικνύει ομοιότητα. Το εσωτερικό γινόμενο συνήθως κανονικοποιείται. Το πιο συνηθισμένο μέτρο ομοιότητας είναι μέσο της μεταβλητής συνημίτονου, όπου μετράται η γωνία απόκλισης ανάμεσα στο διάνυσμα εγγράφου και διάνυσμα ερωτήματος.

6.2.2 Η βάση του Αλγόριθμου

Η εξαγωγή μεταβλητών βασίζεται στην θεωρία του VSM. Πιο συγκεκριμένα, στο πρώτο βήμα επιλέγουμε ένα μικρό κείμενο ως δόμημα, το οποίο αποτελείται από μια απλή ή σύνθετη πρόταση και θεωρούμε την λέξη ως συστατικό του δομήματος. Κατά συνέπεια, σύμφωνα με το δεύτερο βήμα μια μη γλωσσική προσέγγιση υιοθετείται γιατί όλα τα συστατικά του δομήματος χρησιμοποιούνται σε μια διαδικασία ευρετηρίασης που εξαρτάται στην σειρά της θέσης τους. Πιο λεπτομερειακά, κάθε συστατικό λαμβάνει ένα αντίστοιχο βάρος ανεξάρτητα από τον αριθμό εμφάνισής του στο δόμημα.

Στην δική μας περίπτωση, οι τρεις κυρίαρχες μεταβλητές απόδοσης βάρους όρου αντικαθίστανται από τις μεταβλητές του παρακάτω διανύσματος, όπως φαίνεται τύπο (2):

$$\vec{w}_j = [i_j \quad s_j \quad k_j] \quad (6.8)$$

Το i αντιστοιχεί στην σειρά του συστατικού στο μικρό κείμενο, s αποτελεί έναν αριθμό κωδικοποίησης, όπου για λόγους απλοποίησης ορίζεται μέσω της διαδικασίας κωδικοποίησης ASCII (Poulos, Papavasiliou, Chrissikopoulos 2006) και k αντιστοιχεί στον αριθμό χαρακτήρων του όρου.

Έπειτα η εξίσωση 1 μετατρέπεται στην εξίσωση (3):

$$\vec{V}_j = [w_1, w_2, \dots, w_j] \quad (6.9)$$

Για λόγους κανονικοποίησης, το διάνυσμα $\vec{V}_j$ υπολογίζεται από το ισοδύναμο διάνυσμα της εξίσωσης (4):

$$\vec{F}_j = \frac{\vec{V}_j}{\|\vec{V}_j\|} \quad (6.10)$$

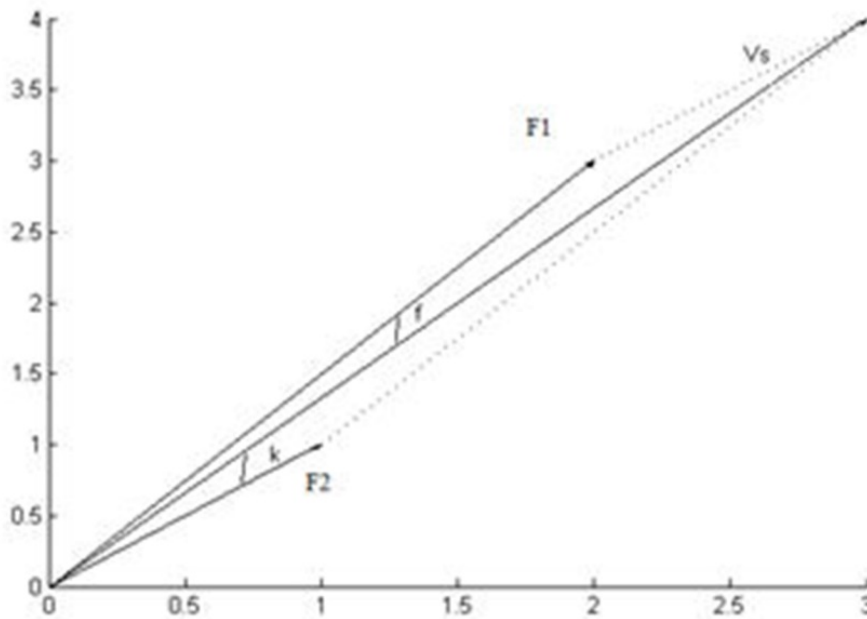
Επιπλέον, ορίζουμε το διάνυσμα $\vec{V}_s$, το οποίο απεικονίζεται ως η συνισταμένη διανύσματος και από εδώ και στο εξής θα αναφέρεται ως διάνυσμα μικρού κειμένου (βλέπε εικ. 1).

$$\vec{V}_s = \sum_j \vec{F}_j \quad (6.11)$$

6.2.3 Κριτήριο ομοιότητας

Ο βαθμός συσχέτισης ανάμεσα σε $\vec{F}_j$ και $\vec{V}_s$ εξάγεται μέσω της εξίσωσης 6 και ειδικότερα από τον συντελεστή r ο οποίος αναπαριστά το εσωτερικό γινόμενο ανάμεσα στο διάνυσμα κειμένου $\vec{V}_s$ και στο διάνυσμα συστατικού $\vec{F}_j$ σύμφωνα με το βήμα 3 (βλέπε ενότητα 2.1 αυτού του άρθρου). Ακόμη, η διαδικασία αυτή εκφράζεται από την γενική θεώρηση σχετικά με την θεωρία ομοιότητας εγγράφων (Harispeetal. 2013)):

$$r^{(j)} = \cos \theta^{(j)} = \frac{\vec{F}_j \cdot \vec{V}_s}{\|\vec{F}_j\| \|\vec{V}_s\|} \quad (6.12)$$



Σχήμα 6.10 Απεικόνιση διανυσμάτων συστατικών $\vec{F}_1, \vec{F}_2$ και διανύσματος συνισταμένης $\vec{V}_s$ στο Ευκλείδειο επίπεδο

6.2.4 Μετατροπή Μεταβλητών

Υιοθετείται το διάνυσμα $\vec{U}_j$ αντί για το $\vec{V}_j$ διάνυσμα, όπου η μεταβλητή i του $\vec{W}_j$ αντικαθίσταται από την μεταβλητή r του $\vec{U}_j$ (βλέπε εξίσωση 7). Ο λόγος για αυτήν την αντικατάσταση είναι διότι η μεταβλητή r θεωρείται ένας πολύ σημαντικός συντελεστής στην θεωρία σημασιολογίας σύμφωνα με (Harispeetal. 2013). Η αντικατάσταση πραγματοποιήθηκε σκοπίμως αφού το διάνυσμα $\vec{U}_j$ αναπαριστάτο βαθμό επηρεασμού μέσω της μεταβλητής r γιατί αυτό θα οδηγήσει σε μια νέα διαδικασία κατάταξης ανάλογα τον επηρεασμό της.

$$\vec{U}_j = \begin{bmatrix} r_j & s_j & k_j \end{bmatrix} \quad (6.13)$$

6.2.5 Στατιστική θεμελίωση

Η στατιστική θεμελίωση της παραπάνω μελέτης ολοκληρώνεται σε τρία στάδια. Στο πρώτο στάδιο υπολογίζεται ο βαθμός επηρεασμού της μεταβλητής r στις άλλες δύο μεταβλητές που ορίζεται στην εξίσωση 7, σε σχέση με τον αριθμό λέξεων του κάθε παραδείγματος. Όλη αυτή η διαδικασία έχει περιγραφεί στην θεωρία (βλέπε ενότητα 2.3 και 2.4 του άρθρου) και υλοποιείται παρακάτω στο στάδιο α. Στη συνέχεια, στο δεύτερο στάδιο β χρησιμοποιείται το κριτήριο χ^2 συσχετισμένο με τον δείκτη συμφωνίας w του Kendall, ώστε στο τρίτο στάδιο γ να υπολογιστεί ο βαθμός πιθανότητας να υπάρξει συμφωνία ή μη συμφωνία με βάση το κριτήριο χ^2 . Αυτά περιγράφονται αναλυτικά παρακάτω, στο στάδιο β και γ αντίστοιχα και συνδέονται με την θεωρία:

1. Η στατιστική επιβεβαίωση του παραπάνω μετασχηματισμού επιτυγχάνεται ελέγχοντας εάν οι τρεις (3) μεταβλητές r, s, k συσχετίζονται. Αυτό πραγματοποιείται εφαρμόζοντας τον δείκτη συμφωνίας w του Kendall (Zar 1999), στον παρακάτω τύπο 8:

$$W = \frac{\sum_{j=1}^n R_j^2 - \frac{\sum_{j=1}^n R_j}{n}}{M^2(n^2 - 1)} \quad (6.14)$$

Το M αναπαριστά τον αριθμό μεταβλητών που συσχετίζονται, δηλαδή σε αυτή την περίπτωση το $M=3$ και το n αναπαριστά τον αριθμό συστατικών στο μικρό κείμενο.

$$\text{Όπου } [d] = \text{rank}(U_j) \quad (6.15)$$

$$[R]_{j=1}^n = \sum_{j=1}^n [d] \quad (6.16)$$

Το d αντιστοιχεί στην κατάταξη του κειμένου του διανύσματος $\vec{U}_j$ (εξίσωση 9) και το R ορίζει την κατάταξη ως το άθροισμα των μεταβλητών κάθε εγγράφου (εξίσωση 10).

1. Σε αυτή την περίπτωση ο στατιστικός έλεγχος βασίζεται στην μηδενική υπόθεση ότι αυτές οι τρεις μεταβλητές δεν συσχετίζονται μεταξύ τους, ώστε να οριστεί ο αριθμός συστατικών που επηρεάζουν το συμμετέχον μικρό κείμενο. Η μηδενική υπόθεση H_0 υποδεικνύει ότι οι τρεις μεταβλητές δεν συσχετίζονται και η εναλλακτική υπόθεση H_A υποδεικνύει το ακριβώς αντίθετο. Κατά συνέπεια, το κριτήριο χ^2 ορίζεται ως εξής (βλέπε ενότητα 4.2.1):

$$\chi_r^2 = 3 * (n - 1) * F(r, s, k) \quad (6.17)$$

2. Η τιμή της αθροιστικής κατανομής του χ^2 εξάγεται χρησιμοποιώντας τον βαθμό ελευθερίας $\nu = n - 1$ και υπολογίζεται από την εξίσωση 12.

$$P = 1 - \int_0^x \frac{t^{\frac{\nu-2}{2}} e^{-\frac{t}{2}}}{2^{\frac{\nu}{2}} \Gamma\left(\frac{\nu}{2}\right)} dt \quad (6.18)$$

Ο έλεγχος είναι μονόπλευρης κατανομής διότι αναζητάμε το κατάλληλο αριθμό συστατικών για ένα μικρό κείμενο. Έπειτα η μηδενική υπόθεση H_0 γίνεται δεκτή/ισχύει όταν $P < 0.001$ και κατά συνέπεια το H_A ικανοποιείται με το $P \geq 0.001$.

