ΗΡΑΚΛΗΣ Γ. ΒΑΡΛΑΜΗΣ

ΣΗΜΑΣΙΟΛΟΓΙΚΟΣ ΧΑΡΑΚΤΗΡΙΣΜΟΣ, ΟΡΓΑΝΩΣΗ ΚΑΙ ΔΙΑΧΕΙΡΙΣΗ ΠΕΡΙΕΧΟΜΕΝΟΥ ΤΟΥ ΠΑΓΚΟΣΜΙΟΥ ΙΣΤΟΥ, ΜΕ ΧΡΗΣΗ ΟΝΤΟΛΟΓΙΩΝ ΚΑΙ ΕΜΦΑΣΗ ΣΤΟ ΡΟΛΟ ΤΩΝ ΥΠΕΡΣΥΝΔΕΣΜΩΝ

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΕΠΤΕΜΒΡΙΟΣ 2003

ΗΡΑΚΛΗΣ Γ. ΒΑΡΛΑΜΗΣ

Τμημα Πληροφορικης

ΣΗΜΑΣΙΟΛΟΓΙΚΟΣ ΧΑΡΑΚΤΗΡΙΣΜΟΣ, ΟΡΓΑΝΩΣΗ ΚΑΙ ΔΙΑΧΕΙΡΙΣΗ ΠΕΡΙΕΧΟΜΕΝΟΥ ΤΟΥ ΠΑΓΚΟΣΜΙΟΥ ΙΣΤΟΥ, ΜΕ ΧΡΗΣΗ ΟΝΤΟΛΟΓΙΩΝ ΚΑΙ ΕΜΦΑΣΗ ΣΤΟ ΡΟΛΟ ΤΩΝ ΥΠΕΡΣΥΝΔΕΣΜΩΝ

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

ΑΘΗΝΑ 2003

Ευχαριστίες

Κατ' αρχή θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου, Επίκουρο Καθηγητή Μιχάλη Βαζιργιάννη, για όσα μου προσέφερε όλα αυτά τα χρόνια. Οι γνώσεις του, οι ιδέες του και οι εμπνεύσεις του αποτέλεσαν τη βάση για αυτή τη διατριβή, ενώ η καθοδήγηση και η βοήθεια που απλόχερα μου παρείχε σε κάθε δύσκολη καμπή της προσπάθειάς μου με βοήθησαν να φτάσω ως εδώ. Θα ήθελα ιδιαίτερα να τον ευχαριστήσω για τη δυνατότητα που μου προσέφερε να συμμετάσχω σε διεθνή συνέδρια και συνεργασίες καθώς και για το άρτιο εργασιακό περιβάλλον που φρόντισε να μου εξασφαλίσει.

Επίσης, θα ήθελα να ευχαριστήσω τα μέλη της ερευνητικής ομάδας DB-NET του ΟΠΑ, Μ. Χαλκίδη, Μ. Ειρηνάκη, Σ. Βαλαβάνη, Ε. Βραχνό, Γ. Μπατιστάκη για τις εποικοδομητικές συζητήσεις και την συνεργασία που είχαμε. Θα ήθελα ακόμη να ευχαριστήσω τους φοιτητές του τμήματος Πληροφορικής του ΟΠΑ με τους οποίους συνεργάστηκα όλα αυτά τα χρόνια και ιδιαίτερα τους: Χ. Λάμπο, Γ. Τσατσαρώνη, Ε. Μπλάντα, Ν. Οικονομάκου, Μ. Θεοδωράκη, Α. Βλάχο, Χ. Αντωνόπουλο, Λ. Ρούσσο, Α. Νταή.

Επιπρόσθετα, θα ήθελα να ευχαριστήσω τους ακόλουθους ανθρώπους για τις δημιουργικές συζητήσεις και την συνεργασία τους σε σχετικά ερευνητικά θέματα: τον Καθηγητή S. Abiteboul, (INRIA Furturs), τον B. Nguyen (Υπ. Διδάκτορας, INRIA), Τους Π. Πούλο, Ι. Κουρούπη, Χ. Πατερίτσα και Γ. Ακρίβα (Υπ. Διδάκτορες, ΕΜΠ).

Παράλληλα θα ήθελα να εκφράσω τις ευχαριστίες μου στους καθηγητές και το προσωπικό του ΟΠΑ για την όποια υποστήριξη που παρείχαν όλο αυτό το διάστημα. Επιπρόσθετα θα πρέπει να αναφερθώ στα ακόλουθα ερευνητικά έργα: ΠΕΝΕΔ (χρηματοδοτούμενο από την Γενική Γραμματεία Έρευνας και Τεχνολογίας), IKUM και FAETHON (χρηματοδοτούμενα από την Ευρωπαϊκή Ένωση στο IST πλαίσιο) σ' αναγνώριση της υλικοτεχνικής υποστήριξης που μου παρείχαν σε όλα αυτά τα χρόνια.

Τέλος, θα ήθελα να ευχαριστήσω την οικογένειά μου για την αγάπη και την υποστήριξη που μου προσφέρει συνεχώς. Ιδιαίτερα θα ήθελα να ευχαριστήσω την Κατερίνα για τη φιλολογική επιμέλεια της παρούσας διατριβής.

Δημοσιεύσεις

Η διατριβή βασίστηκε στις ακόλουθες δημοσιεύσεις σε διεθνή επιστημονικά περιοδικά και συνέδρια με σύστημα κριτών:

- I. Varlamis, B. Nguyen, M. Vazirgiannis, M. Halkidi. "Organizing Web Documents into Thematic Subsets using an Ontology", to appear in the VLDB Journal, special issue on "Semantic Web".
- I. Varlamis, B. Nguyen, M. Halkidi, M. Vazirgiannis. "THESUS: Organising Web Document Collections Based on Semantics & Clustering", to appear in IEEE / TKDE Journal.

- M. Eirinaki, M. Vazirgiannis, I. Varlamis. "SEWeP: Using Site Semantics and a Taxonomy to Enhance the Web Personalization Process", in Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 2003
- B. Nguyen, M. Vazirgianis, I. Varlamis, M. Halkidi. "Organising Web Documents into Thematic Subsets using an Ontology (THESUS)", Journées scientifiques Web sémantique 2002.
- I. Varlamis, M. Vazirgiannis, " Bridging XML-Schema and relational databases. A system for generating and manipulating relational databases using valid XML documents", in the proceedings of ACM Symposium on Document Engineering, Nov. 2001, Atlanta, USA
- I. Varlamis, M. Vazirgiannis, "Web document searching, using enhanced hyperlink semantics based on XML", in the proceedings of IDEAS 2001 conference, Grenoble, France.
- G. Akrivas, S. Ioannou, E. Karakoulakis, K. Karpouzis, Y. Avrithis, A. Delopoulos, S. Kollias, I. Varlamis and M. Vaziriannis. An Intelligent System for Retrieval and Mining of Audiovisual Material Based on the MPEG-7 Description Schemes. in Proc. of the European Symposium on Intelligent Technologies, Hybrid Systems and their implementation on Smart Adaptive Systems (EUNITE), Tenerife, Spain, 12-14 December 2001.
- I. Varlamis, M. Vazirgiannis, P. Poulos, G. Akrivas, S. Ioannou, "X-Database. A middleware for collaborative video annotation, storage and retrieval", ", in the proceedings of the 8th Panhellenic Conference. Cyprus 2001.
- I. Varlamis, M. Vazirgiannis, P. Poulos "Using XML as a medium for describing, modifying and querying audiovisual content stored in relational database systems", to appear at the International Workshop on Very Low Bitrate Video Coding (VLBV). Athens 2001.

ΠΕΡΙΕΧΟΜΕΝΑ

1 Εισαγωγή....1
1.1 Διαχείριση εγγράφων του Παγκόσμιου Ιστού....1
1.2 Τοποθέτηση της διατριβής στο επιστημονικό πεδίο....4
1.3 Η προσέγγιση του THESUS....6
1.4 Συνεισφορά της διατριβής....9
1.5 Οργάνωση της διατριβής....10
2 Σχετιζόμενο επιστημονικό υπόβαθρο....13
2.1 Εξαγωγή και αποθήκευση πληροφορίας ιστο-εγγράφων....13
2.1.1 Τεχνικές αδόμητου κειμένου....13
2.1.2 Τεχνικές ημι-δομημένων εγγράφων....13
2.1.3 Δομημένα έγγραφα - XML....14
2.2 Μοντέλα αναπαράστασης της εξαγόμενης πληροφορίας....15
2.2.1 Ιεραρχικό μοντέλο....15
2.2.2 Αταξινόμητος γράφος Μοντέλο OEM....16
2.2.3 Μοντέλο κατευθυνόμενου γράφου....18
2.2.4 Σχεσιακό Μοντέλο....19
2.3 XML και σχεσιακές βάσεις δεδομένων....19
2.3.1 Αποθήκευση ολόκληρων XML εγγράφων....20
2.3.2 Απεικόνιση των περιεχομένων των XML εγγράφων σε οντότητες της βάσης δεδομένων....21
2.3.3 Περιορισμοί των υπαρχόντων συστημάτων....22
2.4 Υπερσύνδεσμοι....23
2.4.1 Εξαγωγή λέξεων από τους υπερσυνδέσμους....24
2.4.2 Ανάλυση συνδέσμων και ομαδοποίηση ιστοσελίδων....24
2.5 Ερωτήσεις στον Παγκόσμιο Ιστό....26
2.6 Οντολογίες....27
2.6.1 Μια κατηγοριοποίηση των οντολογιών....28
2.6.2 Οντολογίες και Ιστός....29
2.6.3 Το Wordnet....32
2.7 Σημασιολογικός Ιστός - Semantic Web....33
2.8 Ομαδοποίηση εγγράφων – Μέτρα ομοιότητας/απόστασης....34
2.8.1 Γενικές ιδέες σχετικά με τα μέτρα ομοιότητας....35
2.8.2 Υπάρχοντα μέτρα ομοιότητας μεταξύ των συνόλων....36
2.8.3 Ομοιότητα μεταξύ δύο στοιχείων μιας οντολογίας....37
2.8.4 Αλγόριθμοι συσταδοποίησης εγγράφων....38
2.8.5 Οι αλγόριθμοι του THESUS....42
2.8.6 Μέτρα αξιολόγησης της ποιότητας συσταδοποίησης....44
2.8.6.1 Εσωτερικά μέτρα ποιότητας....44
2.8.6.2 Σχετικά μέτρα ποιότητας....45
2.8.6.3 Εξωτερικά μέτρα ποιότητας....45
2.9 Συλλογή εγγράφων - Crawlers....47
2.9.1 Κατηγορίες περιηγητών....48
2.9.2 Αξιολόγηση σημαντικότητας εγγράφου....49
2.9.3 Τρόποι λειτουργίας των περιηγητών....51
2.10 Συμπεράσματα....53
3 Διαχείριση εγγράφων στο THESUS....55
3.1 Συλλογή εγγράφων από τον Ιστό....56

3.1.1 Καθορισμός αρχικών εγγράφων 57
3.1.2 Κριτήρια επιλογής συνδέσμων 58
3.2 Χαρακτηρισμός εγγράφων 60
3.2.1 Εξαγωγή λεξικών χαρακτηρισμών 60
3.2.2 Εμπλουτισμός χαρακτηρισμών 62
3.3 Συσταδοποίηση (αλγόριθμοι clustering) 63
3.3.1 Μέτρο ομοιότητας 63
3.3.2 Αλγόριθμοι συσταδοποίησης 64
3.3.2.1 Ο αλγόριθμος DBSCAN 64
3.3.2.2 Ο αλγόριθμος COBWEB 66
3.4 Απόδοση ετικέτας σε συστάδες (labeling) 70
3.5 Το THESUS XML-Schema 72
3.5.1 Η δομή των XML εγγράφων 74
3.6 Αρχειοθέτηση αποτελεσμάτων – Σημασιολογική διαχείριση ερωτήσεων 75
3.7 Ανασκόπηση της μεθοδολογίας THESUS 76
3.8 Περιορισμοί του συστήματος THESUS 77
4 Χαρακτηρισμός των εγγράφων με βάση τους υπερσυνδέσμους – Η γλώσσα του THESUS 78
4.1 Το μοντέλο πληροφορίας του THESUS 78
4.1.1 Τύποι δεδομένων του μοντέλου πληροφορίας του THESUS 79
4.1.1.1 Ορισμοί του μοντέλου THESUS 80
4.1.1.2 Οντότητες του Παγκόσμιου Ιστού 80
4.2 Η γλώσσα THESUS 80
Θέματα σχεδίασης της γλώσσας 81
Επεκτάσεις της γλώσσας 81
4.2.1 Τελεστές περιήγησης 83
4.2.2 Τελεστές σημασιολογίας των συνδέσμων 84
4.2.2.1 Σύνθετοι τελεστές σημασιολογίας συνδέσμων 87
4.2.3 Τελεστές ανάλυσης συνδεσμολογίας 88
5 Μέτρο ομοιότητας εγγράφων με χρήση οντολογίας 93
5.1 Απεικόνιση λέξεων σε έννοιες μιας οντολογίας 93
5.2 Μέτρο ομοιότητας για σύνολα όρων μιας οντολογίας 96
5.2.1 Το μέτρο ομοιότητας Wu και Palmer 97
5.2.2 Επεκτείνοντας το μέτρο σε σύνολα όρων μιας ιεραρχίας 97
5.2.3 Μέτρο ομοιότητας μεταξύ συνόλων όρων με βάρη 99
5.3 Μελέτη πολυπλοκότητας 103
6 Το σύστημα THESUS 105
6.1 Αρχιτεκτονική – Τεχνολογίες 105
6.1.1 Σύνδεση οντολογίας με το Wordnet 108
6.1.2 Εφαρμογή Χαρακτηρισμού Εγγράφων 109
6.2 Πειραματικά αποτελέσματα 109
6.2.1 Αποδοτικότητα του συστήματος 110
6.2.2 Αξιολόγηση διαδικασίας εξαγωγής λέξεων από τους εισερχόμενους συνδέσμους 111
6.2.2.1 Χαρακτηρισμός εγγράφου 111
6.2.2.2 Χαρακτηρισμός συνόλου εγγράφων 112
6.2.3 Έλεγχος της ικανότητας του THESUS να ανακαλύπτει θεματικά υποσύνολα 116
6.2.4 Παράγοντες που επηρεάζουν την αποτελεσματικότητα του THESUS 118

6.2.4.1 Σύγκριση τεχνικών χαρακτηρισμού....121
6.2.4.2 Σύγκριση του μέτρου ομοιότητας με το μέτρο συνημιτόνου....121
6.2.4.3 Αξιολόγηση της ποιότητας συσταδοποίησης με χρήση εξωτερικών μέτρων ποιότητας....122
6.2.5 Τελεστές ανάλυσης υπερσυνδέσμων....124
6.3 Συμπέρασμα....127
7 Το υποσύστημα X-Database....129
7.1 Απεικόνιση σε σχεσιακό μοντέλο....131
7.1.1 Ολοκλήρωση πολλών αρχείων XML-Schema....132
7.1.2 Απεικόνιση των βασικών στοιχείων του XML-Schema....132
7.1.3 Απεικόνιση του εμφωλιασμού των στοιχείων....134
7.1.4 Κληρονομικότητα τύπων....136
7.1.5 Πολυμορφισμός στοιχείων στην XML....137
7.2 Οι επιτρεπόμενες λειτουργίες....139
7.2.1 Ενημέρωση των δεδομένων....140
7.2.2 Διαχείριση ερωτήσεων....144
7.3 Η αρχιτεκτονική του συστήματος X-Database....147
7.3.1 Ροή δεδομένων....148
7.4 Αξιολόγηση του συστήματος....149
7.4.1 Δημιουργία σχεσιακού σχήματος....149
7.4.2 Μαζική εισαγωγή δεδομένων....151
7.4.3 Ενημέρωση εγγράφων....153
7.4.4 Διαγραφή....154
7.4.5 Επιλογή....156
7.4.6 Συμπεράσματα από τη διαδικασία αξιολόγησης....158
8 Συμπεράσματα....159
8.1 Συνεισφορά του THESUS....160
8.1.1 Για την οργάνωση των εγγράφων....161
8.1.2 Για την απεικόνιση XML σε σχεσιακή βάση....162
8.2 Μελλοντικές βελτιώσεις....162
8.2.1 Για την οργάνωση των εγγράφων....162
8.2.2 Για την απεικόνιση XML σε σχεσιακή βάση....164
Βιβλιογραφία....165
Παράρτημα Α – XML-Schema για τα έγγραφα του THESUS....177
Παράρτημα Β – Σύνοψη κατηγορημάτων του THESUS....178
Παράρτημα Γ – Οι διευθύνσεις ιστού των κεντρικών ιστοσελίδων μουσείων του Λονδίνου....179
Παράρτημα Δ – XML-Schema αρχείο εντολών διαχείρισης....180
ΓΛΩΣΣΑΡΙΟ....182
ΕΥΡΕΤΗΡΙΟ ΣΧΗΜΑΤΩΝ....185
ΕΥΡΕΤΗΡΙΟ ΠΙΝΑΚΩΝ....186

THESUS. Σημασιολογικός χαρακτηρισμός, οργάνωση και διαχείριση περιεχομένου του Παγκόσμιου Ιστού, με χρήση οντολογιών και έμφαση στο ρόλο των υπερσυνδέσμων

Περίληψη

Η παρούσα διατριβή περιγράφει μια νέα μέθοδο συλλογής, χαρακτηρισμού και οργάνωσης εγγράφων του Παγκόσμιου Ιστού (ΠΙ). Η διαφοροποίηση της μεθόδου από τις ήδη υπάρχουσες έγκειται στη χρήση μιας θεματικής οντολογίας σε όλα τα επίπεδα της μεθόδου. Η οντολογία περιγράφει σε κάθε περίπτωση το πεδίο ενδιαφέροντος και συνεπώς τα έγγραφα που συλλέγονται και οργανώνονται αποτελούν ένα Θεματικό Υποσύνολο (THEmatic SUbset) του ΠΙ. Για ένα Θ.Υ. του ΠΙ, η μέθοδος οργάνωσης διακρίνεται σε τρία βασικά στάδια: α) το χαρακτηρισμό των εγγράφων με λεξικά και σημασιολογικά χαρακτηριστικά, β) την οργάνωση των εγγράφων σε ομάδες με κοινά χαρακτηριστικά και γ) τη διαχείριση της συγκεντρωμένης και οργανωμένης πληροφορίας.

Με τον όρο λεξικά χαρακτηριστικά ενός εγγράφου αναφερόμαστε στο σύνολο των λέξεων που το περιγράφουν ενώ με τον όρο σημασιολογικά χαρακτηριστικά στο σύνολο των εννοιών της οντολογίας στις οποίες αντιστοιχούν οι λέξεις αυτές. Η οργάνωση των εγγράφων του ΠΙ σε θεματικά υποσύνολα γίνεται με βάση την ομοιότητα των σημασιολογικών τους χαρακτηριστικών. Για τον υπολογισμό της ομοιότητας δύο εγγράφων εισάγεται ένα νέο μέτρο που λαμβάνει υπόψη την απόσταση – στην οντολογία – των συνόλων εννοιών που περιγράφουν τα δύο έγγραφα. Το μέτρο αυτό δε βασίζεται στην απόλυτη λεξική ομοιότητα μεταξύ των δύο περιγραφών, όπως συμβαίνει στα υπάρχοντα μέτρα ομοιότητας, αλλά στη σημασιολογική ομοιότητα που εμφανίζουν. Για το λόγο αυτό είναι περισσότερο ευέλικτο και δίνει καλύτερα αποτελέσματα.

Οι λέξεις και έννοιες που εξάγονται για κάθε έγγραφο αποθηκεύονται σε ξεχωριστό XML αρχείο, το οποίο μπορεί να χρησιμοποιηθεί από άλλες εφαρμογές αλλά και να περιέχεται στο αρχικό έγγραφο ως αρχείο μετα-δεδομένων. Η δομή του κάθε XML εγγράφου περιγράφεται αυστηρά από ένα αρχείο XML-Schema. Για να διευκολύνεται η επεξεργασία της συγκεντρωμένης πληροφορίας (ερωτήσεις, εξόρυξη γνώσης κτλ.), τα δεδομένα των αρχείων XML αποθηκεύονται σε μια σχεσιακή βάση δεδομένων.

Στα πλαίσια της διατριβής αναπτύχθηκε η γλώσσα THESUS, που ορίζει ένα σύνολο τελεστών για τη διαχείριση των υπερσυνδέσμων και της πληροφορίας που αυτοί φέρουν, και το σύστημα THESUS, που υλοποιεί την προτεινόμενη μέθοδο διαχείρισης των εγγράφων του ΠΙ και ταυτόχρονα χρησιμοποιεί το νέο μέτρο για τον υπολογισμό της ομοιότητας δύο εγγράφων. Επίσης αναπτύχθηκε μια μεθοδολογία απεικόνισης των δομών της XML-Schema στο σχεσιακό μοντέλο καθώς και το σύστημα X-Database, που αναλαμβάνει την αυτόματη δημιουργία σχεσιακής βάσης δεδομένων από το XML-Schema και την αποθήκευση, διαχείριση και ανάκτηση των XML εγγράφων στη σχεσιακή βάση δεδομένων.

1 Εισαγωγή

Το πλήθος και η ετερογένεια των εγγράφων του ΠΙ και η διαρκής αύξηση σε μέγεθος και συχνότητα πρόσβασης, απαιτούν την αποτελεσματική οργάνωση των περιεχομένων του ΠΙ. Η ανάγκη για γρήγορη και ποιοτική αναζήτηση είναι επιτακτικότερη από ποτέ. Προς το παρόν τα έγγραφα του ιστού ταξινομούνται κυρίως με βάση το περιεχόμενό τους και κατά συνέπεια οι αναζητήσεις περιορίζονται σε αυτό. Το γεγονός ότι τα έγγραφα συνδέονται μεταξύ τους με συνδέσμους, οι οποίοι μεταφέρουν σημαντικές πληροφορίες, συχνά αγνοείται κατά το χαρακτηρισμό των εγγράφων, αν και αποτελεί πολύτιμη πληροφορία.

1.1 Διαχείριση εγγράφων του Παγκόσμιου Ιστού

Μέχρι σήμερα, δύο είναι οι σημαντικότερες προσεγγίσεις στο πρόβλημα της διαχείρισης των εγγράφων του ΠΙ: οι μηχανές αναζήτησης και οι δικτυακοί κατάλογοι ή πύλες.

Οι μηχανές αναζήτησης συνεισφέρουν στη διαχείριση των περιεχομένων του ΠΙ, ψάχνοντας τεράστιο πλήθος εγγράφων για να απαντήσουν τις ερωτήσεις των χρηστών που συχνά αποτελούνται από λίστες λέξεων.

Ο βασικός σκοπός των μηχανών αναζήτησης είναι να περιορίσουν τα έγγραφα που επιστρέφουν ως απάντηση στα απολύτως σχετικά με την ερώτηση, να πετύχουν δηλαδή μέγιστο ποσοστό ανάκλησης (recall) και ακρίβειας (precision) [LS69], [Ko97], [R79]. Ως ποσοστό ανάκλησης ορίζεται το πλήθος των σχετικών εγγράφων που επιστρέφονται ως απάντηση ως προς το συνολικό πλήθος των σχετικών εγγράφων, ενώ ως ποσοστό ακρίβειας ορίζεται το πλήθος των σχετικών εγγράφων που επιστρέφονται ως απάντηση ως προς το συνολικό αριθμό απαντήσεων.

Πολλές φορές τα αποτελέσματα των μηχανών αναζήτησης δεν ικανοποιούν τους χρήστες, είτε γιατί είναι πολλά και απαιτείται χρόνος για να τα ελέγξουν οι χρήστες ένα προς ένα (μεγάλο ποσοστό ανάκλησης αλλά μικρό ποσοστό ακρίβειας), είτε γιατί περιορίζονται μόνο σε έγγραφα που περιέχουν ακριβώς τις λέξεις της ερώτησης και αγνοούν ενδιαφέροντα έγγραφα που πραγματεύονται παρόμοια θέματα αλλά χρησιμοποιούν διαφορετική ορολογία (μεγάλο ποσοστό ακρίβειας και μικρό ποσοστό ανάκλησης). Πολλές φορές οι απαντήσεις περιλαμβάνουν άσχετα έγγραφα που περιέχουν τη λέξη του ερωτήματος αλλά με μια διαφορετική σημασία, μειώνοντας έτσι το ποσοστό ακρίβειας.

Τα προβλήματα αυτά οφείλονται στο γεγονός ότι οι αλγόριθμοι ανάκτησης εγγράφων αγνοούν δύο πολύ σημαντικά χαρακτηριστικά που εμφανίζει η φυσική γλώσσα: την πολυσημία (polysemy) και την συνωνυμία (synonymy) των λέξεων [M+93]. Σύμφωνα με τα χαρακτηριστικά αυτά μία λέξη μπορεί να έχει πολλές διαφορετικές σημασίες ανάλογα με το γλωσσικό περιβάλλον στο οποίο εμφανίζεται (πολυσημία), ενώ πολλές διαφορετικές λέξεις μπορεί να σημαίνουν το ίδιο πράγμα (συνωνυμία). Για να αντιμετωπιστούν τέτοιες ιδιαιτερότητες της γλώσσας χρειάζεται να γίνει

σημασιολογική ανάλυση τόσο των εγγράφων του ιστού όσο και των ερωτήσεων που υποβάλουν οι χρήστες μιας μηχανής αναζήτησης.

Κάποιες προσπάθειες που έχουν γίνει έως τώρα στον ΠΙ περιορίζονται στο να επεκτείνουν τις ερωτήσεις των χρηστών με συνώνυμες λέξεις [SJ71],[MW+72],[Sa73] και αγνοούν τη σημασιολογική πλευρά των εγγράφων και των ερωτήσεων.

Οι κατάλογοι διευθύνσεων ιστού και οι θεματικές δικτυακές πύλες προσφέρουν μια διαφορετική προσέγγιση στη διαχείριση των περιεχομένων του ΠΙ. Με τη βοήθεια μιας ομάδας ειδικών συντακτών συγκεντρώνουν τα έγγραφα που οι συντάκτες θεωρούν σχετικά, τα χαρακτηρίζουν και τα ταξινομούν σε μια ιεραρχία κατηγοριών. Η τοποθέτηση των εγγράφων σε κάποια κατηγορία δεν είναι αντικειμενική αφού εξαρτάται από την κρίση των συντακτών. Επίσης, επειδή ο χαρακτηρισμός και η κατηγοριοποίηση δεν γίνονται αυτόματα: α) η διαδικασία είναι χρονοβόρα, β) οι κατάλογοι περιέχουν ένα πολύ μικρό ποσοστό των εγγράφων του ΠΙ και γ) η διαδικασία ενημέρωσής τους είναι αργή.

Τόσο η αυτόματη συγκέντρωση και οργάνωση όσο και η επιλεκτική συλλογή και ταξινόμηση των εγγράφων από ειδικούς εφαρμόζονται συνδυαστικά στις σύγχρονες μηχανές αναζήτησης. Οι υπάρχουσες μηχανές αναζήτησης χρησιμοποιούν ιεραρχικούς καταλόγους διευθύνσεων που ενημερώνονται από ανθρώπους και ταυτόχρονα βάσεις δεδομένων στις οποίες συγκεντρώνουν και χαρακτηρίζουν με αυτόματο τρόπο έγγραφα του ΠΙ.

Όταν αναφερόμαστε σε αναζητήσεις στον ΠΙ, η συνήθης πρακτική είναι να θεωρούμε ότι οι επισκέπτες του ΠΙ υποβάλουν ερωτήσεις στις μηχανές αναζήτησης χρησιμοποιώντας μια ή περισσότερες λέξεις που περιγράφουν την πληροφορία που αναζητούν. Τα αποτελέσματα μιας αναζήτησης δεν είναι πάντα εύχρηστα, καθώς οι μηχανές αναζήτησης επιστρέφουν στους χρήστες τεράστιες λίστες από διευθύνσεις ιστού, πολλές από τις οποίες μπορούν να θεωρηθούν άσχετες με την υποβαλλόμενη ερώτηση. Οι διευθύνσεις είναι ταξινομημένες σε φθίνουσα σειρά σχετικότητας ως προς την ερώτηση, όπου η σχετικότητα καθορίζεται από το πλήθος των κοινών λέξεων μεταξύ της ερώτησης και του κάθε εγγράφου. Η κάθε διεύθυνση συνοδεύεται από μια σύντομη περιγραφή που μπορεί να περιλαμβάνει τον τίτλο του εγγράφου, ένα μικρό απόσπασμα από τα περιεχόμενα του εγγράφου ή μια λίστα λέξεων που έχουν εξαχθεί από τα περιεχόμενα του εγγράφου.

Η "τυφλή" επεξεργασία των περιεχομένων ενός εγγράφου του ΠΙ και η εξαγωγή των συχνότερα εμφανιζόμενων λέξεων σε αυτό δεν αρκούν για να περιγράψουν σωστά τα έγγραφα. Η περιγραφή που εξάγεται αυτόματα από τα περιεχόμενα των εγγράφων δεν είναι αξιόπιστη καθώς βασίζεται στην αξιοπιστία των δημιουργών των εγγράφων και συχνά είναι ελλιπής καθώς πολλά έγγραφα δεν είναι αυτο-περιγραφικά, ενώ σε ορισμένες περιπτώσεις (π.χ. έγγραφα πολυμέσων) δεν είναι δυνατή η εξαγωγή λεξικής πληροφορίας από το περιεχόμενο. Για ορισμένα έγγραφα χρησιμοποιείται η περιγραφή που έχει δώσει η ομάδα συντακτών. Η περιγραφή αυτή είναι πολύ πιο ακριβής από τις αυτόματα εξαγόμενες περιγραφές, το ποσοστό όμως των εγγράφων αυτών είναι πολύ μικρό σε σχέση με το σύνολο των εγγράφων του ΠΙ.

Οι ερωτήσεις που υποβάλλονται στις μηχανές αναζήτησης απλά ψάχνουν για έγγραφα που περιέχουν μια τουλάχιστο ή όλες τις λέξεις μιας λίστας λέξεων και συνεπώς δεν επαρκούν για να καλύψουν τις εξειδικευμένες ανάγκες των χρηστών. Προς την κατεύθυνση αυτή αναπτύσσονται πιο σύνθετοι μηχανισμοί ερωτήσεων που λαμβάνουν υπόψη τους την περιοχή ιστού (domain) των εγγράφων (π.χ. ψάξε για έγγραφα κάτω από τη διεύθυνση aueb.gr), τους συνδέσμους των εγγράφων (π.χ. ψάξε για έγγραφα που έχουν σύνδεσμο προς το έγγραφο Χ) και άλλα χαρακτηριστικά των εγγράφων, όπως τη γλώσσα, την ημερομηνία κτλ.

Καθώς μπαίνουμε πλέον στην εποχή του Σημασιολογικού Ιστού (Semantic Web), η πληροφορία που μπορούμε να εξάγουμε από τα έγγραφα είναι πλουσιότερη και συνεπώς η προσέγγιση που ακολουθούν οι μηχανές αναζήτησης για τη συγκέντρωση, οργάνωση και ανάκτηση των περιεχομένων του ιστού πρέπει να προσαρμοστεί ανάλογα. Όπως αναφέρθηκε και από τους εμπνευστές του Σημασιολογικού Ιστού [BH+01] (Tim Berners-Lee, James Hendler και Ora Lassila):

> *Μια νέα μορφή περιεχομένων του ΠΙ, που θα είναι κατανοητή από τους υπολογιστές, θα προκαλέσει μια επανάσταση με νέες προοπτικές.*

Τα βασικά χαρακτηριστικά του νέου ιστού είναι:

- Σημασιολογικό περιεχόμενο: Τα έγγραφα δεν έχουν μόνο λεξικό αλλά και σημασιολογικό περιεχόμενο, που καθορίζει τη σημασία των περιεχομένων τους σε σχέση με το θεματικό πεδίο στο οποίο αναφέρονται.
- Αναπαράσταση της γνώσης: Αν μέχρι τώρα η γλώσσα HTML [W3a] αρκούσε για να περιγράψουμε την εμφάνιση των εγγράφων, η γλώσσα XML [W3a] χρησιμοποιείται για να περιγράψει τη δομή τους και η γλώσσα RDF [W3c] τη σημασία τους και τις σχέσεις μεταξύ των δομικών στοιχείων ενός εγγράφου. Η αναπαράσταση της γνώσης με τυπικό τρόπο εισάγει μια νέα μορφή επικοινωνίας μεταξύ ανθρώπων και υπολογιστών.
- Οντολογίες [G93]: Οι έννοιες που πραγματεύονται τα έγγραφα συνδέονται μεταξύ τους με σχέσεις, δημιουργώντας έτσι μια βάση γνώσης που μπορεί να φανεί χρήσιμη στις διαδικασίες οργάνωσης και ανάκτησης πληροφορίας από τον ιστό.
- Πράκτορες [Hd01]: Αποτελούν την κινητήρια δύναμη του νέου ιστού. Είναι προγράμματα προσαρμοσμένα στη συλλογή, επεξεργασία και ανταλλαγή περιεχομένων του ΠΙ. Αποτελούν τους "πληροφοριοδότες" του Σημασιολογικού Ιστού, έχουν τη δυνατότητα να αυτο-προσδιορίζονται, περιγράφοντας τις υπηρεσίες που προσφέρουν και τον τρόπο με τον οποίο επικοινωνούν.
- Εξέλιξη της γνώσης: Το σημαντικότερο ίσως χαρακτηριστικό του Σημασιολογικού Ιστού είναι η δυνατότητά του να εξελίσσει τη γνώση που συσσωρεύεται από τους πράκτορες, επεκτείνοντας, συρρικνώνοντας, διασπώντας ή συγχωνεύοντας τις υπάρχουσες οντολογίες, αναλύοντας τις εννοιολογικές τάσεις στα περιεχόμενα του Ιστού κτλ.

Τα θέματα που πρέπει να αντιμετωπίσουν τα συστήματα συλλογής και οργάνωσης πληροφορίας συνοψίζονται στα εξής:

- Συλλογή εγγράφων: Κατά τη διαδικασία συλλογής των εγγράφων πρέπει να καθοριστούν τα κριτήρια με τα οποία θα διαχωρίζονται τα έγγραφα που θα επιλέγονται για συλλογή από τα υπόλοιπα. Πρέπει επίσης να βρεθούν μέθοδοι που θα επιταχύνουν τη διαδικασία συλλογής των εγγράφων.

- Ενημέρωση συλλογής: Καθώς το σύνολο των εγγράφων του ΠΙ αλλάζει διαρκώς, νέα έγγραφα προστίθενται, και υπάρχοντα έγγραφα καταργούνται ή μετακινούνται σε νέες διευθύνσεις, πρέπει να καθοριστεί ο τρόπος και η συχνότητα με την οποία θα ενημερώνεται η συλλογή των εγγράφων.
- Χαρακτηρισμός των εγγράφων: Τα περιεχόμενα των εγγράφων δεν είναι πάντοτε αξιόπιστη πηγή ή δεν επαρκούν για το χαρακτηρισμό των εγγράφων, όπως συμβαίνει για παράδειγμα με αρχεία εικόνας, βίντεο, ήχου κτλ. Συνεπώς απαιτούνται νέες μέθοδοι για την εξαγωγή χαρακτηρισμών για τα έγγραφα.
- Αποσαφήνιση εννοιών [Y00]: Οι λεξικές περιγραφές δεν αντιστοιχίζονται μοναδικά σε σημασιολογικές, απαιτείται αποσαφήνιση των εννοιών που αντιπροσωπεύουν οι λέξεις ενός εγγράφου.
- Οργάνωση εγγράφων: Απαιτούνται αποδοτικοί αλγόριθμοι που θα μπορούν να επεξεργάζονται μεγάλο αριθμό εγγράφων του ΠΙ, θα εξάγουν αντιπροσωπευτικές περιγραφές γι' αυτά και θα καθορίζουν τις ομοιότητες και διαφορές τους.
- Διαχείριση ερωτήσεων: Χρειάζεται να αναπτυχθούν νέοι τρόποι υποβολής ερωτήσεων, που θα λαμβάνουν υπόψη τους υπερσυνδέσμους των εγγράφων και τη σημασιολογική πληροφορία που αυτοί μεταφέρουν, καθώς και αλγόριθμοι που θα τις επεξεργάζονται εξετάζοντας και τη σημασιολογική ομοιότητα μεταξύ εγγράφων και ερωτήσεων και θα εντοπίζουν σχετικά έγγραφα. Πρέπει επίσης να λαμβάνεται υπόψη η συντακτική δομή εγγράφων και ερωτήσεων ώστε να μην υπάρχουν ασάφειες ([KE+01], [KL03]), όπως για παράδειγμα στην φράση «άνθρωπος δάγκωσε σκύλο», όπου μια διαφορετική διάταξη των ουσιαστικών αλλάζει τελείως το νόημα.
- Παρουσίαση αποτελεσμάτων: Ο τρόπος παρουσίασης των αποτελεσμάτων οφείλει να είναι εύχρηστος και να επιτρέπει το γρήγορο εντοπισμό του εγγράφου που ενδιαφέρει και ταυτόχρονα να δίνει μια άποψη άλλων παρόμοιων εγγράφων.

1.2 Τοποθέτηση της διατριβής στο επιστημονικό πεδίο

Σε πολλά διαφορετικά πεδία της επιστήμης, από την κοινωνιολογία ως τη φυσική, μια οντότητα δεν ορίζεται μόνο από τις συστατικές της οντότητες αλλά και από το εννοιολογικό πλαίσιο στο οποίο ανήκει. Παρ' όλα αυτά, όταν αναφερόμαστε σε αναζητήσεις στον ΠΙ, η συνήθης πρακτική είναι να θεωρούμε ότι μια ιστοσελίδα ορίζεται αποκλειστικά από το περιεχόμενό της. Στην πραγματικότητα ο Παγκόσμιος Ιστός αποτελεί ένα γράφο. Ακριβέστερα, είναι ένας *κατευθυνόμενος γράφος*, όπου οι σελίδες αποτελούν τους κόμβους και οι σύνδεσμοι τις ακμές του γράφου. Αυτό το χαρακτηριστικό είναι που διαφοροποιεί τον ΠΙ από μια απλή συλλογή εγγράφων.

Ο παγκόσμιος ιστός και η οργάνωση της πληροφορίας που περιέχει αποτελούν το βασικό πεδίο αυτής της διατριβής. Στα πλαίσια αυτά παρουσιάζεται μια νέα προσέγγιση στο χαρακτηρισμό και τη διαχείριση των εγγράφων του ΠΙ, που βασίζεται σε δύο βασικές αρχές: Πρώτον, στο ότι πολύτιμη πληροφορία για ένα έγγραφο μπορεί να εξαχθεί από τα έγγραφα που *δείχνουν προς* αυτό και δεύτερον, στο ότι η προσθήκη -στις ακμές του γράφου- σημασιολογικών χαρακτηρισμών, που προκύπτουν από μια οντολογία, προσφέρει νέους μηχανισμούς ερωτήσεων, επιταχύνει και βελτιώνει τις αναζητήσεις. Στη συνέχεια χρησιμοποιούνται

εναλλακτικά οι όροι «δικτυακό έγγραφο» και «ιστοσελίδα» για να αναφερθούμε στα περιεχόμενα του ΠΙ.

Η έρευνα που υπάρχει έως σήμερα γύρω από τον ΠΙ και τα περιεχόμενά του καλύπτει και συνδυάζει πολλούς τομείς ενδιαφέροντος. Τα θέματα που πραγματεύεται η παρούσα διατριβή είναι:

Διαχείριση υπερσυνδέσμων: Εφόσον οι υπερσύνδεσμοι θεωρούνται βασικές πηγές πληροφορίας για τα περιεχόμενα του ΠΙ, πρέπει να εξεταστούν με ανάλογη προσοχή. Για το λόγο αυτό μελετάται το υπάρχον επιστημονικό υπόβαθρο στην ανάλυση των συνδέσμων και της εξαγωγής πληροφορίας από αυτούς.

Μοντέλα πληροφορίας ιστοσελίδων: Οι ιστοσελίδες μαζί με τους συνδέσμους είναι τα δομικά στοιχεία του ΠΙ. Η οργάνωση της πληροφορίας των ιστοσελίδων σε αυστηρά δομημένα ή ημι-δομημένα μοντέλα αποτελεί το σημείο έρευνας πολλών σύγχρονων ερευνητών. Αυτή η περισσότερο ή λιγότερο αυστηρή δομή των εγγράφων του ιστού είναι που τα διαφοροποιεί από τα απλά έγγραφα κειμένου.

XML και σχεσιακές βάσεις: Η γλώσσα XML σχεδιάστηκε για την αποθήκευση ημι-δομημένων δεδομένων και υιοθετείται ως πρότυπο από ολοένα και περισσότερους οργανισμούς για την αποθήκευση της πληροφορίας που δημοσιεύεται στο δίκτυο. Η πληροφορία που εξάγεται από τα υπάρχοντα έγγραφα του ιστού και τους υπερσυνδέσμους έχει δομή, και συνεπώς μπορεί να αποθηκευθεί σε δομές XML. Η ανάλυση μεγάλου αριθμού εγγράφων του ιστού συνεπάγεται τη δημιουργία μεγάλων συλλογών εγγράφων XML. Η απευθείας επεξεργασία των συλλογών αυτών δεν είναι εφικτή λόγω του μεγάλου όγκου τους και των περιορισμένων δυνατοτήτων που έχουν τα υπάρχοντα συστήματα διαχείρισης εγγράφων XML σε σύγκριση με τα αντίστοιχα σχεσιακά συστήματα διαχείρισης βάσεων δεδομένων. Η διαχείριση των XML εγγράφων μέσα από τα συνήθη συστήματα διαχείρισης δεδομένων είναι ένα πολύ ενδιαφέρον και ενεργό πεδίο έρευνας που πρέπει να μελετηθεί.

Οντολογίες [G93]: Τα περιεχόμενα του ΠΙ, βρίσκονται σε μια περίοδο μετάβασης: τα στατικά και χωρίς καμιά συγκεκριμένη δομή έγγραφα, αντικαθίστανται από δεδομένα που αντλούνται από βάσεις δεδομένων και συμπλέκονται σύμφωνα με προκαθορισμένα πρότυπα σε δυναμικά έγγραφα. Παράλληλα, στα πλαίσια του Σημασιολογικού Ιστού (Semantic Web) στα έγγραφα προσαρτώνται σημασιολογικές πληροφορίες, που σκοπό έχουν να διευκολύνουν την "κατανόηση" της σημασίας των εγγράφων όχι από ανθρώπους αλλά από εφαρμογές υπολογιστών. Εκτός λοιπόν από το περιεχόμενο, έχουμε και τα σημασιολογικά χαρακτηριστικά των εγγράφων που οργανώνονται σε οντολογίες. Σκοπός των οντολογιών είναι η οργάνωση και αναπαράσταση της σημασιολογικής πληροφορίας που μεταφέρουν τα έγγραφα, με τρόπο που να είναι αναγνώσιμος και αντιληπτός από εφαρμογές υπολογιστών.

Οργάνωση των εγγράφων και μέτρα ομοιότητας: Ένα βήμα για την αποδοτικότερη διαχείριση των εγγράφων μιας συλλογής είναι η οργάνωσή τους σε ομάδες εγγράφων με κοινά χαρακτηριστικά και η εξαγωγή κανόνων για την εύκολη ομαδοποίηση των νέων εγγράφων. Για την οργάνωση των εγγράφων σε ομάδες είναι προτιμότερη η τεχνική της συσταδοποίησης η οποία δεν απαιτεί πρότερη γνώση των χαρακτηριστικών της κάθε ομάδας, σε αντίθεση με την κατηγοριοποίηση [DC00], [KS97]. Η οργάνωση των εγγράφων σε ομάδες είναι δυνατή με την υιοθέτηση

κατάλληλων μέτρων ομοιότητας, ανάμεσα σε έγγραφα, που να στηρίζονται στην σημασιολογική εγγύτητα των περιγραφών των εγγράφων.

Συγκέντρωση εγγράφων του ΠΙ: Ένα πολύ σημαντικό πρόβλημα που καλούνται να αντιμετωπίσουν όσοι ασχολούνται με τη διαχείριση των εγγράφων του ΠΙ είναι η συγκέντρωση και η ενημέρωση της πληροφορίας [AV+01]. Καθώς νέο περιεχόμενο προστίθεται καθημερινά, μεγάλο μέρος του περιεχομένου παράγεται δυναμικά, ενώ έγγραφα μετακινούνται ή αποσύρονται διαρκώς, η διαδικασία συγκέντρωσης γίνεται περισσότερο χρονοβόρα και η ανάγκη για ενημέρωση περισσότερο συχνή. Στην περίπτωση που η συγκέντρωση των εγγράφων είναι επιλεκτική [CB+99], είναι σημαντικό να καθοριστούν τα κριτήρια με τα οποία θα επιλέγονται τα σχετικά έγγραφα.

1.3 Η προσέγγιση του THESUS

Το πρώτο βήμα για τη βελτίωση των μηχανισμών αναζήτησης είναι η βελτίωση των χαρακτηρισμών που εξάγονται για κάθε έγγραφο. Ένας τρόπος για να βελτιωθούν οι εξαγόμενοι χαρακτηρισμοί για ένα έγγραφο είναι να γίνει καλύτερη επεξεργασία των περιεχομένων του εγγράφου. Η "καλύτερη" επεξεργασία περιλαμβάνει [R79]: α) τεχνικές επεξεργασίας κειμένου, όπως αποστελέχωση (*stemming*) των λέξεων με αφαίρεση προθεμάτων και επιθεμάτων, β) στατιστικές τεχνικές ανάλυσης κειμένου των εγγράφων, όπως εύρεση της συχνότητας εμφάνισης των λέξεων σε ένα έγγραφο και της συχνότητας εμφάνισης της λέξης στα έγγραφα μιας συλλογής, γ) αλλά και ευριστικές τεχνικές (*heuristics*), όπως αφαίρεση λέξεων που χαρακτηρίζονται κοινές (*common words*) π.χ. άρθρα, σύνδεσμοι, μόρια κτλ.

Επιπλέον, για να βελτιώσουμε το χαρακτηρισμό ενός εγγράφου, μπορούμε να επεκτείνουμε τη διαθέσιμη πληροφορία που έχουμε για το έγγραφο, συμβουλευόμενοι τις "αναφορές" (*citations*) άλλων εγγράφων προς αυτό. Στην περίπτωση του ΠΙ η αναφορά ενός εγγράφου προς κάποιο άλλο υλοποιείται μέσω των υπερσυνδέσμων.

Οι αναφορές χρησιμοποιούνται στη βιβλιογραφία ως ένα πολύ αξιόπιστο κριτήριο για την ποιότητα των εγγράφων [G72],[G79]. Η ύπαρξη πολλών αναφορών προς ένα έγγραφο, καθώς και οι αναφορές που κάνει το ίδιο το έγγραφο προς άλλα είναι κριτήρια για τη σημαντικότητά του [PB+98]. Η αναφορά προς ένα έγγραφο παρουσιάζεται μέσα στο κείμενο με τρόπο που να προϊδεάζει τον αναγνώστη για το θέμα (χρησιμοποιώντας ρήματα και ουσιαστικά) και την ποιότητα του εγγράφου (χρησιμοποιώντας συνήθως επιθετικούς προσδιορισμούς, π.χ. καλό, κακό, σημαντικό, ενδιαφέρον κτλ.).

Διαβάζοντας κανείς τα περιεχόμενα ενός εγγράφου μπορεί να διαμορφώσει μια υποκειμενική άποψη για το έγγραφο αυτό, επηρεαζόμενος σε σημαντικό βαθμό από τις προσωπικές του γνώσεις και πεποιθήσεις. Από την άλλη πλευρά, μια "μηχανή" μπορεί να εξάγει την περιγραφή από τα περιεχόμενα του εγγράφου, τα οποία επίσης εμπεριέχουν υποκειμενισμό, καθώς παρέχονται από το συγγραφέα του εγγράφου.

Διαβάζοντας κανείς μια αναφορά προς ένα έγγραφο, μπορεί να σχηματίσει μια πρώτη άποψη για το έγγραφο, χωρίς καν να διαβάσει το ίδιο το έγγραφο. Εξετάζοντας το σύνολο των αναφορών προς το έγγραφο, μπορεί να σχηματίσει μια πιο

ολοκληρωμένη άποψη για το έγγραφο αυτό. Η άποψη αυτή θα είναι τόσο πιο αξιόπιστη και πλήρης όσο περισσότερες αναφορές υπάρχουν για τα έγγραφα αυτά.

Αν θεωρήσουμε τον ΠΙ ως μια τεράστια συλλογή εγγράφων που συνδέονται μεταξύ του με υπερσυνδέσμους, μπορούμε για κάθε έγγραφο της συλλογής να διακρίνουμε δύο τύπους συνδέσμων: τους ***εξερχόμενους*** από το έγγραφο και τους ***εισερχόμενους*** σε αυτό. Οι εξερχόμενοι περιέχονται στο έγγραφο, "δείχνουν" προς άλλα έγγραφα και αποτελούν προτάσεις του συγγραφέα του εγγράφου προς τους αναγνώστες. Οι προτάσεις αυτές συνοδεύονται από μια σύντομη περιγραφή που δίνει στους αναγνώστες μια πρώτη πληροφορία για το έγγραφο. Οι εισερχόμενοι σύνδεσμοι ενός εγγράφου περιέχονται σε πολλά, διαφορετικά έγγραφα. Αναμφίβολα οι άνθρωποι που έχουν επισκεφθεί ένα έγγραφο του ΠΙ, και έχουν δημιουργήσει ένα υπερσύνδεσμο προς αυτό, έχουν καλύτερη γνώση των περιεχομένων του εγγράφου από οποιονδήποτε αλγόριθμο επεξεργασίας των περιεχομένων του εγγράφου. Για τον χαρακτηρισμό ενός εγγράφου με βάση τους εισερχόμενους συνδέσμους από μια συλλογή εγγράφων απαιτείται:

- ανάλυση των συνδέσμων όλων των εγγράφων της συλλογής,
- εντοπισμός των συνδέσμων που δείχνουν προς το συγκεκριμένο έγγραφο,
- εξαγωγή των περιγραφών που περιέχονται στους συνδέσμους,
- συγκέντρωση και ανάλυση του συνόλου των περιγραφών.

Οι σύνδεσμοι είναι λοιπόν οι ακρογωνιαίοι λίθοι στο σχεδιασμό του ΠΙ, και κατά συνέπεια πρέπει να λαμβάνονται υπόψη κατά τη συλλογή και οργάνωση της πληροφορίας αλλά και κατά τις αναζητήσεις. Μέχρι τώρα, η χρησιμότητα των συνδέσμων περιοριζόταν στον καθορισμό της σημαντικότητας των εγγράφων. Η θεωρία της ανάλυσης των συνδέσμων (Link Analysis [K99], [H00]) αντιμετωπίζει τον ΠΙ ως ένα γράφο, με κόμβους τα έγγραφα και ακμές τούς συνδέσμους, αγνοώντας το πληροφοριακό περιεχόμενο των συνδέσμων. Με την ανάλυση της συνδεσμολογίας και μόνο του ΠΙ προσδιορίζονται ως "σημαντικά" τα έγγραφα που διαθέτουν πολλούς εισερχόμενους ή εξερχόμενους συνδέσμους και ως "όμοια" τα έγγραφα που έχουν πολλούς κοινούς εισερχόμενους ή εξερχόμενους συνδέσμους.

Στην περίπτωση του ΠΙ, οι περιγραφές που φέρουν οι εισερχόμενοι υπερσύνδεσμοι είναι απαραίτητες για το χαρακτηρισμό των εγγράφων, όπως στην περίπτωση πολυμεσικών εγγράφων (εικόνων, ήχων κτλ) από το περιεχόμενο των οποίων δεν είναι δυνατό να εξαχθεί λεξική πληροφορία.

"***Οι υπερσύνδεσμοι αποτελούν σημαντική πηγή πληροφορίας για το χαρακτηρισμό των εγγράφων του ΠΙ.***"

Ο συνδυασμός της πληροφορίας που φέρουν οι σύνδεσμοι, της πληροφορίας που φέρουν τα ίδια τα έγγραφα αλλά και της σημασιολογίας των συνδέσμων συνδράμει στον καλύτερο χαρακτηρισμό και κατ' επέκταση στην καλύτερη οργάνωση των εγγράφων του ΠΙ, διευκολύνοντας τις αναζητήσεις. Στόχος είναι η ομαδοποίηση των σελίδων του ιστού σε θεματικά υποσύνολα που τα ονομάζουμε THESUs (**The**matic **Su**bsets of the WWW). Η σημασιολογική εγγύτητα των σελίδων ενός THESU δεν καθορίζεται μόνο από τα περιεχόμενα των σελίδων αλλά και από τη σημασιολογική πληροφορία που φέρουν οι σύνδεσμοι που υποδεικνύουν την κάθε σελίδα (*link semantics*). Αυτή η προσέγγιση αγνοείται από τις υπάρχουσες μηχανές αναζήτησης.

Συνήθως οι συγγραφείς των εγγράφων του ΠΙ παρέχουν περιγραφές για τα υποδεικνυόμενα έγγραφα μέσα ή γύρω από την περιοχή του υπερσυνδέσμου. Οι περιγραφές αυτές, που αποτελούνται από μερικές λέξεις, μπορούν να εξαχθούν και να χρησιμοποιηθούν για να χαρακτηρίσουν το έγγραφο. Παρ' όλα αυτά, οι περιγραφές αυτές δεν είναι πάντοτε αρκετές, ειδικά όταν οι λέξεις που τις αποτελούν μπορεί να έχουν πολλές διαφορετικές σημασίες. Για παράδειγμα, όταν η λέξη "ενέργεια" εμφανίζεται εντός ενός συνδέσμου, έχουμε μια ένδειξη ότι το έγγραφο στο οποίο αναφέρεται ο σύνδεσμος μιλά για "*ενέργεια*", χωρίς να μπορούμε να διασαφηνίσουμε αν πρόκειται για κάποια *ενέργεια-πράξη* ή για την *ενέργεια* όπως ορίζεται στη Φυσική. Για το λόγο αυτό είναι σημαντικό να ξεκαθαρίσουμε τις έννοιες που κρύβονται πίσω από τις λέξεις, ώστε να βελτιώσουμε την ακρίβεια των περιγραφών που εξάγουμε από τους υπερσυνδέσμους.

Η αποσαφήνιση των εννοιών μπορεί να επιτευχθεί με τη χρήση μιας οντολογίας, που περιέχει έννοιες σχετικές με κάποιο πεδίο ενδιαφέροντος, και ενός λεξιλογικού θησαυρού, ο οποίος θα χρησιμοποιηθεί για να αντιστοιχηθούν οι λεξικές περιγραφές σε σημασιολογικές.

Η δομή των θησαυρών καθορίζεται με σαφήνεια από διεθνή πρότυπα [ISO-2788] που προσδιορίζουν τις σχέσεις μεταξύ των όρων μονογλωσσικών θησαυρών [ISO 2788:1986], τις επιπλέον σχέσεις για πολυγλωσσικούς θησαυρούς [ISO 5964:1985] και μεθόδους για την ανάλυση εγγράφων, τον καθορισμό του θέματος τους και την επιλογή σημαντικών όρων [ISO 5963:1985]. Προβλήματα όπως η συγχώνευση μονογλωσσικών θησαυρών [SC97], καθώς και ο πλεονασμός πληροφορίας σε αυτούς [YA88] είναι πολύ σημαντικά, παρόλα αυτά εκτός των ορίων της παρούσας διατριβής, η οποία θεωρεί μοναδικό μονογλωσσικό θησαυρό.

Η χρήση της οντολογίας περιορίζει το σημασιολογικό εύρος και το πλήθος διαφορετικών εννοιών της κάθε λέξης και παρέχει έναν ομοιογενή τρόπο περιγραφής των εγγράφων. Ο θησαυρός χρησιμοποιείται για την απεικόνιση των εξαγόμενων λέξεων σε έννοιες της οντολογίας. Το αποτέλεσμα αυτής της διαδικασίας είναι η μετατροπή των λεξικών περιγραφών σε σημασιολογικές περιγραφές. Με τον τρόπο αυτό βελτιώνεται η ποιότητα των χαρακτηρισμών και δημιουργείται η βάση για την ανάπτυξη εξελιγμένων μηχανισμών διαχείρισης γνώσης.

Η ύπαρξη μιας θεματικής οντολογίας και ενός λεξιλογικού θησαυρού επιτρέπει τη μετατροπή των λεξικών περιγραφών που έχουμε για τα έγγραφα σε σημασιολογικές.

Η σημασιολογική πληροφορία που φέρουν οι σύνδεσμοι εμπλουτίζει το σημασιολογικό περιεχόμενο των ιστοσελίδων και μπορεί να αποδειχθεί πολύτιμη όταν πρόκειται για αναζητήσεις στον Παγκόσμιο Ιστό που αποσκοπούν στην πλήρη και απέριττη πληροφορία.

Η παρούσα προσέγγιση αποτελεί μια προσπάθεια διαχείρισης εγγράφων του ΠΙ, που στηρίζεται στον εμπλουτισμό των υπερσυνδέσμων με σημασιολογικά χαρακτηριστικά και εκμετάλλευσης αυτών στα διάφορα στάδια συλλογής, οργάνωσης και διαχείρισης των περιεχομένων του ΠΙ. Το σύστημα THESUS που βασίστηκε στην προσέγγιση αυτή, εξάγει λεξικούς χαρακτηρισμούς από τους υπερσυνδέσμους και τους μετατρέπει σε σημασιολογικούς με τη χρήση μιας οντολογίας. Το αποτέλεσμα της

διαδικασίας για ένα μεγάλο όγκο εγγράφων του ΠΙ είναι μια πλούσια δεξαμενή δεδομένων που μπορούμε να την εκμεταλλευτούμε ποικιλοτρόπως.

1.4 Συνεισφορά της διατριβής

Ο χαρακτηρισμός των εγγράφων του ΠΙ στην παρούσα διατριβή στηρίζεται σε δύο σημαντικές επισημάνσεις.

- Πρώτον, στο ότι οι δημιουργοί των ιστοσελίδων χρησιμοποιούν τους υπερσυνδέσμους για να υποδείξουν μια ιστοσελίδα και ταυτόχρονα για να προσφέρουν μια σύντομη περιγραφή γι' αυτήν.
- Δεύτερον, στο ότι οι εισερχόμενοι σύνδεσμοι περιγράφουν τις ιστοσελίδες καλύτερα και πιο αντικειμενικά απ' ότι τα ίδια τα περιεχόμενα των σελίδων.

Με βάση αυτές τις επισημάνσεις παρουσιάζεται μια μεθοδολογία για το χαρακτηρισμό των εγγράφων του ΠΙ που περιλαμβάνει: εξαγωγή πληροφορίας από το κείμενο των υπερσυνδέσμων, διάκριση των εισερχόμενων και εξερχόμενων συνδέσμων ενός εγγράφου και ορισμό διαφορετικών τρόπων με τους οποίους μπορεί να χαρακτηριστεί ένα έγγραφο ή μια ομάδα εγγράφων με χρήση της σημασιολογίας των υπερσυνδέσμων.

Στα αποτελέσματα της παρούσας διατριβής περιλαμβάνονται:

- Ένα μοντέλο πληροφορίας και μια γλώσσα διαχείρισης που επιτρέπουν: αρχικά την επιλογή θεματικών υποσυνόλων εγγράφων του ΠΙ και τον εμπλουτισμό τους με περιγραφές οι οποίες εξάγονται από τους συνδέσμους που αναφέρονται στα συγκεκριμένα έγγραφα, και στη συνέχεια υποβολή και απάντηση ερωτήσεων οι οποίες βασίζονται στις συνδέσεις μεταξύ των σελίδων και στη σημασιολογική πληροφορία που παρήχθη για αυτές. Τέτοιες ερωτήσεις εξάγουν χρήσιμα αποτελέσματα που δεν είναι δυνατό να εξαχθούν με τους υπάρχοντες μηχανισμούς αναζήτησης.
- Ένας μηχανισμός για την εξαγωγή λέξεων από τους υπερσυνδέσμους και τον εμπλουτισμό τους με σημασιολογικά χαρακτηριστικά. Αυτό γίνεται με την απεικόνιση του συνόλου των λέξεων που χαρακτηρίζουν ένα έγγραφο σε ένα σύνολο από έννοιες (όρους) μιας θεματικής ιεραρχίας εννοιών (οντολογίας). Στην παρούσα υλοποίηση ο μηχανισμός χρησιμοποιεί το Wordnet[M+93] ως θησαυρό και το μέτρο ομοιότητας των Wu και Palmer [WP94] για τον υπολογισμό της ομοιότητας ανάμεσα στους όρους μιας ιεραρχίας.
- Ένα νέο μέτρο ομοιότητας μεταξύ συνόλων από έννοιες (με αντίστοιχα βάρη) μιας οντολογίας. Τα έγγραφα, όπως και οι ερωτήσεις των χρηστών, αντιμετωπίζονται ως σύνολα εννοιών με βάρη. Η ομοιότητα ανάμεσα στα σύνολα δε βασίζεται στον κοινό αριθμό εννοιών των δύο συνόλων (ακριβής λεξική ομοιότητα – exact matching) αλλά στη συνολική σημασιολογική ομοιότητα των εννοιών του ενός συνόλου με όλες τις έννοιες του άλλου συνόλου. Ως αποτέλεσμα, οι ερωτήσεις των χρηστών μπορεί να επιστρέψουν έγγραφα που δεν περιέχουν κάποια από τις λέξεις ή έννοιες που ζητήθηκαν, αλλά ομώνυμες, υπερώνυμες ή υπώνυμες έννοιες αυτών.
- Ένας μηχανισμός ομαδοποίησης εγγράφων που χρησιμοποιεί το προαναφερθέν μέτρο για να εντοπίσει τις ομάδες των εγγράφων. Δύο αλγόριθμοι συσταδοποίησης, που βασίζονται στην πυκνότητα των εγγράφων σε μια συστάδα και κατ' επέκταση στην ομοιότητα μεταξύ εγγράφων, προσαρμόστηκαν στις ανάγκες του συστήματος.

- Ένα πλήρως ανεπτυγμένο σύστημα το οποίο:
  α. Συλλέγει διευθύνσεις ιστού και τα περιεχόμενα τους με βάση ένα προκαθορισμένο σύνολο λέξεων που υπόκεινται σε κοινό θεματικό πεδίο (ορολογία),
  β. Εξάγει λέξεις από τους εισερχόμενους προς τη συλλογή υπερσυνδέσμους (με επεξεργασία του υπερκειμένου στη "γειτονιά" του υπερσυνδέσμου),
  γ. Αντιστοιχεί τις λέξεις σε έννοιες στην οντολογία,
  δ. Ομαδοποιεί τα έγγραφα χρησιμοποιώντας είτε τις λεξικές είτε τις εννοιολογικές περιγραφές που έχουν προκύψει γι' αυτά,
  ε. Αποθηκεύει την πληροφορία σε μια σχεσιακή βάση δεδομένων.
- Ένα σύνολο από επιπλέον υπηρεσίες για το χαρακτηρισμό εγγράφων ή συνόλου εγγράφων με βάση την πληροφορία που περιέχουν οι εισερχόμενοι ή εξερχόμενοι σύνδεσμοί του. Αυτές οι υπηρεσίες επιτρέπουν στους χρήστες:
  - Να δώσουν ένα σύνολο από διευθύνσεις σελίδων και να πάρουν τις λέξεις που εμφανίζονται συχνά στους εισερχόμενους ή εξερχόμενους συνδέσμους,
  - Να δώσουν δύο σύνολα διευθύνσεων σελίδων (Α και Β) και να πάρουν τις λέξεις που εμφανίζονται συχνά στους συνδέσμους από σελίδες του συνόλου Α προς σελίδες του συνόλου Β.

Το σύστημα είναι διαθέσιμο στη διεύθυνση ιστού:
http://www.db-net.aueb.gr/thesus
Παράλληλα βρίσκεται υπό εξέταση η αίτηση για χορήγηση Διπλώματος Ευρεσιτεχνίας με αριθμό 20030100216, που έχει ως αξιώσεις τα αποτελέσματα της διατριβής που αναφέρθηκαν προηγουμένως.

1.5 Οργάνωση της διατριβής

Στη συνέχεια παρουσιάζουμε *τα περιεχόμενα* της διατριβής και τη σχέση μεταξύ τους. Το παρόν κεφάλαιο αναφέρθηκε στην ανάγκη για καλύτερη οργάνωση και αποτελεσματικότερη διαχείριση των εγγράφων του ΠΙ. Παρουσιάστηκαν συνοπτικά οι υπάρχουσες κατευθύνσεις καθώς και οι τεχνολογίες που άπτονται του προβλήματος της περιγραφής, οργάνωσης και διαχείρισης των εγγράφων του ΠΙ. Τέλος, έγινε μια πρώτη εισαγωγή στην προτεινόμενη προσέγγιση, στα κύρια χαρακτηριστικά της, τα εργαλεία και τη μεθοδολογία που χρησιμοποιεί, καθώς και στην προσδοκώμενη συνεισφορά.

Στο δεύτερο κεφάλαιο παρουσιάζεται μια εκτενής ανασκόπηση της διαδικασίας εξαγωγής, εμπλουτισμού, αναπαράστασης και αποθήκευσης της πληροφορίας των εγγράφων του ΠΙ, καθώς και των διαδικασιών συλλογής και οργάνωσης των εγγράφων με βάση τη σημασιολογία τους. Μέσα από την ανασκόπηση αυτή παρουσιάζονται ορισμένες σημαντικές ερευνητικές προσπάθειες στους επιμέρους τομείς.

Το κεφάλαιο τρία παρουσιάζει αναλυτικά την προτεινόμενη μεθοδολογία συγκέντρωσης και διαχείρισης των εγγράφων του ΠΙ. Περιγράφεται η διαδικασία δημιουργίας μιας «θεματικής» συλλογής εγγράφων, ο αλγόριθμος εξαγωγής λέξεων από τους υπερσυνδέσμους των εγγράφων και η διαδικασία απεικόνισής τους σε έννοιες μιας οντολογίας. Παρουσιάζονται επίσης οι δύο βασικοί αλγόριθμοι συσταδοποίησης, το μέτρο ομοιότητας μεταξύ εγγράφων που χρησιμοποιούν και οι τροποποιήσεις που έγιναν στα πλαίσια της προσέγγισης. Παράλληλα, συζητείται η

ακολουθούμενη προσέγγιση για την περιγραφή των συστάδων που προκύπτουν. Τέλος, παρουσιάζεται η προτεινόμενη δομή για την αρχειοθέτηση της εμπλουτισμένης εξαγόμενης πληροφορίας και η μεθοδολογίας διαχείρισης ερωτήσεων.

Στο κεφάλαιο τέσσερα παρουσιάζεται λεπτομερώς το μοντέλο πληροφορίας του THESUS και ορίζονται οι τύποι δεδομένων που υποστηρίζει. Ακολούθως, παρουσιάζονται θέματα σχεδίασης της γλώσσας THESUS καθώς και οι βασικότεροι τελεστές της, και οι σημαντικότερες επεκτάσεις τους.

Το κεφάλαιο πέντε εισάγει ένα νέο μέτρο ομοιότητας μεταξύ εγγράφων που αντιστοιχίζονται ως σύνολα εννοιών μιας οντολογίας. Για το λόγο αυτό περιγράφεται αρχικά η διαδικασία απεικόνισης των λέξεων που εξάγονται για κάθε έγγραφο σε έννοιες μιας οντολογίας. Στη συνέχεια παρουσιάζονται οι ιδιότητες που θα πρέπει να έχει το μέτρο, και οι δύο διαφορετικές μορφές του προτεινόμενου μέτρου (με ή χωρίς βάρη).Το κεφάλαιο ολοκληρώνεται με μελέτη της πολυπλοκότητας της διαδικασίας συσταδοποίησης λαμβάνοντας υπόψη τον υπολογισμό του μέτρου ομοιότητας.

Στο κεφάλαιο έξι παρουσιάζεται το σύστημα THESUS, που αναπτύχθηκε στα πλαίσια της προαναφερθείσας μεθοδολογίας. Η αρχιτεκτονική του συστήματος και οι τεχνολογίες που χρησιμοποιήθηκαν καθώς και οι επιμέρους υποστηρικτικές εφαρμογές που αναπτύχθηκαν περιγράφονται στο ίδιο κεφάλαιο. Ακολουθεί μια εκτεταμένη αξιολόγηση του συστήματος που ελέγχει: τη διαδικασία εξαγωγής λέξεων από τους συνδέσμους, το χαρακτηρισμό ενός εγγράφου ή μιας ομάδας εγγράφων και τον καθορισμό υποσυνόλων μιας συλλογής εγγράφων. Στη συνέχεια παρουσιάζεται μια συγκριτική αξιολόγηση των τεχνικών χαρακτηρισμού, του μέτρου ομοιότητας και των αλγορίθμων συσταδοποίησης που χρησιμοποιεί το THESUS. Το κεφάλαιο ολοκληρώνεται με κάποια παραδείγματα από τη χρήση των τελεστών ανάλυσης της πληροφορίας των υπερσυνδέσμων και με γενικότερα συμπεράσματα από τη διαδικασία αξιολόγησης.

Το κεφάλαιο επτά παρουσιάζει με λεπτομέρεια το σύστημα X-Database που αναπτύχθηκε παράλληλα με το σύστημα THESUS για να υποστηρίξει τις διαδικασίες αρχειοθέτησης των XML εγγράφων, που παράγονται από τη διαδικασία συλλογής και περιγραφής των εγγράφων του ΠΙ, σε μια σχεσιακή βάση δεδομένων. Το σύστημα επιτρέπει την αυτόματη δημιουργία του σχεσιακού σχήματος από το XML-Schema αρχείο που περιγράφει τη δομή των XML εγγράφων. Επιτρέπει επίσης τη διάφανη διαχείριση των αποθηκευμένων εγγράφων και την υποβολή επερωτήσεων χωρίς γνώση της δομής της σχεσιακής βάσης, παρά μόνο με χρήση των XML δομών. Η μεθοδολογία απεικόνισης του XML στο σχεσιακό μοντέλο, οι λύσεις που προτείνονται σε ιδιαίτερα θέματα απεικόνισης των XML δομών και οι επιτρεπόμενες λειτουργίες του συστήματος περιγράφονται αναλυτικά. Το κεφάλαιο ολοκληρώνεται με μια παρουσίαση της αρχιτεκτονικής του συστήματος και μια εκτεταμένη διαδικασία αξιολόγησης των επιδόσεών του.

Η διατριβή καταλήγει στο κεφάλαιο οκτώ με μία σύνοψη της συνεισφοράς του προτεινόμενου συστήματος και της μεθοδολογίας και με καταγραφή των κατευθύνσεων για μελλοντική εργασία.

2 Σχετιζόμενο επιστημονικό υπόβαθρο

2.1 Εξαγωγή και αποθήκευση πληροφορίας ιστο-εγγράφων

Οι τεχνικές που χρησιμοποιούνται για την εξαγωγή χαρακτηρισμών από τα έγγραφα του ΠΙ διακρίνονται σε τρεις μεγάλες κατηγορίες:

- Στις τεχνικές που βασίζονται στο πλήρες και αδόμητο κείμενο
- Σε αυτές που προσαρμόζονται στην ημι-δομημένη μορφή των εγγράφων που είναι γραμμένα σε γλώσσα HTML
- Στις τεχνικές που επεξεργάζονται έγγραφα XML με συγκεκριμένη δομή.

2.1.1 Τεχνικές αδόμητου κειμένου

Οι τεχνικές αυτής της κατηγορίας προέρχονται από έρευνα στην περιοχή της Ανάκτησης Πληροφορίας από κείμενο [R79], [Sa68]. Χρησιμοποιούν αλγόριθμους που επεξεργάζονται όλο το κείμενο του εγγράφου και δημιουργούν λίστες λέξεων που περιέχονται σε αυτό μαζί με τον αριθμό εμφανίσεων της κάθε λέξης. Χρησιμοποιούν επίσης λεξικά και γραμματικούς κανόνες για κάθε γλώσσα για να ομαδοποιούν τις λέξεις που έχουν την ίδια ρίζα [L68], ή που είναι συνώνυμες. Οι τεχνικές αυτές βασίζονται στη δομή της γλώσσας και επεξεργάζονται εξαντλητικά το κείμενο για να εξαγάγουν πληροφορίες.

Σε αυτή την κατηγορία περιλαμβάνονται και οι προσπάθειες για εξαγωγή πληροφορίας από αδόμητο κείμενο (ειδήσεις [L95], μηνύματα, αγγελίες κλπ). Σκοπός τέτοιων συστημάτων είναι η εξαγωγή πληροφορίας από το αδόμητο κείμενο και η πλήρωση των χαρακτηριστικών κάποιου προτύπου με τιμές. Η «χρήσιμες» περιοχές ενός κειμένου προσδιορίζονται είτε μέσω του τρόπου γραφής (π.χ. κεφαλαία γράμματα, αριθμοί, κ.ά.) είτε με τον εντοπισμό μέσα στο κείμενο των ίδιων των χαρακτηριστικών του προτύπου ή συνωνύμων τους [CC+03]. Η πληροφορία αυτή μπορεί να χρησιμοποιηθεί για την ταξινόμηση των κειμένων ([SA+03]) ή για την υποβολή ερωτήσεων.

Οι αλγόριθμοι που επεξεργάζονται το κείμενο είναι αρκετά πολύπλοκοι, καθώς επιτελούν πολλούς ελέγχους για να καλύψουν όλες τις πιθανές εξαιρέσεις, ενώ είναι προσαρμοσμένοι στις ιδιαιτερότητες και τη δομή της γλώσσας για την οποία αναπτύχθηκαν. Στην περίπτωση των εγγράφων του ιστού η πολυγλωσσία αποτελεί σημαντικό περιορισμό για τέτοιες τεχνικές.

2.1.2 Τεχνικές ημι-δομημένων εγγράφων

Σε αντίθεση με τα αρχεία κειμένου, τα έγγραφα του ΠΙ που είναι γραμμένα σε HTML παρουσιάζουν μια "χαλαρή" δομή, στην οποία μπορεί να διακρίνει κανείς στοιχεία όπως ο τίτλος του εγγράφου, επικεφαλίδες, λίστες, πίνακες, εικόνες, μεταδεδομένα κτλ., που περιβάλλονται από συγκεκριμένες ετικέτες της γλώσσας. Συχνά, πολλά έγγραφα του ίδιου δικτυακού τόπου έχουν την ίδια δομή καθώς έχουν παραχθεί με βάση το ίδιο πρότυπο. Στις περιπτώσεις αυτές χρησιμοποιείται ένα πρόγραμμα προσαρμοσμένο στην κοινή δομή των εγγράφων [E99] το οποίο λειτουργεί ως περίβλημα (*wrapper*) που βοηθά την εξαγωγή συγκεκριμένων στοιχείων από αυτά.

Το πρώτο βήμα στη δημιουργία ενός *περιβλήματος* για το έγγραφο είναι η περιγραφή της δομής του εγγράφου. Για να διευκολυνθεί η ανάγνωση μεγάλου πλήθους σελίδων, που έχουν διαφορετικές δομές, είναι απαραίτητη η γρήγορη ανάπτυξη περιβλημάτων που θα προσαρμόζονται στην κάθε δομή. Το σύστημα TSIMMIS [HG+97] επιτρέπει στους χρήστες να καθορίσουν τη δομή ενός εγγράφου HTML και να προσδιορίσουν τα σημεία που τους ενδιαφέρουν, μέσα από ένα σύνολο κανόνων. Στη συνέχεια δημιουργεί αυτόματα ένα περίβλημα προσαρμοσμένο στη δομή αυτή. Το δεύτερο βήμα είναι η ανάγνωση των εγγράφων και η εξαγωγή των στοιχείων που έχουν υποδείξει οι χρήστες. Τα προγράμματα ανάλυσης κειμένων, στην περίπτωση των εγγράφων HTML, εντοπίζουν τα χαρακτηριστικά σημεία του εγγράφου βασιζόμενα στις ετικέτες της γλώσσας. Στα πλαίσια του συστήματος ARANEUS [AM+97], αναπτύχθηκε η γλώσσα EDITOR [MA97] με σκοπό να αυτοματοποιήσει τη διαδικασία εξαγωγής πληροφορίας από ημι-δομημένα έγγραφα.

Τα στοιχεία που εξάγονται από τα έγγραφα, οι σχέσεις και οι ιεραρχίες ανάμεσά τους καθορίζονται με βάση τα χαρακτηριστικά της μορφής τους (π.χ. μέγεθος ή στυλ γραμμάτων, εσοχή κτλ) και ακολούθως, απεικονίζονται σε ένα μοντέλο δεδομένων. Τέλος, μπορούν να εξαχθούν και να αποθηκευθούν σε μια βάση δεδομένων. Εναλλακτικά μπορούμε να δημιουργήσουμε περιβλήματα για διάφορα έγγραφα του ΠΙ που θα εντοπίζουν συγκεκριμένα στοιχεία των εγγράφων, όποτε τους ζητηθεί. Τα δεδομένα που εντοπίζονται στα έγγραφα έχουν διαφορετικές δομές – παρ' όλα αυτά με τη χρήση των περιβλημάτων αλλά και κάποιων μεσαζόντων προγραμμάτων (*mediators*) μπορούμε να τα χειριστούμε σαν να ήταν αποθηκευμένα σε μια βάση δεδομένων [RS97]. Ο ρόλος των μεσαζόντων είναι να δέχονται ερωτήσεις σε μια κοινή γλώσσα ερωτήσεων και να τις προωθούν στις υποκείμενες δομές δεδομένων, δίνοντας έτσι την αίσθηση μιας όμοια δομημένης πληροφορίας.

Η γλώσσα WEBSQL [MM+97] ακολουθεί αυτή την προσέγγιση, παρέχοντας στους χρήστες μια σειρά από τάξεις Java, με τις οποίες μπορούν να εξάγουν συγκεκριμένα στοιχεία από έγγραφα HTML περιγράφοντας τη δομή τους και ακολούθως να κάνουν ερωτήσεις πάνω σ' αυτά.

Η αυτοματοποίηση μιας τέτοιας διαδικασίας στηρίζεται στη συντακτική ορθότητα των εγγράφων και απαιτεί συχνά την ανθρώπινη παρέμβαση για την αντιμετώπιση σφαλμάτων. Οι σχετικές προσπάθειες για την αυτόματη κατασκευή περιβλημάτων για ημιδομημένα έγγραφα [Ku97][MM+01], βασίζονται σε συστήματα που αρχικά προπαιδεύονται με ένα σύνολο εγγράφων καθορισμένης δομής και στη συνέχεια συμπεραίνουν τη δομή νέων εγγράφων επιλέγοντας από ένα σύνολο πιθανών περιβλημάτων

2.1.3 Δομημένα έγγραφα - XML

Καθώς όλο και μεγαλύτερο ποσοστό των εγγράφων του ΠΙ παράγεται δυναμικά από στοιχεία που βρίσκονται σε βάσεις δεδομένων, η HTML δεν είναι αρκετή για να αποτυπώσει την αρχική δομή της πληροφορίας, η οποία είναι δύσκολο να ανακτηθεί. Η γλώσσα XML επεκτείνει τις δυνατότητες της HTML προσθέτοντας νέες ετικέτες και δίνοντας έμφαση στη δομή και όχι στην εμφάνιση των εγγράφων. Με την XML οι δημιουργοί των εγγράφων μπορούν να προσθέσουν σε ένα έγγραφο πληροφορίες για τη δομή του και το περιεχόμενό του. Με τον τρόπο αυτό η αρχική δομή των περιεχομένων μπορεί να ανακτηθεί πολύ εύκολα από το XML έγγραφο. Παράλληλα

με χρήση γλωσσών μορφοποίησης, όπως η XSL ή τα XSL-Formatting Objects [W3d], τα στοιχεία του XML εγγράφου μορφοποιούνται ώστε το τελικό αποτέλεσμα να είναι οπτικά το ίδιο με την HTML.

Η ιδέα πίσω από την χρήση της XML είναι ότι τα χρήσιμα δεδομένα στον Ιστό είναι αυτά που προέρχονται από κάποια βάση δεδομένων. Με την XML μπορούμε να περιγράψουμε πλήρως τη δομή των δεδομένων, να τα ανακτήσουμε και να τα αποθηκεύσουμε σε άλλες βάσεις δεδομένων ή να υποβάλλουμε απευθείας ερωτήσεις σε αυτά. Μια προσέγγιση που χρησιμοποιεί την XML για να καθορίσει τη δομή των εγγράφων περιγράφεται στο [SA99a], [SA99b].

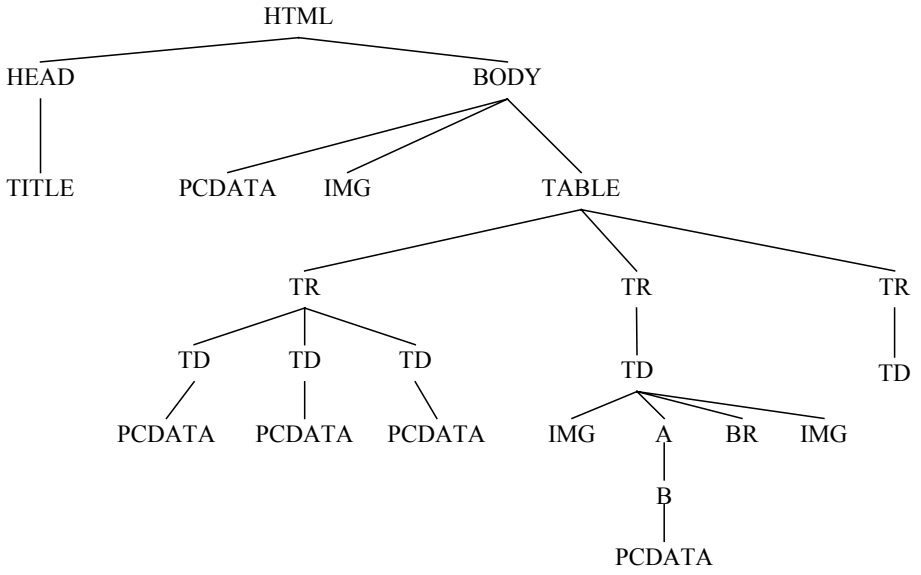
Το πρόβλημα στην περίπτωση του ΠΙ, που παραμένει προς επίλυση ακόμη και μετά τον καθορισμό της δομής των εγγράφων, είναι η διαχείριση των εγγράφων αυτών που διαφέρουν σημαντικά μεταξύ τους σε δομή. Μια πολύ σημαντική προσπάθεια συλλογής των XML εγγράφων του ΠΙ και διαχείρισής τους με μια γλώσσα ερωτήσεων που προσαρμόζεται στις πολλές διαφορετικές δομές αποτελεί το σύστημα XYLEME [AC+00].

2.2 Μοντέλα αναπαράστασης της εξαγόμενης πληροφορίας

Για να γίνει δυνατή η διαχείριση της πληροφορίας του ιστού απαιτείται ένα ενδιάμεσο μοντέλο αναπαράστασής της. Το μοντέλο αυτό θα πρέπει να προσαρμόζεται στη χαλαρή και ανομοιογενή δομή των εγγράφων [FL+98].

2.2.1 Ιεραρχικό μοντέλο

Το μοντέλο αυτό συμβαδίζει με την ιεραρχική δομή των στοιχείων των HTML εγγράφων. Στην περίπτωση του συστήματος W4F [SA99c], τα περιεχόμενα των HTML εγγράφων που θα εξαχθούν περιγράφονται με τη γλώσσα HEL, εξάγονται και αποθηκεύονται σε μια ιεραρχική δομή.



Σχήμα 1. Παράδειγμα δέντρου που αντιστοιχεί σε μια ιστοσελίδα HTML

Κάθε έγγραφο απεικονίζεται σαν ένα δένδρο με μία ρίζα (συνήθως τον τίτλο του εγγράφου), ορισμένους εσωτερικούς κόμβους (π.χ. λίστες) και ορισμένα φύλλα (είτε ετικέτες χωρίς περιεχόμενο, π.χ. ετικέτα αλλαγής γραμμής
, είτε κόμβους με καθαρά λεξικό περιεχόμενο CDATA). Αυτή η δενδρική δομή βασίζεται στο Μοντέλο DOM (Document Object Model) [W3e]. Το NSL (Nested String List – Λίστα Φωλιασμένων Αλφαριθμητικών) [SA99c] είναι μια δομή που αναπαριστά τις δενδρικές δομές ως λίστες από αλφαριθμητικά που περιέχουν με τη σειρά τους άλλες λίστες αλφαριθμητικών (βλέπε Σχήμα 1). Το μοντέλο αυτό περιγράφει αποκλειστικά τη δομή των εγγράφων και δεν ασχολείται με τις σχέσεις μεταξύ των εγγράφων (υπερσύνδεσμοι).

2.2.2 Αταξινόμητος γράφος Μοντέλο OEM

Το Μοντέλο Ανταλλαγής Αντικειμένων OEM (Object Exchange Model) έχει αρκετές ομοιότητες με το αντικειμενοστραφές μοντέλο, αν και είναι πιο ευέλικτο στη διαχείριση ημι-δομημένων δεδομένων. Κάθε αντικείμενο στο OEM αποτελείται από άλλα αντικείμενα που αποτελούν τα χαρακτηριστικά του. Ομάδες όμοιων αντικειμένων αναπαρίστανται ως συλλογές αντικειμένων. Τα αντικείμενα που συνθέτουν ένα αντικείμενο μπορεί να έχουν διαφορετική δομή. Στο μοντέλο δεν διατηρείται η σειρά εμφάνισης των αντικειμένων στο έγγραφο (βλ. Σχήμα 2, Σχήμα 3) [DF+99][AB+98]

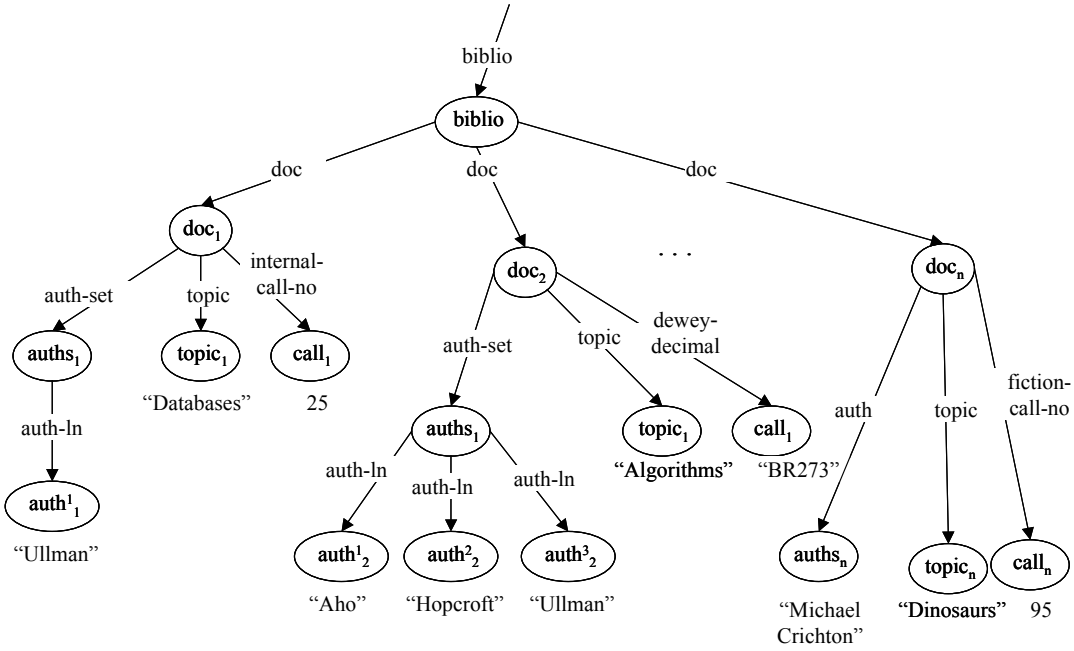
<biblio, set, {doc1 , doc2 , …, doc_n }>
  doc_1 is <doc, set, {$auths_1$, $topic_1$, $call_1$ }>
    $auths_1$ is <auth-set, set, {$auth^1_1$}>
      $auth^1_1$ is <auth-ln, string, “Ullman">
    $topic_1$ is <topic, string, “Databases">
    $call_1$ is <internal-call-no, integer, 25>
  doc_2 is <doc, set, {$auths_2$, $topic_2$, $call_2$ }>
    $auths_2$ is <auth-set, set, {$auth^1_2$, $auth^2_2$,$auth^3_2$}>
      $auth^1_2$ is <auth-ln, string, “Aho">
      $auth^2_2$ is <auth-ln, string, “Hopcroft">
      $auth^3_2$ is <auth-ln, string, “Ullman">
    $topic_2$ is <topic, string, "Algorithms">
    $call_2$ is <dewey-decimal, string, “BR273">
      . . .
  doc_n is <doc, set, {$auths_n$, $topic_n$, $call_n$ }>
    $auths_n$ is <auth, string, “Michael Crichton”>
    $topic_n$ is <topic, string, “Dinosaurs">
    $call_n$ is <fiction-call-no, integer, 95>

Σχήμα 2. Παράδειγμα δομής του OEM

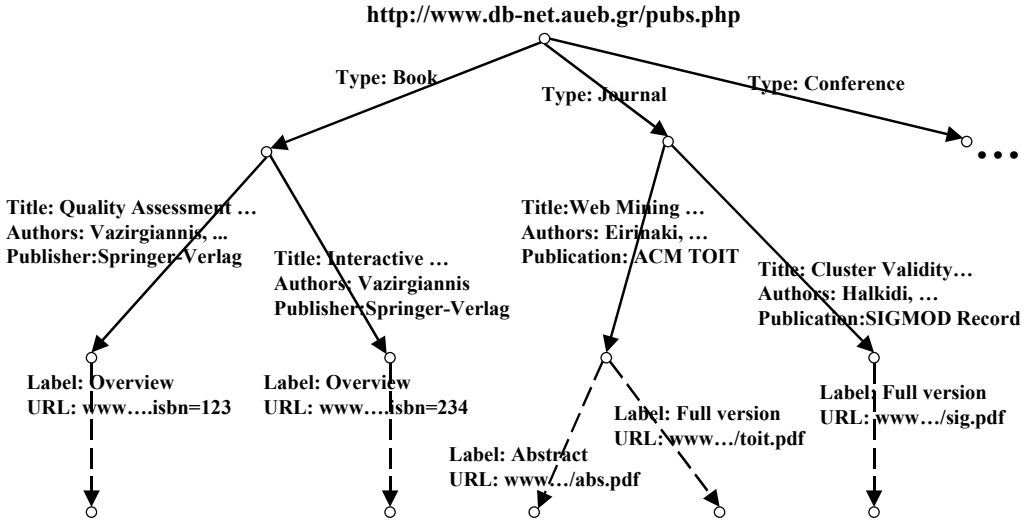
Το μοντέλο OEM αναπαριστά το κάθε έγγραφο ως ένα χαρακτηρισμένο αταξινόμητο κατευθυνόμενο γράφο. Όλοι οι κόμβοι του γράφου είναι αντικείμενα Σ*ύνθετα* ή Α*τομικά* και χαρακτηρίζονται μοναδικά (*oid)*. Τα Ατομικά αντικείμενα περιέχουν τιμές κάποιου Ατομικού τύπου, π.χ. ακέραιος, πραγματικός, κείμενο, εικόνα. Το περιεχόμενο των *Σύνθετων* αντικειμένων είναι ένα σύνολο από ζεύγη (*label, subobject)* όπου η ετικέτα (label) περιγράφει λεκτικά τη σχέση μεταξύ του αντικειμένου και του υπο-αντικειμένου του (subobject). Τα σύνθετα αντικείμενα έχουν εξερχόμενες ακμές προς τα υπο-αντικείμενά τους ενώ τα ατομικά αντικείμενα αποτελούν τερματικούς κόμβους του γράφου [GH+99]. Το μοντέλο OEM χρησιμοποιήθηκε στο σύστημα TSIMMIS[PM+95] για την αποθήκευση της

εξαγόμενης από τα έγγραφα του ιστού πληροφορίας. Ένα υπερσύνολο του OEM είναι και το μοντέλο ODMG που χρησιμοποιήθηκε από το σύστημα Garlic [RS97].