6.2.6 Πειραματικό Μέρος

Αυτό το πείραμα αποτελείται από τα παρακάτω βήματα:

1. Αρχικά συλλέχθηκε ένα δείγμα ώστε να εφαρμοστεί ο αλγόριθμος. Το σύνολο δεδομένων αυτής της έρευνας αποτελείται από 100 μικρά κείμενα τα οποία εξάχθηκαν από περιλήψεις (abstracts) επιστημονικών άρθρων. Τα επιστημονικά αυτά άρθρα ανακτήθηκαν σε μορφή pdf από το Directory of Open Access Journals (DOAJ) και σε σχέση με το θέμα τους όλα προέρχονται από το επιστημονικό πεδίο της βιοιατρικής. Με την χρήση ενός φυλλομετρητή, από τον ιστότοπο του Directory of Open Access Journals (DOAJ) αποθηκεύτηκαν σε τοπικό φάκελο τα 100 άρθρα.
2. Εν συνεχεία, επιλέχθηκαν 100 μικρά κείμενα, όπου κάθε ένα από αυτά είχε μέσο μήκος πάνω από 25 συστατικών. Κάθε μικρό κείμενο επεξεργάστηκε μέχρι το 25^ο συστατικό και όχι πιο πάνω για τους παρακάτω λόγους:
 - ο Καθώς ο αλγόριθμος αυτός βασίζεται στο VSM, υπάρχει περιορισμός σχετικά με τα μεγάλα κείμενα. Η αναπαράσταση τους δεν μπορεί να είναι επιτυχής καθώς σε τέτοια μεγέθη δεν μπορούν να βρεθούν μεταβλητές ομοιότητας σύμφωνα με Salton (1975).
 - ο Ένας ακόμη λόγος για το όριο των 25 συστατικών προέκυψε μέσα από την πειραματική διαδικασία, η οποία έδειξε ότι για μήκος μεγαλύτερο των 25 συστατικών δεν υπήρχε ουσιαστική διαφορά στα αποτελέσματα.
 - ο Τέλος ο περιορισμός αυτός ισχύει για όλο το δείγμα δεδομένων, καθώς πρέπει να υπάρχει ομοιογένεια ώστε να μπορούν να διεξαχθούν έγκυρα και αντικειμενικά συμπεράσματα σε ένα γενικό επίπεδο.

Εν τέλει, η σύνταξη του αλγόριθμου και η διαδικασία εφαρμογής του διεξάχθηκαν μέσω του υπολογιστικού προγράμματος MATLAB.

6.2.7 Εφαρμογή του Αλγόριθμου και Στατιστική Επεξεργασία

Σε αυτήν την ενότητα θα παρουσιαστεί η εφαρμογή του προτεινόμενου αλγόριθμου χρησιμοποιώντας ως παράδειγμα το παρακάτω μικρό κείμενο (βλέπε πίνακα 1):

“Many theorists have suggested that working memory capacity plays a crucial role in reading comprehension however, traditional measures of short-term memory, like digit span and word span, are either not correlated or only weakly correlated with reading ability”

Αρχικά μέσω των εξισώσεων 2-6, εξάγονται τα δεδομένα των μεταβλητών (r, s, k) (βλέπε εικ. 7). Στην διαδικασία μέσω του δείκτη συμφωνίας του Kendall τα δεδομένα των μεταβλητών (r, s, k) κατατάσσονται και υπολογίζονται σε σειρά κατάταξης αθροισμάτων (R_j), για κάθε συστατικό μέσω των εξισώσεων 8-10.

Για παράδειγμα το συστατικό *many*, το οποίο αποτελεί την πρώτη λέξη του παραπάνω παραδείγματος, αντιστοιχεί σε ένα διάνυσμα με γωνία απόκλισης r ίση με 1.6435 σε σχέση με το διάνυσμα – συνισταμένη του μικρού κειμένου. Το s ισούται με την τιμή 4 (αριθμός χαρακτήρων) και το k σύμφωνα με την κωδικοποίηση ASCII αντιστοιχεί στο 405. Με τον ίδιο τρόπο, εξάγονται οι τιμές για όλα τα 37 συστατικά του μικρού κειμένου, όπως μπορεί να δει κανείς στον Πίνακα 6.1.

Κατόπιν της διαδικασίας που αναλύθηκε παραπάνω, η τιμή του χ^2 με $(38-1=37)$ βαθμούς ελευθερίας υπολογίζεται μέσω της εξίσωσης 11 και από την τιμή αυτή λαμβάνεται η αθροιστική πιθανότητα $P = 3.5055e - 06$, μέσω της εξίσωσης 12.

Τέλος, χρησιμοποιώντας το έλεγχο μονόπλευρης κατανομής για πιθανότητα $H_0|P < 0.001$, η μηδενική υπόθεση γίνεται δεκτή, υποδεικνύοντας ότι δεν υπάρχει συσχέτιση μεταξύ των τριών μεταβλητών (r, s, k).

$$\chi_r^2 = 89,06$$

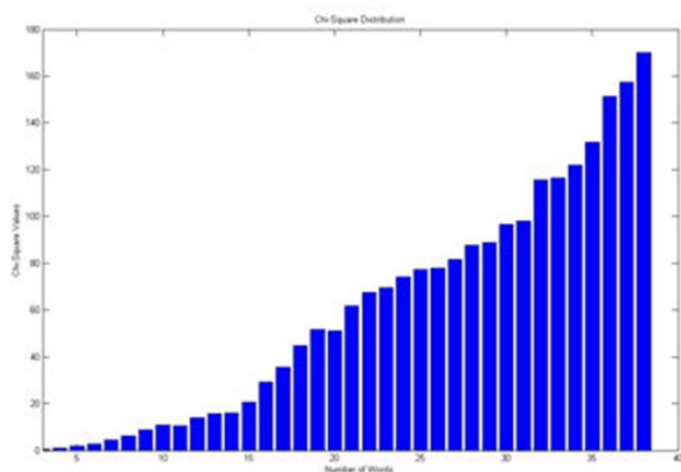
Πίνακας 6.1. Algorithm Implementation-Kendall's Coefficient of Concordance-Decision H_0 (an example short text)							
Word s	r		s		k		Sums of R_i
j	Data	rank	Data	rank	data	rank	
1	1.6435	28	4	12.0	405	8.0	48.0000
2	1.6699	29	9	32.5	997	33.0	94.5000
3	1.3755	23	4	12.0	420	9.0	44.0000
4	1.5487	26	9	32.5	971	32.0	90.5000
5	1.1232	16	4	12.0	433	11.0	39.0000
6	1.3378	22	7	25.5	769	28.0	75.5000
7	1.1818	18	6	21.0	665	22.0	61.0000
8	1.2430	21	8	30.0	846	30.0	81.0000
9	0.8526	11	5	18.0	553	19.0	48.0000
10	4.1013	37	1	1.0	97	1.0	39.0000
11	0.9320	12	7	25.5	739	26.0	63.5000
12	0.2012	4	4	12.0	434	12.5	28.5000
13	1.6753	30	2	3.0	215	3.0	36.0000
14	0.6862	7	7	25.5	730	24.5	57.0000
15	1.1718	17	13	38.0	1402	38.0	93.0000
16	0.6565	6	8	30.0	812	29.0	65.0000
17	0.9587	13	11	37.0	1179	37.0	87.0000
18	0.5982	5	8	30.0	869	31.0	66.0000
19	3.3125	35	2	3.0	213	2.0	40.0000
20	0.6884	8	10	35.0	1045	34.0	77.0000
21	0.0927	1	7	25.5	709	23.0	49.5000
22	1.2065	20	4	12.0	421	10.0	42.0000
23	0.7047	9	5	18.0	529	18.0	45.0000
24	1.3804	24	4	12.0	434	12.5	48.5000
25	2.8707	33	3	6.0	307	5.0	44.0000
26	1.5666	27	4	12.0	444	14.5	53.5000
27	1.4494	25	5	18.0	478	17.0	60.0000
28	3.3433	36	3	6.0	312	6.0	48.0000
29	0.8056	10	6	21.0	641	20.0	51.0000
31	3.3023	34	3	6.0	337	7.0	47.0000
32	0.1112	3	10	35.0	1061	35.5	73.5000
33	6.3095	38	2	3.0	225	4.0	45.0000
34	2.4095	31	4	12.0	450	16.0	59.0000
35	1.1958	19	6	21.0	653	21.0	61.0000
36	0.1046	2	10	35.0	1061	35.5	72.5000
37	2.8506	32	4	12.0	444	14.5	58.5000
38	1.1168	15	7	25.5	730	24.5	65.0000
$F=0,8023 \quad \chi_r^2 = 89,06 \quad P=3,5055e-06 \text{ accept for } P<0,001$							

6.2.8 Επαναληπτική Διαδικασία

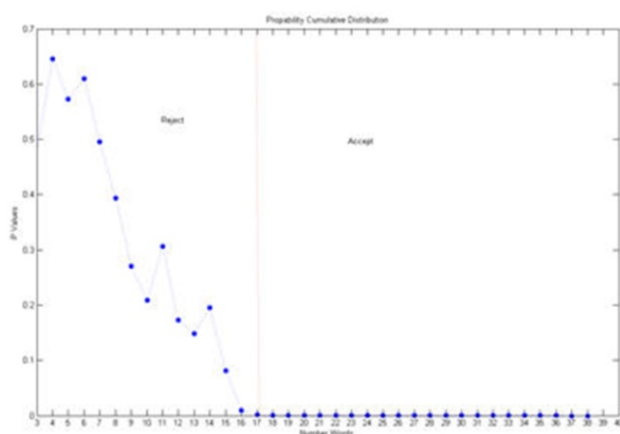
Χρησιμοποιώντας το ίδιο παράδειγμα μικρού κειμένου εφαρμόζεται ο αλγόριθμος $\left[\vec{U}_j\right]_3^{38}$. Η επαναληπτική διαδικασία διεξάγεται με $j=3:1:38$ λέξεις.

Ως επαναληπτική διαδικασία ορίζεται μια υπολογιστική/μαθηματική διαδικασία συνεχούς επανάληψης ενός κύκλου εργασιών/δραστηριοτήτων, ώστε να προσεγγιστεί ένα επιθυμητό αποτέλεσμα ((Everitt 2006, Downinget. al 2009).

Έτσι το αρχικό μικρό κείμενο τμηματοποιείται σε μικρά κείμενα με ποικίλα μήκη ξεκινώντας από μήκος 3 συστατικών έως 38 και αυτή η διαδικασία εκτελείται για κάθε ένα από τα παραπάνω μικρά κείμενα (βλέπε ενότητα 3.1 αυτού του άρθρου). Έπειτα ο έλεγχος χ^2 επαναλαμβάνεται (35) τριάντα πέντε φορές (βλέπε σχήμα 6.11) και υπολογίζονται οι αθροιστικές πιθανότητες και ο έλεγχος υπόθεσης (βλέπε σχήμα 6.12).



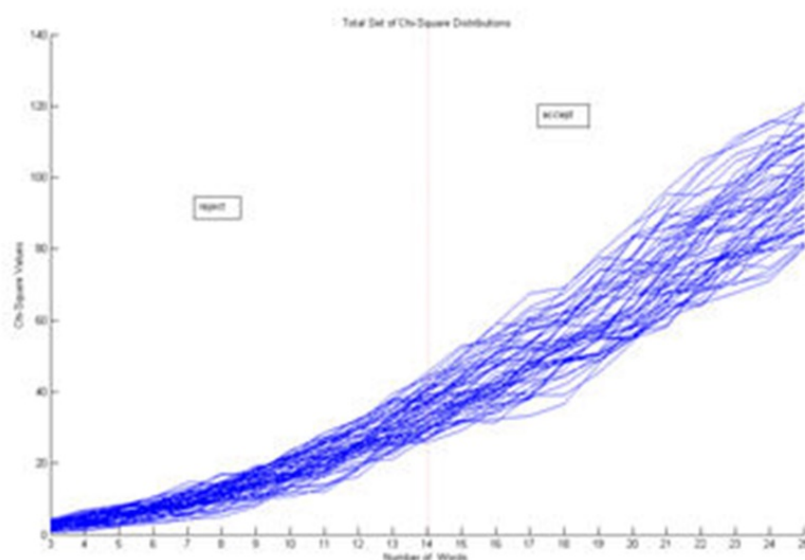
Σχήμα 6.11 Η κατανομή της συνάρτησης χ^2 σύμφωνα με την επαναληπτική διαδικασία του τμηματοποιημένου μικρού κειμένου



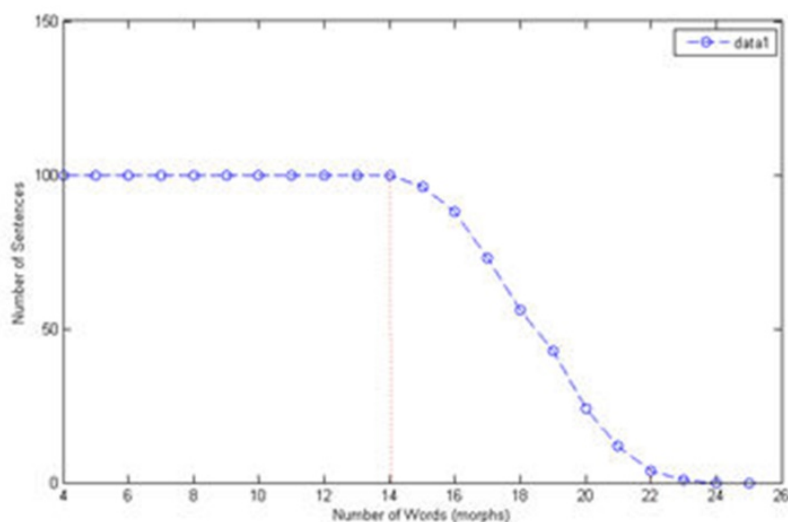
Σχήμα 6.12 Οι αθροιστικές πιθανότητες και ο έλεγχος υπόθεσης

6.2.9 Επαναληπτική Διαδικασία στο Σύνολο των Δεδομένων

Η επαναληπτική διαδικασία εκτελείται για κάθε μικρό κείμενο σε ένα εύρος από 3 έως 25 συστατικά, συνολικά (22) είκοσι δύο φορές. Η διαδικασία εκτελείται για ένα δείγμα 100 συνόλων δεδομένων, συνολικά 2200 υπολογισμοί. Έπειτα ο έλεγχος χ^2 των συνόλων δεδομένων αναπαρίσταται (βλέπε εικ. 4) και οι αθροιστικές πιθανότητες καθώς και ο έλεγχος της υπόθεσης παρουσιάζονται στα σχήματα 6.13-6.14.



Σχήμα 6.13 Αθροιστικές πιθανότητες και έλεγχος υπόθεσης για το σύνολο των δεδομένων



Σχήμα 6.14 Αθροιστική κατανομή πιθανοτήτων για το σύνολο των δεδομένων

Διεξάγοντας αξιολόγηση των παραπάνω στατιστικών αποτελεσμάτων, τα χωρίζουμε σε τρία μέρη και κάνουμε τις παρακάτω παρατηρήσεις:

1. Όπως μπορεί κανείς να δει στην κατανομή στο σχήμα 6.12 «Αθροιστικές πιθανότητες και έλεγχος υπόθεσης για το σύνολο των δεδομένων» μέσω της πειραματικής συνάρτησης, η συνεκτικότητα παρατηρείται για όλο συνολικά το δείγμα των 100 μικρών κειμένων σε μήκος πρότασης αποτελούμενο από 14 συστατικά. Από τα 14 και πάνω μπορεί κανείς να παρατηρήσει έλλειψη συνεκτικότητας στην πειραματική συνάρτηση.
2. Σύμφωνα με το πείραμα, όπως απεικονίζεται στην εικ. 5 «Αθροιστική κατανομή πιθανοτήτων για το σύνολο των δεδομένων», όλες οι υπολογισμένες πιθανότητες με βάση την εξίσωση 12 απορρίπτουν την μηδενική υπόθεση για μικρά κείμενα με μήκος ίσο με 14 συστατικά, με πιθανότητα, $p < 0.01$ ή p περίπου 99%. Συνεπώς αυτό δείχνει ότι στο επιλεγμένο δείγμα ο συσχετισμός των τριών μεταβλητών ισχύει μέχρι το μήκος των 14^{ων} συστατικών. Συνεχίζοντας για μήκος από 15 έως 23 συστατικά οι

συσχετίσεις ασθενούν. Τέλος για μικρά κείμενα με μήκος 25 συστατικών και πάνω οι συσχετισμοί πλέον δεν υφίστανται/καταρρέουν

3. Μέσω της εικ. 6 «*Αθροιστικές πιθανότητες και έλεγχος υπόθεσης*», η προαναφερθείσα διαδικασία απεικονίζει συγκεντρωτικά την μηδενική υπόθεση, η οποία είτε γίνεται δεκτή ή απορρίπτεται. Όπως φαίνεται στο σχήμα 6.14, όλες οι μηδενικές υποθέσεις απορρίπτονται για μήκος μικρών κειμένων έως 14. Στο διάστημα από 15 έως 22 συστατικά μπορεί κανείς να παρατηρήσει αστάθεια σε σχέση με την απόρριψη της μηδενικής υπόθεσης. Εν τέλει για μήκος μικρού κειμένου 23 συστατικών και πάνω, η μηδενική υπόθεση ικανοποιείται πλήρως, συνεπώς γίνεται δεκτή.

Συνοψίζοντας τα παραπάνω αποδεικνύουν ότι:

- Κάθε μικρό κείμενο εξαρτάται από τρεις προκαθορισμένες μεταβλητές, όταν το μικρό αυτό κείμενο αποτελείται από τουλάχιστον 14 συστατικά.
- Κάθε μικρό κείμενο, με μήκος περίπου 14 συστατικά, εξαρτάται από τρεις προκαθορισμένες μεταβλητές.

Αυτές οι δύο αρχές ικανοποιούνται πλήρως με πιθανότητα πάνω από 99% και με βαθμό ελευθερίας που κυμαίνεται από 3 – 13.

6.2.10 Συμπεράσματα

Το βασικό ερώτημα αυτού του άρθρου είναι η μέτρηση της λεξιλογικής συνεκτικότητας στα μικρά κείμενα, ώστε να επιτυγχάνεται η αναγνωσιμότητα και η κατανόηση. Παρότι αυτό το ζήτημα έχει μελετηθεί και στο παρελθόν, για πρώτη φορά με αυτήν την προσέγγιση εφαρμόζεται η εξαγωγή τριών μεταβλητών (r , s , k) που προσδιορίζουν μοναδικά κάθε συστατικό μικρού κειμένου. Αυτές οι μεταβλητές θεωρούνται σημαντικοί παράγοντες σε σχέση με την συνεκτικότητα του μικρού κειμένου. Διεξάχθηκε η διαδικασία υπολογισμού του δείκτη συμφωνίας Kendall ώστε να επιβεβαιωθεί ο συσχετισμός μεταξύ των τριών μεταβλητών, σε σχέση με το μήκος μικρών κειμένων. Όλα τα μικρά κείμενα για το παραπάνω πείραμα, είναι στην αγγλική γλώσσα, αλλά προκειμένου να επιβεβαιωθεί η υπόθεση ότι η γλώσσα του κειμένου δεν παίζει σημαντικό ρόλο, διεξάχθηκε και ένα μικρότερης κλίμακας πείραμα, με μικρά κείμενα στην γερμανική γλώσσα, στο οποίο προέκυψε ελάχιστη διαφοροποίηση, η οποία δεν διαταράσσει την μηδενική υπόθεση.

Σημαντικά στατιστικά αποτελέσματα προέκυψαν σε σχέση με τις τρεις μεταβλητές που επιλέχθηκαν, τα οποία έδειξαν ότι ο συσχετισμός είναι σημαντικός με πιθανότητα που αγγίζει το 95%, για μικρά κείμενα μήκους 17 συστατικών και το 99% για μικρά κείμενα μήκους 14 συστατικών. Συνεπώς, ένα μικρό κείμενο με μήκος που κυμαίνεται από 14-17 συστατικά μπορεί να θεωρηθεί ως ένας δείκτης επιτυχίας του κειμένου σχετικά με την επιτυχημένη μεταφορά του νοήματος του, το οποίο είναι ο λόγος δημιουργίας.