Η γλώσσα WebOQL [AM98] εισάγει την έννοια των υπερ-δένδρων, όπου τα έγγραφα συνδέονται μεταξύ τους με ακμές που αντιστοιχούν στους πραγματικούς συνδέσμους, αλλά έχουν και εσωτερική δομή επίσης δενδρική (βλ. Σχήμα 4). Επίσης η γλώσσα *StruQL*, που χρησιμοποιείται από το σύστημα STRUDEL [FF+97] θεωρεί τον Ιστό ως χαρακτηρισμένο κατευθυνόμενο γράφο.



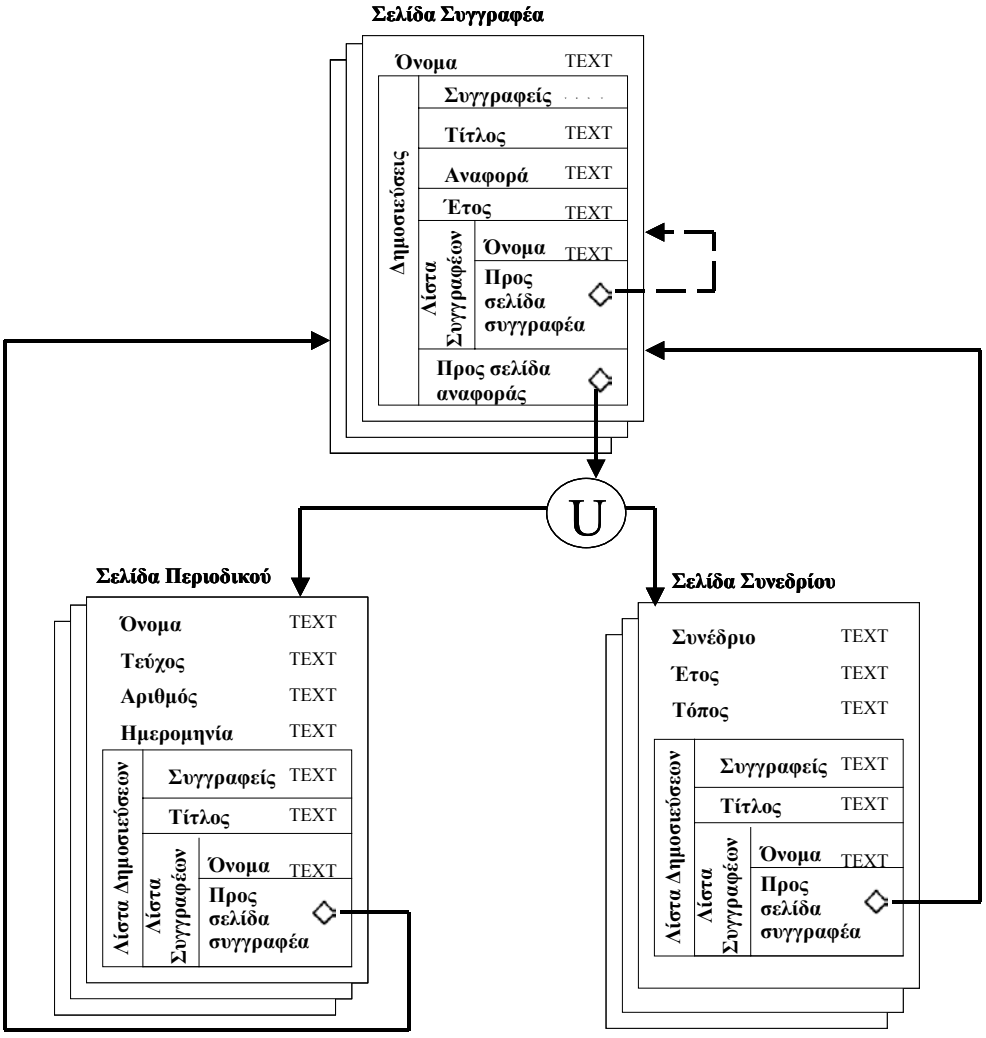
Σχήμα 3. Παράδειγμα αναπαράστασης του OEM



Σχήμα 4. Παράδειγμα υπερ-δένδρου (http://www.db-net.aueb.gr/pubs.php)

2.2.3 Μοντέλο κατευθυνόμενου γράφου

Το μοντέλο αυτό προτείνεται στο σύστημα ARANEUS [AM+97] και μετατρέπει μια συλλογή εγγράφων του ΠΙ σε ένα κατευθυνόμενο πολυ-γράφο (directed multigraph). Κάθε έγγραφο HTML είναι ένα αντικείμενο με προσδιοριστικό τη διεύθυνση ιστού του και χαρακτηριστικά όπως: κείμενο, εικόνες, συνδέσμους (απλά χαρακτηριστικά – με μοναδικές τιμές), πίνακες, λίστες κ.ά. (σύνθετα χαρακτηριστικά – με πολλαπλές τιμές). Τα σύνθετα χαρακτηριστικά εκφράζονται ως λίστες απλών χαρακτηριστικών.



Σχήμα 5. Σύνδεση προτύπων εγγράφων στο σύστημα Araneus

Σύμφωνα με το μοντέλο ARANEUS:

- Ένα Πρότυπο Έγγραφο (*Page Scheme)* αντιπροσωπεύει ένα σύνολο εγγράφων με ίδια δομή. Ακόμη και έγγραφα μοναδικής δομής αντιστοιχούν σε ένα Πρότυπο. Ουσιαστικά το Πρότυπο Έγγραφο (π.χ. Σελίδα περιοδικού, συνεδρίου, συγγραφέα στο Σχήμα 5) είναι το "καλούπι" που θα χρησιμοποιούσαμε για να ξαναφτιάξουμε τα αρχικά έγγραφα από τα δεδομένα που έχουμε εξαγάγει.

- Δύο έγγραφα με κοινά χαρακτηριστικά αλλά διαφορετικές δομές (Πρότυπα) μπορούν να συγχωνευθούν σε ένα Ενωτικό Πρότυπο (Union - συμβολίζεται με U στο Σχήμα 5)
- Χαρακτηριστικά που δεν παίρνουν τιμές σε όλα τα έγγραφα δηλώνονται ως προαιρετικά στο Πρότυπο Έγγραφο.

Σε αντίθεση με άλλα μοντέλα, οι σύνδεσμοι λαμβάνονται υπόψη. Σύνδεσμοι μπορεί να εμφανίζονται μεταξύ δύο Πρότυπων Εγγράφων ή Ενωτικών Προτύπων, αν όλα τα έγγραφα του ενός Προτύπου εμφανίζουν σύνδεσμο προς έγγραφα του άλλου. Δημιουργείται έτσι ένας γράφος όχι σε επίπεδο εγγράφων μόνο, αλλά και σε επίπεδο προτύπων, που μπορεί να δώσει πολύ χρήσιμες πληροφορίες. Το μοντέλο είναι αρκετά αποτελεσματικό σε συλλογές εγγράφων που μοιράζονται ίδιες ή παρόμοιες δομές – για παράδειγμα σε έγγραφα που παράγονται από μια βάση δεδομένων με βιβλιογραφικές πληροφορίες με αυτόματο τρόπο, όπως στη δικτυακή βιβλιογραφική βάση DBLP [DBLP].

2.2.4 Σχεσιακό Μοντέλο

Συχνά τα δεδομένα που εξάγονται από τα έγγραφα απεικονίζονται σε σχεσιακές βάσεις δεδομένων. Στην περίπτωση αυτή μας ενδιαφέρει να απεικονίσουμε τις ετικέτες της HTML και τα περιεχόμενά τους αλλά και τις σχέσεις μεταξύ των εγγράφων της συλλογής. Στο [SS98] ορίζονται τέσσερις τύποι σχέσεων: α) οι βασικές, β) οι σχέσεις ετικέτας-τιμής, γ) οι σχέσεις μεταξύ ετικετών, δ) οι σχέσεις για δευτερογενή δεδομένα που αφορούν το έγγραφο αλλά δεν περιέχονται σε αυτό.

Κάθε λέξη του εγγράφου προσδιορίζεται από κάποιο κωδικό, ομοίως και κάθε έγγραφο. Οι διευθύνσεις ιστού των εγγράφων αναλύονται και κάθε τμήμα τους αποθηκεύεται σε διαφορετική πλειάδα. Κάθε ετικέτα αποθηκεύεται στη σχέση TAG μαζί με το περιεχόμενό της. Στη σχέση LINK αποθηκεύονται πληροφορίες για την πηγή και τον στόχο του κάθε συνδέσμου.

Η γλώσσα ερωτήσεων *WebSQL* [MM+97] αντιμετωπίζει τον ΠΙ ως μια σχεσιακή βάση που αποτελείται από δύο εικονικές σχέσεις, *Έγγραφο* και *Σύνδεσμος*, αγνοώντας όμως πλήρως την εσωτερική δομή των εγγράφων.

Ένας διαφορετικός τρόπος χρήσης του σχεσιακού μοντέλου είναι για την αποθήκευση πληροφορίας που έχει εξαχθεί από τα έγγραφα και έχει ενδιάμεσα αποθηκευθεί σε μια ιεραρχική δομή όπως αυτή που περιγράφει η XML. Μεγάλος αριθμός ερευνητών ασχολείται με την περιοχή αυτή που αντιμετωπίζει τόσο θεωρητικά όσο και πρακτικά προβλήματα. Μέρος της παρούσας διατριβής αποτελεί και ο τομέας της απεικόνισης δομημένων εγγράφων σε σχεσιακές βάσεις δεδομένων.

2.3 XML και σχεσιακές βάσεις δεδομένων

Χάρη στην ιεραρχική, αυτο-περιγραφική δομή της, η XML αποτελεί ένα εύχρηστο και ευέλικτο μοντέλο απεικόνισης δεδομένων. Τα έγγραφα XML μπορεί να έχουν αυθαίρετη ή συγκεκριμένη δομή. Για τον ορισμό και τον έλεγχο της δομής των XML εγγράφων έχουν οριστεί πρότυπα όπως η γλώσσα Ορισμού Τύπου Εγγράφου (Document Type Definition – DTD [W3f]) ή η γλώσσα XMLSchema [W3g].

Η εξόρυξη δεδομένων από τα έγγραφα του ΠΙ και η αποθήκευσή τους σε ένα μοντέλο όπως αυτό της XML μας επιτρέπει να διατηρήσουμε τα δεδομένα αυτά σε μια δομημένη, αυτο-περιγραφική, εύληπτη και ευέλικτη μορφή, να τα οργανώσουμε και να κάνουμε αναζητήσεις σε αυτά. Παρ'όλα αυτά, καθώς η XML και οι δορυφορικές της τεχνολογίες βρίσκονται ακόμη σε ένα πρώιμο στάδιο, είναι δύσκολο να διαχειριστούμε μεγάλο όγκο δεδομένων, να κάνουμε σύνθετες αναζητήσεις πάνω στα XML έγγραφα ή να επιτελέσουμε πολύπλοκες λειτουργίες, όπως η εξόρυξη γνώσης. Γι' αυτό αναγκαζόμαστε να χρησιμοποιήσουμε την XML ως μια ενδιάμεση αποθηκευτική δομή και να απεικονίσουμε τα δεδομένα μας σε ένα καθιερωμένο μοντέλο δεδομένων, όπως το σχεσιακό.

Για τη μετάβαση από το μοντέλο της XML σε κάποιο άλλο μοντέλο, π.χ. το σχεσιακό, απαιτείται μια σειρά από μετασχηματισμούς, που θα απεικονίσουν τις δομές XML σε δομές του νέου μοντέλου. Απαιτείται επίσης ένας μηχανισμός, που θα υλοποιήσει τους μετασχηματισμούς αυτούς και τελικά θα αναλάβει την αποθήκευση των περιεχομένων των XML εγγράφων στη νέα μορφή χωρίς απώλειες και με τρόπο αναστρέψιμο.

Οι κατευθύνσεις που υπάρχουν σήμερα για την ολοκλήρωση μεταξύ της XML και των βάσεων δεδομένων είναι δύο: α) η δημιουργία συστημάτων βάσεων δεδομένων προσαρμοσμένων στη δομή της XML και β) η επέκταση των υπαρχόντων συστημάτων ώστε να υποστηρίζουν τη διαχείριση XML εγγράφων.

Στην πρώτη περίπτωση, τα έγγραφα XML αποθηκεύονται ως αντικείμενα (large objects - LOBs) και αναπτύσσονται οι κατάλληλοι μηχανισμοί για το χειρισμό αυτών των αντικειμένων. Τέτοια συστήματα διαθέτουν ξεχωριστούς μηχανισμούς διαχείρισης σχήματος, ευρετηρίων και αποθήκευσης και ξεχωριστή μηχανή αναζήτησης προσαρμοσμένη στη δομή και τα ιδιαίτερα χαρακτηριστικά της XML. Τα συστήματα αυτά είναι αποτελεσματικά στη διαχείριση ολόκληρων εγγράφων XML, αλλά δυσκολεύονται να χειριστούν τα περιεχόμενα, ή τμήματα των εγγράφων.

Στη δεύτερη κατεύθυνση, τα υπάρχοντα συστήματα βάσεων δεδομένων επενδύουν σε μια ώριμη τεχνολογία όπως αυτή των σχεσιακών βάσεων για να χειριστούν δεδομένα XML. Στόχος τους να απεικονίσουν το ιεραρχικό μοντέλο της XML στο σχεσιακό μοντέλο, επιτρέποντας έτσι την αποθήκευση των περιεχομένων των XML εγγράφων αλλά και την ανάκτηση των αποθηκευμένων στοιχείων σε μορφή XML.

2.3.1 Αποθήκευση ολόκληρων XML εγγράφων

Η κατηγορία αυτή περιλαμβάνει: α) συστήματα που παρέχουν αποθήκευση και διαχείριση XML εγγράφων με τη μορφή αντικειμένων (Υβριδικά συστήματα σχεσιακά με δυνατότητες XML - XML enabled systems) και β) συστήματα που παρέχουν αποτελεσματική αποθήκευση και διαχείριση ημι-δομημένων δεδομένων (Γηγενή συστήματα XML - Native XML systems). Τα πρώτα αναπτύχθηκαν από εταιρίες που ήδη δραστηριοποιούνται στο χώρο των βάσεων δεδομένων, ως επεκτάσεις στα υπάρχοντα συστήματα, ενώ τα δεύτερα προήλθαν κυρίως από ερευνητική εργασία στη διαχείριση ημι-δομημένων εγγράφων.

Στα συστήματα που υποστηρίζουν XML, όπως η Oracle 9i [DL01], ένα ολόκληρο XML έγγραφο αποθηκεύεται σε ένα πεδίο *τύπου XML*. Τα XML έγγραφα που αποθηκεύονται στο ίδιο πεδίο πρέπει να έχουν την ίδια δομή. Τα συστήματα αυτά

επιπλέον προσφέρουν μηχανισμούς επικύρωσης της ορθότητας των αποθηκευμένων XML εγγράφων σύμφωνα με κάποια δομή που περιγράφεται σε DTD ή XML-Schema. Οι χρήστες μπορούν να υποβάλουν ερωτήσεις προς τα έγγραφα της βάσης χρησιμοποιώντας ένα συντακτικό που στηρίζεται στη γλώσσα ερωτήσεων XPATH [W3h].

Τα Γηγενή συστήματα XML (XML-native) είναι συστήματα αρχείων που αποθηκεύουν τα έγγραφα XML φυσικά κοντά στο μέσο αποθήκευσης και πραγματοποιούν φυσικές συνδέσεις, που είναι πολύ ταχύτερες από τις λογικές συνδέσεις που συμβαίνουν στις σχεσιακές ΒΔ. Παρ' όλα αυτά, τέτοια συστήματα δυσκολεύονται ή αργούν στη διαχείριση πολύ μεγάλων XML εγγράφων. Για το λόγο αυτό ερευνητικά συστήματα όπως τα Rufus[SL+93], Lore [GH+99], Strudel [FL+98] και Natix [KM00] αλλά και εμπορικές εφαρμογές όπως τα eXcelon[ODI01], Tamino [SK+00] και GoXML [XGT01] χρησιμοποιούν υβριδικά μοντέλα που υποστηρίζουν σύνθετα αντικείμενα για την πληροφορία των XML εγγράφων αλλά και απλά αντικείμενα για να αποθηκεύουν λιγότερο σύνθετες και λιγότερο σημαντικές XML δομές.

2.3.2 Απεικόνιση των περιεχομένων των XML εγγράφων σε οντότητες της βάσης δεδομένων

Τα συστήματα που επιτρέπουν την απεικόνιση και την ανταλλαγή δεδομένων μεταξύ των XML εγγράφων και βάσεων δεδομένων χωρίζονται σε δύο κατηγορίες [B01]: α) συστήματα που βασίζονται ***σε συγκεκριμένους κανόνες απεικόνισης*** που καθορίζονται από τους χρήστες. Με βάση αυτά τα πρότυπα απεικόνισης αποθηκεύουν συγκεκριμένα μέρη των XML εγγράφων σε ένα προκαθορισμένο σχήμα βάσης και β) συστήματα που ακολουθούν ***γενικούς κανόνες απεικόνισης*** για να αποθηκεύσουν αυτόματα δομές της XML στο σχήμα της βάσης, το οποίο δημιουργείται αυτόματα. Η βάση δεδομένων που χρησιμοποιείται είναι σχεδόν πάντα σχεσιακή.

Τα συστήματα της πρώτης κατηγορίας, που χρησιμοποιούν απεικόνιση βασισμένη σε κάποιο πρότυπο, είναι ευέλικτα καθώς επιτρέπουν οποιοδήποτε τμήμα του XML εγγράφου να αποθηκευθεί σε οποιοδήποτε πεδίο της βάσης. Απαιτούν όμως γνώση της γλώσσας με την οποία περιγράφονται οι κανόνες απεικόνισης. Ένα σύστημα που συνδέει έγγραφα XML και σχεσιακές βάσεις δεδομένων έχει διπλό ρόλο: α) παράγει XML έγγραφα συνθέτοντας σχεσιακά δεδομένα, β) αποθηκεύει τα περιεχόμενα των XML εγγράφων στη σχεσιακή βάση. Στα περισσότερα εμπορικά συστήματα [IBM02], [SQL01], και οι δύο λειτουργίες υποστηρίζονται, με διαφορετικό κάθε φορά συντακτικό των κανόνων απεικόνισης. Ορισμένα μάλιστα συστήματα χειρίζονται μονόπλευρα το πρόβλημα ολοκλήρωσης XML και σχεσιακών βάσεων δεδομένων και επιτρέπουν μόνο την εξαγωγή δεδομένων της βάσης σε XML και όχι το αντίστροφο. Αυτό γίνεται με τη δημιουργία μιας ενδιάμεσης όψης των δεδομένων σε μια απλουστευμένη δομή XML και στη συνέχεια με κάποια γλώσσα ερωτήσεων πάνω στις XML απόψεις (SilkRoute [FT+00], Xperanto[CK+00]). Εναλλακτικά χρησιμοποιούνται επεκτάσεις της SQL που παράγουν απευθείας XML έγγραφα από τα αποτελέσματα των ερωτήσεων [SS+00].

Στην αυτόματη απεικόνιση όλα τα δεδομένα των XML εγγράφων αποθηκεύονται στη βάση δεδομένων ακολουθώντας ένα σύνολο κανόνων που καθορίζονται με βάση ένα προκαθορισμένο μοντέλο. Τα δεδομένα αποθηκεύονται σε ένα σχεσιακό ή αντικειμενοστραφές σχήμα. Στην περίπτωση των σχεσιακών σχημάτων κάθε XML

έγγραφο απεικονίζεται σε ένα σύνολο πινάκων ενώ στα αντικειμενοστραφή σχήματα σε ένα "δέντρο" αντικειμένων. Οι σύνθετες δομές απεικονίζονται σε τάξεις, ενώ οι απλές δομές (απλοί τύποι στοιχείων, χαρακτηριστικά και κείμενο) απεικονίζονται σε μονότιμα χαρακτηριστικά (scalar properties) των τάξεων. Το αντικειμενοστραφές μοντέλο μπορεί να απεικονιστεί στη συνέχεια σε σχεσιακό: οι τάξεις απεικονίζονται σε σχέσεις, τα μονότιμα χαρακτηριστικά σε πεδία και τα χαρακτηριστικά σύνθετου τύπου σε ζεύγη πρωτεύοντος-εξωτερικού κλειδιού [DF+99]. Το LegoDB [BF+02] είναι ένα σύστημα που απεικονίζει τη δομή των XML εγγράφων όπως περιγράφεται σε ένα αρχείο DTD στο οικονομικότερο σε χώρο σχεσιακό σχήμα και αποθηκεύει τα περιεχόμενα των εγγράφων σε αυτό. Οι μέθοδοι απεικόνισης αυτής της κατηγορίας παράγουν συνήθως φωλιασμένα, μη κανονικοποιημένα σχήματα [SA+89].

2.3.3 Περιορισμοί των υπαρχόντων συστημάτων

Τα υπάρχοντα συστήματα που προσφέρουν δυνατότητες αποθήκευσης και διαχείρισης XML εγγράφων πάνω από σχεσιακές βάσεις δεδομένων αντιμετωπίζουν διάφορους περιορισμούς που οφείλονται κυρίως στην ανοικτή και λιγότερο αυστηρή δομή της XML και στην πολυπλοκότητα του σχήματος των XML εγγράφων. Για το λόγο αυτό αναγκάζονται να κάνουν ορισμένες παραδοχές που διευκολύνουν αλλά δεν επιλύουν το πρόβλημα της διασύνδεσης XML και σχεσιακών βάσεων δεδομένων. Πιο συγκεκριμένα:

Η απεικόνιση με βάση συγκεκριμένους κανόνες απαιτεί την ανθρώπινη παρέμβαση: Η απεικόνιση που βασίζεται σε συγκεκριμένους κανόνες είναι πιο ευέλικτη καθώς επιτρέπει να αποθηκεύσουμε στη βάση μόνο τα περιεχόμενα των εγγράφων που επιθυμούμε καθορίζοντας παράλληλα σε ποια πεδία θα αποθηκευθούν. Παράλληλα όμως, η μέθοδος αυτή απαιτεί να προϋπάρχει η βάση και επιπλέον απαιτεί εξοικείωση με τη δομή της βάσης, τη δομή των XML εγγράφων αλλά και τη γλώσσα απεικόνισης.

Η απεικόνιση με βάση συγκεκριμένους κανόνες απαιτεί να προϋπάρχει το σχήμα της βάσης: Τα συστήματα που χρησιμοποιούν απεικόνιση βασισμένη σε κανόνες υποθέτουν ότι το σχήμα βάσεων δεδομένων έχει σχεδιαστεί και δημιουργηθεί εκ των προτέρων. Στην περίπτωση σύνθετων δομών XML (όπως για παράδειγμα τα XML έγγραφα MPEG-7 [ISOa]) είναι δύσκολο να παραχθεί με το χέρι το σχήμα της βάσης δεδομένων και επιπλέον να απεικονιστούν τα περιεχόμενα των XML εγγράφων στη βάση δεδομένων. Στην περίπτωση αυτή προτιμούνται συστήματα που αυτόματα παράγουν το σχήμα της βάσης και καθορίζουν τους κανόνες απεικόνισης του περιεχομένου των εγγράφων.

Το σχήμα της βάσης που παράγεται δεν είναι βέλτιστο: Η αυτόματη διαδικασία απεικόνισης είναι απλή αλλά δεν υποστηρίζει τα συγκεκριμένα χαρακτηριστικά γνωρίσματα βάσεων δεδομένων, όπως ο καθορισμός του τύπου των ιδιοτήτων (μέγεθος, ελάχιστες/μέγιστες τιμές κτλ.), οι μηχανισμοί ερωτήσεων, οι διαδικασίες κανονικοποίησης και βελτιστοποίησης (π.χ. ευρετήρια κτλ).

Το πρότυπο DTD έχει περιορισμένες δυνατότητες στην περιγραφή της δομής των XML εγγράφων: Πολλά από τα συστήματα απεικόνισης [ST+99] βασίζονται στο πρότυπο DTD [W3f] για να περιγράψουν τη δομή των εγγράφων XML. Η χρήση DTDs αντί της XML-Schema διευκολύνει τη διαδικασία απεικόνισης καθώς οι δομές

που ορίζονται με DTD είναι απλούστερες, αλλά έχει ορισμένους περιορισμούς που οφείλονται κυρίως στις περιορισμένες ικανότητες της γλώσσας DTD: δεν επιτρέπουν να ορίσουμε νέους τύπους, δεν επιτρέπουν να χρησιμοποιήσουμε περισσότερα από ένα αρχεία, εγγυώνται ορθότητα αναφορών μόνο εντός του ίδιου XML εγγράφου κτλ.

Το περιεχόμενο βάσεων δεδομένων δεν μπορεί να ενημερωθεί: Η πλειοψηφία των προαναφερθέντων συστημάτων επιτρέπει την αποθήκευση των εγγράφων XML στη βάση δεδομένων και τις αναζητήσεις στη βάση δεδομένων, αλλά δεν αντιμετωπίζουν το πρόβλημα ενημέρωσης των περιεχομένων της βάσης δεδομένων. Ακόμα και όταν επιτρέπουν ενημέρωση των περιεχομένων, όπως π.χ. το [FK99], δεν εγγυώνται την εγκυρότητα των ενημερωμένων εγγράφων (π.χ. επιτρέπουν τη διαγραφή ενός υποχρεωτικού χαρακτηριστικού ή στοιχείου).

2.4 Υπερσύνδεσμοι

Οι υπερσύνδεσμοι και η πληροφορία που μεταφέρουν αποτελούν σημαντικό στοιχείο της παρούσας εργασίας και γι' αυτό θα συζητηθούν εκτενέστερα στη συνέχεια. Θεωρώντας τον ΠΙ ως ένα κατευθυνόμενο γράφο, με κόμβους τα έγγραφα και ακμές τους συνδέσμους, μπορούμε να διακρίνουμε δύο τύπους ακμών (συνδέσμων): α) τους εισερχόμενους προς ένα κόμβο (έγγραφο) και β) τους εξερχόμενους από αυτόν. Στην παρούσα μορφή των συνδέσμων της HTML κάθε σύνδεσμος έχει ένα μόνο κόμβο ως αρχή (πηγαίο έγγραφο) και ένα μόνο κόμβο ως τέλος (τελικό έγγραφο). Οι σύνδεσμοι που εισέρχονται σε ένα έγγραφο έχουν ως αρχή άλλα έγγραφα του ιστού. Οι δημιουργοί των εγγράφων αυτών εκδηλώνουν, με τη χρήση των συνδέσμων, το ενδιαφέρον τους προς το έγγραφο που υποδεικνύουν, ενώ ταυτόχρονα παρέχουν και μια σύντομη περιγραφή (μια ή περισσότερες λέξεις) για το τελικό έγγραφο.

Βέβαια, ένα ποσοστό των συνδέσμων του ιστού, φτιάχνεται για να εξυπηρετεί τις ανάγκες περιήγησης των χρηστών εντός των δικτυακών τόπων (π.χ. σύνδεσμοι προς την κεντρική σελίδα του τόπου, προς προηγούμενη ή επόμενη κτλ.) και δεν προσφέρει πολύ χρήσιμες περιγραφές. Τέτοιοι σύνδεσμοι είναι εύκολο να εντοπιστούν καθώς οι διευθύνσεις του πηγαίου και του τελικού εγγράφου είναι παρόμοιες [S97] (συνήθως βρίσκονται στον ίδιο ιστότοπο). Ένα πολύ μικρότερο ποσοστό συνδέσμων δημιουργείται εσκεμμένα με σκοπό να παραπλανήσει τους αναγνώστες του εγγράφου για τα περιεχόμενα των τελικών εγγράφων. Αυτοί οι σύνδεσμοι είναι δύσκολο να εντοπιστούν, παρ' όλα αυτά, είναι δύσκολο για το δημιουργό ενός εγγράφου να δημιουργήσει πολλούς τέτοιους συνδέσμους προς το έγγραφό του καθώς θα πρέπει να έχει πρόσβαση σε πολλούς διαφορετικούς δικτυακούς τόπους.

Η έρευνα που γίνεται στο πεδίο των υπερσυνδέσμων μπορεί να διακριθεί σε δύο επίπεδα: α) στην ανάλυση της συνδεσμολογίας τους ΠΙ, που λαμβάνει υπόψη της μόνο την "αρχή" και το "τέλος" των συνδέσμων και εντοπίζει πρότυπα στο γράφο του ΠΙ, β) στην εξαγωγή λέξεων από τους υπερσυνδέσμους και στη χρήση της για το χαρακτηρισμό των εγγράφων. Πολύ σημαντικό μέρος της σχετιζόμενης έρευνας καλύπτουν και τα συστήματα ερωτήσεων πάνω στην πληροφορία των υπερσυνδέσμων.

2.4.1 Εξαγωγή λέξεων από τους υπερσυνδέσμους

Ο Kleinberg [K99] δηλώνει πως *"Η δομή ενός περιβάλλοντος υπερμέσων μπορεί να αποτελέσει πλούσια πηγή πληροφορίας για το περιεχόμενο του περιβάλλοντος (...) Όμως για το πρόβλημα της αναζήτησης σε διασυνδεδεμένα περιβάλλοντα όπως ο Παγκόσμιος Ιστός, είναι ξεκάθαρο από τις παρούσες τεχνικές ότι πρέπει να εκμεταλλευθούμε πλήρως την πληροφορία που περιέχεται στους συνδέσμους."*

Οι Zamir και Etzioni [ZE98] εισάγουν την ιδέα των "ισχυρών" συνδέσμων, οι οποίοι περιέχουν περιγραφική πληροφορία για τα έγγραφα στα οποία δείχνουν. Οι ίδιοι ισχυρίζονται πως η πληροφορία μπορεί να περιοριστεί σε πέντε λέξεις, ενώ χρησιμοποιούν εμπειρικά αποτελέσματα για να το αποδείξουν. Στο [CD+98a] τα πειράματα που στοχεύουν στον εντοπισμό του βέλτιστου εύρους για το παράθυρο κειμένου γύρω από τον σύνδεσμο καταλήγουν στο συμπέρασμα ότι 50 χαρακτήρες πριν και μετά το σύνδεσμο είναι αρκετοί για να εντοπίσουμε πληροφορία σχετική με το σύνδεσμο. Τα πειράματα αυτά μετρούν τη συχνότητα εμφάνισης των λέξεων σε μια περιοχή πριν και μετά τον σύνδεσμο.

Στο [VV01] προτείνεται ένα σύστημα που εμπλουτίζει τις αναζητήσεις στον ΠΙ, προσθέτοντας σημασιολογική πληροφορία στα έγγραφα και τους συνδέσμους τους, και μια δομή για την οργάνωση της πληροφορίας αυτής. Τα σημασιολογικά χαρακτηριστικά προέρχονται από μια εννοιολογική ιεραρχία η οποία ποικίλει στα διάφορα θεματικά πεδία και δημιουργείται από τους ειδικούς κάθε πεδίου.

Πιστεύουμε ότι είναι ευκολότερο να χαρακτηρίσουμε μια ιστοσελίδα χρησιμοποιώντας την πληροφορία που παρέχουν οι σελίδες που δείχνουν προς αυτή αντί να χρησιμοποιούμε την πληροφορία που παρέχει η ίδια η σελίδα. Ως αποτέλεσμα, η πληροφορία που φέρει ένας σύνδεσμος δεν περιορίζεται μόνο στο στόχο του συνδέσμου αλλά και στο πώς ο σύνδεσμος χαρακτηρίζει το στόχο.

2.4.2 Ανάλυση συνδέσμων και ομαδοποίηση ιστοσελίδων

Στις περισσότερες περιπτώσεις η ομαδοποίηση ιστοσελίδων βασίζεται στη μεταξύ τους συνδεσιμότητα [CD+98b], [CD+99] και όχι στα σημασιολογικά χαρακτηριστικά που μπορεί να φέρουν οι σύνδεσμοι. Οι τεχνικές εξόρυξης γνώσης από τα περιεχόμενα των ιστοσελίδων, περιορίζονται στην εξόρυξη λέξεων από το κείμενο των σελίδων, αγνοώντας τη δομή των ιστοσελίδων και τη συνδεσιμότητά τους με άλλες σελίδες. Είναι πολύ σημαντικό να λάβουμε υπόψη μας τους συνδέσμους, αλλά και τα περιεχόμενα των σελίδων, ώστε να μπορέσουμε στη συνέχεια να τμηματοποιήσουμε μεγάλες συλλογές εγγράφων του ΠΙ.

Ορισμένες πρώτες ερευνητικές δραστηριότητες που σχετίζονται με την μελέτη της σημαντικότητας των συνδέσμων προέρχονται από το πεδίο των δικτύων υπερμέσων ([BH+99], [PM+97]).

Λόγω του μεγέθους του ΠΙ, πολύ συχνά προτιμάται να εφαρμόζονται οι τεχνικές ομαδοποίησης εγγράφων στο σύνολο των απαντήσεων μιας αναζήτησης και όχι σε όλο το σύνολο του ΠΙ. Στο [ZE98] προτείνεται μια τεχνική ομαδοποίησης – που βασίζεται σε δέντρα καταλήξεων (suffix tree-clustering) – σύμφωνα με την οποία ο αλγόριθμος αναζητά κοινές φράσεις ανάμεσα στα έγγραφα της συλλογής για να δημιουργήσει ομάδες. Τα έγγραφα της συλλογής αρχικά καθαρίζονται (με την

εφαρμογή ενός αλγορίθμου αποκοπής καταλήξεων). Οι βασικές ομάδες προσδιορίζονται με τη χρήση ενός δέντρου καταλήξεων, με αποτέλεσμα τη δημιουργία ενός ανεστραμμένου ευρετηρίου φράσεων για όλη τη συλλογή. Τα τμήματα που παρουσιάζουν αρκετές επικαλύψεις (αρκετά κοινά έγγραφα) συγχωνεύονται. Ο χρόνος εκτέλεσης του αλγορίθμου τμηματοποίησης είναι γραμμικός προς το πλήθος των εγγράφων.

Οι Haveliwala κ.ά [HG+02] προτείνουν μια διαφορετική διαδικασία τμηματοποίησης συνόλων από έγγραφα του ΠΙ. Κάθε έγγραφο του ΠΙ u, αναπαρίσταται από ένα σύνολο όρων που εξάγονται από τα περιεχόμενα του εγγράφου, από τις περιοχές των συνδέσμων που δείχνουν προς το u κτλ. Επίσης τα αντίστοιχα βάρη ενός όρου συμπληρώνουν την περιγραφή ενός εγγράφου. Συνεπώς κάθε έγγραφο αναπαρίσταται ως σύνολο:

$$B_u = \left\{\left(w_u^1, f_u^1\right), \left(w_u^2, f_u^2\right), \ldots, \left(w_u^k, f_u^k\right)\right\}$$

όπου w_u^i είναι οι λέξεις που χαρακτηρίζουν το u και f_u^i τα αντίστοιχα βάρη.
Η ομοιότητα των εγγράφων βασίζεται στις λέξεις που εξάγονται από αυτά και στη συχνότητά τους. Το μέτρο ομοιότητας υπολογίζει το ποσοστό των κοινών όρων προς το πλήθος των κοινών και μη κοινών όρων για δύο έγγραφα. Η σύγκριση ανάμεσα στους όρους γίνεται με ακριβή λεξικά κριτήρια χρησιμοποιώντας τη συντεταγμένη Jaccard (*Jaccard coefficient*) [W88].

Η πληροφορία των συνδέσμων χρησιμοποιείται ήδη από μηχανές αναζήτησης για να ξεκαθαρίσει και να ταξινομήσει τα αποτελέσματα των ερωτήσεων, για παράδειγμα ο αλγόριθμος PageRank [PB+98] του Google δίνει προτεραιότητα σε σελίδες με πολλούς εισερχόμενους ή εξερχόμενους συνδέσμους. Ένα άλλο ενδιαφέρον σύστημα που χρησιμοποιεί τη δομή των συνδέσμων για να ομαδοποιεί τα αποτελέσματα είναι το Kartoo [Kar01]. Το σύστημα Northern Light [Nor01] κατηγοριοποιεί κάθε έγγραφο της συλλογής σε προκαθορισμένες θεματικές κατηγορίες και στις ερωτήσεις των χρηστών επιστρέφει ως απάντηση τις πλησιέστερες κατηγορίες. Τέλος, το σύστημα Vivísimo [Viv01] χρησιμοποιεί την τμηματοποίηση του συνόλου των αποτελεσμάτων για την καλύτερη παρουσίασή τους. Χρησιμοποιεί για κάθε σελίδα την πληροφορία που επιστρέφουν οι μηχανές αναζήτησης (τίτλοι και σύντομες περιγραφές). Το Vivísimo δεν χρησιμοποιεί προκαθορισμένα θέματα και παράγει τις περιγραφές των διαφόρων ομάδων με βάση τις περιγραφές των εγγράφων που τις απαρτίζουν.

Μια μέθοδος για την κατηγοριοποίηση και περιγραφή εγγράφων του ΠΙ παρουσιάζεται στο [GT+02]. Εκεί χρησιμοποιούνται οι εισερχόμενοι σύνδεσμοι και οι λέξεις που αυτοί περιέχουν για να περιγράψουν τα έγγραφα. Ένας ταξινομητής εκπαιδεύεται για να ταξινομεί τα έγγραφα. Προτείνεται επίσης μια μέθοδος για την επιλογή χαρακτηριστικών και για το χαρακτηρισμό συνόλων εγγράφων του ΠΙ που χρησιμοποιεί το μέτρο της εντροπίας. Τα αποτελέσματα της προτεινόμενης προσέγγισης δείχνουν ότι το κείμενο στους εισερχόμενους συνδέσμους έχει μεγαλύτερη περιγραφική ισχύ από τα ίδια τα περιεχόμενα του εγγράφου.

Συστήματα που χρησιμοποιούν θησαυρούς για να βελτιώσουν τα αποτελέσματα ή να επεκτείνουν τις ερωτήσεις των χρηστών εμφανίζονται σε πολλές ερευνητικές προσπάθειες όπως στο [QF94].

Ένα ακόμη σύστημα που ταξινομεί έγγραφα του ΠΙ σε προκαθορισμένες κατηγορίες, αφού πρώτα εκπαιδευτεί σε ένα ταξινομημένο σύνολο εγγράφων, παρουσιάζεται στο [CG+97]. Στην περίπτωση αυτή τα έγγραφα αναπαρίστανται ως διανύσματα λέξεων με αντίστοιχες συχνότητες εμφάνισης.

Μια πολύ ενδιαφέρουσα προσέγγιση παρουσιάζεται στο [KR+99], όπου μια ανάλυση ενός τμήματος του γράφου του ΠΙ οδηγεί στην ανακάλυψη κοινοτήτων εγγράφων που συνδέονται μεταξύ τους με πολύ μεγαλύτερη συχνότητα απ' ότι με τα υπόλοιπα έγγραφα.

2.5 Ερωτήσεις στον Παγκόσμιο Ιστό

Η WebSQL [MM+97] είναι μια δομημένη γλώσσα ερωτήσεων που επιτρέπει την εξαγωγή πληροφορίας από τον Ιστό. Επιτρέπει επίσης την πλοήγηση στον Ιστό είτε με συστηματική επεξεργασία όλων των συνδέσμων μιας σελίδας είτε με τη χρήση μονοπατιών που ικανοποιούν κάποιο πρότυπο, είτε με συνδυασμό των δύο. Η WebSQL θεωρεί τον ιστό ως μια σχεσιακή βάση δεδομένων που αποτελείται από δύο (εικονικές) οντότητες: *Έγγραφο* και *Σύνδεσμο*. Τα χαρακτηριστικά της οντότητας Έγγραφο περιλαμβάνουν: τη διεύθυνση ιστού, τον τίτλο, το κείμενο, το μέγεθος, την ημερομηνία τελευταίας αλλαγής και τον τύπο του εγγράφου, ενώ ο Σύνδεσμος περιλαμβάνει την διεύθυνση του αρχικού εγγράφου, μια ετικέτα και τη διεύθυνση ιστού του τελικού εγγράφου. Με χρήση ενός επεκτάσιμου συνόλου τάξεων, που είναι γραμμένες σε Java, επιτρέπει ερωτήσεις στους δύο εικονικούς πίνακες. Οι ερωτήσεις επιτρέπουν είτε αναζητήσεις (π.χ. βρες το κείμενο όλων των συνδέσμων προς έγγραφα postscript) είτε εργασίες συντήρησης του ΠΙ (π.χ. βρες όλους τους "σπασμένους" συνδέσμους σε ένα έγγραφο, τις φωτογραφίες που λείπουν κτλ.)

Η υπηρεσία "σύνθετης αναζήτησης" του Google [G03] επιτρέπει στους χρήστες να απαιτούν ή να εξαιρούν κάποιους όρους από τον τίτλο των εγγράφων ή από τη διεύθυνσή τους. Επιτρέπει επίσης την αναζήτηση για τα έγγραφα που "δείχνουν" προς ένα συγκεκριμένο έγγραφο ή για έγγραφα όμοια σε περιεχόμενο με ένα συγκεκριμένο έγγραφο. Ο αλγόριθμος σειριακής ταξινόμησης των αποτελεσμάτων, ο PageRank [PB+98], χρησιμοποιείται μόνο για να δώσει υψηλότερη σειρά σε πιο σημαντικές σελίδες αλλά όχι για να ομαδοποιήσει σελίδες με κοινή θεματολογία.

Το σύστημα WebWatcher [AF+95] ανιχνεύει την ύπαρξη συγκεκριμένων λέξεων μέσα στους υπερσυνδέσμους και στη γύρω περιοχή και ταξινομεί τους υπερσυνδέσμους με βάση τη σχετικότητά τους στα ενδιαφέροντα των χρηστών. Το σύνολο των λέξεων κλειδιών παράγεται με τη χρήση ενός συνόλου εγγράφων που χρησιμοποιούνται για την εκπαίδευση του συστήματος και περιορίζεται σε μερικές εκατοντάδες όρων.

Το σύστημα ARC [CD+98a] χρησιμοποιεί το κείμενο γύρω από τους υπερσυνδέσμους για να περιγράψει τα περιεχόμενα των σελίδων στις οποίες αναφέρονται και για να αυξήσει το σημαντικότητα μιας σελίδας ως κόμβου παραπομπών (*hub*) ή κόμβου-αυθεντία (*authority*) σε ένα θέμα [K99]. Πιο συγκεκριμένα, αν ένας σημαντικός κόμβος παραπομπών έχει ένα υπερσύνδεσμο προς μια σελίδα p και χρησιμοποιεί κείμενο σχετικό με κάποιο θέμα, τότε η σελίδα p μπορεί να θεωρηθεί σημαντικός κόμβος-αυθεντία στο θέμα αυτό.

Η βασική υπόθεση σε όλα αυτά τα συστήματα, είναι ότι υπερσύνδεσμοι προς μια συγκεκριμένη σελίδα δρουν ως συστάσεις προς τη σελίδα. Οι συστάσεις αυτές γίνονται από συγγραφείς άλλων σελίδων, μπορούν να θεωρούνται αντικειμενικές και να προσθέτουν στη σημαντικότητα της σελίδας. Σύμφωνα με την Henzinger [Hz01], αυτός είναι ο τρόπος ταξινόμησης που βασίζεται στη συνδεσμολογία (connectivity-based ranking).

2.6 Οντολογίες

Οι οντολογίες αναπτύχθηκαν στο πεδίο της Τεχνητής νοημοσύνης για να διευκολύνουν τη διανομή και την επαναχρησιμοποίηση γνώσης. Αποτελούν ένα δημοφιλές ερευνητικό θέμα για διάφορες ερευνητικές κοινότητες τεχνητής νοημοσύνης, συμπεριλαμβανομένων της επεξεργασίας φυσικής γλώσσας και της διαχείρισης και αναπαράστασης γνώσης.