Εξάλλου, τα αποτελέσματα τα οποία εξάχθηκαν βρίσκονται σε συμφωνία με τα τρέχοντα επιστημονικά ευρήματα του τομέα της γλωσσολογίας, όπου Kornai (2008) αναφέρει ότι η μέσο μήκος (υπο) πρότασης αποτελείται από πάνω από 5 συστατικά, Cutts (2009) ότι το μέσο μήκος (υπο) πρότασης κυμαίνεται ανάμεσα σε 15-20 συστατικά, ώστε να διατηρείται η αναγνωσιμότητα και οι Taskaretal. (2004) διεξάγουν τα πειράματα του θέτοντας περιορισμούς στα μήκη (υπο) προτάσεων ίσο ή μικρότερο από 15 συστατικά. Το αποτέλεσμα του άρθρου αυτού βρίσκονται ακόμη σε συμφωνία με αποτελέσματα προερχόμενα από τις επιστήμες της γνωστικής ψυχολογίας και νευροεπιστήμης, όπου ο Baddeley (2003) αναφέρει πάλι τον περιορισμό των 15 συστατικών ανά (υπό) πρόταση. Με βάση τα πειράματα των Davemans και Carpenters (1980), των Andersonetal. (2001), του Zimny (1987), μπορεί κανείς να υπολογίσει ότι η μέση (υπό) πρόταση θα πρέπει να περιέχει περίπου 16 συστατικά, ώστε να: είναι κατανοητή, ώστε να επικοινωνεί το μήνυμα της με επιτυχία και να διατηρεί την αναγνωσιμότητα του κειμένου. Σε αυτήν την έρευνα παρατηρείται η συνάφεια μεταξύ των επιστημονικών τομέων της γλωσσολογίας, της νευροεπιστήμης και της γνωστικής ψυχολογίας.

Μελλοντικά Σχέδια:

- Η δημιουργία μιας οντολογίας, η οποία θα βασίζεται στις μεταβλητές και την στατιστική ανάλυση σε σχέση με το μήκος μικρών κειμένων που προέκυψαν από αυτήν την έρευνα.
- Η χρήση των τριών μεταβλητών (r , s , k), ώστε να διερευνηθεί πειραματικά πως αυτές επηρεάζουν τον τρόπο σύλληψης μικρών κειμένων από τον ανθρώπινο εγκέφαλο

- Να εκφραστεί η σχέση μεταξύ των τριών μεταβλητών με μια μαθηματική εξίσωση, ώστε να είναι δυνατή η περιγραφή και ο υπολογισμός της λεξιλογικής συνεκτικότητας των μικρών κειμένων με βάση το μήκος τους
- Ένα σημαντικό χάσμα μεταξύ της νευροεπιστήμης και της γλωσσολογίας, το οποίο αναφέρεται ως “OntologicalIncommensurabilityProblem” (OIP) (βλέπε ενότητα 1.2) όπου “τα βασικά στοιχεία της γλωσσολογικής θεωρίας δεν μπορούν να αντιστοιχηθούν με τις βασικές βιολογικές μονάδες που αναγνωρίζονται από την νευροεπιστήμη” όπως αναφέρουν οι Poeppel και Embick (2005). Αυτή η έρευνα μπορεί να αποτελέσει το θεμέλιο λίθο ώστε να μειωθεί το χάσμα μεταξύ των δύο επιστημονικών πεδίων, νευροεπιστήμη και γλωσσολογία, εστιάζοντας στο ζήτημα του μήκους των μικρών κειμένων, καθώς τα αποτελέσματα που προέρχονται και από τις δύο επιστήμες βρίσκονται σε συμφωνία.
- Τα αποτελέσματα αυτής της έρευνας σε συνδυασμό με τα συμπεράσματα από τα επιστημονικά πεδία νευροεπιστήμης και γλωσσολογίας, σε σχέση με το μήκος μικρών κειμένων μπορούν να αποτελέσουν μια νέα αρχή για την θεωρία ελέγχου και την σταθερότητα συστημάτων.
- Έρευνα που θα αφορά στις εφαρμογές της υπόθεσης του άρθρου σε κείμενα διαφόρων γλωσσών. Πιο συγκεκριμένα συγκριτική μελέτη ανάμεσα στις γλώσσες, των οποίων οι λέξεις αποτελούνται από μικρότερου μήκους μορφήματα και μεγαλύτερου μήκους μορφήματα, ώστε να οριστεί ένα σταθερό εύρος.
- Τέλος, λαμβάνοντας υπόψη ότι ο νόμος MA εφαρμόζεται σε γονιδιώματα όπως αναφέρουν και οι Baixeriesetal. (2013), Baixeries, Hernández-Fernández, Ferrer-i-Cancho (2012) και οι Forns et al. (2013) θα μπορούσε να διερευνηθεί επιστημονικά η πιθανή σχέση μεταξύ του αριθμού των μεταβλητών και του αριθμού των συστατικών σε άλλα επιστημονικά πεδία (όπως βιολογία), όπου θα μπορούσαν να θεωρηθούν ως ένας άξονας για την συνεκτικότητα του συστήματος και ως ανερχόμενη επιστημονική προτεραιότητα.

Εφαρμόζοντας τον παραπάνω αλγόριθμο κατασκευάζονται μέσω λογισμικού Matlab (100) εκατό κείμενα των (50) πενήντα λέξεων το καθένα και τότε παρατηρείται η σύγκλιση και των 100 περίεξ τιμών των (14) δεκατεσσάρων λέξεων ([βλέπε video 2](#)).

6.3 Επεξεργασία κειμένου για ανάκτηση και εξαγωγή πληροφορίας

Στην ενότητα αυτή διεξάγεται ένα πείραμα με σκοπό τη σύγκριση του γνωστού σχήματος απόδοσης βαρών $tf-idf$ με τον αλγόριθμο που έχει παρουσιαστεί στις παραπάνω ενότητες.

Αρχικά, για να πραγματοποιηθεί η σύγκριση αυτή, συλλέγεται ένα δείγμα 10 τίτλων επιστημονικών άρθρων. Στη συνέχεια οι τίτλοι αυτοί προ-επεξεργάζονται προκειμένου να αφαιρεθούν οι διακόπτουσες λέξεις, δίνοντας ως αποτέλεσμα προ-επεξεργασμένους τίτλους με πλήθος λέξεων ίσο με εννέα λέξεις για κάθε τίτλο, ώστε να υπάρχει ομοιογένεια στο δείγμα. Στη συνέχεια μέσω του προγράμματος MATLAB, εφαρμόζεται ο αλγόριθμος ώστε να συγκεντρωθούν οι τιμές για το δείγμα και πιο συγκεκριμένα, όλα τα βήματα όπως αναφέρονται στις ενότητες III. Α και III. Β. Έτσι τα αποτελέσματα φαίνονται στον πίνακα 6.2:

Πίνακας 6.2. Αποτελέσματα εφαρμογής αλγόριθμου για το δείγμα

Τίτλοι	Τριάδες αλγόριθμου	Cosθ		
T1	applications nanoparticles assembly	0.0283	0.0319	0.0602
T2	production agriculture tools	0.0254	0.0419	0.0693
T3	advances diagnosis applications	0.0539	0.1102	0.1118
T4	synthesis applications properties	0.0179	0.0343	0.0361

T5	life itself subjectivity	0.0989	0.1062	0.1111
T6	throughput molecular techniques	0.0418	0.0631	0.0776
T7	cells technological opportunities	0.0168	0.0274	0.0376
T8	scientific legislative stem	0.0533	0.0884	0.1409
T9	mammography breast diagnosis	0.0226	0.0887	0.1200
T10	hypoxia implications tumor	0.0278	0.0356	0.0635

Στη συνέχεια πραγματοποιείται ο υπολογισμός των τιμών με βάση το σχήμα απόδοσης βάρους *tf-idf*, με τη βοήθεια του προγράμματος MATLAB. Δημιουργείται η μήτρα όρων εγγράφων και συλλέγονται οι όροι. Έπειτα υπολογίζεται η βαρύτητα κάθε όρου ανά τίτλο, με βάση τους παράγοντες απόδοσης βάρους του *tf-idf*. Στον παρακάτω πίνακα 6.3, φαίνονται τα αποτελέσματα:

*Πίνακας 6.3. Αποτελέσματα εφαρμογής *tf-idf* για το δείγμα*

Τίτλοι	Όροι με βάση τη βαρύτητα	Συχνότητα
T1	FePt, assembly, century, nanobiotechnology, scientific / advances	2,05 / 0,25
T2	agriculture, biotechnology, future, new, pig, production, tools / advances	2,05 / 0,25
T3	accuracy, genetic, improve, preimplantation, range/ advances	2,05 / 0,25
T4	aliphatic, nanotechnology, polyester, polymer, properties, stars/ biomedicine	2,05 / 0,41
T5	century, first, itself, life, politics, power, subjectivity, twenty / biomedicine	2,05 / 0,41
T6	high, molecular, nanobiotechnology, systems, techniques, throughput / advances	2,05 / 0,25
T7	aging, concepts, opportunities, regenerative / advances	2,05 / 0,25
T8	admixed, embryos, ethical, human, legislative, scientific / advances	2,05 / 0,25
T9	aided, breast, cancer, computer, detection, mammography / advances	2,05 / 0,25
T10	assessing, hypoxia, implications, methods, planning, treatment, tumor, vivo / advances	2,05 / 0,25

Τέλος, προχωρώντας στη σύγκριση του αλγόριθμου με το *tf-idf* καταλήγουμε στα εξής αποτελέσματα, όπως συνοψίζονται στον πίνακα 6.4:

Πίνακας 6.4: Κοινοί όροι ανά τίτλο για το δείγμα

Τίτλοι	Κοινοί όροι
T1	assembly
T2	production agriculture tools
T3	advances
T4	properties

T5	life itself subjectivity
T6	throughput molecular techniques
T7	opportunities
T8	scientific legislative
T9	mammography breast
T10	hypoxia implications tumor

Για όλο το δείγμα συνολικά, οι όροι που έχουν εξαχθεί με βάση τον αλγόριθμο εμπεριέχονται στο σύνολο των αποτελεσμάτων σύμφωνα με το *tf-idf*. Πιο συγκεκριμένα:

- 40% των αποτελεσμάτων με βάση το *tf-idf*, εμπεριέχουν και τους τρεις όρους εξαγωγής με βάση τον αλγόριθμο.
- 20% των αποτελεσμάτων με βάση το *tf-idf*, εμπεριέχουν τους δύο όρους εξαγωγής με βάση τον αλγόριθμο.
- 40% των αποτελεσμάτων με βάση το *tf-idf*, εμπεριέχουν και έναν από τους όρους εξαγωγής με βάση τον αλγόριθμο.