Πιο πρόσφατα, η έννοια της οντολογίας διαδίδεται σε τομείς όπως η ευφυής ολοκλήρωση πληροφοριών, τα συνεργατικά συστήματα πληροφοριών, η ανάκτηση πληροφοριών, το ηλεκτρονικό εμπόριο, και η διαχείριση γνώσης. Το γεγονός ότι οι οντολογίες γίνονται τόσο δημοφιλείς οφείλεται σε μεγάλο μέρος σε αυτό που υπόσχονται: μια κοινόχρηστη και απλοϊκή κατανόηση κάποιας περιοχής, που μπορεί να αποτελέσει το σημείο επικοινωνίας μεταξύ των ανθρώπων και των συστημάτων εφαρμογών.

Οι οντολογίες αναπτύσσονται για να παρέχουν μια σημασιολογία των πηγών πληροφοριών που μπορούμε να επεξεργαστούμε μηχανικά και να τη χρησιμοποιήσουμε για την επικοινωνία μεταξύ διαφορετικών πρακτόρων (λογισμικό και άνθρωποι). Πολλοί ορισμοί των οντολογιών έχουν δοθεί στην τελευταία δεκαετία, αλλά αυτός που χαρακτηρίζει καλύτερα την ουσία μιας οντολογίας είναι ο εξής [G93]:

Μια οντολογία είναι μια τυπική, ρητή προδιαγραφή μιας κοινής αντίληψης.

- Ως "αντίληψη" αναφέρεται ένα αφηρημένο πρότυπο κάποιου φαινομένου που προσδιορίζει τις σχετικές έννοιες εκείνου του φαινομένου. Ο τύπος των εννοιών που χρησιμοποιούνται και οι περιορισμοί στη χρήση τους καθορίζονται ρητά.
- Ο όρος "τυπική" δηλώνει ότι η οντολογία πρέπει να είναι αναγνώσιμη από μηχανή. Οι μεγάλες οντολογίες όπως το Wordnet [M+93] ή το Sensus [AA+01] παρέχουν έναν θησαυρό για πάνω από 70.000 όρους φυσικής γλώσσας.
- Τέλος, ο όρος "κοινή" απεικονίζει την έννοια ότι μια οντολογία συλλαμβάνει τη συναινετική γνώση, δηλαδή δεν είναι περιορισμένη σε κάποιο άτομο, αλλά αποδεκτή από μια ομάδα.

Σύμφωνα με τον Gruber [G93], η οντολογία είναι ο καθορισμός ενός συνόλου εννοιών:

"An ontology is a specification of a conceptualisation"

Στο πεδίο του καταμερισμού της γνώσης η οντολογία μπορεί να προσδιοριστεί ως η περιγραφή (η τυπική καταγραφή) των εννοιών και των σχέσεων που μπορεί να υπάρχουν ανάμεσά τους, για έναν άνθρωπο ή μια κοινότητα ανθρώπων.

Ο ρόλος των οντολογιών, στη διαδικασία εφαρμοσμένης μηχανικής γνώσης, είναι να διευκολυνθεί η κατάρτιση ενός προτύπου γνώσης σε κάποια θεματική περιοχή. Μια οντολογία παρέχει ένα λεξιλόγιο των όρων και των σχέσεων, όπως διαμορφώνονται στη συγκεκριμένη περιοχή. Επειδή οι οντολογίες στοχεύουν στη συναινετική γνώση περιοχών, η ανάπτυξή τους είναι συχνά μια συνεταιριστική διαδικασία, που εμπλέκει διαφορετικούς ανθρώπους, ενδεχομένως στις διαφορετικές θέσεις. Οι άνθρωποι που συμφωνούν σε μια οντολογία δεσμεύονται στη χρήση της οντολογίας αυτής.

2.6.1 Μια κατηγοριοποίηση των οντολογιών

Η οντολογία μπορεί να προσδιοριστεί ως μια ανθρώπινη προσπάθεια για οργάνωση της υπάρχουσας γνώσης σε ένα συγκεκριμένο γνωστικό πεδίο. Στην οργάνωση αυτή οι έννοιες απεικονίζονται ως κόμβοι (με συγκεκριμένες ιδιότητες) ενός γράφου και οι μεταξύ τους σχέσεις ως κατευθυνόμενες ακμές του γράφου. Παράλληλα, στο γράφο αυτό έρχονται να προστεθούν στιγμιότυπα των εννοιών, κανόνες και αξιώματα που τις συσχετίζουν. Οι σχέσεις ανάμεσα στις έννοιες μπορεί να αναπαριστούν σύσταση (σχέση "περιέχει"), γενίκευση-ειδίκευση (σχέση "είναι") ή οποιαδήποτε άλλη σχέση μπορεί να ορίσει ένας άνθρωπος· σχηματίζουν έτσι ένα κατευθυνόμενο γράφο με ακμές ποικίλης σημασίας. Απλουστεύοντας αυτή τη λογική, οποιαδήποτε οργάνωση εννοιών που συνδέονται μεταξύ τους με ένα τουλάχιστο τύπο σχέσης μπορεί να θεωρηθεί οντολογία.

Ανάλογα με το επίπεδο γενικότητας διακρίνουμε διαφορετικούς τύπους οντολογιών που εκπληρώνουν διαφορετικούς ρόλους στην διαδικασία οικοδόμησης ενός συστήματος γνώσης. Μεταξύ των άλλων, μπορούμε να διακρίνουμε τους ακόλουθους τύπους οντολογίας [SB+98]:

- Οι οντολογίες περιοχών (domain ontologies), που συλλαμβάνουν τη γνώση ενός συγκεκριμένου τομέα ενδιαφέροντος (ηλεκτρονική, ιατρική, κτλ.).
- Οι οντολογίες μετα-δεδομένων (metadata ontologies), όπως αυτή του Dublin Core [H99], που παρέχουν ένα λεξιλόγιο για την περιγραφή του περιεχομένου ορισμένων πηγών πληροφοριών.
- Οι γενικές οντολογίες (generic ontologies) που στοχεύουν στη σύλληψη της γενικής γνώσης για τον κόσμο, στην παροχή των βασικών αντιλήψεων και εννοιών όπως χρόνος, διάστημα, κατάσταση, γεγονός κτλ. Κατά συνέπεια, ισχύουν σε διάφορες θεματικές περιοχές. Παραδείγματος χάριν, μια οντολογία μερωνύμων (σχέσεων part-of / περιέχει) μπορεί να καλύπτει πολλές θεματικές περιοχές και να αναλύει τη σύσταση των εννοιών που εμπλέκονται σε αυτές.
- Οι αντιπροσωπευτικές οντολογίες (representational ontologies) δεν περιορίζονται σε κάποια ιδιαίτερη περιοχή. Ορίζουν κάποιες οντότητες χωρίς να δηλώνουν τι πρέπει να αντιπροσωπευθεί από αυτές. Μια γνωστή αντιπροσωπευτική οντολογία είναι η οντολογία Frame [M75], η οποία καθορίζει γενικές έννοιες όπως πλαίσια, θέσεις, και τιμές που επιτρέπουν την έκφραση της γνώσης με έναν αντικειμενοστραφή τρόπο [LG01]. Για παράδειγμα δύο έννοιες A και B μιας οντολογίας που σχετίζονται με τη σχέση X ($A \xrightarrow{X} B$), μπορούν να απεικονίστουν σε δύο πλαίσια A και B. Στη θέση του πλαισίου A που αντιστοιχεί στη σχέση X, υπάρχει (ως τιμή) ένας δείκτης προς το πλαίσιο B.

- Άλλοι τύποι οντολογιών, όπως οι οντολογίες μεθόδου και στόχου (method and task ontologies). Οι οντολογίες στόχου παρέχουν όρους για συγκεκριμένες εργασίες επίλυσης προβλημάτων (π.χ. η έννοια "υπόθεση" ανήκει στην οντολογία στόχου "διαγνώσεων"), και οι οντολογίες μεθόδου παρέχουν όρους για συγκεκριμένες μεθόδους επίλυσης προβλημάτων (π.χ. η έννοια "σωστή κατάσταση" ανήκει στην οντολογία μεθόδου "προτείνω-και-αναθεωρώ"). Οι οντολογίες στόχου και μεθόδου παρέχουν μια άποψη αιτιολόγησης της γνώσης θεματικών περιοχών.

Μέρος της έρευνας για τις οντολογίες ενδιαφέρεται για την πρόβλεψη και την οικοδόμηση της τεχνολογίας που απαιτείται για μεγάλης κλίμακας επαναχρησιμοποίηση των οντολογιών σε παγκόσμιο επίπεδο. Προκειμένου να επιτραπεί όσο το δυνατόν περισσότερη επαναχρησιμοποίηση, οι οντολογίες πρέπει να είναι μικρές ενότητες με μια υψηλή εσωτερική συνοχή και ένα περιορισμένο ποσό αλληλεπίδρασης μεταξύ των ενοτήτων.

Συχνά, ο καθορισμός μιας θεματικής οντολογίας δεν είναι αυτοσκοπός. Η ανάπτυξη μιας οντολογίας αφορά τον καθορισμό ενός συνόλου στοιχείων και της δομής τους, με τρόπο που να μπορεί να χρησιμοποιηθεί από άλλα προγράμματα. Προγράμματα που μπορεί να χρησιμοποιήσουν τις οντολογίες ως εργαλείο είναι: οι μέθοδοι επίλυσης προβλήματος, οι εφαρμογές που δεν περιορίζονται σε συγκεκριμένο θεματικό πεδίο, οι πράκτορες λογισμικού κτλ.

Πολύ συχνά οι *Οντολογίες* εξισώνονται με ιεραρχίες τάξεων (*Ταξινομίες*) [R01], με μόνη σχέση αυτή της υπο-τάξης. Συχνά επίσης περιορίζονται σε λίστες ορισμών και μετατρέπονται σε *Ορολογίες* [R01] χωρίς επιπλέον γνώση. Και στις δύο περιπτώσεις έχουμε περιορισμένη χρήση των οντολογιών, καθώς δεν έχουμε ορισμούς αξιωμάτων που να περιορίζουν τις ερμηνείες ή τις τιμές των χαρακτηριστικών των οριζόμενων οντοτήτων. Για παράδειγμα μια ταξινομία μπορεί να συνδέει τις έννοιες «περιοδικό» και «άρθρο» με τη σχέση «περιέχει», δεν μπορεί όμως να περιέχει το αξίωμα ότι δύο άρθρα του ίδιου περιοδικού έχουν ίδιο αριθμό τεύχους και έκδοσης.

2.6.2 Οντολογίες και Ιστός

Τα έγγραφα στον ΠΙ χαρακτηρίζονται κυρίως από τις λέξεις που εξάγουμε γι' αυτά και από κάποιο βαθμό σημαντικότητας που λαμβάνει υπόψη τη συνδεσμολογία όλου του ΠΙ [BP98]. Η ομοιότητα ανάμεσα στα έγγραφα, ή ανάμεσα στις ερωτήσεις των χρηστών και τα έγγραφα, βασίζεται στην απόλυτη λεξική ομοιότητα των όρων που αντιστοιχούν σ' αυτά περιορίζοντας έτσι σημαντικά τις αναζητήσεις. Για παράδειγμα, ένα έγγραφο d_1 που χαρακτηρίζεται από τις λέξεις: d_1={φίδι, έρημος} δε θα θεωρηθεί ποτέ σχετικό με ένα έγγραφο d_2 που χαρακτηρίζεται από τις λέξεις: d_2={οχιά, Σαχάρα}. Είναι προφανές ότι οι δύο λίστες (και άρα και τα αντίστοιχα έγγραφα) σχετίζονται, καθώς μια "οχιά" είναι "φίδι" και η "Σαχάρα" είναι "έρημος" και συνεπώς, το έγγραφο d_2 πραγματεύεται τις ίδιες έννοιες με το έγγραφο d_1, απλά είναι πιο εξειδικευμένο. Με αντικατάσταση των λέξεων με έννοιες και μάλιστα σε μια ιεραρχία εννοιών (π.χ. σε μια οντολογία), έχουμε μια πιο ευέλικτη διαδικασία εύρεσης της ομοιότητας από το δυαδικό ταίριασμα (binary matching), που διαχειρίζεται τις γενικεύσεις και εξειδικεύσεις των εννοιών.

Το μέλλον του Παγκόσμιου Ιστού είναι ο Σημασιολογικός Ιστός [SW03] [W3i], στον οποίο η διαχείριση των περιεχομένων είναι πιο εύκολη, καθώς η χρήση ετικετών XML επιτρέπει το χαρακτηρισμό των περιεχομένων, την προσθήκη σημασιολογίας στα έγγραφα, αλλά και επιπλέον πληροφορίας για τη δομή των εγγράφων. Παρ' όλα αυτά, τα περισσότερα έγγραφα του ΠΙ δεν υποστηρίζουν ακόμη τη σήμανση του Σημασιολογικού Ιστού και δεν περιέχουν σημασιολογική πληροφορία. Καθώς η διαδικασία μετάβασης των παλιών εγγράφων στη νέα μορφή δεν είναι εύκολο να γίνει χειροκίνητα, λόγω του μεγάλου αριθμού εγγράφων του ΠΙ, απαιτούνται μηχανισμοί που θα εξάγουν αυτόματα σημασιολογική πληροφορία από τα έγγραφα και θα την προσθέσουν σε αυτά.

Οι μηχανισμοί επεξεργασίας των εγγράφων αναλαμβάνουν την εξαγωγή λέξεων από τα περιεχόμενα των εγγράφων, τους εισερχόμενους συνδέσμους προς αυτά κτλ. Η αντιστοίχιση των λέξεων, που εξάγονται από τους συνδέσμους, σε έννοιες μιας οντολογίας δεν αποσκοπεί να προσδιορίσει την αντικειμενική σημασία των συνδέσμων. Το κυριότερο όφελος έγκειται στο γεγονός ότι οι λέξεις που εξάγονται από τους συνδέσμους, και που δεν μας ενδιαφέρουν στο σύνολό τους, αντιστοιχίζονται σε ένα στενότερο σύνολο από έννοιες, οργανωμένες σε μια δομή που απεικονίζει τον τομέα ενδιαφέροντός μας (οντολογία). Οι ερωτήσεις των χρηστών με όμοιο τρόπο αντιστοιχίζονται σε έννοιες της οντολογίας. Με τον τρόπο αυτό μειώνονται οι διαστάσεις του προβλήματος εύρεσης εγγράφων που ικανοποιούν τις ερωτήσεις των χρηστών και επιπλέον η απαίτηση για ακριβή λεξική ομοιότητα μετατρέπεται σε προσεγγιστική ομοιότητα εννοιών. Οι έλεγχοι σε επίπεδο συμβολοσειρών, που έπρεπε να συγκρίνουν εξαντλητικά τις λέξεις του ερωτήματος με τις λέξεις του κάθε εγγράφου, μετατρέπονται σε αναζητήσεις για την πλησιέστερη έννοια στο γράφο της οντολογίας.

Η αποσαφήνιση της έννοιας των λέξεων (word sense disambiguation - [IV98], [Y00]) που εξάγονται από ένα έγγραφο είναι ένα πολύ σημαντικό και δύσκολο πρόβλημα. Οι αλγόριθμοι που ασχολούνται με την επίλυση τέτοιων προβλημάτων θα μπορούσαν να διαχωριστούν: α) σε αυτούς που απαιτούν κάποιο σύνολο προ-χαρακτηρισμένων εγγράφων για προπαίδευση και β) σε αυτούς που βασίζονται σε λεξικογραφικές πηγές (π.χ. λεξικά, θησαυρούς κτλ.) και αλγορίθμους που επεξεργάζονται εξ αρχής κάποιο σύνολο εγγράφων ενός εξειδικευμένου τομέα. Τα έγγραφα αυτά χρησιμοποιούν συγκεκριμένη ορολογία και οι λέξεις που εμφανίζονται σε αυτά έχουν περιορισμένο σημασιολογικό εύρος.

Στην πρώτη κατηγορία, της επιβλεπόμενης αποσαφήνισης (supervised disambiguation), ανήκει η δουλειά των Brown κ.ά. [BP+91] που έχουν χρησιμοποιήσει τον αλγόριθμο Flip-Flop για να αποσαφηνίσουν τις έννοιες γαλλικών λέξεων. Στην περίπτωση αυτή αναλύονται τα έγγραφα μιας συλλογής ως προς τη δομή και τη σειρά εμφάνισης των λέξεων σε αυτά και εντοπίζονται τα χαρακτηριστικά εκείνα που καθορίζουν τη σημασία των αμφισβητούμενων λέξεων (indicators). Για παράδειγμα η λέξη ”per” ή η ύπαρξη ενός αριθμού πριν τη λέξη "cent" καθορίζουν τη σημασία που έχει η λέξη (αν είναι ποσοστό ή αριθμός). Στην ίδια κατηγορία της επιβλεπόμενης αποσαφήνισης ανήκει και η προσπάθεια των Gale κ.ά [GC+93] να χρησιμοποιήσουν την προσέγγιση Bayes για να αποσαφηνίσουν έννοιες λέξεων. Αντί να προσδιορίσουν ένα συγκεκριμένο χαρακτηριστικό που οδηγεί στην αποσαφήνιση μιας λέξης, θεωρούν ότι πολλά χαρακτηριστικά

επηρεάζουν τη σημασία μιας λέξης, καθένα σε διαφορετικό βαθμό, ανάλογα με τη συχνότητα εμφάνισής του στο περιβάλλον της λέξης.

Στα πλεονεκτήματα αυτών των αλγορίθμων συγκαταλέγονται τα μεγάλα ποσοστά επιτυχίας, που φτάνουν ακόμη και το 90% σε ορισμένες περιπτώσεις για την Bayesian αποσαφήνιση [GC+93]. Οι αλγόριθμοι αυτοί απαιτούν ένα σύνολο χαρακτηρισμένων εγγράφων (πληροφοριακό περιεχόμενο) για να λειτουργήσουν και επιπλέον δεν εκμεταλλεύονται τα ιεραρχικό δέντρα κατηγοριοποίησης εννοιών όπως αυτό του Wordnet.

Στην δεύτερη κατηγορία, ο αλγόριθμος που έχει προταθεί από τον Lesk [L86] αποσαφηνίζει τις έννοιες κάνοντας χρήση λεξικού. Εντοπίζει δηλαδή την πιθανότερη έννοια με τη βοήθεια του ερμηνευτικού ορισμού της σημασίας κάθε έννοιας, όπως αυτή παρέχεται από κάποιο λεξικό ή άλλη πηγή. Βασίζεται στην απλή ιδέα ότι τα στοιχεία του ορισμού μιας λέξης είναι πιθανόν καλή ένδειξη της χρησιμοποιούμενης σημασίας της λέξης. Παρ' όλα αυτά, δεν επιτυγχάνει αποτελέσματα μεγαλύτερης ακρίβειας από 50-70 %, και γι’ αυτό προτείνονται διάφορες τεχνικές βελτιώσεις.

Στα πλαίσια τέτοιων βελτιώσεων, πολύ σημαντική είναι η δουλειά του Yarowsky [Y94], [Y95] που εστιάζει σε δύο περιορισμούς:

- Μια έννοια ανά περιεχόμενο, σύμφωνα με τον οποίο η σημασία μιας λέξης δεν αλλάζει ιδιαίτερα στα όρια ενός λήμματος, ή ενός εγγράφου.
- Μια έννοια ανά τοποθέτηση, σύμφωνα με τον οποίο η διάταξη (απόσταση, σειρά και συντακτική σχέση) συγκεκριμένων λέξεων γύρω από μια λέξη, παρέχει ισχυρά στοιχεία για την σημασία της λέξης.

Ουσιαστικά, ο Yarowsky εντοπίζει, για καθεμία από τις διαφορετικές σημασίες μιας λέξης, τις λέξεις που εμφανίζονται με μεγάλη συχνότητα στο αντίστοιχο λήμμα ενός λεξικού. Οι λέξεις αυτές μπορούν να χρησιμοποιηθούν για την αποσαφήνιση της έννοιας της λέξης όταν βρεθούν στο περιβάλλον της. Έτσι, περιορίζει σημαντικά το σημασιολογικό εύρος κάθε λέξης. Προϋποθέτει φυσικά ότι το λεξικό διαθέτει τη δυνατότητα διάκρισης των λημμάτων για τις διαφορετικές σημασίες της λέξης.

Αν για παράδειγμα, για τη λέξη wind ένα ερμηνευτικό λεξικό μας δίνει τις ακόλουθες δύο ερμηνίες (απόσπασμα από το Wordnet):

- *a musical instrument in which the sound is produced by an enclosed column of air that is moved by the breath.*
- *air moving from an area of high pressure to an area of low pressure.*

τότε οι λέξεις *musical*, *instrument*, *sound* κλπ. αποτελούν διακριτικό της πρώτης σημασίας και αντίστοιχα οι λέξεις *moving*, *area*, *pressure* αποτελούν διακριτικό της δεύτερης σημασίας. Αντίθετα η λέξη *air* δεν αποτελεί διακριτικό καθώς υπάρχει και στις δύο ερμηνίες.

Ο μη-επιβλεπόμενος αλγόριθμος του Schütze [S92] χρησιμοποιεί τεχνικές συσταδοποίησης στα συμφραζόμενα μιας πολύσημης λέξης. Οι συμφραζόμενες λέξεις αναπαρίστανται με διανύσματα και συσταδοποιούνται με τη βοήθεια ενός μέτρου ομοιότητας διανυσμάτων και ενός ιεραρχικού αλγορίθμου. αναπαριστά με διανύσματα τις λέξεις που εντοπίζει συστάδες συμφραζόμενες λέξεων.

Η χρήση αλγορίθμων που δεν εκμεταλλεύονται το Wordnet, όπως του Yarowsky, με υποσχόμενα αποτελέσματα 91 - 96 %, και οι μη-επιβλεπόμενοι αλγόριθμοι, όπως του

Schütze, με υποσχόμενα αποτελέσματα 85 %, στη διαδικασία αποσαφήνισης πρέπει να εξετασθεί. Η πολυπλοκότητα των αλγορίθμων είναι σημαντική και επιπλέον σε πολλές περιπτώσεις επιβάλλεται η χρήση πληροφοριακού υλικού (συλλογές εγγράφων, θησαυροί) που δεν είναι πάντοτε διαθέσιμο.

2.6.3 Το Wordnet

Το Wordnet είναι ένα λεξικολογικό σύστημα αναφοράς, στο οποίο τα αγγλικά ουσιαστικά, τα ρήματα, τα επίθετα και τα επιρρήματα οργανώνονται σε ομάδες συνωνύμων. Κάθε ομάδα αντιπροσωπεύει μια έννοια στο λεξικό. Διαφορετικές σχέσεις συνδέουν τις έννοιες μιας ομάδας συνωνύμων, όπως η γενικό-ειδικό για τα ρήματα και τα ουσιαστικά, το μέρος-σύνολο για τα ουσιαστικά κτλ. Κάθε λέξη στο Wordnet ανήκει σε ένα ή περισσότερα ***synsets*** (synonym set). ***Κάθε synset αντιστοιχεί σε μια έννοια*** της λέξης ως ρήμα, ουσιαστικό, επίθετο ή επίρρημα, ενώ συνοδεύεται από μια σύντομη περιγραφή και ένα σύνολο από συνώνυμες λέξεις. Η βασική οντότητα στο Wordnet είναι η ***έννοια*** και όχι η *λέξη*. Η *έννοια* παίζει και το ρόλο του κόμβου στον κατευθυνόμενο γράφο του Wordnet. Ακμές αυτού του γράφου είναι οι σχέσεις μεταξύ των εννοιών (υπερώνυμο, υπώνυμο, συνώνυμο, τροπώνυμο κτλ.). Ο σύνθετος αυτός γράφος μπορεί να αντιμετωπιστεί πιο απλοϊκά, καθώς οι έννοιες ρημάτων και ουσιαστικών οργανώνονται ως θεματικές ιεραρχίες που διαμορφώνουν ένα "δάσος" 25 δέντρων (με σχέση γενικό-ειδικό) των ουσιαστικών και 15 των ρημάτων. Για τα επίθετα και τα επιρρήματα το Wordnet παρέχει μόνο συνώνυμα. Αντίστοιχες ιεραρχίες ορίζονται και από τη σχέση «*περιέχει*» μεταξύ ουσιαστικών.

Το πλεονέκτημα του Wordnet, συγκριτικά με άλλους θησαυρούς, είναι ότι οι όροι που περιέχει ανήκουν σε κάποια ιεραρχία και δεν αποτελούν μόνο μέλη μιας ομάδας συνωνύμων όρων. Η απλουστευμένη αντιμετώπιση του γράφου, ως συνόλου από ιεραρχίες, και η χρήση των ιεραρχιών, αντί των συνόλων συνωνύμων, επιτρέπει τον καθορισμό νέων πρακτικών: για τον υπολογισμό της ομοιότητας δύο όρων, δύο συνόλων όρων, για τη διαχείριση ερωτήσεων κτλ.

Η ιεραρχική διάταξη των εννοιών στο Wordnet μας επιτρέπει να καθορίσουμε ένα διαφορετικό μέτρο της ομοιότητας μεταξύ δύο όρων που δεν είναι συνώνυμοι αλλά βρίσκονται στην ίδια ιεραρχία. Ποικίλα μέτρα ομοιότητας μπορούν να υιοθετηθούν για τον καθορισμό της ομοιότητας σε μια ιεραρχία [R99], [L98], [WP94].

Συνοψίζοντας, μπορούμε να διακρίνουμε κάποιους λόγους για τους οποίους όσοι ασχολούνται με τη Σημασιολογική Οργάνωση της γνώσης ισχυρίζονται ότι το Wordnet δεν είναι οντολογία και να αντιπαραθέσουμε τους λόγους για τους οποίους θεωρούμε ότι το Wordnet μπορεί να θεωρηθεί οντολογία και να χρησιμοποιηθεί ως τέτοια.

Γιατί το Wordnet δεν είναι οντολογία:

- ***δεν είναι περιορισμένο σε ένα γνωστικό πεδίο***, όπως συμβαίνει με τις περισσότερες υπάρχουσες οντολογίες. Το μεγάλο εύρος γνώσης που καλύπτει το καθιστά αδύναμο σε σύγκριση με πολύ μικρότερες οντολογίες που αναπτύσσονται για κάποιο συγκεκριμένο πεδίο.
- ***δεν ορίζει στιγμιότυπα***. Βασικό στοιχείο του Wordnet είναι οι έννοιες των διαφόρων λέξεων που περιέχει. Έννοιες όπως "συνθέτης" και στιγμιότυπα όπως "Mozart" και "Beethoven" θεωρούνται ισοδύναμα, προκαλώντας σύγχυση.

- ***δεν ορίζει αξιώματα***. Η λογική των αξιωμάτων, ως περιορισμών που τίθενται στη δομή της οντολογίας, δεν υποστηρίζεται στο Wordnet.
- ***χωρίζει τους όρους που περιέχει σε 4 διακριτές κατηγορίες*** (4 μέρη του λόγου: ουσιαστικά, ρήματα, επίθετα, επιρρήματα) κάτι που δεν είναι απαραίτητο σε μια οντολογία.

Γιατί το Wordnet μπορεί να χρησιμοποιηθεί ως οντολογία

- ***ορίζει έννοιες και σχέσεις μεταξύ τους***. Με τον τρόπο αυτό ορίζει ένα γράφο μεταξύ των εννοιών που μας επιτρέπει να εντοπίσουμε "κοντινές" και "μακρινές" έννοιες.
- ***δεν περιορίζεται σε ένα τύπο σχέσης***. Σε αντίθεση με τις ταξινομίες δεν περιορίζεται σε ένα μόνο τύπο σχέσης μεταξύ των εννοιών (συνώνυμα, γενίκευση-εξειδίκευση), αλλά ορίζει διαφορετικούς τύπους σχέσεων δημιουργώντας έτσι ένα "πολύχρωμο" γράφο εννοιών. Σε πολλές περιπτώσεις ο γράφος συνδέει διαφορετικά μέρη του λόγου, όπως για παράδειγμα η σχέση «σχετίζεται» (pertains to) που συνδέει ένα επίθετο με ένα ουσιαστικό.

2.7 Σημασιολογικός Ιστός - Semantic Web

Ο Σημασιολογικός Ιστός είναι ένα δίκτυο πληροφοριών που συνδέεται με τρόπο που είναι εύκολα επεξεργάσιμος από τις μηχανές, σε παγκόσμια κλίμακα. Είναι ένας αποδοτικός τρόπος αναπαράστασης δεδομένων του ΠΙ, με τη μορφή μιας καθολικά συνδεμένης βάσης δεδομένων.

Ο Σημασιολογικός Ιστός βασίζεται σε μια ιδέα του Tim Berners-Lee [BH+01], εμπνευστή του ΠΙ, των διευθύνσεων URI, του πρωτοκόλλου HTTP, και της γλώσσας HTML. Έχουν ήδη διαμορφωθεί πολλές γλώσσες, και εργαλεία σχετικά με το Σημασιολογικό Ιστό. Εντούτοις, οι τεχνολογίες είναι ακόμα πρώιμες και υπάρχει αρκετός προβληματισμός για τη μετάβαση προς το Σημασιολογικό Ιστό αλλά και για τα επόμενα βήματα σε αυτόν.

Το πρόβλημα με την πλειοψηφία των δεδομένων του ΠΙ είναι ότι στην παρούσα μορφή τους δεν μπορούν να χρησιμοποιηθούν σε μια μεγάλη κλίμακα, καθώς δεν υπάρχει ένα ενιαίο πλαίσιο επεξεργασίας τους. Για παράδειγμα, ο ΠΙ περιέχει τις πληροφορίες για τοπικές αθλητικές εκδηλώσεις, καιρικές πληροφορίες, δρομολόγια αεροπλάνων, τηλεοπτικούς οδηγούς κτλ. Όλες αυτές οι πληροφορίες παρουσιάζονται σε πολυάριθμους δικτυακούς τόπους, σε διάφορες γλώσσες, με διαφορετικές ορολογίες. Για να μπορέσουμε να χειριστούμε την πληροφορία αυτή με μηχανικό τρόπο, πρέπει να την ομογενοποιήσουμε σε δομή και συντακτικό.

Ο Σημασιολογικός Ιστός στηρίζεται γενικά σε ένα συντακτικό που αντιστοιχεί τα δεδομένα σε μοναδικούς ενιαίους προσδιοριστές πόρων (URIs - Uniform Resource Identifier). Τα έγγραφα του Σημασιολογικού Ιστού ακολουθούν το πρότυπο RDF (Resource Description Framework [W3c]). Έτσι κάθε έγγραφο αποτελείται από τριάδες δηλώσεων RDF. Κάθε τριάδα περιέχει ένα θέμα (υποκείμενο), ένα αντικείμενο και ένα κατηγόρημα που συνδέει το θέμα με το αντικείμενο. Και τα τρία στοιχεία της τριάδας είναι μοναδικά URIs. Το παράδειγμα:

<http://www.db-net.aueb.gr/hercules>
<http://love.example.org/terms/reallyLikes>
<http://www.w3.org/People/Berners-Lee/Weaving/>

δηλώνει ότι: στο αντικείμενο Ηρακλής (“Hercules”) αρέσει (“really likes”) το βιβλίο "Weaving the Web" του Tim Berners-Lee.

Το RDF είναι ιδανικό πρότυπο για τη δημοσίευση των περιεχομένων των βάσεων δεδομένων στον Ιστό. Οτιδήποτε τοποθετεί κανείς στον ιστό έχει το δικό του URI, που δεν αντιστοιχεί απαραίτητα σε μια υπαρκτή διεύθυνση ιστού, έτσι ώστε να μπορεί οποιοσδήποτε θέλει να αναφερθεί σε αυτό. Επίσης η λογική των τριάδων “*υποκείμενο – σχέση – αντικείμενο*” επιτρέπει να προσθέσουμε σημασιολογικά χαρακτηριστικά στην πληροφορία, τα οποία στη συνέχεια μπορούν να επεξεργαστούν άλλα προγράμματα.

Εφόσον οι τριάδες του RDF μπορεί να περιέχουν οποιαδήποτε αντικείμενα κι υποκείμενα και να ορίζουν οποιεσδήποτε σχέσεις, καταλαβαίνει κανείς ότι είναι σχεδόν αδύνατο για μια μηχανή να κατανοήσει τη σημασία της κάθε τριάδας και ακόμη πιο δύσκολο να ομογενοποιήσει την πληροφορία του Σημασιολογικού Ιστού. Για το λόγο αυτό απαιτείται ένα ***σχήμα*** που θα περιγράφει τη δομή των τριάδων σε ένα έγγραφο, και μια ***οντολογία*** που θα καθορίζει τα επιτρεπόμενα αντικείμενα, υποκείμενα και τις μεταξύ τους σχέσεις. Δύο συστήματα που αναπτύχθηκαν για το σκοπό αυτό είναι τα RDF Schemas [W3j] και η γλώσσα περιγραφής οντολογιών DAML+OIL [W3k] (DARPA Agent Markup Language with Ontology Inference Layer). Ο οργανισμός W3C έχει ξεκινήσει τη λειτουργία του Web Ontology Working Group[W3l], μιας ομάδας που σκοπό έχει να καθορίσει μια ενιαία γλώσσα καθορισμού οντολογιών.

2.8 Ομαδοποίηση εγγράφων – Μέτρα ομοιότητας/απόστασης

Οι δύο συνηθέστερες τεχνικές εξόρυξης γνώσης που χρησιμοποιούνται για την ομαδοποίηση εγγράφων του ιστού είναι η *Κατηγοριοποίηση* και η *Συσταδοποίηση*.

Η *Κατηγοριοποίηση* προϋποθέτει μια συγκεκριμένη κατηγοριοποίηση των εγγράφων και μια προπαίδευση του συστήματος με ένα σύνολο εγγράφων [DC00], [KS97]. Κατά τη διάρκεια της προπαίδευσης, τα έγγραφα τοποθετούνται στις υπάρχουσες κατηγορίες. Στη συνέχεια επεξεργάζονται ώστε να παραχθούν οι κανόνες (classifiers – κατηγοριοποιητές) με τους οποίους θα αποφασίζεται η κατηγοριοποίηση των νέων εγγράφων . Κατά τη *Συσταδοποίηση* θεωρούμε ότι δεν υπάρχει πρότερη γνώση για τις κατηγορίες στις οποίες εμπίπτουν τα έγγραφα και συνεπώς και για τους κανόνες ομαδοποίησης. Καθώς δεν υπάρχει διαδικασία προπαίδευσης, η Συσταδοποίηση περιλαμβάνεται στις μη-επιβλεπόμενες τεχνικές ομαδοποίησης. Στην περίπτωση των εγγράφων του ΠΙ, δεν μπορούμε να θεωρήσουμε γνωστό εκ των προτέρων το περιεχόμενο των εγγράφων και κατ’ επέκταση τις κατηγορίες στις οποίες ανήκουν τα έγγραφα. Για το λόγο αυτό προτιμάται η τεχνική της Συσταδοποίησης.

Η Συσταδοποίηση είναι μια από τις βασικότερες τεχνικές ανάλυσης δεδομένων που σκοπό έχει τη οργάνωση μιας συλλογής αντικειμένων σε συμπαγείς ομάδες που ονομάζονται συστάδες. Κάθε συστάδα περιέχει αντικείμενα που είναι πολύ όμοια μεταξύ τους και διαφέρουν σημαντικά από τα αντικείμενα των υπολοίπων συστάδων [R92]. Η Συσταδοποίηση είναι μια μορφή μη-επιβλεπόμενης κατηγοριοποίησης, που δεν προϋποθέτει οι ομάδες στις οποίες χωρίζονται τα αντικείμενα να είναι εκ των προτέρων γνωστές.

Η τεχνική της συσταδοποίησης στον ΠΙ έχει χρησιμοποιηθεί σε συστήματα, για την παρουσίαση μιας συλλογής εγγράφων [CK+92], για την οργάνωση των αποτελεσμάτων μιας μηχανής αναζήτησης σε μια ερώτηση [ZE+97], καθώς και για την αυτόματη δημιουργία ιεραρχικών συστάδων εγγράφων [KS97]. Η αυτόματη δημιουργία μιας ταξινομίας εγγράφων του ιστού, σαν αυτή που παρέχουν οι δικτυακοί κατάλογοι, είναι ένας από τους στόχους παρόμοιων συστημάτων. Μια διαφορετική προσέγγιση, που συνδυάζει συσταδοποίηση και κατηγοριοποίηση, ξεκινά ανακαλύπτοντας συστάδες σε μια υπάρχουσα κατηγοριοποίηση, όπως αυτή του Yahoo, και στη συνέχεια τις χρησιμοποιεί για να εξάγει κανόνες κατηγοριοποίησης για νέα έγγραφα [AG+99].

Για να χωρίσουμε μια συλλογή εγγράφων σε συστάδες πρέπει:

- να προσδιορίσουμε τα χαρακτηριστικά των εγγράφων (π.χ. λέξεις, φράσεις, σύνδεσμοι, αν πρόκειται για έγγραφα του ιστού) και
- να αναπαραστήσουμε τα έγγραφα αυτά σε κάποιο μοντέλο. Το πιο συνηθισμένο μοντέλο αναπαράστασης εγγράφων είναι το Μοντέλο του Χώρου Διανυσμάτων (Vector Space Model) [SW+75]. Κάθε έγγραφο αναπαρίσταται ως διάνυσμα χαρακτηριστικών με μέγεθος όσο το πλήθος των διαφορετικών χαρακτηριστικών των εγγράφων. Κάθε στοιχείο του διανύσματος έχει ένα βάρος που δηλώνει τη σημαντικότητά του στο χαρακτηρισμό του εγγράφου. Στην περίπτωση μιας συλλογής εγγράφων το βάρος συνήθως παίρνει διακριτές τιμές 0 ή 1, δηλώνοντας αν το αντικείμενο διαθέτει ή όχι το συγκεκριμένο χαρακτηριστικό ή ενδιάμεσες τιμές που δηλώνουν τη σημαντικότητα του χαρακτηριστικού. Για παράδειγμα όταν το χαρακτηριστικό είναι μια λέξη που εξάγεται από τα περιεχόμενα του εγγράφου το βάρος αντιπροσωπεύει τη συχνότητα της λέξης στο έγγραφο ή την σπανιότητα της λέξης σε ολόκληρη τη συλλογή,
- πρέπει επίσης να καθορίσουμε ένα μέτρο για τον υπολογισμό της ομοιότητας μεταξύ δύο εγγράφων, ή δύο συστάδων. Συχνά χρησιμοποιούμενα μέτρα είναι τα μέτρα Συνημιτόνου, Jaccard και Dice [R79], [W88] και [SJ+00].
- τέλος, πρέπει να επιλεγεί ο κατάλληλος αλγόριθμος συσταδοποίησης που θα ενσωματώνει το μέτρο ομοιότητας και θα λαμβάνει υπόψη του τα χαρακτηριστικά των αντικειμένων.

2.8.1 Γενικές ιδέες σχετικά με τα μέτρα ομοιότητας

Για να ομαδοποιήσουμε σημεία στο χώρο αρκεί να γνωρίζουμε την ακριβή θέση των σημείων και χρησιμοποιώντας ένα μέτρο απόστασης (π.χ. ευκλείδεια απόσταση) να υπολογίσουμε τις αποστάσεις μεταξύ όλων των σημείων. Γνωρίζοντας τις αποστάσεις μεταξύ των σημείων και χρησιμοποιώντας έναν αλγόριθμο συσταδοποίησης που βασίζεται στην πυκνότητα των σημείων μπορούμε να εντοπίσουμε συστάδες σημείων τα οποία βρίσκονται σε μικρή απόσταση μεταξύ τους (μεγάλη εσωτερική πυκνότητα) και σε μεγάλη απόσταση από τα υπόλοιπα (μικρή εξωτερική πυκνότητα). Στην περίπτωση των εγγράφων του ΠΙ δεν έχουμε πλέον ένα χώρο με άξονες, αρχή και συντεταγμένες, και συνεπώς η χρήση της ευκλείδειας απόστασης δεν είναι δυνατή. Κάθε έγγραφο περιγράφεται από ένα σύνολο λέξεων ή εννοιών μιας οντολογίας οι οποίες μπορεί να συνοδεύονται και από αντίστοιχους βαθμούς σημαντικότητας (βάρη). Χρειαζόμαστε λοιπόν ένα νέο μέτρο που θα υπολογίζει την απόσταση ή ομοιότητα δύο εγγράφων με βάση τις περιγραφές αυτές και τον κατάλληλο αλγόριθμο για τον εντοπισμό συμπαγών και διακριτών συστάδων εγγράφων.

Εκτός από το μοντέλο του Χώρου Διανυσμάτων που προαναφέρθηκε, υπάρχει και το μοντέλο των Ευρετηρίων Λανθάνουσας Σημασιολογίας (Latent Semantic Indexing) [D94],[BC+92],[DD+90] που αντικαθιστά το διάνυσμα λέξεων για ένα έγγραφο με ένα διάνυσμα σύνθετων εννοιών. Οι σύνθετες αυτές έννοιες ανήκουν σε ένα πεπερασμένο σύνολο εννοιών (ευρετήριο), που έχει προκύψει από το αρχικό σύνολο λέξεων όλων των εγγράφων μιας συλλογής έπειτα από συγχώνευση λέξεων που είναι σχετικές μεταξύ τους (συνώνυμες, συν-εμφανίζονται συχνά κτλ.). Οι τεχνικές που χρησιμοποιούν αυτό το μοντέλο αναπαράστασης βασίζονται σε ένα αντιπροσωπευτικό δείγμα του συνόλου εγγράφων για να ορίσουν το ευρετήριο, κάτι που δεν είναι διαθέσιμο στην περίπτωση των εγγράφων του ΠΙ. Επιπλέον, η ομοιότητα των εγγράφων βασίζεται και πάλι στην ακριβή ομοιότητα (εσωτερικό γινόμενο, ή μέτρο συνημιτόνου) των όρων των διανυσμάτων που αντιστοιχούν στα έγγραφα.

2.8.2 Υπάρχοντα μέτρα ομοιότητας μεταξύ των συνόλων

Διάφορα μέτρα για τον υπολογισμό της ομοιότητας/απόστασης μεταξύ συνόλων στοιχείων υπάρχουν ήδη στη βιβλιογραφία [EM97], [GH+96]. Δύο από τα πιο απλά μέτρα που έχουν χρησιμοποιηθεί ευρέως είναι τα μέτρα ***Jaccard*** και ***Συνημιτόνου***.

Αν Α και Β δύο σύνολα στοιχείων η ομοιότητά τους σύμφωνα με το μέτρο Συνημιτόνου ορίζεται ως εξής:

$$S_C(A,B) = \frac{|A \cap B|}{\sqrt{|A|}\sqrt{|B|}}$$ ομοιότητα Συνημιτόνου Εξ.1

Κατ' αντιστοιχία η ομοιότητα Jaccard ορίζεται ως το πλήθος των κοινών τους στοιχείων προς το πλήθος των κοινών και μη κοινών στοιχείων τους:

$$S_J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$ ομοιότητα Jaccard Εξ.2

Στο [EM97] εξετάζεται το πρόβλημα υπολογισμού της ομοιότητας ή της απόστασης μεταξύ δύο πεπερασμένων συνόλων σημείων σε ένα μετρικό διάστημα. Εξετάζονται μερικές από τις μεθόδους υπολογισμού της απόστασης, όπως η ελάχιστη απόσταση σύνδεσης (minimum distance link). Όλοι οι αλγόριθμοι είναι πολυωνυμικού χρόνου. Στο [GH+96] μελετάται η ιδέα του υπολογισμού των ομοιοτήτων μεταξύ των συνόλων, με χρήση της συντεταγμένης Jaccard. Το θέμα της δημιουργίας ευρετηρίου των "γειτονικών" εγγράφων ως προς την απόσταση αντιμετωπίζεται στο [GG+01], που επίσης χρησιμοποιείται η συντεταγμένη Jaccard στον υπολογισμό της ομοιότητας. Οι Bidault κ.ά. [BFF02] κάνουν χρήση ευρετηρίων για τον εντοπισμό των πλησιέστερων εγγράφων σε κάποιο έγγραφο.

Το παραδοσιακό μέτρο συνημιτόνου (cosine measure) προέρχεται από το πεδίο ανάκτησης πληροφοριών (βλ. [SM83]) και έχει παρόμοια συμπεριφορά με τη συντεταγμένη Jaccard.

Στην περίπτωση που στη θέση των σημείων έχουμε έγγραφα που χαρακτηρίζονται από λέξεις, τα έγγραφα μπορούν να αναπαρασταθούν ως διανύσματα σε ένα χώρο N-διαστάσεων (όσες και οι διαφορετικές λέξεις). Στην περίπτωση αυτή οι λέξεις θεωρούνται ανεξάρτητες μεταξύ τους και γι' αυτό το μέτρο Jaccard (ή το μέτρο

συνημιτόνου) λαμβάνει υπόψη του μόνο την απόλυτη ομοιότητα μεταξύ των λέξεων. Αν οι λέξεις σχετίζονται μεταξύ τους το πρόβλημα διαφοροποιείται και τα μέτρα αυτά δεν επαρκούν.

Στην περίπτωση που ένα έγγραφο χαρακτηρίζεται από τους όρους της οντολογίας, η ομοιότητα μεταξύ δύο συνόλων με τέτοιους όρους μπορεί να υπολογιστεί κάνοντας χρήση της θέσης των όρων στην οντολογία. Αυτό έχει ως αποτέλεσμα, έγγραφα που χαρακτηρίζονται από διαφορετικά σύνολα λέξεων να εμφανίζουν ένα σημαντικό βαθμό ομοιότητας. Παραδείγματος χάριν, η ομοιότητα για δύο έγγραφα που χαρακτηρίζονται από τα ζεύγη όρων {"γάτα", "τρόφιμα"} και {"αιλουροειδές", "διατροφή"} θα είχε τιμή 0 σύμφωνα με το συντελεστή Jaccard, ενώ θα ήταν μεγαλύτερη από 0 αν λαμβάναμε υπόψη την ομοιότητα των όρων που μπορεί να υπολογιστεί από την απόσταση των όρων σε μια ιεραρχία, όπως το Wordnet ή μια οποιαδήποτε άλλη ταξινομία.

2.8.3 Ομοιότητα μεταξύ δύο στοιχείων μιας οντολογίας

Προκειμένου να υπολογιστεί η απόσταση μεταξύ δύο συνόλων όρων μιας οντολογίας, πρέπει πρώτα να υπολογιστεί η ομοιότητα στην απλούστερη περίπτωση μεταξύ δύο δεδομένων όρων της οντολογίας. Στα [RS+94], [LC+98], [JC97] και [R95], [R99] προτείνονται διαφορετικά μέτρα υπολογισμού της ομοιότητας σε μια ταξινομία όπως το Wordnet, ενώ στο [L98] υπάρχει μια σύγκριση μεταξύ αυτών των μέτρων και άλλων, όπως το μέτρο Wu και Palmer [WP94] και αυτό των Miller και Charles [MC91]. Στο [DJ02] επίσης προτείνεται η χρήση του μέτρου Wu και Palmer για τον υπολογισμό της ομοιότητας μεταξύ όρων μιας οντολογίας.

Από τα προτεινόμενα μέτρα, αυτά των Jiang-Conrath [JC97], Resnik [R95], και Lin [L98], υπολογίζουν την ομοιότητα ανάμεσα σε δύο έννοιες με βάση: α) τον ελάχιστο κοινό πρόγονο των εννοιών στην ιεραρχία του Wordnet και β) το πληροφοριακό περιεχόμενο της κάθε έννοιας. Ως πληροφοριακό περιεχόμενο ορίζεται η συχνότητα με την οποία εμφανίζεται η έννοια της ιεραρχίας μέσα σε μια προκαθορισμένη συλλογή εγγράφων. Για το λόγο αυτό τα συγκεκριμένα μέτρα απαιτούν προπαίδευση του συστήματος με έτοιμες συλλογές εγγράφων και το αποτέλεσμά τους εξαρτάται από την ποιότητα και πληρότητα της αρχικής συλλογής.

Αντίθετα, τα μέτρα Leacock-Chodorow [LC+98] και Wu-Palmer υπολογίζουν την ομοιότητα δύο εννοιών ως το ελάχιστο μονοπάτι ανάμεσα στα δέντρα των υπωνύμων των εννοιών. Εξάλλου, διευκρινίζεται ότι ενώ τα μέτρα των Resnik, Lin και Leacock-Chodorow υπολογίζουν τη σημασιολογική ομοιότητα δύο εννοιών, το μέτρο Jiang-Conrath υπολογίζει τη σημασιολογική απόσταση των εννοιών.

Το μέτρο Wu και Palmer είναι γρήγορο να υπολογιστεί, και δίνει εξίσου καλά αποτελέσματα με τα υπόλοιπα (βλ. [L98]), ενώ παράλληλα δεν απαιτεί κάποια προπαίδευση καθώς δε χρησιμοποιεί πληροφοριακό περιεχόμενο. Λαμβάνοντας υπόψη ένα δέντρο, και δύο κόμβους a,b αυτού του δέντρου, βρίσκουμε καταρχήν το βαθύτερο (από την άποψη του βάθους στο δέντρο) κοινό πρόγονό τους το c. Το μέτρο ομοιότητας υπολογίζεται ως εξής:

$$S_{W\&P}(a,b) = \frac{2.Depth(c)}{Depth(a)+Depth(b)}$$, μέτρο Wu and Palmer Εξ.3

2.8.4 Αλγόριθμοι συσταδοποίησης εγγράφων

Οι αλγόριθμοι συσταδοποίησης για έγγραφα μπορούν να ταξινομηθούν στις ακόλουθες κατηγορίες: *μεριστικοί* (*partitional)*, *ιεραρχικοί (hierarchical)*, αλγόριθμοι που βασίζονται σε γράφους (graph based), σε νευρωνικά δίκτυα (neural network-based), και *πιθανοτικοί (probabilistic)*. Επιπλέον, σύμφωνα με τον τρόπο που χειρίζεται την αβεβαιότητα στην επικάλυψη συστάδων, ένας αλγόριθμος μπορεί να είναι είτε *διακριτός* (crisp), που θεωρεί κάθε έγγραφο να ανήκει αποκλειστικά σε μια συστάδα, ή *συγκεχυμένος* (fuzzy), όταν θεωρεί ότι ένα έγγραφο μπορεί να ταξινομηθεί σε περισσότερες από μια συστάδες. Οι περισσότεροι από τους υπάρχοντες αλγορίθμους παράγουν διακριτές συστάδες. Στις παραγράφους που ακολουθούν παρουσιάζουμε τους σημαντικότερους αλγορίθμους σε κάθε κατηγορία.

Μεριστικοί - Partitional

Οι μεριστικές ή μη ιεραρχικές προσεγγίσεις επιτυγχάνουν έναν μονο-επίπεδο, μη ιεραρχικό, χωρισμό της συλλογής των εγγράφων σε έναν προκαθορισμένο αριθμό συστάδων. Οι αλγόριθμοι της κατηγορίας αυτής διαιρούνται σε επαναληπτικούς και μη επαναληπτικούς. Οι περισσότεροι είναι επαναληπτικοί και χρησιμοποιούν μη επαναληπτικές διαδικασίες μόνο στο πρώτο βήμα. Για παράδειγμα, επιλέγουν αυθαίρετα κάποια έγγραφα ως κέντρα (seeds) των συστάδων στο πρώτο βήμα και σε κάθε επανάληψη αναθεωρούν τα κέντρα αυτά.

Χρησιμοποιούν έναν πίνακα διανυσμάτων χαρακτηριστικών (feature vector matrix) και παράγουν συστάδες που βελτιστοποιούν ένα κριτήριο (π.χ. μεγιστοποιούν το άθροισμα των μέσων ομοιοτήτων συνημιτόνου για όλα τα ζεύγη εγγράφων μιας συστάδας, ελαχιστοποιούν την ομοιότητα συνημιτόνου του κέντρου κάθε συστάδας προς το κέντρο ολόκληρης της συλλογής κτλ.). Κάθε γραμμή του πίνακα διανυσμάτων αντιστοιχεί σε ένα έγγραφο και κάθε στήλη σε ένα χαρακτηριστικό. Η θέση ij περιέχει το βάρος του χαρακτηριστικού j για το έγγραφο i.

Ο πιο συνηθισμένος μεριστικός αλγόριθμος είναι ο K-means [BL96], που βασίζεται στην έννοια του κέντρου της συστάδας, ενός αντιπροσωπευτικού εγγράφου για όλα τα έγγραφα της συστάδας. Ο αλγόριθμος αρχίζει με την επιλογή Κ αντιπροσωπευτικών εγγράφων (κέντρα) και την ανάθεση των υπολοίπων εγγράφων στη συστάδα με το πλησιέστερο κέντρο. Υπολογίζει επαναληπτικά νέα κέντρα συστάδων και ανακατανέμει τα έγγραφα μέχρις ότου ικανοποιηθεί κάποιο κριτήριο τερματισμού. Παραλλαγές του αλγορίθμου k-means είναι οι αλγόριθμοι ISODATA [JM+99] και bisecting k-means [KS+00].

Το πλεονέκτημα αυτών των αλγορίθμων είναι ότι είναι απλοί και έχουν χαμηλές υπολογιστικές απαιτήσεις. Το μειονέκτημα είναι ότι η συγκέντρωση είναι μάλλον αυθαίρετη, δεδομένου ότι εξαρτάται από πολλές παραμέτρους όπως, τις τιμές των παραμέτρων εισαγωγής, την επιλογή των αρχικών κέντρων συστάδων και τη σειρά επεξεργασίας των εγγράφων.

Ιεραρχικοί - Hierarchical

Οι ιεραρχικοί αλγόριθμοι συσταδοποίησης παράγουν μια ακολουθία φωλιασμένων συστάδων. Συνήθως η ομοιότητα μεταξύ κάθε ζεύγους εγγράφων αποθηκεύεται σε ένα πίνακα ομοιότητας διαστάσεων nxn, όπου n ο αριθμός των εγγράφων. Σε κάθε στάδιο, ο αλγόριθμος είτε συγχωνεύει δύο συστάδες (***συσσωρευτικές μέθοδοι***-agglomerative) είτε διασπά μια συστάδα σε δύο (***διαχωριστικές μέθοδοι***-divisive).

Το αποτέλεσμα της συσταδοποίησης έχει δενδρική μορφή, αποκαλούμενη *δενδρόγραμμα*, με μια κορυφαία συστάδα που περιέχει όλα τα έγγραφα της συλλογής και πολλές συστάδες στο κατώτατο σημείο με ένα έγγραφο η καθεμία. Με την επιλογή του κατάλληλου επιπέδου του δενδρογράμματος παίρνουμε έναν χωρισμό σε τόσες συστάδες όσες επιθυμούμε. Το δενδρόγραμμα είναι μια αναπαράσταση που χρησιμοποιείται και για διαδικασίες ανάκτησης από ένα σύνολο εγγράφων, δεδομένου ότι καθορίζει την πορεία που πρέπει να ακολουθήσουμε για την ανάκτηση ενός εγγράφου [R92].

Σχεδόν όλοι οι ιεραρχικοί αλγόριθμοι που χρησιμοποιούνται για συσταδοποίηση εγγράφων είναι συσσωρευτικοί (Hierarchical Agglomerative Clustering algorithms). Ένας χαρακτηριστικός συσσωρευτικός αλγόριθμος αρχίζει με την ανάθεση κάθε εγγράφου της συλλογής σε μια νέα συστάδα. Στη συνέχεια, υπολογίζει την ομοιότητα μεταξύ όλων των ζευγαριών των συστάδων και την αποθηκεύει σε ένα πίνακα ομοιότητας (στη θέση ij αποθηκεύεται η ομοιότητα μεταξύ των συστάδων i και j). Κατόπιν, οι δύο πιο όμοιες συστάδες συγχωνεύονται και ο πίνακας ομοιότητας ενημερώνεται για να περιέχει την ομοιότητα μεταξύ της νέας συστάδας και των υπολοίπων συστάδων. Αυτή η διαδικασία επαναλαμβάνεται, έως ότου παραμένει μόνο μια συστάδα ή έως ότου φτάσουμε στον επιθυμητό αριθμό συστάδων.

Οι ιεραρχικές συσσωρευτικές μέθοδοι συγκέντρωσης διαφέρουν ανάλογα με τον τρόπο που υπολογίζουν την ομοιότητα μεταξύ δύο συστάδων. Άλλοτε λαμβάνεται υπόψη η ομοιότητα των πιο όμοιων εγγράφων των δύο συστάδων (Απλή σύνδεση [V86],[R79],[Si73], [R92]), η ομοιότητα των λιγότερο όμοιων εγγράφων (Πλήρης σύνδεση [V86][D77]), ή η ομοιότητα των κέντρων των δύο συστάδων (centroids similarity [KS+00]) και άλλοτε η μέση ομοιότητα των εγγράφων των δύο συστάδων (Μέσος όρος ομάδας [V86],[KS+00]) ή η μετακίνηση του κέντρου της νέας συστάδας ως προς τα κέντρα των συστατικών της (μέθοδος Ward [M83],[HW86]).

Υπάρχουν αρκετά πειράματα στη βιβλιογραφία που συγκρίνουν τις διάφορες μεθόδους ιεραρχικής συσταδοποίησης. Τα περισσότερα συγκλίνουν στην άποψη ότι η μέθοδος απλής σύνδεσης, αν και είναι η μόνη που μπορεί να εφαρμοστεί για μεγάλες συλλογές, δε δίνει καλά αποτελέσματα [HW89], [W88], [KS+00]. Η μέθοδος μέσου όρου ομάδας δίνει τα καλύτερα αποτελέσματα, είναι όμως και η πιο πολύπλοκη ($>O(n^2)$) [HW89], [KS+00], [ZK02].

Αλγόριθμοι γράφων – Graph Based

Τα έγγραφα της συλλογής αντιμετωπίζονται ως κόμβοι και οι μεταξύ τους ομοιότητες ως ακμές του γράφου με συγκεκριμένα βάρη. Οι αλγόριθμοι αυτής της κατηγορίας βασίζονται στο διαχωρισμό του γράφου σε υπο-γράφους. Σε πολλές υλοποιήσεις όσες ακμές δεν έχουν μεγάλο βάρος (σύμφωνα με κάποιο κατώφλι) αφαιρούνται. Με βάση τις ακμές που απομένουν, ορίζονται συστάδες εγγράφων. Το άθροισμα των βαρών στις εσωτερικές ακμές μιας συστάδες (μεταξύ των εγγράφων τις ίδιας συστάδας) είναι μεγαλύτερο από το άθροισμα των βαρών των εξωτερικών ακμών της συστάδας (προς έγγραφα εκτός συστάδας). Ως εκ τούτου, οι προκύπτουσες συστάδες περιέχουν ιδιαίτερα σχετικά έγγραφα.

Ο Chameleon [KH+99] και ο Association Rule Hypergraph Partitioning [BG+99] είναι δύο αλγόριθμοι γράφων. Διαφέρουν στον τρόπο που παράγουν το γράφο και

στον αλγόριθμο που χρησιμοποιούν. Ο Chameleon βασίζεται στην προσέγγιση "κ-γειτόνων" (k-nearest neighbor graph), ενώ ο Association Rule Hypergraph Partitioning βασίζεται στους υπερ-γράφους. Ο Chameleon χρησιμοποιεί έναν ιεραρχικό συσσωρευτικό αλγόριθμο στη συνέχεια και γι'αυτό μπορεί να χαρακτηριστεί ως υβριδικός αλγόριθμος που συνδυάζει γράφους και ιεραρχικούς αλγορίθμους συσταδοποίησης. Ο ARHP χρησιμοποιείται στο πρόγραμμα WebACE [HB+98] για να ομαδοποιήσει τα αποτελέσματα μιας μηχανής αναζήτησης.

Μια άλλη προσέγγιση αποτελεί ο αλγόριθμος του Dhillon [D01] που χρησιμοποιεί την επαναληπτική τμηματοποίηση με χρήση διμερή γράφου (iterative bipartite graph partitioning) για να ομαδοποιήσει έγγραφα ή λέξεις.

Οι αλγόριθμοι γράφων δείχνουν να ταιριάζουν πολύ καλά με τη φύση του προβλήματος συσταδοποίησης εγγράφων του ΠΙ, καθώς τα έγγραφα μιας συλλογής αποτελούν από μόνα τους ένα γράφο με ακμές τους μεταξύ τους υπερσυνδέσμους. Μια υβριδική προσέγγιση που θα συνδυάζει τον πραγματικό γράφο που σχηματίζουν οι υπερσύνδεσμοι με το γράφο εννοιολογικής ομοιότητας που μπορούμε να καθορίσουμε σε μια συλλογή ίσως αποτελεί την ιδανική λύση στο πρόβλημα συσταδοποίησης των εγγράφων του ΠΙ και αποτελεί ένα από τα επόμενα ερευνητικά βήματα για το σύστημα THESUS. Παρ' όλα αυτά, πρέπει να ληφθούν υπόψη η πολυπλοκότητα των αλγορίθμων γράφων και οι απαιτήσεις που έχουν σε μνήμη καθώς είναι προτιμότερο ο γράφος να εγκαθίσταται στην μνήμη.

Αλγόριθμοι που χρησιμοποιούν νευρωνικά δίκτυα – Based on neural networks
Οι *αυτο-οργανωνόμενοι χάρτες χαρακτηριστικών* (self-organizing feature maps-SOMs) του Kohonen [K98] είναι ένα ευρέως χρησιμοποιημένο μη-επιβλεπόμενο μοντέλο νευρωνικών δικτύων. Αποτελείται από δύο επίπεδα: το επίπεδο εισόδου με n κόμβους εισόδου, ένα για κάθε έγγραφο, και το επίπεδο εξόδου με k κόμβους εξόδου, ένα για κάθε περιοχή απόφασης. Οι μονάδες εισόδου λαμβάνουν τα δεδομένα και τα διαδίδουν στις μονάδες εξόδου. Σε κάθε μια από τις μονάδες εξόδου ορίζεται ένα διάνυσμα βάρους. Σε κάθε βήμα εκμάθησης, ένα έγγραφο από τη συλλογή ανατίθεται στην έξοδο με το πιο όμοιο διάνυσμα βάρους. Το διάνυσμα βάρους αυτής της εξόδου τροποποιείται, ώστε να πλησιάσει ακόμη περισσότερο στο συγκεκριμένο έγγραφο. Η διαδικασία επαναλαμβάνεται, μέχρις ότου να μην υπάρχουν άλλες αλλαγές στα διανύσματα βάρους των κόμβων εξόδου. Το αποτέλεσμα του αλγορίθμου είναι η διάταξη των εγγράφων σε ένα διδιάστατο χώρο, κατά τέτοιο τρόπο ώστε η ομοιότητα μεταξύ των εγγράφων να αντιστοιχίζεται σε τοπογραφική απόσταση μεταξύ των περιοχών απόφασης k. Αν θεωρήσουμε ότι οι μονάδες εξόδου είναι τοποθετημένες σε ένα διδιάστατο πλέγμα, τα έγγραφα συσσωρεύονται γύρω από αυτές σχηματίζοντας έτσι μια διδιάστατη αναπαράσταση (χάρτης) του συνόλου.

Μια άλλη προσέγγιση που προτείνεται στη βιβλιογραφία είναι το μοντέλο *ιεραρχικού χάρτη χαρακτηριστικών* [M98], που είναι βασισμένο σε μια ιεραρχική οργάνωση περισσότερων του ενός SOMs. Ο στόχος αυτής της προσέγγισης είναι να ξεπεράσει τους περιορισμούς που επιβάλλονται από το διδιάστατο πλέγμα εξόδου του μοντέλου SOM. Έτσι παράγει μια πολυεπίπεδη ιεραρχία στην οποία κάθε διδιάστατος χάρτης ενός επιπέδου προστίθεται στο επόμενο επίπεδο. Έχουμε έτσι πολλά επίπεδα λεπτομέρειας, αν επιλέξουμε να αυξήσουμε τη λεπτομέρεια γύρω από μια μονάδα εξόδου, βλέπουμε στο επόμενο επίπεδο πως κατανέμονται τα έγγραφα της σε ένα νέο πλέγμα.

Τα νευρωνικά δίκτυα είναι συνήθως χρήσιμα σε περιβάλλοντα με πολύ θόρυβο, με δεδομένα με σύνθετη εσωτερική δομή και συχνές αλλαγές. Το πλεονέκτημα αυτής της προσέγγισης είναι η δυνατότητα να δοθούν υψηλής ποιότητας αποτελέσματα χωρίς υψηλή υπολογιστική πολυπλοκότητα. Τα μειονεκτήματα είναι η δυσκολία εξήγησης των αποτελεσμάτων και το γεγονός ότι το διδιάστατο πλέγμα εξόδου μπορεί να περιορίσει την απεικόνιση και να οδηγήσει σε απώλεια πληροφοριών. Επιπλέον, η επιλογή των αρχικών βαρών μπορεί να επηρεάσει το αποτέλεσμα [JM+99].

Αλγόριθμοι ασάφειας – Fuzzy algorithms
Όλες οι προηγούμενες προσεγγίσεις θεωρούν πως κάθε έγγραφο ανήκει σε μια και μόνο μια συστάδα. Οι αλγόριθμοι ασάφειας δεν είναι αποκλειστικοί, υπό την έννοια ότι κάθε έγγραφο μπορεί να ανήκει σε περισσότερες από μία συστάδες. Και πάλι η καλύτερη συσταδοποίηση είναι αυτή που βελτιστοποιεί κάποιο κριτήριο. Το γεγονός ότι ένα έγγραφο μπορεί να ανήκει σε περισσότερες της μιας συστάδες περιγράφεται από μια *συνάρτηση μέλους*. Η συνάρτηση μέλους υπολογίζει για κάθε έγγραφο ένα διάνυσμα στο οποίο το στοιχείο στη θέση i δείχνει το βαθμό ιδιότητας μέλους του εγγράφου στη συστάδα i.

Ο συχνότερα χρησιμοποιημένος ασαφής αλγόριθμος συσταδοποίησης είναι ο Fuzzy C-means [BE+84], μια παραλλαγή του αλγορίθμου K-means. Ο βαθμός ιδιότητας μέλους ενός εγγράφου σε κάθε συστάδα εξαρτάται από την ομοιότητα του εγγράφου με το αντιπροσωπευτικό έγγραφο κάθε συστάδας.

Πιθανοτικοί αλγόριθμοι – Probabilistic algorithms
Ένας άλλος τρόπος χειρισμού της ασάφειας είναι να χρησιμοποιηθούν οι πιθανοτικοί αλγόριθμοι συσταδοποίησης. Αυτοί οι αλγόριθμοι χρησιμοποιούν στατιστικά μοντέλα, αντί για κάποια μέτρα ομοιότητας, για να υπολογίσουν την ομοιότητα μεταξύ των εγγράφων. Η βασική ιδέα είναι ο καθορισμός της πιθανότητας ενός εγγράφου να ανήκει σε μια συστάδα. Κάθε έγγραφο μπορεί να ανήκει σε περισσότερες από μια συστάδες με διαφορετικές πιθανότητες. Οι πιθανοτικές προσεγγίσεις [EH81] υποθέτουν ότι τα έγγραφα μπορούν να χωριστούν στις συστάδες σύμφωνα με μια συνάρτηση κατανομής πιθανότητας (ΣΚΠ). Η ΣΚΠ μιας συστάδας δίνει την πιθανότητα της παρατήρησης ενός εγγράφου με συγκεκριμένα βάρη στα χαρακτηριστικά του, σε εκείνη την συστάδα. Δεδομένου ότι η ιδιότητα μέλους ενός εγγράφου σε κάθε συστάδα δεν είναι γνωστή εκ των προτέρων, τα στοιχεία χαρακτηρίζονται από μια κατανομή, η οποία είναι το μίγμα όλων των διανομών συστάδων. Δύο ευρέως χρησιμοποιούμενοι πιθανοτικοί αλγόριθμοι είναι ο αλγόριθμος μεγιστοποίησης προσδοκίας (Expectation Maximization) [DL+77] και ο AutoClass [CS96]. Η έξοδος των πιθανοτικών αλγορίθμων είναι η πιθανότητα κάθε εγγράφου να ανήκει σε κάθε συστάδα.

Αλγόριθμοι ανάλυσης συνδέσμων - Link Analysis algorithms
Η χρήση της δομής των συνδέσμων για τη συσταδοποίηση μιας συλλογής είναι βασισμένη στην ανάλυση αναφορών (citation analysis [G72]) στον τομέα της βιβλιομετρίας. Η ανάλυση αναφορών υποθέτει ότι εάν ένα πρόσωπο που δημιουργεί ένα έγγραφο αναφέρει δύο άλλα έγγραφα, τότε τα έγγραφα πρέπει να σχετίζονται μεταξύ τους. Κατ' αυτό τον τρόπο, ο αλγόριθμος συσταδοποίησης προσπαθεί να ενσωματώσει την ανθρώπινη κρίση κατά το χαρακτηρισμό των εγγράφων. Δύο μέτρα

ομοιότητας μεταξύ δύο εγγράφων p και q που βασίζονται στην ανάλυση αναφορών είναι: η συν-αναφορά (co-citation), η οποία είναι ο αριθμός των εγγράφων που αναφέρουν ταυτόχρονα τα p και q και η σύζευξη (coupling), η οποία είναι ο αριθμός εγγράφων στα οποία «δείχνουν» και το p και το q. Μεγαλύτερες τιμές των δύο μέτρων συνεπάγονται μεγαλύτερη ομοιότητα μεταξύ των εγγράφων.

Οι Botafogo & Shneiderman [BS91] προτείνουν έναν αλγόριθμο, βασισμένοι σε έναν αλγόριθμο γράφων, που βρίσκει έντονα συνδεδεμένους κόμβους στο γράφο που σχηματίζει ένα υπερκείμενο. Ο αλγόριθμος χρησιμοποιεί ένα μέτρο πυκνότητας, που δείχνει την ενδο-συνοχή του υπερκειμένου, υπολογίζοντας τη μέση απόσταση συνδέσεων μεταξύ των κόμβων. Ο αλγόριθμος προσδιορίζει τις συστάδες ως ιδιαίτερα συνδεδεμένους υπο-γράφους. Στο [B93] επεκτείνεται η ιδέα που περιλαμβάνει στον υπολογισμό της πυκνότητας και τον αριθμό των διαφορετικών μονοπατιών που συνδέουν δύο κόμβους. Αυτός ο εκτεταμένος αλγόριθμος παράγει καλύτερες συστάδες, με λογικό μέγεθος και ιδιαίτερα σχετικούς κόμβους.

Ο αλγόριθμος του Larson [L96], εφάρμοσε την ανάλυση αναφορών σε μια συλλογή των εγγράφων ΠΙ. Η ανάλυση αναφορών αρχίζει με την κατασκευή ενός πίνακα κοινών αναφορών, στη θέση ij του οποίου περιέχεται η συχνότητα κοινών αναφορών των εγγράφων i και j. Μια συνάρτηση συσχέτισης εφαρμόζεται για να μετατρέψει τις συχνότητες σε συντεταγμένες συσχέτισης. Το τελευταίο βήμα είναι η απεικόνιση του πίνακα κοινών αναφορών σε ένα διδιάστατο χάρτη. Η ερμηνεία του χάρτη μπορεί να αποκαλύψει ενδιαφέροντες σχέσεις και σχηματισμούς ομάδων εγγράφων. Η πολυπλοκότητα του αλγορίθμου είναι $O\left(\frac{n^2}{2} - n\right) \cong O(n^2)$.

2.8.5 Οι αλγόριθμοι του THESUS

Είναι προφανές από τα προηγούμενα ότι το πρόβλημα της συσταδοποίησης των εγγράφων έχει απασχολήσει τους ερευνητές που ασχολούνται με την ανάκτηση πληροφορίας [F87], [AG+99]. Το πρόβλημα γίνεται πιο συγκεκριμένο για τα έγγραφα του ΠΙ [ZE98], καθώς λαμβάνονται υπόψη οι συνδέσεις μεταξύ των εγγράφων [K99]. Δεδομένου ότι εργαζόμαστε με κατηγορικά δεδομένα χρειαζόμαστε ένα μέτρο απόστασης ή ομοιότητας και έναν αλγόριθμο που να βασίζεται στην πυκνότητα των εγγράφων. Δύο αλγόριθμοι που έχουν υλοποιηθεί και ενταχθεί στο σύστημα THESUS είναι ο αλγόριθμος COBWEB [F87] για κατηγορικά δεδομένα και ο αλγόριθμος DBSCAN [EK+96],[EK+98]. Και οι δύο αλγόριθμοι έχουν τροποποιηθεί ώστε να μπορούν να χρησιμοποιούν το μέτρο ομοιότητας μεταξύ εγγράφων που ορίζεται στο THESUS.

Όπως προαναφέρθηκε:

Το πρόβλημα της συσταδοποίησης είναι εντελώς διαφορετικό όταν πρόκειται για σημεία ενός Ευκλείδειου χώρου. Στην περίπτωσή των εγγράφων του ΠΙ δεν υπάρχει διάταξη μεταξύ τους, και τα αντικείμενα που πρόκειται να ομαδοποιηθούν είναι σύνολα λέξεων (με βάρη) που αντιστοιχούν σε έννοιες μιας θεματικής οντολογίας. Διαθέτουμε μόνο ένα μέτρο ομοιότητας ανάμεσα σε τέτοια σύνολα.

Ο πρώτος αλγόριθμος που χρησιμοποιεί το σύστημα THESUS είναι ο δημοφιλής αλγόριθμος πυκνότητας DBSCAN [EK+96] με μια τροποποίηση στα κριτήρια

πυκνότητας. Ένας αυξητικός ιεραρχικός αλγόριθμος ο COBWEB [F87], παραλλαγμένος για να δέχεται έγγραφα με τη δομή που ορίζονται στο THESUS, έχει επίσης υλοποιηθεί.

Η επιλογή του αλγορίθμου DBSCAN στηρίχθηκε σε δύο κυρίως λόγους:

- δεν απαιτεί πρότερη γνώση του αριθμού των παραγόμενων συστάδων,
- έχει τη δυνατότητα να ανιχνεύει συστάδες που δεν έχουν σφαιρικό σχήμα

Επιπλέον είναι ένας αλγόριθμος πυκνότητας που μπορεί να υλοποιηθεί με χρήση ενός μέτρου ομοιότητας χωρίς να απαιτεί διάταξη στο χώρο των εγγράφων.

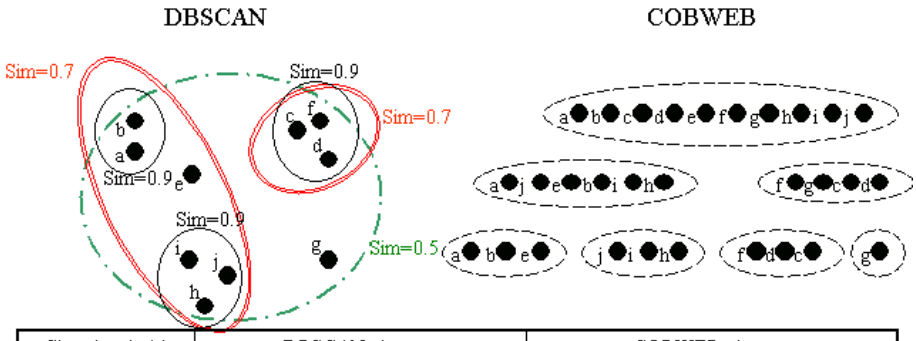
Ο αλγόριθμος COBWEB είναι ένας δημοφιλής ιεραρχικός αλγόριθμος για κατηγορικά δεδομένα που προτιμήθηκε από τους υπολοίπους επειδή είναι αυξητικός και συνεπώς πολύ χρήσιμος στην περίπτωση που νέα έγγραφα του ιστού προστίθενται σε ένα σύνολο.

Ο τρόπος λειτουργίας των δύο αλγορίθμων είναι διαφορετικός όπως και η παραγόμενη ομαδοποίηση. Οι λεπτομέρειες τροποποίησης αλλά και οι αλγόριθμοι περιγράφονται στην παράγραφο 3.3.2. Στα πλαίσια της υλοποίησης δόθηκε έμφαση στον αλγόριθμο DBSCAN, καθώς στα πρώτα πειράματα που πραγματοποιήθηκαν φάνηκε να έχει καλύτερες επιδόσεις. Στη συνέχεια η αποδοτικότητα του αλγορίθμου βελτιώθηκε σημαντικά με χρήση διάφορων τεχνικών αποθήκευσης πληροφορίας στη μνήμη (caching).

Ο αλγόριθμος COBWEB είναι ένας ιεραρχικός αλγόριθμος, που τοποθετεί ένα έγγραφο σε μία ομάδα, σε αντίθεση με τον DBSCAN, που θεωρεί ορισμένα έγγραφα που δε μοιάζουν με τα υπόλοιπα ως θόρυβο και δεν τα τοποθετεί σε ομάδες. Σύμφωνα με τον COBWEB κάθε έγγραφο τοποθετείται στο πρώτο επίπεδο στην ίδια ομάδα με όλα τα υπόλοιπα και στη συνέχεια σε κάποια από τις υποομάδες αυτής στο αμέσως επόμενο επίπεδο. Με συνεχείς συγχωνεύσεις και διασπάσεις ομάδων, ο αλγόριθμος στοχεύει στην αύξηση της συνοχής των παραγόμενων ομάδων σε κάθε επίπεδο. Ως αποτέλεσμα κάθε επίπεδο στην ιεραρχία που παράγεται περιέχει όλα τα έγγραφα του συνόλου, ο αριθμός όμως των ομάδων αυξάνει καθώς κινούμαστε προς τα κατώτερα επίπεδα της ιεραρχίας.

Ο αλγόριθμος DBSCAN παράγει μονοεπίπεδη (όχι ιεραρχική) ομαδοποίηση στην οποία δεν περιέχονται τα έγγραφα που έχουν θεωρηθεί θόρυβος. Μειώνοντας το κατώφλι ομοιότητας για δύο έγγραφα που ανήκουν στην ίδια ομάδα (μια παράμετρος του DBSCAN), το μέγεθος των ομάδων αυξάνεται, ορισμένες ομάδες συγχωνεύονται και λιγότερα έγγραφα λαμβάνονται ως θόρυβος. Επαναλαμβάνοντας τον DBSCAN για το ίδιο σύνολο εγγράφων για μειούμενες τιμές του κατωφλίου ομοιότητας, παράγεται μια ψευδο-ιεραρχική, πολυεπίπεδη ομαδοποίηση εγγράφων. Αξίζει να τονιστεί ότι ο αλγόριθμος DBSCAN σχεδιάστηκε αρχικά για γεωγραφικές βάσεις δεδομένων. Με τον καθορισμό ενός σημασιολογικού μέτρου ομοιότητας, είμαστε σε θέση, με λίγη τροποποίηση, να εφαρμόσουμε τον αλγόριθμο και να πετύχουμε καλά αποτελέσματα.

Τα δύο σχήματα ομαδοποίησης που προκύπτουν απεικονίζονται στο Σχήμα 6.



Sim. threshold (DBSCAN only)	DBSCAN clusters			COBWEB clusters			
0.5	{a,b,c,d,e,f,g,h,i,j}			{a,b,c,d,e,f,g,h,i,j}			
0.7	{a,b,e,h,i,j}		{c,d,f}	{a,j,e,b,i,h}		{f,g,c,d,}	
0.9	{a,b}	{h,i,j}	{c,d,f}	{a,b,e}	{j,i,h}	{f,d,c}	{g}

Σχήμα 6. Ιεραρχίες ομάδων που παράγει το THESUS

2.8.6 Μέτρα αξιολόγησης της ποιότητας συσταδοποίησης

Για να αξιολογήσουμε την ποιότητα της συσταδοποίησης που παράγει ένας αλγόριθμος [KS+00] έχουμε τρεις δυνατότητες:
α) να χρησιμοποιήσουμε ένα *εσωτερικό μέτρο ποιότητας* που θα αξιολογήσει την ποιότητα της συσταδοποίησης χρησιμοποιώντας κριτήρια όπως, το πόσο συμπαγείς (πόσο όμοια είναι μεταξύ τους τα έγγραφα μιας συστάδας) και το πόσο διακριτές είναι οι συστάδες (πόσο ανόμοια είναι τα έγγραφα μιας συστάδας με αυτά των υπολοίπων),
β) να χρησιμοποιήσουμε ένα *σχετικό μέτρο ποιότητας* που θα μας βοηθήσει να συγκρίνουμε δύο η περισσότερες διαφορετικές συσταδοποιήσεις του συνόλου και να επιλέξουμε την καλύτερη. Τα εσωτερικά και σχετικά μέτρα ποιότητας λειτουργούν χωρίς αναφορά σε κάποια πρότερη γνώση του τρόπου με τον οποίο ομαδοποιούνται τα έγγραφα,
γ) τέλος, μπορούμε να χρησιμοποιήσουμε κάποιο *εξωτερικό μέτρο ποιότητας* που θα αξιολογήσει την ποιότητα της συσταδοποίησης με βάση κάποια πρότερη γνώση για την ομαδοποίηση των εγγράφων.

Υπάρχουν πολλά και διαφορετικά μέτρα ποιότητας τα οποία μπορεί να δώσουν εντελώς διαφορετική αξιολόγηση για διάφορα σχήματα και αλγορίθμους συσταδοποίησης, [TK99], [HB+01], [HV01]. Παρ' όλα αυτά, αν ένα σχήμα εμφανίζει καλή απόδοση στα περισσότερα από αυτά τα μέτρα, είναι μια ένδειξη της ποιότητας του συγκεκριμένου σχήματος.

2.8.6.1 Εσωτερικά μέτρα ποιότητας

Όπως προαναφέρθηκε τα εσωτερικά μέτρα ποιότητας μπορούν να αξιολογήσουν ένα σχήμα συσταδοποίησης χωρίς να έχουν πρότερη γνώση για τα έγγραφα και τον τρόπο με τον οποίο πρέπει να ομαδοποιηθούν.

Τα μέτρα αυτής της κατηγορίας υπολογίζουν μια εσωτερική ομοιότητα σε κάθε συστάδα (*Internall similarity*$_{\chi}$) ως τη μέση τιμή της ομοιότητας για κάθε ζεύγος εγγράφων της συστάδας. Με τον τρόπο αυτό έχουν μια ένδειξη για το πόσο συμπαγής είναι η κάθε συστάδα.

$$Internall\ similarity_x = \frac{1}{n_x^{\ 2}} \sum_{d,d' \in x} sim(d', d) \qquad \text{Εξ.4}$$

όπου n_x το πλήθος των στοιχείων της συστάδας x, και d, d' δύο στοιχεία της συστάδας x.

Οι εσωτερικές ομοιότητες των συστάδων χρησιμοποιούνται για τον υπολογισμό της καθολικής εσωτερικής ομοιότητας του σχήματος (*Internall similarity*$_{CS}$). Η εσωτερική ομοιότητα του σχήματος συσταδοποίησης *CS* υπολογίζεται ως το άθροισμα των εσωτερικών ομοιοτήτων των συστάδων που παράγονται σταθμισμένων ως προς το μέγεθος της κάθε συστάδας.

$$Internall\ similarity_{CS} = \frac{\sum_{x=1}^{m} n_x * IS_x}{n} \qquad \text{Εξ.5}$$

όπου m ο αριθμός των συστάδων του σχήματος, n_x το πλήθος των στοιχείων της συστάδας x, IS_x η εσωτερική ομοιότητα της συστάδας και n ο συνολικός αριθμός εγγράφων του συνόλου.

Σε ορισμένα μέτρα, κριτήριο αποτελεί και η εξωτερική ανομοιότητα της κάθε συστάδας, δηλαδή το αντίστροφο της ομοιότητας (similarity) των εγγράφων μιας συστάδας από τα έγγραφα όλων των άλλων συστάδων. Με παρόμοιο τρόπο υπολογίζεται και η καθολική ανομοιότητα των συστάδων του σχήματος CS. Το γινόμενο των δύο μεγεθών αποτελεί το μέτρο ποιότητας του αλγορίθμου. Όσο μεγαλύτερο είναι το γινόμενο αυτό, τόσο καλύτερο είναι το σχήμα συσταδοποίησης.

2.8.6.2 Σχετικά μέτρα ποιότητας

Τα σχετικά μέτρα ποιότητας βασίζονται στα προηγούμενα. Στόχος είναι η επιλογή του καλύτερου σχήματος συσταδοποίησης από ένα σύνολο ορισμένων σχημάτων σύμφωνα με ένα προκαθορισμένο κριτήριο. Για το λόγο αυτό εκτελούν πολλές φορές τον αλγόριθμο συσταδοποίησης με διαφορετικές τιμές των παραμέτρων εισόδου (σε ένα συγκεκριμένο εύρος και με καθορισμένο βήμα) και σε κάθε περίπτωση υπολογίζουν το δείκτη ποιότητας της συσταδοποίησης, χρησιμοποιώντας ένα εσωτερικό μέτρο.

Οι καλύτερες τιμές του δείκτη τοποθετούνται σε γράφημα ως συνάρτηση της κάθε παραμέτρου εισόδου. Με τον τρόπο αυτό δίνουν μια ένδειξη για το πώς η ποιότητα της συσταδοποίησης μεταβάλλεται με κάθε παράμετρο. Σημεία στα οποία το γράφημα του μέτρου ποιότητας για κάποια παράμετρο εμφανίζει σημαντική τοπική αλλαγή, όπως και σημεία στα οποία εμφανίζεται μέγιστο ή ελάχιστο, αποτελούν ενδείξεις για την επιλογή των καλύτερων παραμέτρων εισόδου.

2.8.6.3 Εξωτερικά μέτρα ποιότητας

Τα εξωτερικά μέτρα ποιότητας μας επιτρέπουν να αξιολογήσουμε ένα σχήμα συσταδοποίησης μόνον εφόσον γνωρίζουμε τις ομάδες στις οποίες ανήκουν τα στοιχεία μας. Τα μέτρα ποιότητας αυτής της κατηγορίας ακολουθούν δύο διαφορετικές προσεγγίσεις: α) είτε προσπαθούν να υπολογίσουν την επικάλυψη του

παραγόμενου σχήματος συσταδοποίησης με το αρχικό σχήμα ομαδοποίησης, π.χ. μετρώντας τα ζεύγη των στοιχείων της ίδιας ομάδας που τοποθετούνται στην ίδια και σε διαφορετικές συστάδες, β) είτε βασίζονται στις αρχές της ανάκλησης και ακρίβειας και μετρούν ένα βαθμό ανάκλησης και ακρίβειας για κάθε συστάδα ξεχωριστά.

Το μέτρο F-measure

Το μέτρο F-measure προέρχεται από την περιοχή της ανάκτησης πληροφορίας [RL94] αλλά συχνά χρησιμοποιείται και στον υπολογισμό της ποιότητας ιεραρχικών σχημάτων συσταδοποίησης [LA99]. Μπορεί, παρ' όλα αυτά, να χρησιμοποιηθεί και για την αξιολόγηση επίπεδων σχημάτων συσταδοποίησης. Το μέτρο στηρίζεται στον υπολογισμό των βαθμών ανάκλησης και ακρίβειας κάθε συστάδας ως προς κάθε τάξη.

Η ανάκληση και η ακρίβεια για μια συστάδα j ως προς μια τάξη i ορίζονται ως εξής:

$$\text{Ανάκληση} - \text{recall(i,j)} = \frac{n_{ij}}{n_i} \qquad \text{Εξ.6}$$

$$\text{Ακρίβεια} - \text{precision(i,j)} = \frac{n_{ij}}{n_j} \qquad \text{Εξ.7}$$

όπου n_{ij} είναι ο αριθμός των στοιχείων της τάξης i που περιέχονται στη συστάδα j, n_j είναι το πλήθος στοιχείων της συστάδας j και n_i το πλήθος στοιχείων της τάξης i.

Αν θεωρήσουμε εξίσου σημαντικές την ανάκληση και ακρίβεια (βλ. [RL94]), τότε το μέτρο F για τη συστάδα j ως προς την τάξη i υπολογίζεται από τη σχέση:

$$F(i,j) = \frac{2 \cdot recall(i,j) \cdot precision(i,j)}{recall(i,j) + precision(i,j)} \qquad \text{Εξ.8}$$

Για ένα ολόκληρο ιεραρχικό σχήμα συσταδοποίησης *CS*, το μέτρο F κάθε τάξης είναι η μέγιστη τιμή που επιτυγχάνει σε οποιοδήποτε κόμβο του δέντρου. Η συνολική τιμή του F-measure, σύμφωνα με το [KS+00], υπολογίζεται ως το σταθμισμένο μέσο των τιμών του F-measure για όλες τις τάξεις.

$$F_{CS} = -\sum_i \frac{n_i}{n} \max\{F(i,j)\} \qquad \text{Εξ.9}$$

όπου n ο συνολικός αριθμός στοιχείων και n_i ο αριθμός στοιχείων της τάξης i.

Το μέτρο Rand Statistic

Το μέτρο αυτό [TK99] εξετάζει όλα τα δυνατά ζεύγη (m το πλήθος) που προκύπτουν από τα n στοιχεία του συνόλου. Μετρά τα ζεύγη των στοιχείων που βρέθηκαν ταυτόχρονα στην ίδια τάξη και την ίδια συστάδα (SS) ή σε διαφορετικές τάξεις και διαφορετικές συστάδες (DD). Το μέτρο ποιότητας για το σχήμα συσταδοποίησης δίνεται από τη σχέση:

$$R = \frac{SS + DD}{m}, \text{ όπου } m = \frac{n(n-1)}{2} \qquad \text{Εξ.10}$$

Εντροπία

Η εντροπία [S48] είναι ένα εξωτερικό μέτρο ποιότητας που παρέχει μια ένδειξη για την ποιότητα συστάδων για μη ιεραρχικά σχήματα συσταδοποίησης, ή για ένα

επίπεδο του σχήματος συσταδοποίησης. Γενικότερα η εντροπία υπολογίζει το βαθμό ανομοιογένειας των συστάδων, η οποία κυμαίνεται από 1 όταν όλα τα στοιχεία μιας συστάδας είναι ετερόκλητα (από διαφορετικές ομάδες) έως 0 όταν μια συστάδα περιέχει στοιχεία μιας μόνο ομάδας. Βέλτιστο είναι το σχήμα που ελαχιστοποιεί την εντροπία των συστάδων. Το βασικό μειονέκτημα της εντροπίας εντοπίζεται στο γεγονός ότι μια συστάδα με ένα μόνο στοιχείο έχει εντροπία 0 και κατ' επέκταση ένα σχήμα που τοποθετεί όλα τα στοιχεία σε διαφορετικές συστάδες μπορεί να θεωρηθεί βέλτιστο.

Έστω *CS* ένα σχήμα συσταδοποίησης. Ο υπολογισμός της εντροπίας του σύμφωνα με το [KS+00] έχει ως εξής:
Αρχικά υπολογίζουμε για κάθε συστάδα την κατανομή των στοιχείων κάθε τάξης. Για παράδειγμα, για τη συστάδα j υπολογίζουμε την πιθανότητα p_{ij}, ένα στοιχείο της συστάδας j να ανήκει στην τάξη i.

$$p(i \mid j) = p_{ij} = \frac{n_{ij}}{n_j} \qquad \text{Εξ.11}$$

όπου n_{ij} το πλήθος των στοιχείων τάξης i που βρέθηκαν στη συστάδα j, και n_j ο συνολικός αριθμός στοιχείων της συστάδας j.