6.4 Αξιολόγηση Βιβλιοθήκης με χρήση Νευρωνικού Δικτύου

Σκοπός της εργασίας που πρόσφατα δημοσιεύθηκε (VThimiakouetall 2015) είναι η δημιουργία ενός συστήματος αξιολόγησης βιβλιοθηκών, το οποίο χρησιμοποιείται με στόχο την μέτρηση αποδοτικότητας των λειτουργιών μιας βιβλιοθήκης. Οι παρακάτω μετρήσεις πραγματοποιήθηκαν με βάση τους καθιερωμένους δείκτες του ΜΟΠΑΒ. Οι δείκτες που επιλέγηκαν σύμφωνα με ορισμένα κριτήρια ώστε να αφορούν όλα τα είδη των βιβλιοθηκών και όχι μόνο κάποιες ειδικές ή κάποιο συγκεκριμένο είδος. Έτσι επιλέχθηκαν δείκτες υπηρεσιών που καλύπτουν τις δύο από τις τρεις κύριες λειτουργίες μιας βιβλιοθήκης οι οποίες είναι πρόσκτηση και δανεισμός και επιπρόσθετα είναι εύχρηστοι (δηλαδή εύκολα εξαγόμενοι από τη στατιστική υπηρεσία της βιβλιοθήκης) και επεξεργάσιμοι για την εξαγωγή αποτελεσμάτων. Τα εξαγόμενα αποτελέσματα από την πραγματοποιηθείσα επεξεργασία και κρίση πρέπει αν θεωρηθούν έγκυρα. Η διαδικασία βασίστηκε στη δημιουργία γνώμης ενός ειδικού κριτή υπηρεσιών βιβλιοθήκης σχετικά με την αποδοτικότητα συνολικά των δεικτών. Στη συνέχεια με χρήση νευρωνικού δικτύου τύπου Elman έγινε εξαγωγή γνώσης του ειδικού ώστε αυτή να χρησιμοποιηθεί αντικειμενικά σε κάθε άλλη υπηρεσία βιβλιοθήκης με ένα blind ακριβοδίκαιο σύστημα. Προς αυτό το σκοπό η εργασία χωρίστηκε σε τέσσερα στάδια.

- Στο πρώτο στάδιο έγινε η επιλογή των δεικτών βλ. Πίνακας 6.5.

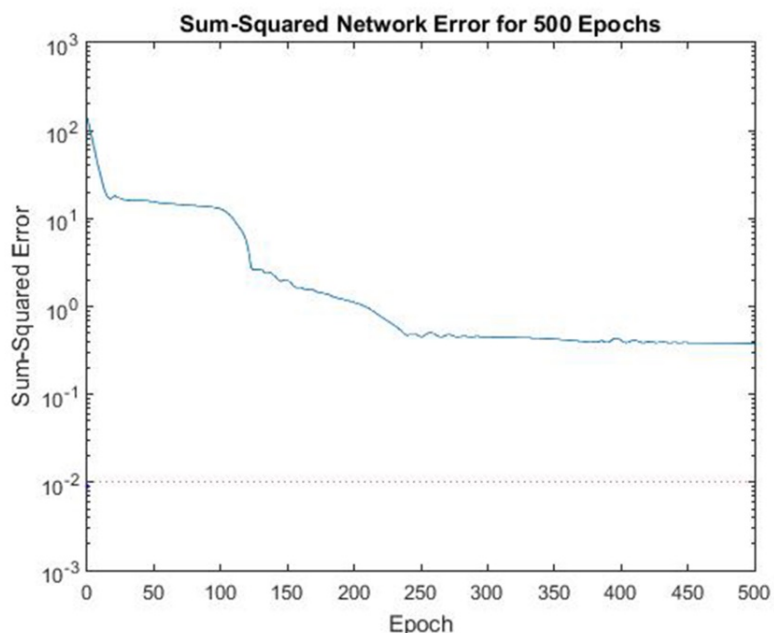
Πίνακας 6.5 Υποσύνολο δεικτών ΜΟΠΑΒ		
P32	Ποσοστό ενεργών εγγεγραμμένων χρηστών (Μελών της Ακαδημαϊκής Κοινότητας) της Βιβλιοθήκης επί του Συνολικού Πληθυσμού	(D2.2/D1)*100
P33	Χρήση συλλογής βιβλίων	D3/D4
P34	Δανεισμοί υλικού συλλογής κατά κεφαλή	D3/D1
P36	Αριθμός τεκμηρίων συλλογής Βιβλιοθήκης κατά κεφαλή	D5/D1
P31	Εικονικές επισκέψεις (visit –sessions) του Website της Βιβλιοθήκης μηνιαίως	D11/D12
P47	Ώρες ημερήσιας λειτουργίας της Βιβλιοθήκης	P47=D20
P48	Διαθέσιμος κατά κεφαλή χώρος Βιβλιοθήκης σε m2	D21/D1
P49	Διαθέσιμες θέσεις αναγνωστηρίου για τους χρήστες στους χώρους της Βιβλιοθήκης κατά κεφαλή	D22/D1
P51	Προσωπικό κατά κεφαλή	D24/D1
P50	Διαθέσιμες θέσεις εργασίας σε Η/Υ για τους χρήστες στους χώρους	D23/D1

	της	
P4	Βιβλιοθηκονόμοι/άριοι ως προς το διοικητικό προσωπικό της Βιβλιοθήκης	D25/D26
P16	Ποσοστό δαπανών βιβλιοδεσίας επί τις συνολικές λειτουργικές δαπάνες	(D36/D37)*100
P64	Συνολικές δαπάνες Βιβλιοθήκης κατά κεφαλή.	D39/D1

- Στο δεύτερο στάδιο συλλέχθηκαν τα δεδομένα και βαθμολογήθηκαν από τον ειδικό με τη διαδικασία που εμφανίζεται στον Πίνακα 6.5.

Πίνακας 6.5 Βαθμολόγηση των των δεικτών			
Δείκτες	ΑΡΙΣΤΗ (HIGH)	METΡΙΑ (GOOD)	ΑΔΥΝΑΜΗ (LOW)
D1	10*	10*	10*
D11	1 (1- 0.3)	0.2 (0.2- 0.1)	0.1 (0.1- 0)
D12	10*	10*	10*
D20	24 (24- 11)	10 (10-7)	6 (6-0)
D21	0.4 (0- 0.4)	0.5 (0.5- 0.6)	0.7 (0,7- 1)
D22	0,4 (0- 0.4)	0,5 (0.5- 0.6)	0.7 (0,7- 1)
D23	1 (1- 0.6)	0.5 (0.5- 0.3)	0.2 (0.2- 0)
D24	0.3 (1- 0,3)	0.2 (0.2- 0.06)	0.05 (0.05- 0)
D25	0.15*	0.05*	0.0075*
D26	0.15*	0.15*	0.006375*
D3	1,00 (1- 0,60)	0.50 (0.50- 0.30)	0.2 (0.2- 0)
D37	0.1*	0.5 (0.5- 0,6)*	1 (0,7- 1)*
D39	0.1*	0.5 (0.5- 0,6)*	1(0,7- 1)*
D4	0.5*	0.3 (0.3- 0.2)*	0.1 (0.1- 0)*
D5	1.0 (1- 0,6)	0.5 (0.5- 0.30)	0.2 (0.20- 0)
D2.2	1.0 (1- 0,3)	0.2 (0.2- 0,1)	0.1 (0.01-0)
D36	0.1 (0- 0.3)	0.3 (0.4- 0,6)	0.7 (0.7- 1)

- Στο τρίτο στάδιο κατασκευάστηκε ένα Elman Νευρωνικό Δίκτυο και ακολούθησε η εκπαίδευση του νευρωνικού δικτύου. Κατά το στάδιο αυτό δημιουργήθηκαν τρεις κατηγορίες εισόδου (Αριστη -Α, καλή-Β, ανεπαρκής -C). Η κάθε είσοδος κατασκευάστηκε να δέχεται 7 διανύσματα εισόδου διανυσμάτων (μεγέθους 1x13) που αντιστοιχούν στις κατηγορίες Α, Β, C (βλ. Πίνακας 6. Στη συνέχεια εκπαιδεύουμε τις τρεις κατηγορίες διανυσμάτων σε δύο στάδια. Στο πρώτο στάδιο χρησιμοποιούμε ένα κρυφό επίπεδο 5 νευρώνων και κατά το οποίο δημιουργούνται βάρη πρώτου επιπέδου. Η διαδικασία αυτή υλοποιήθηκε μέσω της ρουτίνας «initelm» της γλώσσας Matlab. Τέλος η διαδικασία ολοκληρώνεται με το δεύτερο στάδιο εκπαίδευσης κατά το οποίο οι τρεις κατηγορίες διανυσμάτων επιχειρούν να προσεταρισθούν τοπολογικά τα δεδομένα. Η διαδικασία τερματίζει σε 500 εποχικές προσπάθειες (epochs) (βίντεο 3) με ελαχιστοποίηση σφάλματος που αφορά την ιδεατή τοπολογική εκχώρηση σε κατηγορίες. Η διαδικασία αυτή υλοποιήθηκε μέσω της ρουτίνας «trainelm» της γλώσσας Matlab (Zhang, Y.H., κα. 2011) και παρουσιάζεται στο [βίντεο 3](#) και σχήμα 6.15.



Σχήμα 6.15 Εκπαίδευση Νευρωνικού Δικτύου Elman σε αριθμό 500 δοκιμών

Πίνακας 6.6	Δεδομένα διανυσμάτων εισόδου και η κατασκευή συντελεστών βαρύτητας κατά τη διάρκεια εκπαίδευσης ενός TNA Elman (V Thimiakou et all 2015)		
Ένα διάνυσμα εισόδου με βαθμολογία από τον κριτή «καλή»	Δείκτες ΜΟΠΑΒ	W1 (βάρη συσχέτισης)	W2 (βάρη συσχέτισης)
0.0200	P32	0.0100	0.0100
1.3300	P33	2.0000	1.0000
0.0400	P34	0.0200	0.0100
0.0400	P36	0.0100	0.0100
0.0200	P31	0.0100	0.0100
0.2900	P47	0.1300	0.1700
0.0600	P48	0.0800	0.0800
0.0600	P49	0.0800	0.0900
0.0200	P51	0.0100	0.0100
0.0400	P50	0.0200	0.0100
0.3300	P4	1.1800	1.1800
0.9000	P16	0.7000	0.7000
0.0500	P64	0.1000	0.1000

Στο τέταρτο στάδιο υλοποιείται η αξιολόγηση του Νευρωνικού Δικτύου. Η αξιολόγηση του νευρωνικού πραγματοποιείται με τη χρήση σιγμοειδούς συναρτήσεως “y” και εξάγεται από τον υπολογισμό εφαρμογής της σχέσης 6.19 στην 6.20. Τότε τα εξαγόμενα αποτελέσματα τιμών αποτελούν τον κύριο παράγοντα κατάταξης σε μια από τις τρεις κατηγορίες ενός αταξινόμητου για το σύστημα διανύσματος (VThimiakouetall 2015).

$$y = \frac{1}{1 + e^{-x}} \quad (6.19)$$

Τα αποτελέσματα της κατηγοριοποίησης υπολογίζονται από την παρακάτω σχέση:

$$\begin{aligned} y \leq 1.5 &\rightarrow \text{πολύ_καλή} \\ 1.5 < y \leq 2.5 &\rightarrow \text{καλή} \\ y > 2.5 &\rightarrow \text{ανεπαρκής} \end{aligned} \quad (6.20)$$

Ο παραπάνω αλγόριθμος υλοποιήθηκε χρησιμοποιώντας τη ρουτίνα «*simuelm*» της γλώσσας Matlab. Τα αποτελέσματα κατάταξης των υπολοίπων 9 διανυσμάτων (3 από κάθε κατηγορία) εμφανίζονται στον Πίνακα 6.7.