Με βάση την κατανομή των τάξεων, η εντροπία για κάθε συστάδα υπολογίζεται ως:

$$E_j = -\sum_i p_{ij} \cdot \log\left(p_{ij}\right) \qquad \text{Εξ.12}$$

για όλες τις τάξεις i, στοιχεία των οποίων περιέχονται στο j.

Η συνολική εντροπία του σχήματος υπολογίζεται ως το σταθμισμένο άθροισμα των εντροπιών όλων των συστάδων με βάση το μέγεθός τους.

$$E_{CS} = \sum_{j=1}^{m} \frac{n_j \cdot E_j}{n} \qquad \text{Εξ.13}$$

όπου n_j το μέγεθος της συστάδας j, m ο αριθμός των συστάδων και n ο συνολικός αριθμός στοιχείων.

2.9 Συλλογή εγγράφων - Crawlers

Απαραίτητο βήμα για την οργάνωση της πληροφορίας του ιστού και την εξυπηρέτηση των αναζητήσεων είναι η συλλογή της πληροφορίας που βρίσκεται διάσπαρτη σε χιλιάδες εξυπηρετητές του ΠΙ. Ο όγκος της πληροφορίας είναι πολύ μεγάλος και ο ρυθμός ανανέωσής της πολύ ταχύς. Ο ρυθμός ανανέωσης μάλιστα είναι διαφορετικός σε κάθε δικτυακό τόπο – για παράδειγμα τα περιεχόμενα ενός ειδησεογραφικού τόπου ενημερώνονται κάθε μέρα ή και συχνότερα, ενώ τα περιεχόμενα ενός προσωπικού δικτυακού τόπου μπορεί να ενημερώνονται πολύ πιο σπάνια. Για το λόγο αυτό οι εφαρμογές που συλλέγουν περιεχόμενο από τον ιστό θα πρέπει να εξάγουν και να αποθηκεύουν για κάθε έγγραφο του ιστού όσο το δυνατό λιγότερα και χαρακτηριστικότερα στοιχεία. Πρέπει επίσης να ελέγχουν το ρυθμό με τον οποίο ανανεώνονται τα περιεχόμενα των διαφόρων δικτυακών τόπων και να τους επισκέπτονται ανά διαστήματα [CM00].

Τα προγράμματα που αναλαμβάνουν αυτή τη διαδικασία ονομάζονται ***crawlers (περιηγητές)*** [AV+01], καθώς επισκέπτονται διαδοχικά τα έγγραφα του ΠΙ. Ο μόνος

τρόπος που έχουν για να εντοπίσουν νέα έγγραφα είναι αναλύοντας τους συνδέσμους των εγγράφων που έχουν επισκεφθεί. Ο τρόπος λειτουργίας τους προσομοιάζει την κίνηση των χρηστών του ιστού που επισκέπτονται ένα έγγραφο και ακολουθούν στη συνέχεια τους συνδέσμους του. Οφείλουν λοιπόν οι περιηγητές, ξεκινώντας από ένα ή περισσότερα έγγραφα, να "επισκεφτούν" χιλιάδες έγγραφα (δηλαδή να φέρουν τοπικά τα περιεχόμενά τους) μέσα σε πολύ λίγο χρόνο, να τα αναλύσουν και να εξάγουν τα χαρακτηριστικά τους.

Στις περισσότερες πληροφορίες ο περιηγητής αποθηκεύει σε μια βάση δεδομένων την πληροφορία που εξάγει από τα έγγραφα. Η πληροφορία που αποθηκεύει ο περιηγητής για ένα έγγραφο που έχει επισκεφθεί μπορεί να περιλαμβάνει: τη διεύθυνση ιστού, το μέγεθος τους αρχείου, την ημερομηνία επίσκεψης, τους συνδέσμους που περιέχει, χαρακτηριστικές λέξεις και πολλές φορές και το ίδιο το έγγραφο.

2.9.1 Κατηγορίες περιηγητών

Ανάλογα με το στόχο που έχουν οι περιηγητές, εμφανίζουν διαφορετική συμπεριφορά στον τρόπο επιλογής των συνδέσμων που θα ακολουθήσουν: άλλοτε επιλέγουν όλους τους συνδέσμους σε κάθε έγγραφο και άλλοτε τον καλύτερο ή τους καλύτερους σύμφωνα με κάποια κριτήρια.

Τα διάφορα μοντέλα περιηγητών διαφέρουν στην αρχιτεκτονική, στη λογική που ακολουθείται και στη λειτουργία η οποία επιτελείται. Διακρίνουμε τις παρακάτω κατηγορίες προγραμμάτων περιήγησης του ΠΙ:

1) Γενικοί περιηγητές - generic crawlers: Τα προγράμματα αυτής της κατηγορίας επισκέπτονται όλα τα έγγραφα που υποδεικνύονται από τους συνδέσμους χωρίς κανένα περιορισμό. Ο περιηγητής ακολουθεί όλους τους συνδέσμους, χωρίς περαιτέρω ανάλυση των περιεχομένων, και αποθηκεύει τοπικά είτε τα ίδια τα έγγραφα είτε τις δικτυακές τους διευθύνσεις. Το αποτέλεσμα της χρήσης τους είναι ένα πολυπληθές και πολυ-θεματικό σύνολο εγγράφων ή διευθύνσεων ιστού.

2) Επιλεκτικοί περιηγητές - selective crawlers: Στην κατηγορία αυτή ανήκουν οι περιηγητές που επισκέπτονται μόνο τους συνδέσμους εκείνους που ικανοποιούν συγκεκριμένα κριτήρια. Τα κριτήρια αυτά μπορεί να αφορούν μόνο τη διεύθυνση ιστού στην οποία δείχνει ο σύνδεσμος, π.χ. όταν θέλουμε να επισκεφθούμε έγγραφα ενός συγκεκριμένου τομέα (gr, com, net domain), ή να εξετάζουν το κείμενο του συνδέσμου. Ένα τέτοιο πρόγραμμα αναλύει την ιδιαίτερη σύνταξη των εγγράφων που επισκέπτεται, όπως τη μορφοποίηση, τη συνδεσμολογία, τα μεταδεδομένα, και περιορίζεται σε έγγραφα που είναι γραμμένα σε συγκεκριμένη γλώσσα (π.χ. ελληνικά, αγγλικά κτλ.), ή έγγραφα που χαρακτηρίζονται με συγκεκριμένες λέξεις.

3) Εστιασμένοι περιηγητές - focused crawlers [CB+99]**.** Τα προγράμματα αυτής της κατηγορίας παίζουν το ρόλο των μηχανών αναζήτησης του ιστού, λειτουργούν όμως χωρίς βάση δεδομένων. Ξεκινούν έχοντας ως δεδομένο μια ερώτηση και ένα συγκεκριμένο αρχικό έγγραφο. Σε κάθε βήμα αναλύουν τους συνδέσμους του εγγράφου και επιλέγουν για το επόμενο βήμα το έγγραφο που είναι πιθανότερο να ικανοποιεί την ερώτηση.

Η διάκριση μεταξύ των επιλεκτικών και των εστιασμένων περιηγητών δεν είναι πάντοτε εύκολη. Η βασικότερη διαφορά τους είναι η σειρά με την οποία επισκεπτόμαστε τα έγγραφα. Στην πρώτη περίπτωση ακολουθούνται όλοι οι σύνδεσμοι που θεωρούνται σχετικοί, με τη σειρά που εμφανίζονται στο έγγραφο. Στη δεύτερη περίπτωση ο περιηγητής ταξινομεί τα υποψήφια για επίσκεψη έγγραφα και τα επισκέπτεται με σειρά σημαντικότητας. Στην περίπτωση αυτή μπορεί η περιήγηση σε κάθε επίπεδο να περιορίζεται στο καλύτερο ή τα καλύτερα Ν έγγραφα.

Μια ειδική κατηγορία περιηγητών είναι αυτοί που προσαρμόζονται σε έγγραφα με δυναμικό περιεχόμενο. Τα έγγραφα αυτά παράγονται με δυναμικό τρόπο από τα περιεχόμενα βάσεων δεδομένων ανάλογα με τις παραμέτρους χρήσης και αποτελούν τον "Κρυφό ή Βαθύ Ιστό" [IG02]. Σκοπός των συγκεκριμένων περιηγητών είναι να ανακτήσουν όσο το δυνατόν μεγαλύτερο ποσοστό της πληροφορίας του "κρυφού ιστού", δίνοντας διάφορες παραμέτρους στα δυναμικά έγγραφα και αποθηκεύοντας τα παραγόμενα αποτελέσματα.

Σε κάποιες περιπτώσεις [AP+03] έχει γίνει προσπάθεια κατά τη διάρκεια της συλλογής των εγγράφων να εκτιμηθεί και η σημαντικότητα των εγγράφων όπως αυτή δηλώνεται με το πλήθος και τη σημαντικότητα των εισερχόμενων συνδέσμων. Με τον τρόπο αυτό γίνεται μια απλούστευση του αλγορίθμου [K99] που μπορεί όμως να δώσει μια πρώτη ιδέα για την τελική σημαντικότητα του εγγράφου, όπως υπολογίζεται από αλγορίθμους που αναλύουν το γράφο των εγγράφων της συλλογής [PB+98]

2.9.2 Αξιολόγηση σημαντικότητας εγγράφου

Όπως προαναφέρθηκε, σε μια διαδικασία επιλεκτικής περιήγησης του ΠΙ, είναι επιθυμητό να προσπελάζουμε πρώτα τα έγγραφα που έχουν μεγαλύτερη σημαντικότητα. Για να επιτευχθεί κάτι τέτοιο θα πρέπει να δώσουμε έναν ορισμό για τη σημαντικότητα κάθε εγγράφου. Η σημαντικότητα ενός εγγράφου μπορεί να οριστεί ειδικότερα σε μια κλειστή συλλογή διασυνδεδεμένων εγγράφων, αλλά και γενικότερα σε ολόκληρο τον ΠΙ. Αυτό μπορεί να επιτευχθεί με έναν από τους παρακάτω τρόπους [CM+98]:

Βαθμός ομοιότητας με κάποιο ερώτημα Q. Σε αυτήν την προσέγγιση θεωρούμε ένα έγγραφο ή ένα ερώτημα Q, σύμφωνα με το οποίο καθορίζουμε τη σημαντικότητα κάθε εγγράφου. Πιο αναλυτικά, η σημαντικότητα του εγγράφου είναι η λεξική ομοιότητα του εγγράφου προς εξέταση (P) με το ερώτημα (Q). Τα P και Q είναι διανύσματα μεγέθους m. Κάθε θέση του διανύσματος περιέχει μια λέξη ενός "λεξικού" που θεωρούμε ότι περιέχει όλες τις λέξεις. Αν η ι-οστή λέξη περιέχεται στο έγγραφο τότε στην θέση του πίνακα σημειώνουμε την αξία της, αλλιώς σημειώνουμε 0. Η αξία κάθε λέξης μπορεί να υπολογιστεί ως το γινόμενο του αριθμού εμφανίσεών της στο έγγραφο (term frequency-tf) επί τη συχνότητα εμφάνισής της σε όλα τα έγγραφα (inverse document frequency-idf). Είναι κατανοητό ότι για να υπολογίσουμε το idf πρέπει να έχουμε εξετάσει όλα τα έγγραφα της συλλογής [R79]. Επίσης η αξία μιας λέξης μπορεί να καθορίζεται όχι μόνο από το γινόμενο tf·idf αλλά επίσης από την σχετική της θέση στο έγγραφο. Είναι λογικό μια λέξη που βρίσκεται στον τίτλο του εγγράφου και το περιγράφει να είναι πιο σημαντική από μια άλλη που

εμφανίζεται απλά στο κείμενο. Τελικά, η σημαντικότητα του εγγράφου P είναι το εσωτερικό γινόμενο των δύο διανυσμάτων P και Q.

Ο συμβολισμός αυτής της τεχνικής αξιολόγησης εγγράφων είναι IS(P). Αν όμως δεν έχουμε συνολική πληροφορία για το idf όλων των λέξεων τότε προσπαθούμε να το εκτιμήσουμε και χρησιμοποιούμε τον συμβολισμό IS'(P).

Πλήθος εισερχόμενων συνδέσμων - BackLink count. Σύμφωνα με αυτήν την τεχνική η σημαντικότητα ενός εγγράφου προσδιορίζεται από τον αριθμό των συνδέσμων που δείχνουν σε αυτό. Αυτό είναι λογικό καθώς, αν ένα έγγραφο έχει πολλούς εισερχόμενους συνδέσμους, τότε σημαίνει ότι είναι αξιόλογο και το προτείνουν πολλοί. Δηλώνει επίσης ότι κάποιος έχει μεγάλη πιθανότητα να το επισκεφτεί μέσω κάποιου άλλου εγγράφου. Αυτός ο τρόπος χρησιμοποιείται για να αξιολογήσουμε αποτελέσματα αναζήτησης. Ο συμβολισμός του είναι IB(P) αλλά στην περίπτωση που δεν έχουμε επισκεφτεί όλα τα έγγραφα, οπότε δεν ξέρουμε τον συνολικό αριθμό των συνδέσμων, προσπαθούμε να τον εκτιμήσουμε, και γι' αυτό το σημειώνουμε σαν IB'(P).

Βαθμός PageRank[PB+98]. Αυτός ο αλγόριθμος συμπληρώνει και βελτιώνει τον προηγούμενο. Η διαφορά είναι ότι εδώ δε θεωρούμε ότι όλοι οι υπερσύνδεσμοι έχουν ίδια σημασία, αλλά η αξία τους εξαρτάται αναδρομικά από την αξία του εγγράφου από το οποίο προέρχονται. Αυτό σημαίνει ότι ένας σύνδεσμος από ένα σημαντικό έγγραφο είναι πολύ πιθανόν να οδηγεί επίσης σε ένα σημαντικό έγγραφο, κάτι που συμβαίνει και στην πραγματικότητα.

Ο μαθηματικός τύπος είναι:

$$\mathrm{IR(P)}=(1-d)+d\left(\frac{IR(T_1)}{c_1}+\ldots+\frac{IR(T_n)}{c_n}\right) \qquad \text{Εξ.14}$$

όπου $T_1 \ldots T_N$ τα έγγραφα που δείχνουν στην P, $c_1 \ldots c_n$ ο αριθμός των εξερχόμενων συνδέσμων των εγγράφων $T_1 \ldots T_N$ και d μια σταθερά (παράγοντας απόσβεσης) [P01].

Όπως φαίνεται, ο αλγόριθμος προσπαθεί να προσωμοιώσει τη συμπεριφορά ενός τυχαίου επισκέπτη του ΠΙ και να μετρήσει την πιθανότητα να επισκεφθεί το έγγραφο P. Το δεύτερο μέρος του αθροίσματος δίνει την πιθανότητα να επιλέξουμε το P, δεδομένου ότι φτάνουμε αρχικά σε κάποιο από τα T_1, …, T_n, ενώ το πρώτο είναι η πιθανότητα να φτάσουμε κατευθείαν στο P.

Ο αλγόριθμος είναι αναδρομικός και η επίλυση ξεκινάει με όλα τα έγγραφα της συλλογής να έχουν IR(P)=1, ενώ μετά από ορισμένο αριθμό επαναλήψεων οι τιμές για τα διάφορα έγγραφα συγκλίνουν. Χωρίς την ύπαρξη του παράγοντα απόσβεσης (damping factor), ο τύπος θα αντιμετώπιζε προβλήματα στην εύρεση της αξίας ενός εγγράφου. Αν για παράδειγμα είχαμε ένα κύκλο στον κατευθυνόμενο γράφο π.χ. δυο έγγραφα που δείχνουν το ένα στο άλλο τότε οι τιμές τους θα απόκλιναν σε κάθε επανάληψη.

Το $IR(T_i)/c_i$ δείχνει την αξία και την πιθανότητα να προσπελάσουμε το έγγραφο P ενώ βρισκόμαστε στο έγγραφο T_i, με δεδομένο ότι μπορούμε τυχαία να επιλέξουμε έναν από τους συνδέσμους (c_i) του εγγράφου $T_ι$. Για να γίνει πιο κατανοητός ο τύπος, να πούμε ότι ο αλγόριθμος συμπεριφέρεται σαν ένας γενικός περιηγητής του ΠΙ. Προσπελάζει τα έγγραφα ακολουθώντας τους συνδέσμους. Όταν βρεθεί σε έγγραφο χωρίς συνδέσμους τότε πάει σε κάποιο άλλο τυχαία. Η πιθανότητα να βρεθούμε κάποια στιγμή στο έγγραφο P αντιπροσωπεύει την αξία του εγγράφου IR(P). Επίσης κάθε χρονική στιγμή με πιθανότητα d θα πάμε σε κάποια άλλη τυχαία σελίδα. Αυτό μπορεί να συμβεί για 2 λόγους: είτε βρεθήκαμε σε κάποια σελίδα άσχετη με το θέμα του εγγράφου είτε είδαμε αυτό που θέλαμε και πάμε σε κάτι άλλο. Ο παράγοντας d καθορίζεται μεταξύ 0.85 και 0.9.

΄Ενα άλλο θέμα είναι τα έγγραφα που δεν έχουν καθόλου εξερχόμενους συνδέσμους. Τα έγγραφα αυτά ονομάζονται *"dangling documents"* και η αξία τους υπολογίζεται στο τέλος καθώς δεν είναι καίριος παράγοντας στην αξία των άλλων εγγράφων.

Επίσης να τονίσουμε ότι αν και φαίνεται πως η εύρεση της αξίας ενός εγγράφου λόγω της αναδρομής θα κοστίσει αρκετά, πειραματικές μελέτες έχουν δείξει ότι οι αξίες για 26 εκατομμύρια έγγραφα υπολογίζονται σε μερικές ώρες σε ένα συμβατικό υπολογιστή γραφείου.

Πλήθος εξερχόμενων συνδέσμων – Forward Link Count. Εδώ θεωρούμε ότι η αξία ενός εγγράφου είναι τόσο μεγαλύτερη όσο περισσότερους εξερχόμενους υπερσυνδέσμους έχει. Παραδείγματα τέτοιων εγγράφων αποτελούν οι δικτυακοί κατάλογοι, που φυσικά είναι πολύ σημαντικοί κόμβοι του ΠΙ. Σε αντίθεση με το πλήθος εισερχόμενων συνδέσμων και το βαθμό PageRank, η αξία του εγγράφου βρίσκεται μόλις το επισκεφτούμε και δε χρειαζόμαστε συνολική πληροφορία για όλα τα έγγραφα της συλλογής. Ο συμβολισμός της είναι IF(P).

Μετρική θέσης - Location Metric. Το τελευταίο κριτήριο αξιολόγησης εξαρτά την αξία ενός εγγράφου από τη φυσική του θέση στον ΠΙ. Δηλαδή είναι λογικό ένα έγγραφο που είναι σε μικρό βάθος σε ένα δικτυακό τόπο (λίγα '/') να θεωρείται πιο σημαντικό από ένα άλλο που είναι σε αρκετά μεγάλο βάθος. Ένα άλλο κριτήριο είναι αν η διεύθυνση ενός δικτυακού τόπου τελειώνει σε com, org, edu κτλ., υποδηλώνοντας έτσι τον τύπο των εγγράφων και παράλληλα δίνοντας μια ένδειξη για το αν μας ενδιαφέρουν ή όχι. Γενικότερα, η αξία ενός εγγράφου είναι συνάρτηση της διεύθυνσης ιστού του και έτσι μπορούμε να υπολογίσουμε την αξία του πριν ακόμα το προσπελάσουμε. Ο συμβολισμός της είναι IL(P).

Υπάρχουν περιπτώσεις που είναι συμφέρον να συνδυάσουμε κάποια ή και όλα τα μέτρα σημαντικότητας που προαναφέρθησαν, έτσι ώστε να δημιουργήσουμε ακριβώς αυτό που ζητάμε. Αν, για παράδειγμα, μας ενδιαφέρουν τα έγγραφα που ανήκουν σε εταιρίες και έχουν πολλούς εισερχόμενους συνδέσμους, πρέπει να συνδυάσουμε τη BackLink count και Location metrics: $I(P)=w_1\ IL(P) + w_2\ IB(P)$, όπου τα w_1 και w_2 υποδηλώνουν το ενδιαφέρον μας για κάθε ένα από τα μέτρα.

2.9.3 Τρόποι λειτουργίας των περιηγητών

Ένας περιηγητής έχει ως στόχο να επισκεφτεί κάποιο αριθμό εγγράφων με σειρά αξίας. Δηλαδή, πρώτα να επισκεφθεί αυτό με την μεγαλύτερη αξία και να συνεχίσει

προς τα άλλα με φθίνουσα σειρά αξίας. Στην περίπτωση που δεν έχει συνολικές πληροφορίες για τα έγγραφα, που είναι και το πιθανότερο, πρέπει να "μαντέψει" την αξία των εγγράφων. Πρέπει δηλαδή να υπολογίσει τη "φαινόμενη" αξία (I'(P)) χρησιμοποιώντας κάποιο από τα μέτρα που προαναφέρθηκαν στην περιορισμένη του μορφή.

Σύμφωνα με τα παραπάνω έχουμε τρεις τρόπους λειτουργίας ενός περιηγητή:

Απλή περιήγηση - Crawl & Stop. Σύμφωνα με αυτήν την τεχνική, ο περιηγητής ξεκινά από ένα έγγραφο P_0 και επισκέπτεται Κ έγγραφα $R_1 \ldots R_k$, τα λεγόμενα "σημαντικά" έγγραφα. Για έναν ιδανικό περιηγητή θα ίσχυε ότι $I(R_1) > I(R_2) > \ldots > I(R_k)$. Στην πραγματικότητα, ο περιηγητής υπολογίζει τη φαινόμενη αξία και έτσι μόνο για Μ από τα Κ έγγραφα η πραγματική αξία τους θα είναι μεγαλύτερη από την αξία της R_k. Έτσι, η απόδοση του περιηγητή σε αυτήν την περίπτωση θα είναι $P_{CS}=(M \cdot 100)/K$, δηλαδή το ποσοστό των πραγματικά "σημαντικών" εγγράφων που επισκέπτεται (Μ) προς όλα τα έγγραφα που επισκέπτεται (Κ). Όπως περιμέναμε ο ιδανικός περιηγητής έχει απόδοση 100%.

Ας θεωρήσουμε την περίπτωση του τυχαίου περιηγητή, που επισκέπτεται εντελώς τυχαία κάποια από τα Τ έγγραφα του ΠΙ χωρίς να ελέγχει αν τα έχει επισκεφθεί ήδη. Η ανεξάρτητη πιθανότητα να επισκεφθεί ένα σημαντικό έγγραφο σε μια προσπέλαση είναι Κ/Τ. Συνεπώς αν επισκεφθεί Κ έγγραφα η απόδοσή του υπολογίζεται $P_{CSrandom}=(K^2 \cdot 100)/T$.

Περιήγηση με κατώφλι - Crawl & Stop with Threshold. Ο τρόπος είναι παρόμοιος με τον προηγούμενο. Η μόνη διαφορά είναι ότι θεωρούμε "σημαντικό" ένα έγγραφο Ρ, όταν I(P) >=G για κάποια σταθερά G. Αν θεωρήσουμε ότι τα σημαντικά έγγραφα είναι Η και ένας ιδανικός περιηγητής συγκεντρώνει Κ έγγραφα: αν K<H η απόδοση είναι $P_{ST}=(K \cdot 100)/H$, διαφορετικά $P_{ST}=100\%$. Η απόδοση δηλαδή είναι το ποσοστό των "σημαντικών" εγγράφων που έχουμε προσπελάσει όταν τελειώσει η συλλογή προς το συνολικό αριθμό των "σημαντικών" εγγράφων.

Η απόδοση του τυχαίου crawler με αυτήν την τεχνική υπολογίζεται: $P_{STrandom}=(K \cdot H \cdot 100)/T$.

Περιήγηση με περιορισμένο αποθηκευτικό χώρο - Limited Buffer Crawl. Εδώ θεωρούμε τη ρεαλιστική περίπτωση, όπου ο αποθηκευτικός χώρος δεν επαρκεί για να αποθηκεύσουμε όλα τα έγγραφα (παρά μόνο Β στο πλήθος έγγραφα), οπότε πρέπει να επιλέξουμε μια τεχνική αντικατάστασης. Ιδανικά, θα μπορούσαμε να διώχναμε το έγγραφο με το μικρότερο I(P), αλλά αυτό δεν είναι εφικτό γιατί δεν έχουμε δει ακόμα όλα τα έγγραφα. Έτσι, πρέπει να εκτιμήσουμε την αξία των εγγράφων και να διώξουμε κάποια. Στην περίπτωση που επισκεφθούμε τα Τ έγγραφα όλου του ΠΙ και θεωρήσουμε ότι έχουμε χώρο για Β έγγραφα η απόδοση θα είναι το ποσοστό των σχετικών εγγράφων που επισκεφθήκαμε στο τέλος της περιήγησης προς τον αριθμό των εγγράφων που μπορεί να κρατήσει ο αποθηκευτικός μας χώρος συνολικά. Έτσι για ένα ιδανικό περιηγητή η απόδοση είναι ίδια με πριν: αν B<H η απόδοση είναι $P_{LB}=(B \cdot 100)/H$, διαφορετικά $P_{LB}=100\%$. Για έναν τυχαίο περιηγητή $P_{LBrandom}=(B \cdot H \cdot 100)/T$.

2.10 Συμπεράσματα

Η ποιότητα του χαρακτηρισμού των εγγράφων του ιστού δεν είναι επαρκής. Η αντιμετώπιση των εγγράφων ως αδόμητα κείμενα, που δεν έχουν περαιτέρω πληροφορία, οδηγεί σε φτωχά αποτελέσματα, δυσκολεύοντας ταυτόχρονα τις διαδικασίες οργάνωσης, ανάκτησης πληροφορίας και εξαγωγής γνώσης.

Η πληροφορία που φέρουν οι σύνδεσμοι ενός εγγράφου χρησιμοποιείται σε ορισμένες προσεγγίσεις για το χαρακτηρισμό των εγγράφων. Σε καμιά περίπτωση όμως δεν εξετάζεται συλλογικά η πληροφορία όλων των εισερχόμενων ή εξερχόμενων συνδέσμων ενός εγγράφου ή μιας ομάδας εγγράφων. Μια τέτοια προσέγγιση πέραν του ότι ισχυροποιεί τους εξαγόμενους χαρακτηρισμούς μπορεί να δώσει μια νέα προοπτική στο χαρακτηρισμό υποσυνόλων εγγράφων του ΠΙ.

Η πληροφορία που υπάρχει στον Ιστό δεν ακολουθεί κάποια πρότυπα δόμησης. Αν και το τεχνολογικό υπόβαθρο για το πέρασμα στο Σημασιολογικό Ιστό υπάρχει (με τη βοήθεια των γλωσσών XML, RDF, και των προτύπων DAML+OIL, RDF-Schema) χρειάζεται μεγάλη προσπάθεια τόσο για την τυποποίηση της υπάρχουσας γνώσης (Οντολογίες, Ταξινομίες, Ορολογίες) όσο και για τη μετατροπή των υπαρχόντων εγγράφων σε έγγραφα με προφανές σημασιολογικό περιεχόμενο.

Οι υπάρχουσες διαδικασίες συλλογής, οργάνωσης, διαχείρισης και ανάκτησης των εγγράφων του ΠΙ δε λαμβάνουν υπόψη τη σημασιολογία των εγγράφων και των χαρακτηρισμών τους. Οι περισσότεροι αλγόριθμοι ομαδοποίησης εγγράφων, απάντησης ερωτήσεων και συλλογής εγγράφων βασίζονται στις λέξεις που εξάγονται από τα έγγραφα ή τους υπερσυνδέσμους. Χρησιμοποιούν μέτρα ομοιότητας που βασίζονται στη λεξική ομοιότητα μεταξύ των λέξεων και όχι στη σημασιολογική τους ομοιότητα.

3 Διαχείριση εγγράφων στο THESUS

Η ανασκόπηση των ερευνητικών προσπαθειών στον τομέα της διαχείρισης της πληροφορίας του ιστού βοήθησε στη διάγνωση των ελλείψεων των υπαρχόντων συστημάτων και των αναγκών των χρηστών του ΠΙ. Με τα συμπεράσματα που προέκυψαν, τέθηκαν οι βάσεις για μια νέα προσέγγιση του προβλήματος.

Απαραίτητη προϋπόθεση για τη σωστή διαχείριση της πληροφορίας είναι η δημιουργία μιας γλώσσας και ενός συστήματος, που επιτρέπουν τον ορισμό και τη διαχείριση υποσυνόλων του ΠΙ. Τα υποσύνολα αυτά πρέπει να εμφανίζουν κοινά χαρακτηριστικά, όπως συνδέσμους ή θεματολογία. Κατά την επεξεργασία των εγγράφων του ΠΙ δίνεται έμφαση στη χρήση των υπερσυνδέσμων για την εξαγωγή αξιόπιστων χαρακτηρισμών αλλά και στον εμπλουτισμό της εξαγόμενης λεξικής πληροφορίας με σημασιολογικά χαρακτηριστικά.

Στην παρούσα διατριβή παρουσιάζεται μια μέθοδος και το υλοποιημένο σύστημα, που διευκολύνουν τη δημιουργία συλλογών σελίδων από τον ΠΙ και επιτρέπουν τη διάκριση υποσυνόλων στις συλλογές αυτές. Τα υποσύνολα διακρίνονται με βάση τη σημασιολογική ομοιότητα των εγγράφων και τα χαρακτηριστικά της μεταξύ τους συνδεσμολογίας. Τα σημαντικότερα χαρακτηριστικά του THESUS είναι:

- Δημιουργία συλλογών ιστοσελίδων που εντάσσονται σε ένα συγκεκριμένο πεδίο ενδιαφέροντος, με χρήση μιας οντολογίας που περιγράφει το συγκεκριμένο πεδίο.
- Αυτόματος χαρακτηρισμός συνόλων ιστοσελίδων σε δύο στάδια: α) εξαγωγή λέξεων από τους εισερχόμενους συνδέσμους προς τις σελίδες του συνόλου, β) αντιστοίχιση λέξεων σε έννοιες στο σημασιολογικό επίπεδο, με χρήση οντολογίας και θησαυρού.
- Μειωμένες απαιτήσεις για αποθήκευση πληροφορίας, σε μια αποθήκη εγγράφων, καθώς για κάθε ιστοσελίδα αποθηκεύονται μόνο οι έννοιες που της προσδίδουν οι εισερχόμενοι σύνδεσμοι και όχι απαραίτητα το περιεχόμενό της,
- Οργάνωση των συλλογών σε υποσύνολα, καθένα από τα οποία περιέχει σελίδες με παρόμοια σημασιολογικά χαρακτηριστικά ή παρόμοια συνδεσμολογία.

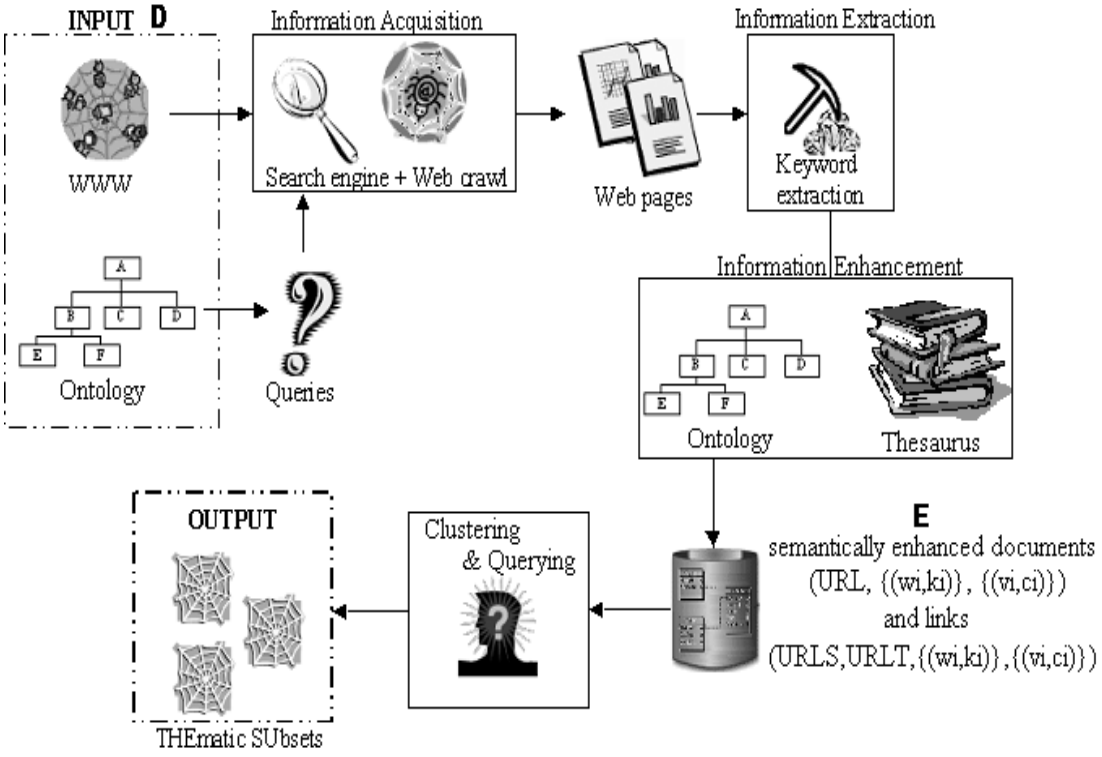
Σημείωση: Το όνομα του συστήματος είναι THESUS. Με τον όρο THESU αναφερόμαστε σε ένα θεματικό υποσύνολο. Με τους όρους σελίδα ή έγγραφο αναφερόμαστε στη βασική οντότητα του συστήματός μας, που αποτελεί έναν κόμβο στο γράφο του ΠΙ.

Η μεθοδολογία δημιουργίας και διαχείρισης ενός THESU, η αρχιτεκτονική του συστήματος και τα βασικά χαρακτηριστικά των διαφόρων τμημάτων του συστήματος παρουσιάζονται στη συνέχεια. Παρουσιάζονται επίσης παραδείγματα των προσφερόμενων υπηρεσιών, που επιτρέπουν το χαρακτηρισμό και την οργάνωση συλλογών ιστοσελίδων σε θεματικές περιοχές με ένα διαφορετικό τρόπο απ' ό,τι γίνεται μέχρι σήμερα.

Το THESUS ΔΕΝ είναι μια μηχανή αναζήτησης αλλά ένα σύστημα που εμπλουτίζει τη σημασιολογική οργάνωση ενός συνόλου εγγράφων και διευκολύνει τους χρήστες να ψάξουν μέσα σε αυτές. Είναι μια προσωπική υπηρεσία που διευκολύνει τους χρήστες στη δημιουργία θεματικών συλλογών από έγγραφα και στην υποβολή

ερωτήσεων σε αυτές. Βέβαια, οι έννοιες που αναπτύσσονται στο παρόν κείμενο μπορούν να επεκταθούν σε ολόκληρο τον Ιστό.

Πιο συγκεκριμένα στα υποκεφάλαια που ακολουθούν παρουσιάζουμε "μια μέθοδο για τη συγκέντρωση, περιγραφή, οργάνωση και διαχείριση εγγράφων με βάση σημασιολογικά χαρακτηριστικά που πηγάζουν από μια θεματική οντολογία". Η μέθοδος αυτή παρουσιάζεται σχηματικά στο ακόλουθο διάγραμμα (Σχήμα 7).



Σχήμα 7. Η μέθοδος συγκέντρωσης, χαρακτηρισμού και διαχείρισης εγγράφων στο THESUS

3.1 Συλλογή εγγράφων από τον Ιστό

Απαραίτητη προϋπόθεση, για ένα σύστημα συλλογής εγγράφων του ιστού, είναι η ύπαρξη ενός ή περισσοτέρων εγγράφων από τα οποία θα ξεκινήσει η συλλογή. Όπως αναφέρθηκε και στην ενότητα 2, τα προγράμματα αναλύουν τους συνδέσμους των εγγράφων της αρχικής συλλογής, ακολουθούν όλους (γενικοί περιηγητές) ή κάποιους από αυτούς (επιλεκτικοί και εστιασμένοι περιηγητές), εφόσον δεν τους έχουν ήδη επισκεφθεί, συλλέγουν και αναλύουν με τη σειρά τους τα έγγραφα στα οποία δείχνουν οι επιλεγμένοι σύνδεσμοι.

Η μέθοδος συλλογής εγγράφων που προτείνεται στα πλαίσια του THESUS θα μπορούσε να καταταγεί στις μεθόδους επιλεκτικής ή εστιασμένης συλλογής. Ο **σκοπός** ενός επιλεκτικού περιηγητή εγγράφων είναι να αναζητήσει επιλεκτικά έγγραφα που ικανοποιούν ορισμένα κριτήρια. Τα κριτήρια συνήθως στηρίζονται, είτε

σε λέξεις κλειδιά που πρέπει να υπάρχουν στους συνδέσμους ή στα έγγραφα στα οποία αυτοί δείχνουν, είτε στη θεματική ομοιότητα των εγγράφων με άλλα έγγραφα υποδείγματα. Ο επιλεκτικός περιηγητής αναζητά τους συνδέσμους που είναι όσο το δυνατό πιο σχετικοί και αγνοεί τους υπόλοιπους. Δημιουργεί έτσι μια λίστα με υποψήφιους συνδέσμους τους οποίους και επισκέπτεται σε επόμενο βήμα, αφού πρώτα επιβεβαιώσει ότι δεν τους έχει ήδη επισκεφθεί.

Σε κάθε βήμα εξάγεται από τους συνδέσμους πληροφορία για το έγγραφο στο οποίο αναφέρονται. Η πληροφορία που εξάγεται χρησιμοποιείται στη συνέχεια για να αξιολογήσει την ομοιότητα του εγγράφου με το θέμα. Για να επιτευχθεί μια καλά προσανατολισμένη διαδικασία συλλογής, απαιτούνται δύο βήματα:

- Το πρώτο βήμα αφορά την ανάλυση των χαρακτηριστικών του συνδέσμου. Τα χαρακτηριστικά μπορεί να αφορούν τη διεύθυνση στην οποία δείχνει, τις λέξεις που χρησιμοποιεί, αν πηγάζει από κείμενο ή εικόνα, τον τύπο ή το όνομα του αρχείου στο οποίο αναφέρεται κτλ.
- Το δεύτερο βήμα αφορά την αξιολόγηση της ομοιότητας του δεικτοδοτούμενου εγγράφου ως προς τα κριτήρια με τα οποία γίνεται η συλλογή. Στο βήμα αυτό υπολογίζεται η αξία του εγγράφου αλλά και ο βαθμός «κεντρικότητας» του υποδεικνυόμενου εγγράφου (*centrality*) [CB+99], δηλαδή σε τι βαθμό το έγγραφο εξυπηρετεί το σκοπό της αναζήτησης. Ουσιαστικά καθορίζεται αν το έγγραφο είναι σημαντικός ***π-κόμβος*** (hub), με πολλούς εξερχόμενους συνδέσμους.

Αν το σύστημα διατηρεί πληροφορίες για το ιστορικό του εγγράφου (πότε το επισκεφθήκαμε τελευταία, κάθε πότε αλλάζει κτλ), μπορεί επίσης να καθορίσει και την προτεραιότητα με την οποία θα το ξανα-επισκεφθεί στο μέλλον.

3.1.1 Καθορισμός αρχικών εγγράφων

Στην περίπτωση του THESUS ενδιαφερόμαστε να συλλέξουμε έγγραφα που σχετίζονται με κάποιο συγκεκριμένο θέμα. Το δεδομένο που έχουμε είναι η οντολογία που περιγράφει το συγκεκριμένο θέμα. Η οντολογία μας θα αποτελέσει το κριτήριο καθορισμού της ομοιότητας ενός εγγράφου κατά τη διαδικασία συλλογής. Για να είμαστε σε θέση να ξεκινήσουμε τη συλλογή, χρειαζόμαστε ένα ή περισσότερα έγγραφα με μεγάλο αριθμό εξερχόμενων συνδέσμων για να αυξήσουμε την πιθανότητα να βρούμε "σχετικούς" συνδέσμους μέσα σε αυτά.

Ένας τρόπος για να συγκεντρώσουμε τα έγγραφα, αν οι χρήστες δεν είναι σε θέση να τα παρέχουν, είναι να ξεκινήσουμε από τις κεντρικές ιστοσελίδες γνωστών δικτυακών καταλόγων που συγκεντρώνουν μεγάλο πλήθος συνδέσμων σε μια ευρύτατη θεματολογία [Dmo],[Goo],[Yah] και να επιλέξουμε να ακολουθήσουμε τους εξερχόμενους συνδέσμους που «σχετίζονται» με το θέμα της οντολογίας, για ορισμένα επίπεδα του γράφου (περίπτωση α, ενότητα 3.1.2)

Ένας εναλλακτικός τρόπος καθορισμού του αρχικού συνόλου εγγράφων είναι με χρήση μιας μηχανής αναζήτησης του ΠΙ (περίπτωση β, ενότητα 3.1.2). Αν θεωρήσουμε ως μόνο δεδομένο ότι η οντολογία Ο που περιγράφει τα ενδιαφέροντα των χρηστών διαθέτει τη σχέση ειδίκευσης-γενίκευσης (ISA) ανάμεσα στις έννοιες, τότε έχουμε μια ιεραρχία εννοιών (Ταξινομία). Η ιεραρχία ξεκινά από τις πιο γενικές έννοιες και διακλαδίζεται προς τις πιο ειδικές. Το σύστημα δημιουργεί όλα τα δυνατά μονοπάτια, από τη ρίζα προς κάθε κόμβο της οντολογίας. Στη συνέχεια, ρωτά μια

μηχανή αναζήτησης για κάθε ένα μονοπάτι εννοιών (κόμβος$_1$ AND κόμβος$_2$ AND … AND κόμβος$_K$) και συλλέγει τις πρώτες απαντήσεις που επιστρέφει η μηχανή σε κάθε ερώτηση.

Αν η οντολογία περιέχει Ν έννοιες, τότε θα δημιουργηθούν Ν μονοπάτια και αντίστοιχα Ν ερωτήσεις προς τη μηχανή αναζήτησης. Αν η μηχανή αναζήτησης επιστρέφει το πολύ Μ απαντήσεις σε κάθε ερώτηση, τότε ο μέγιστος συνολικός αριθμός διευθύνσεων σελίδων (σύνολο D) θα είναι MxN. Τα αποτελέσματα παρουσιάζουν μεγάλη επικάλυψη και για το λόγο αυτό ο τελικός αριθμός διευθύνσεων σελίδων στο D μετά την αφαίρεση όσων σελίδων εμφανίζονται πολλές φορές θα είναι κατά πολύ μικρότερος του MxN.

3.1.2 Κριτήρια επιλογής συνδέσμων

Το στάδιο της εξαγωγής χαρακτηριστικών από τους συνδέσμους ενός εγγράφου επικεντρώνεται στην εξαγωγή λέξεων από τους συνδέσμους, με μια μέθοδο που περιγράφεται αναλυτικότερα στην ενότητα 3.2. Για κάθε σύνδεσμο εξάγονται λέξεις και μετράται η συχνότητα εμφάνισής τους στο σύνδεσμο [R79]: όσο περισσότερες, από τις λέξεις που εξήχθησαν, σχετίζονται με την οντολογία τόσο μεγαλύτερη είναι η αξία του συνδέσμου. Ο βαθμός ομοιότητας μιας λέξης με την οντολογία υπολογίζεται με χρήση ενός θησαυρού (Wordnet). Αν η λέξη περιέχεται ακριβώς στη θεματική οντολογία τότε ο βαθμός ομοιότητάς της είναι μέγιστος (ίσος με 1), αν στο θησαυρό υπάρχει μια συνώνυμη, γενικότερη ή ειδικότερη λέξη της λέξης που εξήχθη, τότε ο βαθμός ομοιότητας της λέξης είναι μικρότερος. Αν η λέξη δε βρίσκεται στην ίδια ιεραρχία με κάποια από τις λέξεις της οντολογίας τότε η ομοιότητά της είναι ελάχιστη και ίση με 0.

Καθορισμός αρχικού συνόλου

Θεωρούμε ένα αρχικό σύνολο σελίδων D = $\{d_i\}$, που σχετίζονται με κάποιο θεματικό πεδίο, και μια οντολογία Ο στο ίδιο πεδίο (π.χ. Τέχνες, Τεχνολογία, κτλ). Το υποσύστημα θεματικής συλλογής παράγει το βασικό σύνολο του THESUS με δύο τρόπους, ανάλογα με την ύπαρξη ή όχι ενός αρχικού συνόλου σελίδων D.

α. Δεδομένο αρχικό σύνολο D

Αν δοθεί το σύνολο D, η διαδικασία συλλογής ξεκινά από αυτό το σύνολο και λειτουργεί αυτόνομα, χωρίς να χρησιμοποιεί υπηρεσίες άλλων μηχανών αναζήτησης του ΙΙΙ. Το σύστημα συλλογής επισκέπτεται όλα τα έγγραφα (διευθύνσεις ιστού) του D, εξάγει πληροφορία από τους εξερχόμενους συνδέσμους τους και κρατά μόνο τους υπερσυνδέσμους που χαρακτηρίζονται από τουλάχιστον μία έννοια της οντολογίας. Στην περίπτωση αυτή εντάσσεται και η περίπτωση που αναφέρθηκε προηγούμενα, στην οποία ξεκινάμε από τις πρώτες σελίδες ορισμένων δικτυακών καταλόγων.