Πίνακας 6.7	Αποτελέσματα κατάταξης του Elman TNA (VThimiakou et al 2015)		
ΚΑΤΑΤΑΞΗ	A ≤ 1.5	1.5 < B ≤ 2.5	C > 2.5
A	3	0	0
B	0	2	1 (2.6)
C	0	0	3

Η αξιολόγηση των αποτελεσμάτων έδειξαν ότι τα διανύσματα της A και C αναγνωρίστηκαν με επιτυχία (100%) από το σύστημα του Τεχνικού Νευρωνικού Δικτύου. Μία λανθασμένη κατάταξη παρουσιάστηκε μόνο στη B ομάδα διανυσμάτων που οριακά με αποτέλεσμα (2.6) κατατάχτηκε στη C κατηγορία. Η μέθοδος εκπαίδευσης του τεχνητού νευρωνικού δικτύου διεξήχθη με επιτυχία και τα αποτελέσματα που απέφερε το σύστημα ήταν σε συμφωνία με αυτά του ειδικού, χωρίς στατιστικά λάθη που να θέτουν σε αμφισβήτηση την αποτελεσματικότητά του.

Βιβλιογραφία

- Altmann, G. (1980). Prolegomena to Menzerath's Law. *Glottometrika* 2, 1–10. Bochum: Brock-meyer
- Anderson J. R., Budson R. and Reder L. M. (2001). A theory of sentence memory as part of a General Theory of Memory. *Journal of Memory and Language*. 45. 337-336.
- Baddeley A. (2003). Working memory: looking back and looking forward. *Nature reviews - neuroscience*. 5. 829-839.
- Baixeries J. et al. (2013). The parameters of the Menzerath-Altmann Law in genomes. *Journal of Quantitative Linguistics*. 20 (2). 94–104.
- Baixeries J., Hernández-Fernández A., Ferrer-i-Cancho R. (2012), Random models of Menzerath–Altmann law in genomes. *Biosystems*. 107 (3), 167–173.
- Carstens W. (2001). Text Linguistics: relevant linguistics? Poetics and linguistics; discourses of war and conflict Conference. 588-595.
- Cutts, M. (2009). Oxford guide to plain English. 3rded. Oxford: Oxford University Press.
- Daveman M., Carpenter P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*. 19. 450-466.
- Eroglu S. (2013). Menzerath–Altmann law for distinct word distribution analysis in a large text. *Physica A*. 392 (12). 2775–2780
- Fahnestock J. (1983). Semantic and Lexical Coherence. *College Composition and Communication*. 34 (4). 400-416. National Council of Teachers of English.
- Forns N. et al. (2013). The challenges of statistical patterns of language: The case of Menzerath 's law in genomes. *Complexity*. 18 (3). 11–17.
- Ge Song et al (2014). Short Text Classification: A Survey. *Journal of Multimedia*. 9 (5). 635-643. Academy Publisher.
- Halliday, M.A.K., Hasan R. (1976). Cohesion in English. London: Longman.
- Harispe S. et al. (2013). Semantic measures for the comparison of units of language, concepts or entities from text and knowledge base analysis. *Arxiv*. 1310. 1285. 1-159.
- Harispe, S. et al. (2013). The semantic measures library and toolkit: fast computation of semantic similarity and relatedness using biomedical ontologies. *Bioinformatics*. 30 (5). Oxford: Oxford University Press.
- Kornai A. (2008). Mathematical linguistics. [online]. Advanced Information and Knowledge Processing. London: Springer.
- Luděk Hřebíček (2002). Zipf's Law and Text. *Glottometrics*. 3. 27-38. Ram Verlag.
- Poeppel, D., Embick, D. (2005). Defining the relation between linguistics and neuroscience. *Twenty-first Century Psycholinguistics: Four Cornerstones*. 103–118.
- Poulos M., Papavaslopoulos S., Chrissikopoulos V. (2006). A text categorization technique based on a numerical conversion of a symbolic expression and an onion layers algorithm . *Journal of Digital Information*. 6 (1).
- Raghavan V. V., Wong S. K. M. (1986). A critical analysis of vector space model for information retrieval. *Journal of the American Society for Information Science*. 37 (5). 279-287.
- Richards J. C., Platt J., Platt H. (1992). Longman dictionary of language teaching and applied linguistics. London: Longman
- Salton G., Wong A., and Yang C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*. 18 (11). 613-620.

- Sparks J.R. (2012). Language/discourse comprehension and understanding. *Encyclopedia of the Learning Sciences*. 1713-1717. Springer.
- Taskar B. et al. (2004). Max-Margin Parsing. *Proceedings of EMNLP 2004*. 1-8.
- Těšitelová M. (1992). Quantitative linguistics. *Linguistics and literary studies in Eastern Europe (LLSEE)*. 37. Philadelphia: John Benjamins
- V Thimiakou, S Kalampokh, F Kargioti, A Tsichlaki and Poulos (2015) Αξιολόγηση Υπηρεσιών Βιβλιοθηκών μέσω Τεχνικού Νευρωνικού Δικτύου. *Hellenic Academic Libraries Journal* 2 (2015) 47–60
- Turney P. D., Pantel P. (2010). From frequency to meaning: vector space models of semantics. *Journal of Artificial Intelligence Research*. 37 (1). 141-188.
- Xiaojun Quan, Gang Liu et al. (2009). Short text similarity based on probabilistic topics. *Knowl Inf Syst*. 473-491.
- Zar Jerrold H. (1999). *Biostatistical Analysis*. 4th ed. Upper Saddle River, N.J.: Prentice Hall.
- Zimny, S. T. (1987). Recognition memory for sentences from a discourse. Unpublished doctoral dissertation, Boulder: University of Colorado.
- Zhang, Y.H., Yuen, C.W.M., Wong, W.K., Kan, C., 2011. An intelligent model for detecting and classifying color-textured fabric defects using genetic algorithms and the Elman neural network. *Textile Research Journal* 81, 1772–1787.

ΣΥΝΟΛΙΚΗ ΒΙΒΛΙΟΓΡΑΦΙΑ

- Altmann, G. (1980). Prolegomena to Menzerath's Law. *Glottometrika* 2, 1–10.
- Ahlsen, E. (2006). *Introduction to Neurolinguistics*. Amsterdam/Philadelphia: John Benjamins Publishing
- Akobeng, A. K. (2007). Understanding diagnostic tests 1: sensitivity, specificity and predictive values. *Acta Paediatr.* 96 (3). P. 338-41.
- Altmann, G. (2002). Zipfian linguistics. *Glottometrics*. 3. P. 19-26.
- Altmann, G. (1980). Prolegomena to Menzerath's law. *Glottometrika*. 2. P. 1–10.
- Anderson J. R., Budiu R. and Reder L. M. (2001). A theory of sentence memory as part of a General Theory of Memory. *Journal of Memory and Language*. 45. 337-336.
- Baddeley A. (2003). Working memory: looking back and looking forward. *Nature reviews - neuroscience*. 5. 829-839.
- Baixeries J. et al. (2013). The parameters of the Menzerath-Altmann Law in genomes. *Journal of Quantitative Linguistics*. 20 (2). 94–104.
- Baixeries J., Hernández-Fernández A., Ferrer-i-Cancho R. (2012), Random models of Menzerath–Altmann law in genomes. *Biosystems*. 107 (3), 167–173.
- Baeza-Yates, R. and Ribeiro-Neto, B. (2011). *Modern Information Retrieval*. 2nd Ed. New York: Addison Wesley
- Baixeries, J. et al. (2013). The parameters of the Menzerath-Altmann Law in genomes. *Journal of Quantitative Linguistics*. 20 (2). P. 94–104.
- Bendersky, M. and Croft, B. W. (2012). Modeling Higher-Order Term Dependencies in Information Retrieval using Query Hypergraphs. In *SIGIR '12 Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. P. 941-950.
- Blair, D. C. and Kimbrough, S. O. (2002). Exemplary Documents: A Foundation for Information Retrieval Design. *Information Processing and Management*. 38 (3). P. 363-379
- Bolshakov, I. A. and Gelbukh, A. (2004). *Computational linguistics: Models, Resources, Applications*. Mexico: Universidad Nacional Autonoma.
- Brants, T. (2004). Natural Language Processing in Information Retrieval. In *Proceedings of the 14th Meeting of Computational Linguistics in the Netherlands*. P. 1-13
- Buettcher, S., Clarke, C. L. A. and Cormack G. V. (2010). *Information Retrieval: Implementing and Evaluating Search Engines*. Cambridge, Mass.: MIT Press
- Buk, S., Rovenchak, A. (2008). Menzerath – Altmann law for syntactic structures in Ukrainian. *Glottology*. 1(1). P. 10-17
- Canfora, G. and Cerulo, L. (2004). A Taxonomy of Information Retrieval Models and Tools. *Journal of Computing and Information Technology(CIT)*. 12 (3). p. 175–194
- Carstens, W. (2001). Text linguistics: relevant linguistics? *Poetics, Linguistics and History: Discourses of War and Conflict. PALA Conference Papers 1999*. Potchefstroom University, South Africa, p. 588-595.
- Ceri, S. et al. (2013). *Web Information Retrieval*. New York: Springer
- Crestani, F. and Wu, S. (2006). Testing the cluster hypothesis in distributed information retrieval. *Information Processing and Management: an International Journal archive*. 42 (5). P. 1137 – 1150
- Kornai A. (2008). Mathematical linguistics. [online]. Advanced Information and Knowledge Processing. London: Springer.

- Croft, B. W., Metzler D. and Strohman, T. (2010). *Search Engines: information retrieval in practice*. Boston: Pearson.
- Dale, E. and Chall, J. S. (1949). The concept of readability. *Elementary English*. 26 (23). P. 19-26
- De Beaugrande, R. A. and Dressler, W. U. (1981). *Introduction to text linguistics*. New York: Longman.
- De Saussure, F. (1966). *Course in General Linguistics*. 3rd ed. New York: McGraw-Hill Book Company
- Diggle, P. J. and Chetwynd A. G. (2011). *Statistics and Scientific Method: an Introduction for Students and Researchers*. Oxford: Oxford University Press
- Dubay, W. H. (2004). *The Principles of Readability*. Costa Mesa, CA: Impact Information.
- Dubin, D. (2004). The most influential paper Gerard Salton never wrote. *LibraryTrends*. 52 (4). P. 748-764.
- Eroglu, S. (2013). Menzerath – Altmann law for distinct word distribution analysis in a large text. *Physica A*. 392 (12). P. 2775–2780
- Espunya i Prat, A. (1994). Computational Linguistics: a Brief Introduction. *Links and Letters*. 1. P. 9-23.
- Everitt, B. S. (2006). *The Cambridge Dictionary of Statistics*. 3rd ed. Cambridge: Cambridge University Press.
- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*. 27. P. 861–874
- Fisher, R. A. (1950). *Statistical methods for research workers*. 11th ed. Rev. London: Oliver and Boyd.
- Friedman, M. (1937). The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. *Journal of the American Statistical Association*. 32 (200). P. 675 - 701.
- Carstens W. (2001). Text Linguistics: relevant linguistics? Poetics and linguistics; discourses of war and conflict Conference. 588-595.
- Galvez, C., de Moya-Anegon, F. and Solana, V. H. (2005). Term conflation methods in information retrieval: Non-linguistic and linguistic approaches. *Journal of Documentation information (JDOC)*. 61 (4). P. 520-547
- Gries, S. Th. (2013). *Statistics for linguistics with R*. 2nd rev. and ext. ed. Berlin: De Gruyter Mouton
- Gries, S. (to appear in). Quantitative methods in linguistics. In Wright J. D. (ed.) *International Encyclopedia of the Social and Behavioral Sciences*. Amsterdam: Elsevier. Retrieved 10 November 2014 from: http://www.linguistics.ucsb.edu/faculty/stgries/research/ToApp_STG_QuantMethInLing_IESBS2.pdf
- Grossman, D.A. and Frieder, O. (2000). *Information Retrieval: Algorithms and Heuristics*. The Kluwer International Series in Engineering and Computer Science (SECS). 461. Boston: Kluwer
- Haitao, L., Wei, H. (2012). Quantitative Linguistics: State of the Art, Theories and Methods. *Journal of Zhejiang University (Humanities and Social Science)*. 43(2). P. 178-192.
- Harispe, S. et al. (2013). Semantic measures for the comparison of units of language, concepts or entities from text and knowledge base analysis. *Arxiv*. 1310 (1285). P. 1-159
- Hauser, M.D., Chomsky, N. and Fitch, W.T. (2002). The faculty of language: what is it, who has it, and how did it evolve? *Science*. 22 (298 (5598)). Pp.1569-79
- Hyafil, L., & Rivest, R. L. (1976). Constructing optimal binary decision trees is NP-complete. *Information Processing Letters*, 5(1), 15–17.
- Hayes, B. (2010). *Introductory linguistics*. Los Angeles: Department of Linguistics University of California. Retrieved 13/12/2015 from: <http://www.linguistics.ucla.edu/people/hayes/20/Text/HayesIntroductoryLinguistics2010.pdf>
- Haykin, S.S., 2009. Neural networks and learning machines. Pearson Education Upper Saddle River.
- Hempel, C. G. (1964). *Fundamentals of Concept Formation in Empirical Science*. International Encyclopedia of Unified Science. 2 (7). Chicago: University of Chicago Press

- Hoey, M. (2003). *Lexical priming and the properties of text*. University of Liverpool. Retrieved 14/09/2014 from: <http://www.monabaker.com/tsresources/LexicalPrimingandthePropertiesofText.htm>
- Hogan, P. C. (ed.) (2014). *The Cambridge Encyclopedia of the Language Sciences*. Cambridge: Cambridge University Press
- Hrebicek, L. (2002). Zipf's Law and Text. *Glottometrics*. 3. P. 27-38.
- Ingwersen, P. (1992). *Information Retrieval Interaction*. London: Taylor Graham Publishing
- Johnson, K. (2008). *Quantitative Methods in Linguistics*. Malden (USA): Blackwell.
- Joho, H. and Jose, J. M. (2006). A Comparative Study of the Effectiveness of Search Result Presentation on the Web. *Advances in Information Retrieval: Lecture Notes in Computer Science*. 3936. P. 302-313
- Jurafsky, D. and Martin, H. J. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall series in artificial intelligence. Upper Saddle River, NJ: Prentice Hall PTR
- Karmeshu, J (2003). *Entropy measures, maximum entropy principle and emerging applications* (Vol. 119). Springer Science & Business Media. Retrieved from <https://www.google.com/books?hl=el&lr=&id=zCoIvQTNOOIC&oi=fnd&pg=PA1&dq=Maximum+Entropy+Principle+and+Emerging+Applications&ots=szlmSmBCt&sig=yDz74WwFS3beemTXBvasEoE2qwM>
- Kargupta, H., Han, J., Philip, S. Y., Motwani, R., & Kumar, V. (2008). *Next generation of data mining*. CRC Press. Retrieved from https://www.google.com/books?hl=el&lr=&id=YTIVP6OnfrC&oi=fnd&pg=PP1&dq=Next+Generation+of+Data+Mining&ots=HQsIYw0N84&sig=3fxwpFU8TbpPdZzI_6cTPjP4OtM
- Κολυβά, Μαριάννα, από τις Διδακτικές Σημειώσεις των Μαθημάτων Αρχειονομία I-II, Ιόνιο Πανεπιστήμιο, Τμήμα Αρχειονομίας-Βιβλιοθηκονομίας, Ακαδημαϊκό έτος 2001-2002.
- Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 79–86.
- Kamps, J. (2009). Presenting Structured Text Retrieval Results. *Encyclopedia of Database Systems*. P. 2130-2134
- Karaman, B. I. (2003). *Polysemy in Natural Language: Case Studies on the Structural Description of Polysemous Lexemes in English, German and Turkish*. PhD thesis. University of Surrey. Retrieved 22/09/2014 from <http://epubs.surrey.ac.uk/2816/1/412061.pdf>
- Kennedy, C. (2009). Ambiguity and Vagueness: an Overview. In: Maienborn, C., Heusinger, K. and Portner, P. (Eds.). *Semantics: an International Handbook of Natural Language Meaning*. Handbooks of Linguistics and Communication Science (HSK). 33 (1). Berlin: De Gruyter Mouton
- Kovacs, E. (2011). Polysemy in Traditional vs Cognitive Linguistics. *Eger Journal of English Studies*. XI. P. 3–19
- Kowalski, G. (2011). *Information Retrieval Architecture and Algorithms*. New York: Springer.
- Kracht, M. (2007). *Introduction to linguistics*. Los Angeles: Department of Linguistics, UCLA. Retrieved 08/12/2014 from: <http://www.homes.uni-bielefeld.de/mkracht/html/ling-intro.pdf>
- Kulacka, A. and Macutek, J. (2007). A discrete formula for the Menzerath - Altmann law. *Journal of Quantitative Linguistics*. 14 (1). P. 23-32
- Kuropka, D. (2004). *Modelle zur Repräsentation natürlichsprachlicher Dokumente: Ontologie-basiertes, Information-Filtering und -Retrieval mit relationalen Datenbanken*. Advances in Information Systems and Management Science. 10. Berlin: Logos.
- Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer. Retrieved from <http://link.springer.com/content/pdf/10.1007/978-1-4614-6849-3.pdf>

- Lalkhen, A.G. and McCluskey, A. (2008). Clinical tests: sensitivity and specificity. *Continuing Education in Anaesthesia, Critical Care and Pain*. 8. P. 221–223.
- Lan, M., Tan, C. L. and Su, J. (2009). Supervised and Traditional Term Weighting Methods for Automatic Text Categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 31(4). P. 721-735
- Legendre, P. (2005). Species Associations: The Kendall Coefficient of Concordance Revisited. *Journal of Agricultural, Biological, and Environmental Statistics*. 10 (2). P. 226–245.
- Legendre, P. (2010). Coefficient of concordance. *Encyclopedia of Research Design*. 1. P. 164-169.
- Lewis, D. D. and Sparck Jones, K. (1996). Natural language processing for information retrieval. *Communications of the ACM*. 39 (1). P. 92-101.
- Maimon, O., & Rokach, L. (2008). *Data mining with decision trees: theory and applications*. USA: World Scientific Publishing.
- Manning, C., Raghavan, P. and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge: Cambridge University Press.
- Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.
- Metzler, D. and Croft, B. W. (2005). Modeling Query Term Dependencies in Information Retrieval with Markov Random Fields. *SIGIR '05 Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*. P. 472-479.
- Mikk, J. (2005). Text comprehensibility (Textverständlichkeit). In Köhler, R., Altmann, G. and Piotrowski, R. G. (eds.) *Quantitative Linguistik - Quantitative Linguistics. Ein Internationales Handbuch*. Berlin: Walter de Gruyter
- Mikros, G. and Milicka, J. (2014). Distribution of the Menzerath's law on the syllable level in Greek texts. In Altmann, G. et al. (eds.) *Empirical Approaches to Text and Language Analysis*. Lüdenscheid: RAM-Verlag.
- Miller, G.A. (2003). The cognitive revolution: a historical perspective. *Trends in Cognitive Science*. 7 (3). P. 141–144
- Moens, M. F. (2006). *Information Extraction: Algorithms and Prospects in a Retrieval Context*. The Information Retrieval Series. Dordrecht: Springer.
- Nadel L. (2005). *Encyclopedia of Cognitive Science*. Wiley
- Neideen, T. and Brasel K. (2007). Understanding Statistical Tests. *Journal of Surgical Education*. 64 (2). P. 93-96
- Onwuchekwa, E. O. (2011). Information Retrieval Methods in Libraries and Information Centers. *An International Multidisciplinary Journal, Ethiopia*. 5(6), P. 108-120.
- Panik M. J. (2012). *Statistical Inference: a short course*. Hoboken, NJ: Wiley
- Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: a critical review and future directions. *Psychonomic Bulletin & Review*. 21 (5). P. 1112-1130
- Powers, D. M. W. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation. *Journal of Machine Learning Technologies*. 2 (1). P. 37-63
- Power law: definitions of power law in Oxford Dictionary (British and World English)* (2015). Retrieved 08/01/2015 from: <http://www.oxforddictionaries.com/definition/english/power-law>
- Poeppel, D., Embick, D. (2005). Defining the relation between linguistics and neuroscience. *Twenty-first Century Psycholinguistics: Four Cornerstones*. 103–118.