Λέμε ότι ένας σύνδεσμος χαρακτηρίζεται από μια έννοια, όταν το υπερκείμενο γύρω και μέσα στο σύνδεσμο περιέχει μια λέξη που αντιστοιχίζεται στη συγκεκριμένη έννοια της οντολογίας (η ομοιότητά της με την έννοια είναι μεγαλύτερη από ένα προκαθορισμένο κατώφλι). Όσο ψηλότερο είναι το κατώφλι τόσο λιγότερες λέξεις αντιστοιχίζονται σε έννοιες της οντολογίας. Οι στόχοι των συνδέσμων που χαρακτηρίζονται από σχετικές έννοιες προτείνονται για το επόμενο βήμα συλλογής. Όταν η διαδικασία συλλογής προτείνει μια σελίδα d_i που ήδη υπάρχει στο σύνολο (που ήδη την έχουμε επισκεφθεί), η διαδικασία σταματά για τη σελίδα αυτή, για να

αποφευχθούν επαναλήψεις. Η διαδικασία σταματά πλήρως μετά από συγκεκριμένο αριθμό βημάτων ή όταν όλες οι σελίδες που προτείνονται για το επόμενο βήμα έχουν ήδη προσπελαστεί.

Λόγω της χαλαρής συνδεσιμότητας των σελίδων του ΠΙ (μόνο 28% των σελίδων είναι πυκνά συνδεδεμένο σύμφωνα με το [BK+00]) είναι πιθανό με μια μικρή οντολογία (με λιγότερο από 20 έννοιες) και ένα μικρό αρχικό σύνολο, η διαδικασία συλλογής να σταματήσει σε μερικά βήματα. Από την άλλη μια ευρεία οντολογία (με μερικές εκατοντάδες έννοιες) και ένα μεγαλύτερο αρχικό σύνολο, θα χρειαστεί περισσότερα βήματα για να ολοκληρωθεί.

β. Χωρίς αρχικό σύνολο D

Αν το D δεν δοθεί, θεωρούμε ότι δεν υπάρχει προηγούμενη γνώση στο συγκεκριμένο θέμα και η αναζήτηση, για έγγραφα που πιθανόν σχετίζονται με το θέμα, ξεκινά από την αρχή με χρήση μιας μηχανής αναζήτησης. Με αυτόματο τρόπο κατασκευάζονται ερωτήσεις προς τη μηχανή αναζήτησης για έγγραφα που περιέχουν έναν ή περισσότερους όρους της οντολογίας. Σχηματίζονται αρχικά όλα τα μονοπάτια όρων της οντολογίας (σύνολο S). Στη συνέχεια για κάθε μονοπάτι s ($s \in S$) δημιουργείται η αντίστοιχη ερώτηση σε μια μηχανή αναζήτησης. Η ένωση των διαφόρων συνόλων απαντήσεων αποτελεί το αρχικό σύνολο D για το σύστημα συλλογής εγγράφων. Η διαδικασία συνεχίζεται όπως στην πρώτη περίπτωση.

Συλλογή εγγράφων

Κατά τη διαδικασία συλλογής εγγράφων από τον Ιστό, από κάθε σελίδα του αρχικού συνόλου εξάγονται όλοι οι υπερσύνδεσμοι που αυτή περιέχει και για κάθε υπερσύνδεσμο, εξάγονται μία ή περισσότερες λέξεις που εντοπίζονται στο υπερκείμενο (hypertext) του συνδέσμου, αλλά και σε μια περιορισμένη περιοχή γύρω από αυτό. Οι λέξεις αυτές αποτελούν τη λεξική περιγραφή που φέρει ο υπερσύνδεσμος.

Η λεξική περιγραφή αντιστοιχίζεται σε ένα σύνολο εννοιών της οντολογίας, με χρήση της γνώσης που μεταφέρει η οντολογία και ένας λεξιλογικός θησαυρός. Το σύνολο των εννοιών αποτελεί τη σημασιολογική περιγραφή του υπερσυνδέσμου. Οι σελίδες στόχοι των συνδέσμων για τους οποίους έχει προκύψει σημασιολογική περιγραφή εξετάζονται και όσοι θεωρούνται σχετικοί με την οντολογία σημειώνονται για επίσκεψη. Οι σελίδες-στόχοι που σημειώθηκαν για επίσκεψη στο πρώτο βήμα αποτελούν τη βάση για το επόμενο βήμα συλλογής σελίδων.

Το σύστημα συλλογής επισκέπτεται και επεξεργάζεται διαδοχικά τα έγγραφα, εξάγει τους υπερσυνδέσμους και τις λεξικές περιγραφές που αυτοί περιέχουν, παράγει τις σημασιολογικές περιγραφές και σημειώνει τους στόχους του επόμενου βήματος. Παράλληλα αποθηκεύει σε ξεχωριστά XML αρχεία τη λεξική και σημασιολογική πληροφορία των συνδέσμων κάθε εγγράφου. Αν ο στόχος ενός συνδέσμου έχει ήδη αποθηκευθεί, η επιπλέον εννοιολογική περιγραφή προστίθεται στην υπάρχουσα, αυξάνοντας τη βαρύτητα της περιγραφής, ενώ δε σημειώνεται για επίσκεψη.

Η διαδικασία επαναλαμβάνεται, μέχρις ότου κανένας από τους στόχους ενός βήματος δεν έχει σημειωθεί για επίσκεψη στο επόμενο βήμα. Το αποτέλεσμα της διαδικασίας είναι η δημιουργία μιας συλλογής XML εγγράφων που περιέχουν τη σημασιολογική και λεξική πληροφορία των υπερσυνδέσμων των αντίστοιχων εγγράφων του ΠΙ. Οι

λεξικοί και σημασιολογικοί χαρακτηρισμοί των εισερχόμενων και εξερχόμενων συνδέσμων που έχουν προκύψει για τα έγγραφα αυτά είναι αποθηκευμένοι σε XML έγγραφα. Τα XML έγγραφα έχουν καθορισμένη δομή και μπορούν είτε να υπάρξουν ως έχουν συνοδεύοντας τα πραγματικά έγγραφα ως έγγραφα μεταδεδομένων (ως XML νησιά δεδομένων [XDI] – data islands – εντός των HTML εγγράφων, ιδανική λύση για την περίπτωση του σημασιολογικού ιστού), είτε να αποτελέσουν μια συλλογή XML εγγράφων για την υποβολή απλών ερωτήσεων (π.χ. εύρεση συνδέσμων με συγκεκριμένους χαρακτηρισμούς), είτε να αποθηκευθούν σε μια σχεσιακή βάση δεδομένων επιτρέποντας σύνθετες ερωτήσεις και προηγμένη επεξεργασία δεδομένων (π.χ. συσταδοποίηση εγγράφων). Στην περίπτωση του THESUS τα XML έγγραφα αποθηκεύονται σε μια σχεσιακή βάση δεδομένων η οποία εξυπηρετεί όλες τις αναζητήσεις και την περαιτέρω επεξεργασία.

3.2 Χαρακτηρισμός εγγράφων

Για να εμπλουτίσουμε την πληροφορία για το σύνολο εγγράφων, επισκεπτόμαστε σε κάθε βήμα της συλλογής τα υποψήφια έγγραφα και συλλέγουμε την πληροφορία των υπερσυνδέσμων. Το πρώτο βήμα είναι η εξαγωγή λέξεων από τους συνδέσμους του κάθε εγγράφου. Η εξαγωγή των σωστών λέξεων από τους συνδέσμους είναι πολύ σημαντική.

3.2.1 Εξαγωγή λεξικών χαρακτηρισμών

Για κάθε υπερσύνδεσμο στα έγγραφα της συλλογής εξάγεται πληροφορία: α) από τα περιεχόμενα του συνδέσμου (π.χ. η συμβολοσειρά ανάμεσα στις ετικέτες <a> και) και β) από δύο συμβολοσειρές (των 100 χαρακτήρων), μια πριν την αρχική ετικέτα του συνδέσμου και μια μετά την τελική ετικέτα. Το "παράθυρο" περιορίζεται από την εμφάνιση συγκεκριμένων ετικετών της HTML όπως <a>, <li>, <tr>, <td> κτλ., που υποδηλώνουν το λογικό τέλος της περιοχής γύρω από το σύνδεσμο. Τελικά εξάγονται κατά μέσο όρο 5 λέξεις ανά σύνδεσμο. Η τιμή για το εύρος του παραθύρου επιλέχθηκε έπειτα από πειράματα, ώστε ο μέσο αριθμός λέξεων ανά σύνδεσμο να είναι πέντε [PW00]. Ο αλγόριθμος επεξεργασίας των συνδέσμων κάθε εγγράφου συνοψίζεται στα εξής:

Αλγόριθμος Επεξεργασίας Υπερσυνδέσμων

```
1: hyperlink offsets {(bi,ei)} ← getOffsets(source)
2: For each pair (bi,ei) in hyperlink offsets
3:hyperlink_text = substring(source, bi, ei)
4:img_offsets {(imbj, imej )}← getImgOffsets (hyperlink_text)
5:  For each pair (imbj, imej )
6:  img= substring(hyperlink_text, imbj, imej)
7:   altText=getAlt(img)
8:   getKeys (altText)
9: hyperlink_text = removetags(hyperlink_text)
10: getkeys (hyperlink_text)
11:pre-hyperlink_text = substring(source, bi -100, bi)
12: pre-hyperlink_text = trimleft (pre-hyperlink_text, specials)
13: pre-hyperlink_text = removetags (pre-hyperlink_text)
14:  getkeys (pre-hyperlink_text)
15: post-hyperlink_text = substring(source, ei, ei +100)
16:  post-hyperlink_text = trimright (post-hyperlink_text, specials)
17:  post-hyperlink_text = removetags (post-hyperlink_text)
18:    getkeys (post-hyperlink_text)
```

Η μέθοδος ***getOffsets*** (γραμμή 1) επιστρέφει ζεύγη ακεραίων (b_i,e_i) που αντιστοιχούν στις θέσεις εμφάνισης των ετικετών ανοίγματος και κλεισίματος των υπερσυνδέσμων στον HTML κώδικα του εγγράφου. Παρόμοια, η μέθοδος ***getImgOffsets*** (γραμμή 4) επιστρέφει τις θέσεις εμφάνισης των ετικετών που αντιστοιχούν σε εικόνες (<img>) εντός του κειμένου του υπερσυνδέσμου. Η μέθοδος ***substring***(source, start, end) επιστρέφει το τμήμα του *source* μεταξύ των θέσεων *start* και *end*. Οι μέθοδοι ***trimleft*** και ***trimright*** (γραμμές 12 και 16) περιορίζουν τις περιοχές κειμένου πριν και μετά το σύνδεσμο αντίστοιχα στην πρώτη εμφάνιση κάποιας από τις ειδικές ετικέτες που υποδηλώνουν αλλαγή περιοχής συνδέσμου (γραμμή πίνακα, στήλη πίνακα, άλλος σύνδεσμος, αντικείμενο λίστας κτλ.). Η μέθοδος ***removetags*** (γραμμές 9,13,17) αφαιρεί όλες τις ετικέτες HTML από ένα κείμενο (hyperlink_text). Τελικά, η μέθοδος ***getkeys*** αφαιρεί τα σημεία στίξης και τις κοινές λέξεις (stopwords), εξάγει λέξεις από το υπόλοιπο κείμενο και μετρά αριθμούς εμφανίσεων κάθε λέξης.

Στο απόσπασμα κώδικα HTML που ακολουθεί οι λέξεις εξάγονται από ένα σύνδεσμο προς τη σελίδα του ερευνητικού κέντρου CERN:

```
<A HREF="http://cts-hp.cts.iisc.ernet.in/"> Centre for Theoretical
Studies</A><D>Indian Institute of Science, Bangalore, India<DT></LI> <LI><A
HREF="http://www.cern.ch/">CERN</A><DD> European Organization for Nuclear
Research/European Laboratory for Particle Physics, Geneva, Switzerland (See also
<A
```

Για το σύνδεσμο εξάγουμε φράσεις:

- Από την περιοχή εντός του συνδέσμους (έντονα σκιασμένο κείμενο με λευκά γράμματα) παίρνουμε τη φράση "CERN",
- Από τους 100 χαρακτήρες που προηγούνται του συνδέσμου (ελαφρώς σκιασμένη περιοχή). Η περιοχή περιορίζεται στη θέση που εμφανίζεται η ετικέτα <LI> και συνεπώς δεν απομένει τίποτα.
- Από τους 100 χαρακτήρες που ακολουθούν το σύνδεσμο (ελαφρώς σκιασμένη περιοχή). Η τελευταία λέξη (“Swi”) αγνοείται γιατί είναι πιθανά ημιτελής, η ετικέτα <DD> αφαιρείται, η λέξη "for" αφαιρείται ως συνηθισμένη. Απομένει η φράση: "European Organization Nuclear Research/European Laboratory Particle Physics, Geneva".

Έτσι για το σύνδεσμο εξάγονται τα εξής ζεύγη λέξεων και αριθμών εμφανίσεων αντίστοιχα: {(1,cern), (2,european), (1,organization), (1,nuclear), (1,research), (1,laboratory), (1,particle), (1,physics), (1,geneva)}.

Συνεπώς για κάθε σύνδεσμο h_j των εγγράφων της συλλογής εξάγουμε μια *λεξική περιγραφή*, ένα σύνολο από λέξεις με τον αντίστοιχο αριθμό εμφανίσεων:

$$\textit{Λεξική περιγραφή συνδέσμου } K_{hj} = \{(w_{jx} , k_{jx})\}$$

όπου k_{jx} μια λέξη που εμφανίστηκε τουλάχιστον μία φορά στο σύνδεσμο και w_{jx} ένας ακέραιος που αντιστοιχεί στον αριθμό εμφανίσεων της λέξης k_{jx} στο κείμενο του υπερσυνδέσμου h_j.

Πρέπει εδώ να τονιστεί η σημαντικότητα του αριθμού εμφανίσεων μιας λέξης στο κείμενο ενός υπερσυνδέσμου, ιδιαίτερα σε συνδυασμό με το πλήθος των υπολοίπων λέξεων του συνδέσμου. Όπως θα εξηγηθεί και στη συνέχεια ο αριθμός εμφανίσεων της λέξης λαμβάνεται υπόψη κατά τον υπολογισμό της σημαντικότητας της λέξης

στην περιγραφή ενός συνδέσμου, και κατά συνέπεια στον υπολογισμό της σημαντικότητας της έννοιας της οντολογίας στην οποία αντιστοιχίζεται η λέξη αυτή.

Στην περίπτωση των εγγράφων της συλλογής χρησιμοποιούμε τους εισερχόμενους συνδέσμους κάθε εγγράφου για να εξάγουμε μια συλλογική περιγραφή. Η *συλλογική λεξική περιγραφή* για κάθε έγγραφο d_i προκύπτει από την ένωση των συνόλων των λέξεων $\{k_{jx}\}$ που αντιστοιχούν στην λεξική περιγραφή κάθε εισερχόμενου συνδέσμου h_j και τους αντίστοιχους αριθμούς εμφανίσεων.

Συλλογική λεξική περιγραφή εγγράφου $K_{di} = \{(w_{iy}, k_{iy})\}$

Ο αριθμός εμφανίσεων w_{iy} μιας λέξης k_{iy} στη συλλογική λεξική περιγραφή ενός εγγράφου d_i είναι ένας ακέραιος αριθμός που αντιστοιχεί στο πλήθος των εισερχόμενων συνδέσμων που χρησιμοποιούν τη λέξη k_{iy} για να χαρακτηρίσουν το έγγραφο d_i. Όπως παρουσιάζεται και στην ενότητα των πειραματικών αποτελεσμάτων αυτός ο τρόπος υπολογισμού του w_{iy} είναι πιο αξιόπιστος από την άθροιση του αριθμού εμφανίσεων της κάθε λέξης σε κάθε σύνδεσμο καθώς επηρεάζεται δυσκολότερα από συνδέσμους που εσκεμμένα χρησιμοποιούν πολλές φορές την ίδια λέξη για να αυξήσουν τη σημαντικότητά της.

3.2.2 Εμπλουτισμός χαρακτηρισμών

Θεωρούμε ότι υπάρχει μια οντολογία O που αντιπροσωπεύει τα σημασιολογικά χαρακτηριστικά ενός πεδίου ενδιαφέροντος. Για κάθε έγγραφο d_i της συλλογής το σύνολο λέξεων K_{di} αντιστοιχίζεται σε ένα σύνολο από όρους $\{c_j\}$ της οντολογίας O, με τη χρήση ενός θησαυρού, στην περίπτωσή μας του Wordnet. Χρησιμοποιούμε ένα μέτρο ομοιότητας, αυτό των Wu και Palmer (βλ. ενότητα 2.8.3) για να υπολογίσουμε την ομοιότητα μεταξύ λέξεων και κατηγοριών. Το μέτρο προϋποθέτει ότι λέξεις και όροι της οντολογίας ανήκουν σε μια ιεραρχία. Η χρήση του Wordnet επιτρέπει τη χρήση του μέτρου Wu και Palmer για όσες λέξεις και όρους της οντολογίας περιέχει.

Το αποτέλεσμα αυτής της διαδικασίας είναι κάθε έγγραφο να εμπλουτιστεί με:

- Λέξεις και βάρη (που αντιστοιχούν στον αριθμό εισερχόμενων συνδέσμων προς το έγγραφο που περιέχουν κάθε λέξη)
- Όρους/κατηγορίες της οντολογίας με τις οποίες θεωρούμε ότι σχετίζεται το έγγραφο και τα αντίστοιχα βάρη.

Χρησιμοποιούμε την ακόλουθη σημειολογία για να ορίσουμε τα ***εμπλουτισμένα έγγραφα***:

Ορισμός: εμπλουτισμένο έγγραφο ονομάζουμε την τριάδα (Doc, K, C), όπου Doc είναι η ταυτότητα του εγγράφου (π.χ. η διεύθυνση ιστού του), K (αντίστοιχα C) είναι το σύνολο των ζευγών $\{(w_i,k_i)\}$ (αντίστοιχα $\{(v_j,c_j)\}$) των λέξεων με τα βάρη (αντίστοιχα κατηγοριών με βάρη) που καθορίζουν το έγγραφο. Το w_i είναι ακέραιος (αντίστοιχα το v_j είναι πραγματικός), το k_i (αντίστοιχα c_j) είναι συμβολοσειρά). Να σημειώσουμε ότι τα w_i και v_j δεν είναι απαραίτητα ίσα όταν i=j.

Το βάρος v_j είναι πραγματικός αριθμός στο διάστημα [0;1]. Η τιμή 1 δείχνει απόλυτη σχέση, το 0 καμία σχέση. Μια τιμή 0.2 δείχνει μικρή σχέση.

Σημείωση: αυτές οι λέξεις αποτελούν θετικό χαρακτηρισμό ενός εγγράφου. Δεν χαρακτηρίζουμε αρνητικά ένα έγγραφο (με λέξεις που είναι απόλυτα άσχετες ή αντίθετες με το περιεχόμενό του).

Προκειμένου να παρασχεθούν οι πιο αξιόπιστες λεξικές περιγραφές για τα έγγραφα ΠΙ, στηριζόμαστε περισσότερο στα δεδομένα που προέρχονται από τους εισερχόμενους συνδέσμους κάθε εγγράφου και λιγότερο στο περιεχόμενο του εγγράφου. Οι εισερχόμενοι σύνδεσμοι ενός εγγράφου δημιουργούνται σε άλλα έγγραφα και συνήθως από άλλους συντάκτες. Οι εισερχόμενοι σύνδεσμοι είναι προτάσεις για τους χρήστες ΠΙ και αμερόληπτες ενδείξεις της ποιότητας και των περιεχομένων των εγγράφων. Είναι δυσκολότερο για τους συντάκτες ενός εγγράφου να επηρεάσουν τους χαρακτηρισμούς των εισερχόμενων συνδέσμων, ειδικά όταν αυτοί εξετάζονται και επεξεργάζονται με τρόπο συλλογικό.

Για παράδειγμα, οι συντάκτες ενός δικτυακού τόπου μπορούν να δημιουργήσουν αναφορές στις άλλες σελίδες μέσα στην περιοχή τους και να επηρεάσουν τις περιγραφές που παράγονται. Η ισχύς αυτών των περιγραφών μειώνεται σημαντικά όταν προστίθενται οι χαρακτηρισμοί από εξωτερικές αναφορές, από έγγραφα άλλων δικτυακών τόπων, και οι εισερχόμενοι σύνδεσμοι επεξεργάζονται στο σύνολό τους.

Προκειμένου να βρεθούν όλοι οι εισερχόμενοι σύνδεσμοι ενός εγγράφου του ΠΙ, πρέπει ιδανικά να επεξεργαστούμε τον πλήρη γράφο του ΠΙ. Εναλλακτικά, μπορούμε να χρησιμοποιήσουμε μια "υπηρεσία εισερχόμενων συνδέσμων", όπως αυτές που παρέχουν οι μηχανές αναζήτησης (π.χ το Google στη διεύθυνση http://www.google.com/advanced_search), το οποίο επιστρέφει όλα τα έγγραφα που δείχνουν ένα δεδομένο έγγραφο. Ο μόνος περιορισμός στη χρήση αυτής της υπηρεσίας είναι ότι επιστρέφει ένα μέγιστο αριθμό 100 εισερχόμενων συνδέσμων για κάθε έγγραφο ΠΙ.

3.3 Συσταδοποίηση (αλγόριθμοι clustering)

Στη φάση αυτή το υποσύστημα συσταδοποίησης επεξεργάζεται τις περιγραφές των εγγράφων. Για να λειτουργήσει ο αλγόριθμος συσταδοποίησης εγγράφων χρειάζεται ένα μέτρο ομοιότητας μεταξύ εγγράφων. Η σχετική έρευνα που έχει γίνει στο πεδίο αυτό συνοψίζεται στο [HB+01]. Μια βασική συνεισφορά του THESUS στο σημείο αυτό είναι η χρήση ενός νέου μέτρου ομοιότητας ανάμεσα στα μέλη δύο συνόλων (όροι της οντολογίας που χαρακτηρίζουν δύο έγγραφα) για τον πληρέστερο υπολογισμό της ομοιότητας των συνόλων (ομοιότητα εγγράφων).

3.3.1 Μέτρο ομοιότητας

Το μέτρο ομοιότητας χρησιμοποιείται κατά τη συσταδοποίηση των εγγράφων αλλά και κατά την απάντηση ερωτήσεων. Η συχνή χρήση του μέτρου, απαιτεί ο υπολογισμός του να είναι γρήγορος ακόμη και σε μεγάλα σύνολα εγγράφων. Για το λόγο αυτό, λαμβάνουμε υπόψη την πολυπλοκότητα υπολογισμού της ομοιότητας και φροντίζουμε να είναι ανεξάρτητη από το πλήθος των εγγράφων στη βάση.

Δεν πρέπει να ξεχνάμε ότι πρόκειται για ομοιότητα μεταξύ συνόλων λέξεων με βάρη και όχι απλά μεταξύ λέξεων. Η έρευνα που έχει γίνει ως σήμερα για τον υπολογισμό ομοιότητας μεταξύ συνόλων στοιχείων ενός χώρου χωρίς διάταξη αλλά με μέτρο ομοιότητας είναι περιορισμένη (βλ. [EM97], [N87]).

Το μέτρο ομοιότητας που εφαρμόζεται στο THESUS είναι μια επέκταση του μέτρου των Wu και Palmer και αναφέρεται στη συνέχεια. Το μέτρο είναι μια προσέγγιση ενός ακριβέστερου μέτρου ομοιότητας, και συμπεριφέρεται σωστά στις περισσότερες περιπτώσεις. Μια λεπτομερής ανάλυση των ιδιοτήτων του μέτρου παρουσιάζονται στην ενότητα 5.2.

3.3.2 Αλγόριθμοι συσταδοποίησης

Στην περίπτωση μιας συλλογής εγγράφων του ιστού δεν υπάρχει καμία πρότερη γνώση για τον τρόπο με τον οποίο συσταδοποιούνται τα έγγραφα της συλλογής, για το πλήθος των ομάδων στις οποίες πρέπει να χωριστεί η συλλογή ή για το πλήθος των εγγράφων σε κάθε ομάδα. Επίσης δεν υπάρχει κάποια διάταξη των εγγράφων στο χώρο, και το μόνο που γνωρίζουμε είναι οι ομοιότητες του κάθε εγγράφου με όλα τα υπόλοιπα [GR+99].

Για τους λόγους αυτούς επιλέγουμε να χρησιμοποιήσουμε αλγορίθμους συσταδοποίησης και όχι κατηγοριοποίησης. Οι αλγόριθμοι συσταδοποίησης που απαιτούν ως είσοδο των αριθμό των συστάδων στις οποίες θα δημιουργηθούν (π.χ. K-Means [Q67]) απορρίπτονται, καθώς δεν μπορούμε εκ των προτέρων να γνωρίζουμε τον αριθμό των διαφορετικών συστάδων εγγράφων στη συλλογή. Αλγόριθμοι που βασίζονται στην πυκνότητα των παραγόμενων συστάδων ([EK+96], [AB+99], [HK98]) είναι προτιμότεροι, καθώς μας ενδιαφέρει οι τελικώς παραγόμενες συστάδες να περιέχουν αρκετά όμοια μεταξύ τους έγγραφα.

Ο βασικός αλγόριθμος που χρησιμοποιεί το σύστημα THESUS είναι μια επέκταση του αλγορίθμου DBSCAN [EK+96]. Ο DBSCAN είναι αλγόριθμος πυκνότητας και σε περίπτωση μετρικών χώρων [GR+99] έχει την ικανότητα να ανιχνεύει μη-σφαιρικές συστάδες. Στην περίπτωση εγγράφων συμπεριλαμβάνει στην ίδια συστάδα έγγραφα που μοιάζουν σημαντικά με ένα ελάχιστο αριθμό εγγράφων της συστάδας. Εκτελώντας με διαφορετικούς συνδυασμούς αρχικών παραμέτρων τον αλγόριθμο DBSCAN το σύστημα επιστρέφει μια ιεραρχική συσταδοποίηση των εγγράφων σε κάθε επίπεδο της οποίας βρίσκονται όλα τα έγγραφα.

Εναλλακτικά το σύστημα υλοποιεί και τον αλγόριθμο COBWEB που δημιουργεί μια ιεραρχία συστάδων ελέγχοντας σε κάθε επίπεδο της ιεραρχίας αν μεγιστοποιείται η πυκνότητα των παραγόμενων συστάδων.

3.3.2.1 Ο αλγόριθμος DBSCAN

Στην ενότητα αυτή παρατίθονται ορισμένα βασικά χαρακτηριστικά του αλγορίθμου καθώς και οι τροποποιήσεις που έγιναν στον αλγόριθμο για να προσαρμοστεί στις ανάγκες του συστήματος.

Ο αλγόριθμος σχεδιάστηκε αρχικά για τον εντοπισμό συστάδων με ακανόνιστα (μη σφαιρικά) σχήματα σε μετρικούς χώρους, με συντεταγμένες και διάταξη. Στην περίπτωση των εγγράφων έχουμε στη διάθεσή μας μόνο ένα μέτρο που μας δίνει την μεταξύ τους ομοιότητα (ή απόσταση). Δεν έχουμε πληροφορία ούτε για τις συντεταγμένες των εγγράφων σε κάποιο χώρο, ούτε για τον τρόπο με τον οποίο διατάσσονται. Απαιτείται λοιπόν μια προσαρμογή του αλγορίθμου DBSCAN στις

συνθήκες του προβλήματος και η ενσωμάτωση του μέτρου ομοιότητας εγγράφων σε αυτόν.

Τροποποιούμε ελαφρώς τους τέσσερις ορισμούς που δίνονται στο [EK+96] για τον αλγόριθμο DBSCAN:

Ορισμός 1ος: Τα εμπλουτισμένα έγγραφα di(URL,Ki,Ci) και dj(URL,Kj,Cj) είναι **γείτονες** αν ικανοποιούν το ακόλουθο κριτήριο:

$$\zeta\,(Ci, Cj) \geq \text{MinSim}$$

όπου MinSim, είναι μια ελάχιστη τιμή, που μπορούμε να θέσουμε, για την ομοιότητα δύο εγγράφων και ζ(Ci, Cj) το μέτρο ομοιότητας μεταξύ εγγράφων που αναπαρίστανται ως σύνολα όρων με βάρη.

Ορισμός 2ος: Η **γειτονιά** ενός εγγράφου di ορίζεται ως εξής:

$$\text{Neighborhood}\,(d_i(URL_i,K_i,C_i)) = \{\, d_k(URL_k,K_k,C_k) \mid \zeta(C_i, C_k) > \text{MinSim}\}$$

Ορισμός 3ος: Ένα έγγραφο d_j είναι **προσιτό με βάση την πυκνότητα** (**Density-reachable**) από κάποιο άλλο έγγραφο d_i για ορισμένες τιμές των MinSim και MinDocs, όπου MinDocs ένας ελάχιστος αριθμός εγγράφων, όταν:

Το έγγραφο d_j ανήκει στη γειτονιά του d_i :

$$d_j \in \text{Neighborhood}(d_i)$$

Ο αριθμός των εγγράφων στη γειτονιά του d_j είναι τουλάχιστον MinDocs:

$$|\,\text{Neighborhood}(d_j)\,| \geq \text{MinDocs}$$

Δεν αρκεί λοιπόν τα δύο έγγραφα να βρίσκονται κοντά (σύμφωνα με το MinSim) αλλά να έχουν κοντά τους και ένα αριθμό άλλων εγγράφων (MinDoc). Αυτό εξασφαλίζει την εσωτερική πυκνότητα των παραγόμενων συστάδων.

Ορισμός 4ος: Δύο έγγραφα d_j και d_i είναι **συνδεδεμένα με βάση την πυκνότητα** (**density connected**) για ορισμένες τιμές των MinSim και MinDocs αν υπάρχει έγγραφο d_k τέτοιο ώστε τα d_j και d_i να είναι ταυτόχρονα *προσιτά* από το d_k με τις δεδομένες τιμές των MinSim και MinDocs.

<u>Ο αλγόριθμος</u>

Οι συστάδες ορίζονται ως σύνολα εγγράφων συνδεδεμένων βάση πυκνότητας. Για να βρει ένα σύνολο συνδεδεμένων εγγράφων, ο αλγόριθμος ξεκινά με ένα τυχαίο έγγραφο d_i και εντοπίζει όλα τα έγγραφα που είναι προσιτά από αυτό με ορισμένα MinSim και MinDocs. Το d_i καλείται *πυρήνας* αν βρεθεί για αυτό ένα σύνολο συνδεδεμένων εγγράφων. Τότε ορίζεται μια συστάδα. Σε αντίθετη περίπτωση (αν για κάποιο έγγραφο d_i δεν οριστεί ένα σύνολο συνδεδεμένων εγγράφων, το d_i θεωρείται θόρυβος και δεν ανήκει σε καμία συστάδα (σύμφωνα με άλλη θεώρηση αποτελεί ξεχωριστή συστάδα). Η διαδικασία επαναλαμβάνεται για κάθε έγγραφο d_i της συλλογής που δεν έχει τοποθετηθεί σε κάποια συστάδα ή δεν έχει χαρακτηριστεί ώς θόρυβος.

Ο αλγόριθμος δέχεται ως είσοδο:

- σύνολα λέξεων ή εννοιών με βάρη που χαρακτηρίζουν κάθε έγγραφο,
- μια ελάχιστη τιμή ομοιότητας, MinSim, για τον καθορισμό της γειτονιάς ενός εγγράφου,
- τον ελάχιστο αριθμό εγγράφων στη γειτονιά ενός εγγράφου προσιτού βάση πυκνότητας, MinDocs.

Το αποτέλεσμα του αλγορίθμου είναι:

- ένα σύνολο από συστάδες εγγράφων και
- ένα σύνολο εγγράφων που θεωρήθηκαν θόρυβος.

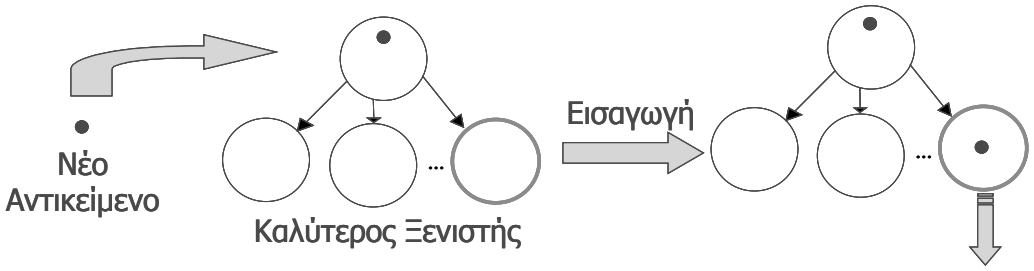
3.3.2.2 Ο αλγόριθμος COBWEB

Ο αλγόριθμος COBWEB είναι ένας ιεραρχικός αυξητικός αλγόριθμος που χωρίζει ένα σύνολο αντικειμένων σε ομάδες (τάξεις) δημιουργώντας ένα δέντρο κατηγοριοποίησης. Σε αντίθεση με τους αλγορίθμους κατηγοριοποίησης οι ομάδες των εγγράφων στον COBWEB δεν είναι γνωστές από πριν. Σε κάθε επίπεδο του δέντρου περιέχονται όλα τα αντικείμενα του συνόλου σε διαφορετική κάθε φορά συσταδοποίηση. Κάθε κόμβος του δέντρου αποτελεί μια τάξη που περιέχει αντικείμενα με κοινά χαρακτηριστικά. Τα παιδιά ενός κόμβου περιέχουν όλα τα αντικείμενα του πατρικού κόμβου χωρισμένα σε υποομάδες.

Ο αλγόριθμος είναι ιεραρχικός καθώς προσφέρει τη δυνατότητα τοποθέτησης των αντικειμένων σε ομάδες και υποομάδες. Είναι παράλληλα αυξητικός καθώς η προσθήκη ενός νέου αντικείμενου στο αρχικό σύνολο δεν απαιτεί εξ αρχής δημιουργία του σχήματος. Αντίθετα το αντικείμενο ενσωματώνεται στο υπάρχον δένδρο. Η ενσωμάτωσή του γίνεται αρχικά στη ρίζα (στο πρώτο επίπεδο με μία ομάδα που περιέχει όλα τα αντικείμενα) και στη συνέχεια σε κάθε επίπεδο επιλέγεται η καλύτερη υποομάδα. Η διαδικασία αυτή προχωρά σε όλα τα επίπεδα μέχρι το αντικείμενο να εισαχθεί σε κάποιο φύλλο του δένδρου.

Κάθε φορά ενημερώνεται ένα μόνο επίπεδο, οπότε και ενημερώνονται τα παιδιά του γονικού κόμβου. Στην περίπτωση που φτάνουμε σε φύλλο του δένδρου, η εισαγωγή γίνεται στο φύλλο και δημιουργούνται δύο νέα φύλλα απόγονοί του (ένα με όλα τα αντικείμενα του φύλλου και ένα με το νέο αντικείμενο). Η ενσωμάτωση του αντικειμένου μπορεί να προκαλέσει την αναδιαμόρφωση του υπόδεντρου του γονικού κόμβου στο επίπεδο που ενημερώνουμε. Οι διαδικασίες ενημέρωσης περιλαμβάνουν:

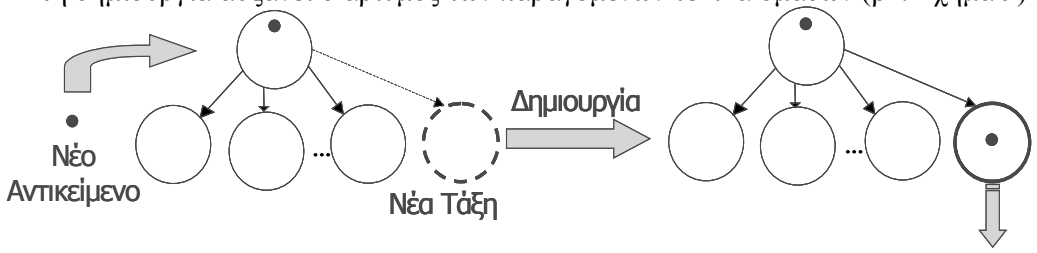
- Εισαγωγή: Στην περίπτωση αυτή το αντικείμενο εισάγεται στην υποομάδα στην οποία "ταιριάζει" περισσότερο. Η διαδικασία στο επόμενο επίπεδο θα αφορά τα παιδιά αυτής της υποομάδας (βλ. Σχήμα 8).



Σχήμα 8. Εισαγωγή αντικειμένου σε υπάρχουσα υποομάδα

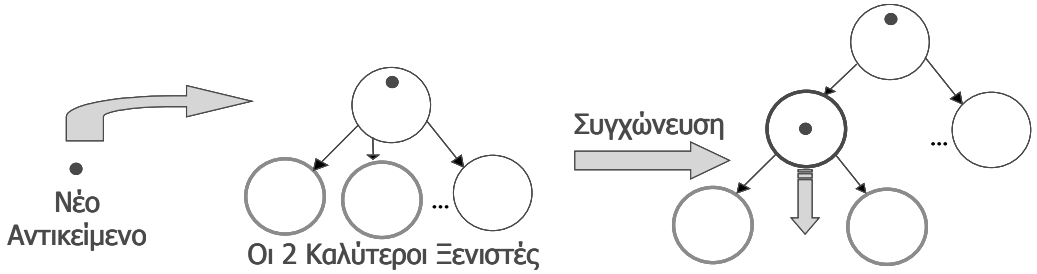
- Δημιουργία: Όταν το αντικείμενο δεν ταιριάζει σε καμία από τις υπάρχουσες υποομάδες, τότε δημιουργείται μια νέα υποομάδα και το αντικείμενο εισάγεται σε

αυτή. Η διαδικασία στο επόμενο επίπεδο συνεχίζεται γι' αυτή την υποομάδα. Με τη δημιουργία αυξάνει ο αριθμός των παραγόμενων τελικά ομάδων (βλ. Σχήμα 9)



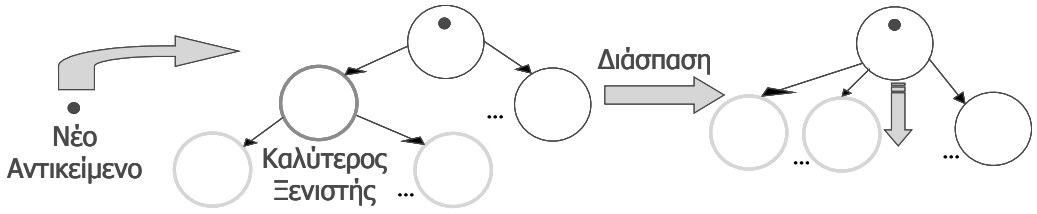
Σχήμα 9. Δημιουργία νέας υποομάδας

- Συγχώνευση: Το αντικείμενο ταιριάζει σε δύο ή περισσότερες υποομάδες οπότε ο αλγόριθμος εξετάζει την πιθανότητα συγχώνευσης των δύο πλησιέστερων υποομάδων. Οι δύο υποομάδες συγχωνεύονται, το αντικείμενο εισάγεται στη νέα ομάδα και η διαδικασία συνεχίζεται στο επόμενο επίπεδο στις υποομάδες αυτής (βλ. Σχήμα 10). Η συγχώνευση είναι μια από τις τεχνικές που χρησιμοποιεί ο αλγόριθμος για να αντιμετωπίσει το πρόβλημα της σειράς με την οποία συσταδοποιούνται τα αντικείμενα.



Σχήμα 10. Συγχώνευση των δύο καλύτερων υποομάδων

- Διάσπαση: Το αντικείμενο πρόκειται να εισαχθεί στην υποομάδα στην οποία ταιριάζει περισσότερο, αυξάνει όμως σημαντικά την ανομοιογένειά της οπότε μια τέτοια εισαγωγή πρέπει να αποφευχθεί. Στην περίπτωση αυτή η υποομάδα αντικαθίσταται από τις ομάδες που βρίσκονται στο επόμενο επίπεδο (από τα παιδιά της) και επανεξετάζεται η διαδικασία εισαγωγής σε κάποια από αυτές. Η διαδικασία συνεχίζεται πλέον από τον γονικό κόμβο (βλ. Σχήμα 11). Με τη διάσπαση αντιμετωπίζεται το πρόβλημα της σειράς με την οποία εξετάζονται τα αντικείμενα για συσταδοποίηση.



Σχήμα 11. Διάσπαση της καλύτερης υποομάδας

Για να επιλεγεί η σωστή διαδικασία, προηγείται μια αξιολόγηση των δέντρων που προκύπτουν. Τελικά επιλέγεται η διαδικασία που δίνει την καλύτερη ποιότητα συσταδοποίησης σε κάθε επίπεδο.

Συνάρτηση αξιολόγησης

Η Συνάρτηση αξιολόγησης (Categorization Utility) μετρά την ποιότητα της διαμέρισης μιας ομάδας σε υποομάδες (γονέας - παιδιά). Πρέπει να εκφράζει την ιδέα μας για την ιδανική συσταδοποίηση στην εφαρμογή που αναπτύσσουμε. Κάθε κόμβος περιγράφει πλήρως το υπόδεντρό του. Έτσι η συνάρτηση αξιολόγησης χρειάζεται δεδομένα μόνο από τους κόμβους του επιπέδου που βρισκόμαστε (γονικός κόμβος & κόμβοι παιδιά). Η συνάρτηση αξιολόγησης χρησιμοποιείται κατά τη δοκιμή των τελεστών προκειμένου να επιλεχθεί ο καταλληλότερος.

Στόχος του αλγορίθμου είναι να επιτύχει υψηλή Συνεκτικότητα και χαμηλή Εξωτερική Ομοιογένεια (δηλαδή υψηλή Εξωτερική Ανομοιογένεια) για τις ομάδες. Αυτά σταθμίζονται με το βαθμό Σημαντικότητας της κάθε ομάδας στη διαμέριση.

Ορίζουμε λοιπόν τρία μέτρα: Σημαντικότητα (Importance), Συνεκτικότητα (Cohesion) και Εξωτερική Ομοιογένεια (External Similarity) που είναι καθοριστικά για την ποιότητα της συσταδοποίησης σε κάθε επίπεδο του COBWEB. Με τα μέτρα αυτά επιδιώκουμε να εντοπίσουμε ***συμπαγείς*** και ***διακριτές*** ομάδες.

Αν θεωρήσουμε ότι εξετάζουμε τη συσταδοποίηση του κόμβου y που χωρίζεται σε m υποομάδες. Τα μέτρα για κάθε μια υποομάδα x ορίζονται ως εξής.

Σημαντικότητα - Importance: Η σημασία της υπό εξέταση ομάδας για τον κόμβο του οποίου εξετάζουμε τη διαμέριση.

$$\text{Importance}_x = \frac{|x|}{|y|} \qquad \text{Εξ.15}$$

όπου |x| είναι ο αριθμός αντικειμένων της υπο εξέταση ομάδας και |y| ο αριθμός αντικειμένων του κόμβου του οποίου εξετάζουμε τη διαμέριση (γονική ομάδα της x).

Συνεκτικότητα - Cohesion: Η συνεκτικότητα μιας ομάδας ως προς τα αντικείμενα που περιέχει. Ο μέσος όρος της ομοιότητας του κάθε αντικειμένου της ομάδας με το αντιπροσωπευτικό αντικείμενο της ομάδας. Το αντιπροσωπευτικό αντικείμενο μιας ομάδας δεν είναι απαραίτητα υπαρκτό μπορεί να είναι ένα ιδεατό αντικείμενο με το μέσο όρο των χαρακτηριστικών των αντικειμένων της ομάδας.

$$\text{Cohesion}_x = \frac{\sum_{i=1}^{|x|} Sim(d_i, d'_x)}{|x|} \qquad \text{Εξ.16}$$

όπου |x| είναι το πλήθος αντικειμένων της ομάδας x, d_i ένα αντικείμενο της ομάδας και d'_x το αντιπροσωπευτικό αντικείμενο της ομάδας.

Εξωτερική ομοιογένεια - External Similarity: Δηλώνει την ομοιότητα μιας υποομάδας με τις υπόλοιπες υποομάδες μιας ομάδας στο δένδρο. Οι τάξεις συγκρίνονται μέσω των αντιπροσωπευτικών αντικειμένων τους, οπότε η εξωτερική

ομοιογένεια μιας υποομάδας ορίζεται ως ο μέσος όρος της ομοιότητας του αντιπροσωπευτικού αντικειμένου της υπό εξέταση υποομάδας με τα αντιπροσωπευτικά αντικείμενα των υπολοίπων υποομάδων.

$$\text{External Similarity}_x = \frac{\sum_{\substack{j=1..m \\ j \neq x}} sim(d'_x, d'_j)}{m-1} \qquad \text{Εξ.17}$$

Όπου d'_x το αντιπροσωπευτικό αντικείμενο της υποομάδας x, το d'_j αντιστοιχεί στα αντιπροσωπευτικά αντικείμενα όλων των άλλων υποομάδων εκτός της x και m είναι το πλήθος των υποομάδων του κόμβου που εξετάζεται (γονική ομάδα).