- Poulos M., Papavaslopoulos S., Chrissikopoulos V. (2006). A text categorization technique based on a numerical conversion of a symbolic expression and an onion layers algorithm . *Journal of Digital Information*. 6 (1).
- Poulos, M. et al. (2007). Specific Selection of FFT Amplitudes from Audio Sports and News Broadcasting for Classification Purposes. *Journal of Graph Algorithms and Applications*. 11 (1). P. 277–307
- Poulimenou S. et al. (2014). Keywords Extraction from Articles' Title for Ontological Purposes. In *Proceedings of the 2014 International Conference on Pure Mathematics, Applied Mathematics, Computational Methods (PMAMCM 2014)*. P. 120-125.
- Poulimenou, S. et al. (to appear in). Short text coherence hypothesis. *Journal of Quantitative linguistics*. 22 (3).
- Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1(1), 81–106.
- Quinlan, J. R. (1987). Simplifying decision trees. *International Journal of Man-Machine Studies*, 27(3), 221–234.
- Quinlan, J. R. (1996). Improved use of continuous attributes in C4. 5. *Journal of Artificial Intelligence Research*, 77–90.
- Quinlan, J. R. (2014). *C4. 5: programs for machine learning*. Elsevier. Retrieved from https://www.google.com/books?hl=el&lr=&id=b3ujBQAAQBAJ&oi=fnd&pg=PP1&dq=C4.5:+programs+for+machine+learning&ots=sP6pXNJIAa&sig=UIPiHH6x1NHG_6Ke8q31hSblErM
- Pulvermuller, F. (2003). *The Neuroscience of Language: On Brain Circuits of Words and Serial Order*. Cambridge: Cambridge University Press.
- Raghavan, V. V. and Wong, S. K. M. (1986). A critical analysis of vector space model for information retrieval. *Journal of the American Society for Information Science*. 37 (5). P. 279-287.
- Richards, J. C. and Schmidt, R. (2002). *Longman dictionary of language teaching and applied linguistics*. 3rd ed. London: Longman
- Robertson, S. (2004). Understanding Inverse Document Frequency: On theoretical arguments for IDF. *Journal of Documentation*. 60 (5). P. 503–520.
- Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*. 24 (5). P. 513-523. Ithaca: Department of Computer Science, Cornell University.
- SAS Institute (1999). *SAS/STAT User's Guide*, Version 8. Cary, NC: SAS Institute, 1999. Retrieved 10/12/2014 from <http://www.okstate.edu/sas/v8/saspdf/stat/chap13.pdf>
- Singhal, A. (2001). Modern Information Retrieval: a Brief Overview. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*. 24 (4). P. 35–43
- Sorell, J. C. (2012). Zipf's Law and Vocabulary. In Chapelle, C. A. (ed.). *The Encyclopedia of Applied Linguistics*. Chisester, UK: Wiley-Blackwell
- Sparck Jones, K. and Kay, M. (1977). Linguistics and information science: a postscript. In Walker, D., Karlgren, H., Kay, M. (Eds) *Natural Language in Information Science - Perspectives and Directions for Research*. Stockholm: Skriptor
- Spoerri A. (1995). *InfoCrystal: a Visual Tool for Information Retrieval*. PhD Thesis. Massachusetts: MIT. Retrieved 10/12/2014 from <http://comminfo.rutgers.edu/~aspoerri/InfoCrystal/InfoCrystal.htm>
- Squire, L. et al (eds.) (2008). *Fundamental Neuroscience*. 3rd ed. Boston: Elsevier / Academic Press.
- Strzalkowski, T. et al. (1999). Evaluating natural language processing techniques in information retrieval: a TREC perspective. In Strzalkowski, T. (Ed.) *Natural Language Information Retrieval*. Dordrecht: Kluwer Academic Publishers.
- Tanskanen, S. K. (2006). *Collaborating towards coherence*. Amsterdam: John Benjamins Publishing

- Tesitelova, M. (1992). *Quantitative linguistics*. Linguistic and Literary Studies in Eastern Europe, 37. Amsterdam: John Benjamins.
- Thimiakou, V., Kalampokh, S., Kargioti, F., Tsihlaki, A., & Poulos, M. S. (2015). Αξιολόγηση Υπηρεσιών Βιβλιοθηκών μέσω Τεχνητού Νευρωνικού Δικτύου. *HealJournal*, 1(2), 47–60.
- Τσίμπος, Κ. και Γεωργιακόδης, Φ. (1999). *Περιγραφική και διερευνητική στατιστική ανάλυση δεδομένων*. 1. Αθήνα: Εκδόσεις Αθαν. Σταμούλης.
- Turney, P. D. and Pantel, P. (2010). From frequency to meaning: vector space models of semantics. *Journal of Artificial Intelligence Research*. 37 (1). P. 141-188.
- Van Rijsbergen, C. J. (1979). *Information Retrieval*. 2nd ed. London: Butterworth.
- Woods, W. A. et al. (2000). Linguistic knowledge can improve information retrieval. In *ANLC '00 Proceedings of the sixth conference on applied natural language processing*. P. 262-267
- Wong, H. B., Gek, Lim G. H. (2011). Measures of Diagnostic Accuracy: Sensitivity, Specificity, PPV and NPV. In *Proceedings of Singapore Healthcare*. 20:4. P. 316-318
- Wu, S., Bi, Y. and Zeng, X. (2010). Retrieval Result Presentation and Evaluation. In *Proceeding KSEM'10 of the 4th international conference on Knowledge science, engineering and management*. P. 125-136.
- Wyllis, R. E. (1981). Empirical and theoretical bases of Zipf's law. *Library Trends*. 30 (1). P. 53-64.
- Yao, Y. Y. (2003). Information-theoretic measures for knowledge discovery and data mining. In *Entropy Measures, Maximum Entropy Principle and Emerging Applications* (pp. 115–136). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-3-540-36212-8_6
- Yeung, R. W. (2008). *Information theory and network coding*. Springer Science & Business Media. Retrieved from https://www.google.com/books?hl=el&lr=&id=KzoE59t_basC&oi=fnd&pg=PA2&dq=Information+Theory+and+Network+Coding&ots=KC9jnBIYVP&sig=8V95MiFyEdxPPZI_HUwzbL1yzYc
- Yuan, Y., & Shaw, M. J. (1995). Induction of fuzzy decision trees. *Fuzzy Sets and Systems*, 69(2), 125–139.
- Zar, J. H. (2010). *Biostatistical Analysis*. 4th ed. Upper Saddle River, NJ: Prentice Hall
- Zamanian, M. and Heydari, P. (2012). Readability of Texts: State of the Art. *Theory and Practice in Language Studies*. 2 (1). P. 43-53.
- Zamir, O. and Etzioni O. (1999). Grouper: a dynamic clustering interface to Web search results. *Proceeding WWW '99 Proceedings of the eighth international conference on World Wide Web archive*. P. 1361-1374. New York, NY: Elsevier
- Zar, J. H. (2010). *Biostatistical Analysis*. 4th ed. Upper Saddle River, NJ: Prentice Hall
- Zhu, W., Zeng N. and Wang N. (2010). Sensitivity, Specificity, Accuracy, Associated Confidence Interval and ROC Analysis with Practical SAS Implementations. In *Proceedings of the NESUG Health Care and Life Sciences*. P. 1-9.
- Zimny, S. T. (1987). Recognition memory for sentences from a discourse. Unpublished doctoral dissertation, Boulder: University of Colorado.
- Zhang, Y.H., Yuen, C.W.M., Wong, W.K., Kan, C., 2011. An intelligent model for detecting and classifying color-textured fabric defects using genetic algorithms and the Elman neural network. *Textile Research Journal* 81, 1772–1787.

EYPETHPIO

A

Agents, 62, 89, 104

D

DOM, 55, 98

H

Hopfield Network, 123
HTML, 7, 33, 53, 58, 63, 89, 90, 91, 94, 97, 103, 105

K

Kendall, **141**

L

LINGUIST, 69
LVQ, 7, 119, 123, 124, 125, 126

M

Menzerath-Altmann, 37, 39, 76, 81, 136, 151, 153

O

OWL, 7, 13, 43, 95, 96, 99, 101, 102, 103

R

Radial Basis Function, 123, 124
RDF, 8, 13, 43, 50, 51, 52, 53, 56, 60, 62, 86, 93, 94, 95, 96, 98, 99, 101, 102, 103, 104
RDFS, 8, 56, 60, 95, 99, 103

S

SHOE, 91, 92, 94, 99
Standard Generalized Markup Language (SGML), 13
supervised, 119
SUPPORT VECTOR, 115

V

VSM, 8, 17, 21, 22, 23, 24, 74, 138, 141

X

XML, 8, 13, 33, 43, 50, 51, 52, 53, 55, 57, 58, 60, 62, 63, 86, 87, 94, 95, 97, 98, 101, 102, 103, 104, 105

Z

Zar, 41, 85, 111, 112, 113, 114, 128, 140, 152, 158

A

Ανάλυση Κειμένων, 136
αναπαράστασης γνώσης, 46, 101
Ανταγωνιστικό τρόπο, 123

B

βάρος, 13, 25, 26, 103, 107, 122, 123, 138

Γ

Γλώσσες Σήμανσης, 10, 42
Γλωσσολογία, 64, 66, 68, 74, 75, 76
Γραμμική μέθοδος, 116
γραμμικό μετασχηματισμό, 124

Δ

Δέντρα αποφάσεων, 115
δημιουργίας ευρετηρίου, 34
διανύσματα των βαρών, 124
διανύσματος εισόδου, 123, 126
Διατύπωση υποθέσεων, 113

E

εκμάθησης, 73, 116, 122, 125
Έλεγχος Στατιστικών Υποθέσεων, 112, 113
Εντροπία, 10
εξόρυξης γνώσης, 14, 21
Επαγωγικής Στατιστικής, 13, 107, 113
Επεξεργασία Φυσικής Γλώσσας, 68, 71, 74

H

H εντροπία του Shannon, 11

K

Κειμενογλωσσολογίας, 65
κλασσικά μοντέλα ΑΠ, 23

M

Μέθοδοι με επίβλεψη και χωρίς επίβλεψη, 118
μέρη του λόγου, 72, 73
μετασχηματισμό Fourier, 126
Μη γραμμική μέθοδος, 117
Μηχανές αναζήτησης, 58, 59
Μηχανικής Μάθησης, 14
ΜΟΠΑΒ, 148, 149
μορφήματα, 71, 72, 145

N

Νευρογλωσσολογία, 69
νευρώνες, 117, 120, 121, 122, 124
Νόμος του Zipf, 77

O

Ο Ανθρώπινος Εγκέφαλος, 119
Οντολογία, 86, 87

Σ

ΣΑΠ., 16, 17, 25, 26, 27, 28, 30, 33, 66
συσσωρευτή, 122

T

Τεχνητά Νευρωνικά Δίκτυα, 119
Τεχνητού Νευρωνικού Δικτύου, 150, 152, 158
ΤΝΔ, 122, 123, 124, 149, 150

Υ

Υπολογιστική Γεωμετρία, 129
υπολογιστική Γλωσσολογία, 64