Οι τιμές των παραμέτρων αυτών υπολογίζονται για τις υποομάδες του κάθε κόμβου. Ο λόγος της συνεκτικότητας προς την εξωτερική ομοιογένεια είναι μια ένδειξη της ποιότητας συσταδοποίησης για την υποομάδα x. Η σημαντικότητα αποτελεί το δείκτη βαρύτητας της υποομάδας x για τον κόμβο y.

Η συνάρτηση αξιολόγησης του σχήματος συσταδοποίησης για τον κόμβο y υπολογίζεται ως ο σταθμισμένος μέσος όρος της ποιότητας των υποομάδων του.

$$CU_y = \frac{\sum_{j=1}^{m} \text{importance}_j \frac{\text{Cohesion}_j}{\text{ExternalSimilarity}_j}}{m} \qquad \text{Εξ.18}$$

όπου ο όρος εντός του αθροίσματος μπορεί να θεωρηθεί ως η Συνάρτηση Αξιολόγησης (CU) κάθε υποομάδας ξεχωριστά.

Όταν έρθει ένα νέο έγγραφο σε κάποιο κόμβο εξετάζεται η τιμή της συνάρτησης αξιολόγησης του σχήματος υποομάδων που προκύπτει για κάθε μια από τις τέσσερεις περιπτώσεις που αναφέρθηκαν (εισαγωγή, δημιουργία, συγχώνευση, διάσπαση) και επιλέγεται η διαδικασία που δίνει τη μέγιστη τιμή.

<u>Προσαρμογή του αλγορίθμου στο THESUS</u>

Για να είμαστε σε θέση να χρησιμοποιήσουμε τον αλγόριθμο COBWEB για την συσταδοποίηση εγγράφων του ΠΙ, πρέπει να τον προσαρμόσουμε στα μέτρα του προβλήματός μας.

Ο αλγόριθμος έχει περιγραφεί στη βιβλιογραφία ως ένας αλγόριθμος για ιεραρχική συσταδοποίηση αντικειμένων με πολλά κατηγορικά και αριθμητικά χαρακτηριστικά. Στην περίπτωση του THESUS τα αντικείμενα είναι έγγραφα που περιγράφονται ως σύνολα ζευγών λέξεων ή εννοιών με τα αντίστοιχα βάρη: $D_i=\{(k_j,v_j), (c_m,v_m)\}$

Χρησιμοποιώντας τις εννοιολογικές περιγραφές των εγγράφων πρέπει να ορίσουμε τα χαρακτηριστικά του αλγορίθμου, όπως το αντιπροσωπευτικό αντικείμενο μιας ομάδας και τα μέτρα αξιολόγησης της συσταδοποίησης.

Αν υποθέσουμε ότι τα έγγραφα $d_{1k}=\{(c_{11},w_{11}),.., (c_{m1},w_{m1})\}$, …,$d_{nk}=\{(c_{1n},w_{1n}),.., (c_{mn},w_{mn})\}$ αποτελούν μια ομάδα (ομάδα k) ,όπου $w_{11}, ..w_{i1}, …, w_{1n}, .., w_{in} \in [0,1]$

Το αντιπροσωπευτικό έγγραφο της ομάδας αυτής ορίζεται ως εξής:

$$D_k = \{ \left(c_1, \frac{\sum_{x=1..n} w_{1x}}{n} \right), .., \left(c_i, \frac{\sum_{x=1..n} w_{ix}}{n} \right) \} \quad \text{Εξ.19}$$

Τα υπόλοιπα μέτρα ορίζονται ως εξής:

$$\text{Σημαντικότητα} - \textbf{Importance}_\textbf{k} = \frac{n_k}{N} \quad \text{Εξ.20}$$

όπου n_k το πλήθος εγγράφων στην ομάδα k και N ο συνολικός αριθμός εγγράφων.

$$\text{Συνεκτικότητα} - \textbf{Cohesion}_\textbf{k} = \frac{\sum_{i=1..k} sim(d_{ik} D_k)}{n_k} \quad \text{Εξ.21}$$

όπου d_{ik} τα έγγραφα της ομάδας k και D_k το αντιπροσωπευτικό έγγραφο της k.

$$\text{Εξωτερική ομοιογένεια} - \textbf{External Similarity}_\textbf{k} = \frac{\sum_{j=1..m, j \neq k} sim(D_k, D_j)}{m-1} \quad \text{Εξ.22}$$

όπου m το πλήθος των διαφορετικών υποομάδων.

Ο αλγόριθμος συμπεριφέρεται διαφορετικά αν αλλάξουμε τον τρόπο ορισμού της συνάρτησης αξιολόγησης για κάθε μια υποομάδα. Σε όλες τις περιπτώσεις θεωρούμε ότι η συνεκτικότητα αυξάνει το μέτρο CU ενώ αντίθετα η εξωτερική ομοιότητα το μειώνει για το λόγο αυτό θεωρήσαμε το μέτρο CU ανάλογο του λόγου των δύο μέτρων. Στην πρώτη δοκιμή που έγινε το μέτρο αξιολόγησης ορίστηκε ως το γινόμενο της σημαντικότητας επί το λόγο συνεκτικότητα/εξωτερική ομοιογένεια. Στην περίπτωση αυτή δημιουργήθηκαν λίγες και μεγάλες ομάδες, μειώνοντας σημαντικά την ευαισθησία του συστήματος. Ενδεικτικά για να αυξήσουμε την ευαισθησία του συστήματος υψώνουμε το λόγο των δύο μέτρων στο τετράγωνο.

1η Δοκιμή:

$$\text{CU} = \text{Importance} \frac{\text{Cohesion}}{\text{External Similarity}} \quad \text{Εξ.23}$$

2η Δοκιμή:

$$\text{CU} = \text{Importance} \left(\frac{\text{Cohesion}}{\text{External Similarity}} \right)^2 \quad \text{Εξ.24}$$

3.4 Απόδοση ετικέτας σε συστάδες (labeling)

Ο καθορισμός μιας ετικέτας για τις συστάδες που προκύπτουν (π.χ. η ανάθεση ενός συνόλου από όρους της οντολογίας σε κάθε συστάδα) είναι πολύ σημαντικός καθώς διευκολύνει την περιήγηση στις παραγόμενες ομάδες και την απάντηση ερωτήσεων. Η ετικέτα αυτή πρέπει να είναι όσο το δυνατό πλησιέστερα στο αντιπροσωπευτικό έγγραφο κάθε ομάδας λαμβάνοντας βέβαια υπόψη δύο παράγοντες: α) ο αριθμός των εννοιών στην ετικέτα πρέπει να είναι περιορισμένος· μια ετικέτα που περιλαμβάνει όλες τις έννοιες της οντολογίας με διάφορα βάρη ενδέχεται να μπερδέψει και όχι να

βοηθήσει την περιήγηση, β) κατά τον περιορισμό των εννοιών της ετικέτας πρέπει να χαθεί όσο το δυνατόν λιγότερη πληροφορία, κάνοντας χρήση της ιεραρχικής σχέσης μεταξύ των εννοιών.

Η συσταδοποίηση των εγγράφων αποτελεί σημαντικό βήμα στον σημασιολογικό εμπλουτισμό της πληροφορίας, επιδιώκουμε όμως και την εύρεση κατάλληλων ετικετών για τις ομάδες για τους εξής λόγους:

- Η συσταδοποίηση των εγγράφων δεν συνεπάγεται και χαρακτηρισμό των συστάδων.
- Χρειαζόμαστε ένα τρόπο υπολογισμού της απόστασης μεταξύ μιας ομάδας και ενός ερωτήματος, ώστε να εντοπίζουμε την πλησιέστερη ομάδα εγγράφων στο ερώτημα. Με τον τρόπο αυτό μπορούμε να περιορίσουμε την αναζήτηση, για τα έγγραφα που ικανοποιούν την ερώτηση, στα έγγραφα της πλησιέστερης ομάδας, συγκρίνοντας σε πρώτο στάδιο το ερώτημα με τις ετικέτες των ομάδων.
- Οι ετικέτες των παραγόμενων ομάδων θα διευκολύνουν την περιήγηση στο σύνολο της συλλογής.

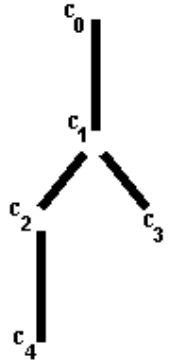
Η διαδικασία περιγραφής των συστάδων συνοψίζεται στα εξής:

- Δημιουργούμε το σύνολο L ως την ένωση όλων των εννοιών (c_{ki}) που εμφανίζονται στα έγγραφα d_k μιας συστάδας S.

$$L= \bigcup_{\forall d_k \in S} c_{ki} \qquad \text{Εξ.25}$$

- Για κάθε έννοια c_j στο L, υπολογίζουμε τον αριθμό ή το ποσοστό των εγγράφων της συστάδας στα οποία εμφανίζεται (βλ. Εξ. 19). Στο σημείο αυτό μπορούμε να θεωρήσουμε ότι έχουμε ένα αντιπροσωπευτικό έγγραφο D της συστάδας S: $D=\{(w_j,c_j)\}$

Καθώς το πλήθος των εννοιών που μπορεί να εμφανιστούν στο D μπορεί να είναι μεγάλο, προσπαθούμε να το περιορίσουμε, χωρίς απώλεια πληροφορίας, χρησιμοποιώντας την οντολογία. Μια παρόμοια προσπάθεια έχει αναφερθεί στο [CC+03]. Στη διαδικασία αυτή, οι λιγότερο σημαντικές έννοιες (αυτές με τις λιγότερες εμφανίσεις, ή αυτές που είναι πολύ χαμηλά στην οντολογία μας και συνεπώς πολύ ειδικές) μιας ομάδας αντικαθίστανται από τις αμέσως γενικότερες στην οντολογία και στη συνέχεια οι επαναλαμβανόμενες έννοιες συγχωνεύονται (αθροίζοντας τις εμφανίσεις). Η αντικατάσταση αυτή περιορίζει το πλήθος των εννοιών που περιγράφουν μια συστάδα στον αριθμό που επιθυμούμε.



Σχήμα 12. Παράδειγμα ταξινομίας

Ας θεωρήσουμε για παράδειγμα την ταξινομία του πιο πάνω σχήματος (Σχήμα 12) και θεωρήσουμε το έγγραφο D αντιπροσωπευτικό της συστάδας S, όπου:

$$D=\{(w_0,c_0), (w_2,c_2),(w_3, c_3), (w_4,c_4)\} \text{ με } w_4<w_1,w_2,w_3$$

και θεωρήσουμε ότι ο μέγιστος αριθμός εννοιών στην ετικέτα της συστάδας είναι 3, θα επιλέξουμε να συγχωνεύσουμε την έννοια c_4.

Αντί απλά να διαγράψουμε την έννοια από την ετικέτα, με τη βοήθεια της ιεραρχίας την ανάγουμε στην αμέσως γενικότερη έννοια που είναι η c_2. Το βάρος της έννοιας c_4 προστίθεται σε αυτό της c_2, που ήδη υπάρχει αφού πρώτα πολλαπλασιαστεί με ένα παράγοντα $k_{c2,c4}$ που εκφράζει την απώλεια λόγω γενίκευσης. Ο παράγοντας αυτός θα μπορούσε να οριστεί ως το αντίστροφο της ομοιότητας κατά Ως αποτέλεσμα η συμπτυγμένη ετικέτα θα έχει τη μορφή:

$$D'=\{(w_0,c_0), (w_2+ w_4',c_2),(w_3, c_3), (w_4,c_4)\} \text{ με } w_4'= k_{c2,c4}\cdot w_4$$

Το αποτέλεσμα τη διαδικασίας συσταδοποίησης και χαρακτηρισμού είναι ένα σύνολο από εμπλουτισμένα κείμενα που επιπλέον περιέχουν τον αριθμό της συστάδας στην οποία ανήκουν. Παράγεται επίσης μια ετικέτα για κάθε συστάδα εγγράφων.

Η παραπάνω διαδικασία περιγραφής των συστάδων ενδέχεται να εμφανίσει προβλήματα όταν σε πολλές συστάδες εμφανίζονται οι ίδιες, πολύ γενικές, έννοιες. Οι λιγότερο γενικές που ταυτόχρονα έχουν και λίγες εμφανίσεις είναι αυτές που οδήγησαν στη διάκριση μεταξύ των συστάδων. Κατά τη διαδικασία περιορισμού των εννοιών όμως αντικαθίστανται με κάποιες γενικότερες έννοιες μειώνοντας έτσι τη διακριτότητα των ετικετών. Σε αυτές τις περιπτώσεις συνίσταται να αγνοούμε την έννοια στη ρίζα ή στο πρώτο επίπεδο της ιεραρχίας.

Ένα πλεονέκτημα της παρούσας προσέγγισης είναι ότι όσο ανεβαίνουμε προς τα ψηλότερα επίπεδα της ιεραρχίας το κόστος γενίκευσης είναι μεγαλύτερο και συνεπώς η επίδραση των όρων που προκύπτουν από πολλές γενικεύσεις εξασθενεί. Έτσι, όταν πολλές έννοιες από χαμηλά επίπεδα της ιεραρχίας συγχωνευθούν σε μια έννοια στα ψηλότερα επίπεδα της ιεραρχίας, το βάρος της προκύπτουσας έννοιας θα είναι μικρό.

Τα προηγούμενα είναι αποτέλεσμα μιας προκαταρκτικής έρευνας στο θέμα της παραγωγής περιγραφών για τις παραγόμενες συστάδες με χρήση της οντολογίας γι΄αυτό και στην συνέχεια δεν θα αναφερθούμε περαιτέρω σε αυτά. Η εύρεση μιας αποτελεσματικότερης διαδικασίας εύρεσης της περιγραφής των ομάδων αποτελεί άμεσο στόχο του συστήματος THESUS.

3.5 Το THESUS XML-Schema

Σε μια προσπάθεια να διευκολυνθεί η διαδικασία χαρακτηρισμού των περιεχομένων του ΠΙ το σύστημα THESUS χρησιμοποιεί την πληροφορία που μεταφέρουν οι υπερσύνδεσμοι για να εξάγει αξιόπιστες περιγραφές για τα έγγραφα του ιστού. Οι περιγραφές εμπλουτίζονται με σημασιολογικά χαρακτηριστικά, με τη χρήση μιας θεματικής οντολογίας και ενός μηχανισμού απεικόνισης λέξεων σε όρους της οντολογίας. Το σύστημα παράγει πλούσια πληροφορία για τα έγγραφα του ιστού που μπορεί να αποτελέσει τη βάση για τη διαχείριση των περιεχομένων του Σημασιολογικού Ιστού.

Η μετάβαση από λέξεις σε έννοιες αποτελεί την ουσία του Σημασιολογικού Ιστού. Ο Hotho κ.ά [HM+01] δηλώνει την ανάγκη για εννοιολογική αναπαράσταση των εγγράφων του ΠΙ και για απεικόνιση των διανυσμάτων λέξεων ενός εγγράφου σε διανύσματα εννοιών. Πολλά εργαλεία έχουν αναπτυχθεί για να βοηθήσουν τη διαδικασία χαρακτηρισμού κειμένων [LU+02], πολυμέσων [SB+02], υπερσυνδέσμων [VV01]. Τα εργαλεία αυτά συνήθως απαιτούν ανθρώπινη παρέμβαση και δεν μπορούν να χρησιμοποιηθούν σε ευρεία κλίμακα. Το αποτέλεσμα αυτών των συστημάτων είναι η παραγωγή περιεκτικών περιγραφών για κάθε έγγραφο που είτε αποθηκεύονται ως μεταδεδομένα στο ίδιο το έγγραφο, είτε αποθηκεύονται σε ξεχωριστά αρχεία, είτε συλλέγονται κεντρικά σε βάσεις δεδομένων για περαιτέρω επεξεργασία.

Στο THESUS ο χαρακτηρισμός ενός εγγράφου βασίζεται στην πληροφορία των εισερχόμενων συνδέσμων. Επιπλέον, το σύστημα επεξεργάζεται τους εξερχόμενους συνδέσμους και αποθηκεύει την εξαγόμενη λεξική και σημασιολογική πληροφορία σε ένα XML αρχείο. Η δομή του αρχείου XML που παράγεται για κάθε έγγραφο του ΠΙ είναι συγκεκριμένη και περιγράφεται από ένα XML-Schema αρχείο. Για μια συλλογή εγγράφων του ΠΙ παράγεται μια αντίστοιχη συλλογή XML αρχείων που μπορούν να αποτελέσουν τη βάση για επιπλέον αναζητήσεις.

Τα πλεονεκτήματα από την ενδιάμεση αποθήκευση της πληροφορίας σε XML έγγραφα είναι αρκετά:

- Καταρχήν η πληροφορία, που παράγεται αυτόματα για κάθε έγγραφο του ιστού και περιέχεται στα XML έγγραφα, μπορεί οποιαδήποτε στιγμή να αναθεωρηθεί και να εμπλουτιστεί. Ο έλεγχος και η ενημέρωση των εγγράφων μπορεί να γίνει είτε αυτόματα από το σύστημα, είτε χειροκίνητα μέσα από κατάλληλη διεπαφή. Η αυτόματη ενημέρωση μπορεί να συμβεί κατά τη διάρκεια συλλογής των εγγράφων, όταν ένας νέος εισερχόμενος σύνδεσμος βρεθεί για ένα υπάρχον έγγραφο. Στην περίπτωση αυτή η πληροφορία που μεταφέρει ο σύνδεσμος προστίθεται πολύ εύκολα στο υπάρχον XML αρχείο. Μέσω της διεπαφής, μπορεί κάποιος άνθρωπος να επέμβει στα αποτελέσματα του συστήματος για ένα έγγραφο και να διορθώσει ή να εμπλουτίσει τους χαρακτηρισμούς.
- Τα XML έγγραφα αποτελούν μια συλλογή δεδομένων στην οποία μπορούμε να κάνουμε ερωτήσεις με χρήση της κατάλληλης γλώσσας ερωτήσεων, XPATH [W3h] ή XQUERY [W3m], π.χ. να εντοπίσουμε έγγραφα του ιστού με πολλούς εισερχόμενους ή εξερχόμενους συνδέσμους, έγγραφα που χαρακτηρίζονται από συγκεκριμένες λέξεις ή έννοιες.
- Η χρήση της XML επιτρέπει στους δημιουργούς των εγγράφων να καθορίσουν με ακρίβεια τη σημασιολογία συνδέσμων και εγγράφων.
- Τα δεδομένα των XML εγγράφων μπορεί να χρησιμοποιηθούν ως μεταδεδομένα των υπαρχόντων εγγράφων. Τα έγγραφα του ιστού μπορούν να περιέχουν εξολοκλήρου την XML πληροφορία ή αναφορές προς τα XML έγγραφα και να είναι έτσι διαθέσιμα σε μηχανές αναζήτησης αλλά και στους χρήστες του ΠΙ.

Η επιπλέον πληροφορία που εξάγεται από τους υπερσυνδέσμους μπορεί να χρησιμοποιηθεί με διάφορους τρόπους, όπως αναλύεται στα επόμενα κεφάλαια:

- Οι μηχανές αναζήτησης μπορούν να παρέχουν προηγμένους τρόπους αναζήτησης, όπως αναζήτηση σελίδων με συγκεκριμένη σημασιολογία και τρόπο σύνδεσης.
- Τα αποτελέσματα των αναζητήσεων ταξινομούνται καλύτερα καθώς εκτός από το πλήθος συνδέσμων για ένα έγγραφο γνωρίζουμε πλέον και τη σημασία τους.

- Τα αποτελέσματα συσταδοποιούνται καλύτερα με βάση την εννοιολογική περιγραφή των εγγράφων και τον συνδέσμων τους.
- Η ύπαρξη μιας οντολογίας διευκολύνει την πλοήγηση στα αποτελέσματα μιας αναζήτησης και στον ταχύτερο εντοπισμό των εγγράφων που μας ενδιαφέρουν.
- Η περιήγηση στα έγγραφα του ΠΙ γίνεται με περισσότερη γνώση καθώς τα προγράμματα πλοήγησης θα μπορούν να παρουσιάζουν την επιπλέον σημασιολογική πληροφορία των συνδέσμων στους χρήστες [KM99].

Όπως δηλώνεται και στο [KE+99] το κέρδος του αναγνώστη από την πληροφορία που ενσωματώνεται στους υπερσυνδέσμους είναι ότι το νέο δίκτυο εγγράφων διευκολύνει την πλοήγηση και προσφέρει γνώση για το έγγραφο στο οποίο δείχνει ένας σύνδεσμος προτού καν τον επισκεφθεί.

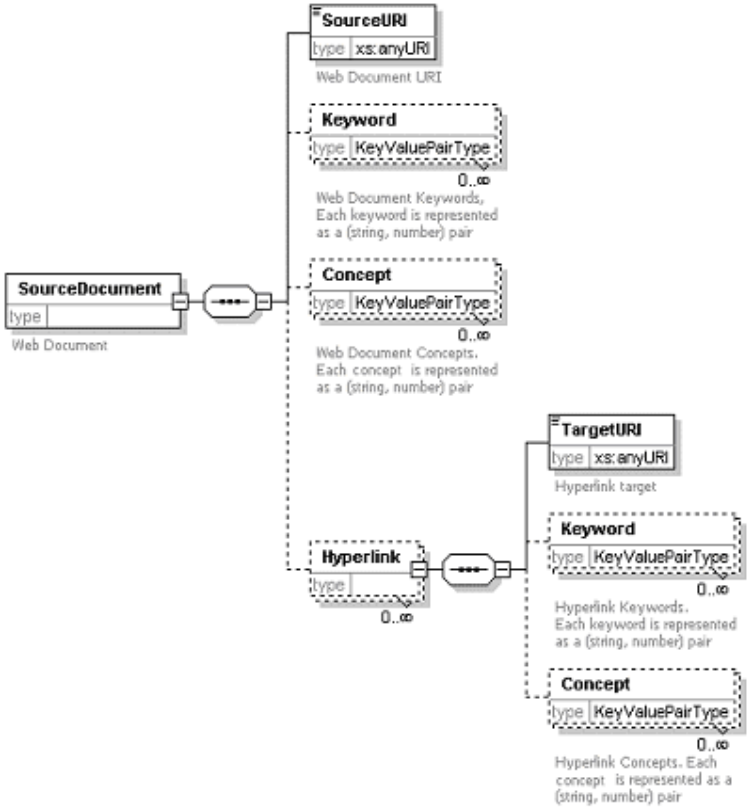
Πιο σύνθετες τεχνικές επεξεργασίας της συγκεντρωμένης πληροφορίας, όπως εξόρυξη γνώσης, στατιστικές αναλύσεις, προσωποποιημένη πληροφόρηση, χάρτες αναπαράστασης του ΠΙ [K98] κτλ, μπορούν να προσαρμοστούν στην υπάρχουσα δομή των XML εγγράφων. Στο THESUS τα XML αρχεία αποθηκεύονται εναλλακτικά σε μια σχεσιακή βάση δεδομένων που δημιουργείται αυτόματα βάσει του XML-Schema αρχείου ώστε να είναι διαθέσιμα για πιο σύνθετες διαδικασίες επεξεργασίας. Το σύστημα X-Database αναπτύχθηκε στα πλαίσια της διατριβής για να εξυπηρετήσει της ανάγκες απεικόνισης της XML-Schema και των XML εγγράφων στο σχεσιακό μοντέλο.

3.5.1 Η δομή των XML εγγράφων

Το αρχείο XML-Schema (Παράρτημα Α) καθορίζει τη δομή των XML εγγράφων. Το XML-Schema είναι γενικό και μπορεί να προσαρμοστεί σε οποιαδήποτε εννοιολογική ιεραρχία ή οντολογία, που αντιπροσωπεύει διάφορα πεδία ενδιαφέροντος.

Κεντρικό στοιχείο του XML εγγράφου όπως φαίνεται και στο Σχήμα 13 είναι το `SourceDocument`,που περιέχει:

- ένα στοιχείο `SourceURI`, που περιέχει το URI του εγγράφου,
- 0 ή περισσότερα στοιχεία `Keyword` και `Concept`. Τα στοιχεία αυτά περιέχουν λέξεις και έννοιες που έχουν εξαχθεί από τους εισερχόμενους συνδέσμους του εγγράφου, ή από τα περιεχόμενα του εγγράφου αν το έγγραφο δεν έχει καθόλου εισερχόμενους συνδέσμους. Και τα δύο στοιχεία είναι ίδιου τύπου και περιέχουν δύο χαρακτηριστικά, ένα αλφαριθμητικό που αντιστοιχεί στη λέξη ή την έννοια και ένα δεκαδικό που αντιστοιχεί στη σημαντικότητα της λέξης ή στο βάρος της έννοιας αντίστοιχα.
- 0 ή περισσότερα στοιχεία `Hyperlink` που αντιστοιχούν στους εξερχόμενους συνδέσμους του εγγράφου. Κάθε στοιχείο `Hyperlink` καθορίζει το URI του εγγράφου στο οποίο δείχνει ο σύνδεσμος εντός του στοιχείου `TargetURI` και τη λεξική και σημασιολογική πληροφορία του συνδέσμου εντός 0 ή περισσοτέρων στοιχείων `Keyword` και `Concept`.



Σχήμα 13. Αναπαράσταση του XML-Schema των εγγράφων του THESUS

3.6 Αρχειοθέτηση αποτελεσμάτων – Σημασιολογική διαχείριση ερωτήσεων

Τα μέτρα υπολογισμού της ομοιότητας μεταξύ εγγράφων βασίζονται στη λεξική ομοιότητα μεταξύ των όρων που εμφανίζονται στις περιγραφές των εγγράφων, κάνοντας στην ουσία δυαδικό ταίριασμα (binary matching). Για παράδειγμα, αν ένα έγγραφο d_1 χαρακτηρίζεται από τη λίστα λέξεων: d_1={μοτοσικλέτα, γρήγορος} θα θεωρηθεί εντελώς άσχετο με ένα έγγραφο d_2 που χαρακτηρίζεται από τη λίστα: d_2={αυτοκίνητο, ταχύς}. Παρ' όλα αυτά, είναι εμφανές ότι οι δύο λίστες, κατά συνέπεια και τα έγγραφα, σχετίζονται σημαντικά καθώς και η "μοτοσικλέτα" και το "αυτοκίνητο" είναι "οχήματα" αλλά και τα επίθετα "γρήγορος" και "ταχύς" είναι συνώνυμα. Συνεπώς τα d_2 και d_1 ασχολούνται με όμοιες έννοιες.

Καθώς το σύστημα THESUS αντιστοιχίζει λέξεις στις αντίστοιχες έννοιες μιας οντολογίας, προσφέρει έναν πιο ευέλικτο μηχανισμό εύρεσης σχετικών εγγράφων, που λαμβάνει υπόψη του τις εξειδικεύσεις και γενικεύσεις των εννοιών. Αυτός ο μηχανισμός εύρεσης σχετικών εγγράφων είναι πολύτιμος για την απάντηση ερωτήσεων.

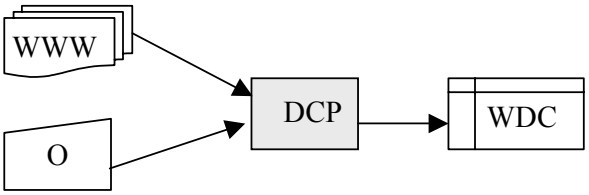
Κάνοντας χρήση του μέτρου ομοιότητας, μεταξύ συνόλου εννοιών με βάρη, που προσφέρει το THESUS, μπορούμε να υπολογίσουμε την ομοιότητα μεταξύ εγγράφων

ή μεταξύ εγγράφων και ερωτημάτων, μπορούμε να συσταδοποιήσουμε έγγραφα του ΠΙ χωρίς την ύπαρξη διάταξης στο χώρο κτλ.

3.7 Ανασκόπηση της μεθοδολογίας THESUS

Η μέθοδος που παρουσιάστηκε, βασίζεται στην ύπαρξη μιας θεματικής οντολογίας (Ο), και του ΠΙ (WWW) και αποσκοπεί στη συγκέντρωση και οργάνωση διαδικτυακών εγγράφων που εντάσσονται στο θεματικό πεδίο που ορίζει η οντολογία. Η μέθοδος αποτελείται από τις επιμέρους διαδικασίες συλλογής, χαρακτηρισμού και εμπλουτισμού των εγγράφων, οργάνωσης των εγγράφων και διαχείρισης ερωτήσεων.

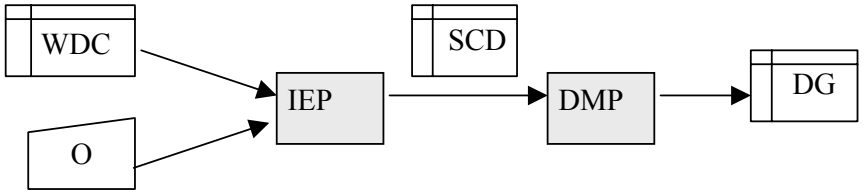
Η διαδικασία συλλογής διαδικτυακών εγγράφων (Document Collection Process – DCP, βλ. Σχήμα 14) αναλύει τους συνδέσμους που παρέχουν γνωστές δικτυακές πύλες και επιλέγει μόνο τους συνδέσμους που χαρακτηρίζονται από έννοιες της οντολογίας· οι σύνδεσμοι αυτοί θεωρούνται σχετικοί με την οντολογία. Αποθηκεύει τα έγγραφα που υποδεικνύονται από τους συνδέσμους αυτούς, αναλύει τους συνδέσμους τους και εντοπίζει όσους σχετίζονται με την οντολογία, κ.ο.κ. Η διαδικασία επαναλαμβάνεται μέχρι ότου όλα τα έγγραφα που υποδεικνύονται από σχετικούς με την οντολογία συνδέσμους προστεθούν στη συλλογή. Η διαδικασία συλλογής μπορεί να παραλειφθεί εφόσον υπάρχει διαθέσιμη συλλογή εγγράφων (Web Document Collection – WDC).



Σχήμα 14. Διαδικασία συλλογής διαδικτυακών εγγράφων

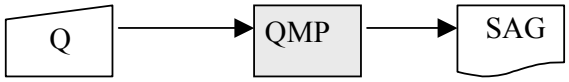
Ακολουθεί η διαδικασία εξαγωγής και εμπλουτισμού της πληροφορίας των εγγράφων (Information Enhancement Process – IEP, βλ. Σχήμα 15). Για το χαρακτηρισμό των εγγράφων της συλλογής (WDC) εξάγονται οι συχνότερα εμφανιζόμενες λέξεις από τα περιεχόμενά τους ή οι λέξεις που εμφανίζονται στους περισσότερους από τους εισερχόμενους συνδέσμους προς τα έγγραφα (όταν πρόκειται για έγγραφα του ΠΙ). Για τον εμπλουτισμό των εγγράφων με σημασιολογικά χαρακτηριστικά (έννοιες μιας οντολογίας) χρησιμοποιείται μια οντολογία, ένας λεξιλογικός θησαυρός και μια διαδικασία εύρεσης των πλησιέστερων εννοιών της οντολογίας για τις λέξεις που εξάγονται. Από τη διαδικασία αυτή προκύπτει ένα σύνολο σημασιολογικά χαρακτηρισμένων εγγράφων (Semantically Characterized Documents, Σχήμα 15, SCD).

Η διαδικασία οργάνωσης των εγγράφων της συλλογής (Document Management Process – DMP, βλ. Σχήμα 15), οργανώνει τα έγγραφα που έχουν προηγουμένως χαρακτηριστεί με σημασιολογικές περιγραφές σε ομάδες εγγράφων (Document Groups, DG) που εμφανίζουν σημασιολογική εγγύτητα. Κάθε έγγραφο τοποθετείται σε μία ομάδα, ενώ κάθε ομάδα έχει τη δική της σημασιολογική περιγραφή.



Σχήμα 15. Διαδικασίες Εξαγωγής και Εμπλουτισμού πληροφορίας, Οργάνωσης Εγγράφων

Η διαδικασία διαχείρισης ερωτήσεων των χρηστών (Query Management Process – QMP, βλ. Σχήμα 16) δέχεται μια λίστα λέξεων ως ερώτημα, το μετατρέπει σε λίστα εννοιών της οντολογίας, επιλέγει τις ομάδες με τις πλησιέστερες σημασιολογικές περιγραφές και ταξινομεί στην απάντηση τα έγγραφά τους με φθίνουσα σειρά εγγύτητας προς την σημασιολογική ερώτηση.



Σχήμα 16. Διαδικασία διαχείρισης ερωτήσεων

3.8 Περιορισμοί του συστήματος THESUS

Το σύστημα THESUS είναι προσανατολισμένο στην οργάνωση θεματικών υποσυνόλων του ΠΙ. Για το λόγο αυτό απαιτεί τον προσδιορισμό ενός θεματικού πεδίου με τη μορφή μιας οντολογίας.

Καθώς το σύστημα βασίζεται κυρίως στους εισερχόμενους συνδέσμους ενός εγγράφου του ΠΙ για να το χαρακτηρίσει περισσότερο αξιόπιστα, χρησιμοποιεί μια υπηρεσία του ΠΙ για να εντοπίζει τους συνδέσμους αυτούς. Η διατήρηση αντιγράφου στο ίδιο το σύστημα του γράφου του ΠΙ θα το έκανε ανεξάρτητο της συγκεκριμένης υπηρεσίας.

Για την εξαγωγή των σημασιολογικών περιγραφών το σύστημα THESUS βασίζεται στη χρήση του εννοιολογικού θησαυρού WordNet. Η απευθείας χρήση του θησαυρού είναι εφικτή μόνο για γλώσσες στις οποίες είναι διαθέσιμο το WordNet. Για τις υπόλοιπες γλώσσες απαιτείται ενδιάμεση μετάφραση με χρήση λεξικού.

Η κλιμάκωση του συστήματος σε ολόκληρο το εύρος του ΠΙ και κατά συνέπεια σε ένα πολυθεματικό πεδίο εφαρμογής εισάγει πολλά ενδιαφέροντα προβλήματα, όπως η εύρεση μιας «παγκόσμιας» οντολογίας, η συγχώνευση των επιμέρους οντολογιών [SC97], η βελτιστοποίηση της αρχιτεκτονικής για την αποτελεσματική διαχείριση μεγάλου όγκου δεδομένων.

4 Χαρακτηρισμός των εγγράφων με βάση τους υπερσυνδέσμους – Η γλώσσα του THESUS

4.1 Το μοντέλο πληροφορίας του THESUS

Στην ενότητα αυτή παρέχεται μια άτυπη εισαγωγή στο μοντέλο πληροφορίας του THESUS, η οποία μας προετοιμάζει για τους ορισμούς της γλώσσας του THESUS που θα ακολουθήσουν. Παρέχονται επίσης και οι τυπικοί ορισμοί των τύπων δεδομένων του THESUS.

Στα πλαίσια του THESUS, θεωρούμε ότι ο Παγκόσμιος Ιστός είναι μια συλλογή από:

- ***έγγραφα*** που προσδιορίζονται με τρόπο μοναδικό από τη διεύθυνση ιστού τους και περιέχουν κείμενο. Στη συνέχεια χρησιμοποιούμε εναλλακτικά τους όρους σελίδα, κόμβος και έγγραφο για να αναφερθούμε στα έγγραφα του ΠΙ. Να υπογραμμίσουμε ότι τα έγγραφα μπορεί να μην περιέχουν κείμενο (π.χ. όταν πρόκειται για εικόνες) αλλά να "αναφέρονται" από άλλα έγγραφα.
- ***συνδέσμους*** που συνδέουν τα έγγραφα. Ένας σύνδεσμος ορίζεται αποκλειστικά από τον αρχικό και τον τερματικό κόμβο. Οι σύνδεσμοι ανάμεσα σε δύο έγγραφα θεωρούνται ως "αναφορές", του αρχικού εγγράφου προς το τελικό, που φέρουν σημασιολογική πληροφορία. Ενδιαφερόμαστε λοιπόν για την ένωση της πληροφορίας που φέρουν όλοι οι σύνδεσμοι από ένα έγγραφο προς ένα άλλο, και συνεπώς δε μας απασχολεί η ακριβής θέση του συνδέσμου στο έγγραφο.

Σημασιολογία συνδέσμων

Ας θεωρήσουμε δύο έγγραφα S και T και το σύνολο των συνδέσμων $\{l_i\}$ του S προς το T. Ας θεωρήσουμε επίσης μια μέθοδο που για κάθε σύνδεσμο l_i επιστρέφει ένα σύνολο από λέξεις $\{k_j\}$ που χαρακτηρίζουν το σύνδεσμο. Όταν οι συγγραφείς ενός εγγράφου του ιστού δημιουργούν ένα σύνδεσμο προς ένα άλλο έγγραφο, χρησιμοποιούν ένα σύνολο λέξεων για να περιγράψουν το έγγραφο αυτό. Αυτές οι λέξεις εμφανίζονται στον πηγαίο κώδικα, του εγγράφου που περιέχει το σύνδεσμο, είτε εντός του συνδέσμου (είναι το κείμενο που αποτελεί το σύνδεσμο), είτε σε μια περιοχή γύρω από αυτόν. Στην περίπτωση εικόνων που λειτουργούν ως σύνδεσμοι, οι συγγραφείς συνήθως χρησιμοποιούν το χαρακτηριστικό ***alt*** για να περιγράψουν το έγγραφο στο οποίο δείχνει ο σύνδεσμος. Αυτή η πληροφορία αποτελεί το λεξικό χαρακτηρισμό του συνδέσμου και την αποκαλούμε **λέξεις συνδέσμων**.

Η σημασιολογία είναι ο κλάδος της σημειολογίας που εξετάζει τις σχέσεις σημείου (π.χ. λέξης) και σημασίας [M86], ή πιο απλά μελετά τη σημασία των λέξεων. Αν λοιπόν θεωρήσουμε το σύνολο των λέξεων που χαρακτηρίζουν ένα έγγραφο, ή ένα σύνδεσμο, και υποθέσουμε μια οντολογία και ένα μηχανισμό απεικόνισης λέξεων σε έννοιες, τότε μπορούμε να ορίσουμε ως *σημασιολογικό χαρακτηρισμό* ενός εγγράφου ή ενός συνδέσμου τις έννοιες στις οποίες αντιστοιχίζονται οι λέξεις που τα χαρακτηρίζουν. Στο THESUS αντιστοιχίζουμε τις εξαγόμενες λέξεις σε έννοιες, χρησιμοποιώντας μια οντολογία και έναν εννοιολογικό θησαυρό (το Wordnet), μετατρέποντας έτσι τις λέξεις των συνδέσμων σε **σημασιολογικά χαρακτηριστικά συνδέσμων** (link semantics).

Χαρακτηρισμός ενός κόμβου με βάση τα σημασιολογικά χαρακτηριστικά των εισερχόμενων συνδέσμων
Σύμφωνα με όσα προαναφέρθηκαν, ένα έγγραφο T χαρακτηρίζεται αποτελεσματικά από τις λέξεις που φέρουν οι εισερχόμενοι σύνδεσμοι. Γι' αυτό το σύνολο των λέξεων που προκύπτει για κάθε σύνδεσμο $\{l_i, \{k_j\}\}$ χαρακτηρίζει το περιεχόμενο του εγγράφου T όπως το βλέπει το έγγραφο S, και κατ' επέκταση η ένωση όλων των συνόλων λέξεων $\{k_j\}$ αποτελεί το λεξικό χαρακτηρισμό που το έγγραφο S προσδίνει στο T.

Αν λάβουμε υπόψη όλους τους εισερχόμενους συνδέσμους ενός εγγράφου T από διάφορα έγγραφα $\{S_i\}$ το αποτέλεσμα θα είναι ένας καθολικός χαρακτηρισμός, μια συλλογική άποψη, του συνόλου $\{S_i\}$ για το έγγραφο T.

Επεκτείνοντας το παράδειγμα, στη θέση του μοναδικού εγγράφου T θεωρούμε ένα σύνολο εγγράφων $\{T_i\}$. Το αποτέλεσμα της επεξεργασίας των συνδέσμων από το $\{S_i\}$ στο $\{T_i\}$ θα είναι ο συλλογικός χαρακτηρισμός που το σύνολο $\{S_i\}$ προσδίδει στο σύνολο $\{T_i\}$. Η προηγούμενη παρατήρηση δικαιολογεί τον ορισμό ενός τελεστή `groupKeywords`, που παίρνει ως είσοδο ένα σύνολο από αρχικές σελίδες $\{S_i\}$ και ένα δεύτερο σύνολο από τελικές σελίδες $\{T_i\}$ και επιστρέφει το σύνολο των λέξεων που προκύπτουν από την επεξεργασία όλων των συνδέσμων από σελίδες του $\{ S_i \}$ προς οποιαδήποτε σελίδα του $\{T_i\}$.

Ανάγκη για περιήγηση του ΠΙ και συλλογή σελίδων
Αν τα σύνολα S και T δεν είναι προκαθορισμένα, ο τελεστής που περιγράφηκε προηγουμένως προϋποθέτει τη δυνατότητα διάσχισης του γράφου του ΠΙ έτσι ώστε: α) να βρεθούν τα έγγραφα που δείχνονται από τους εξερχόμενους συνδέσμους ενός εγγράφου S και β) να βρεθούν τα πηγαία έγγραφα των εισερχόμενων συνδέσμων ενός εγγράφου T για να εξαχθεί η πληροφορία των συνδέσμων. Η διαδικασία περιήγησης μπορεί να λάβει χώρα σε πολλά επίπεδα του γράφου ξεκινώντας από κάποιον αρχικό κόμβο ενδιαφέροντος.

4.1.1 Τύποι δεδομένων του μοντέλου πληροφορίας του THESUS

Στην ενότητα αυτή καθορίζονται λεπτομερώς οι οντότητες που αποτελούν τον πυρήνα της γλώσσας του THESUS. Ο χώρος αναφοράς είναι ο Παγκόσμιος Ιστός που θεωρείται ως συλλογή εγγράφων (`w_docs`) που συνδέονται με συνδέσμους (`w_links`).

Για μια θεματική περιοχή θεωρούμε ότι ένα υποσύνολο των εγγράφων του ΠΙ, που σχετίζονται με ένα συγκεκριμένο θέμα, και των συνδέσμων τους (`docs` και `links` αντίστοιχα) αποτελεί ένα θεματικό υποσύνολο (THESU). Για λόγους πληρότητας θεωρούμε τους τύπους `URL` και `keyword`. Με τον όρο `URL` ορίζουμε μοναδικά ένα έγγραφο του ΠΙ και με τον όρο `keyword` μια συμβολοσειρά που χαρακτηρίζει ένα έγγραφο. Θεωρούμε επίσης τα σύνολα από `URL` και `keyword` ως τύπους του THESUS.

Στη συνέχεια ορίζουμε τις βασικές οντότητες, απαραίτητες για τη δημιουργία και διαχείριση των θεματικών υποσυνόλων του ΠΙ.

4.1.1.1 Ορισμοί του μοντέλου THESUS

Ορισμός 1: *Μια συλλογή από έγγραφα*, `docs`, είναι ένα σύνολο από πλειάδες της μορφής `(URL, {keyword}, other info)`, όπου α) το `URL` προσδιορίζει μοναδικά ένα έγγραφο του ΠΙ, β) `{keyword}` είναι ένα σύνολο από λέξεις που χαρακτηρίζουν ένα έγγραφο και γ) `other info` είναι επιπλέον πληροφοριακά στοιχεία που μπορούμε να αναθέσουμε σε ένα έγγραφο, όπως το περιεχόμενό του σε μορφή κειμένου, ή η ημερομηνία τελευταίας αλλαγής κτλ.

Ορισμός 2: *Μια συλλογή από συνδέσμους*, `links`, είναι ένα σύνολο από πλειάδες της μορφής `(URLS, URLT, {keyword})`, όπου: α) το `URLS` είναι το URL του εγγράφου από το οποίο ξεκινά ο σύνδεσμος, β) το `URLT` είναι το URL του εγγράφου στο οποίο δείχνει ο σύνδεσμος και γ) `{keyword}` είναι το σύνολο των λέξεων που περιέχονται στο έγγραφο και στη γύρω περιοχή.

4.1.1.2 Οντότητες του Παγκόσμιου Ιστού

Για να κατασκευάσουμε έγγραφα (`docs`) πρέπει να εξαγάγουμε πληροφορία από τις ιστοσελίδες. Τα σύνολα `w_docs` και `w_links` χρησιμοποιούνται για το λόγο αυτό και επιτρέπουν τη θεώρηση του ΠΙ ως μια εικονική οντότητα στο σύστημα THESUS. Τμήματα αυτής της οντότητας υλοποιούνται αν αποθηκεύσουμε τοπικά πληροφορίες για τα έγγραφα του ΠΙ, όπως κάνουν οι μηχανές αναζήτησης.

Ορισμός 3: Τα *έγγραφα του ΠΙ*, `w_docs`, είναι ένα σύνολο από πλειάδες της μορφής `(URL,text)`, όπου: α) το `URL` χαρακτηρίζει μοναδικά τα έγγραφα στον Ιστό και β) `text` είναι το περιεχόμενο του εγγράφου σε μορφή κειμένου.

Ορισμός 4: Οι *σύνδεσμοι του ΠΙ*, `w_links`, είναι ένα σύνολο από πλειάδες της μορφής `(URLS, URLT)`, όπου: α) το `URLS` είναι το URL του εγγράφου από το οποίο ξεκινά ο σύνδεσμο και β) το `URLT` είναι το URL του εγγράφου στο οποίο δείχνει ο σύνδεσμος. Στο σημείο αυτό θεωρούμε ότι όλοι οι σύνδεσμοι από μια σελίδα προς μια άλλη συγχωνεύονται σε μια πλειάδα στο σύνολο `w_links`.

4.2 Η γλώσσα THESUS

Στην ενότητα αυτή παρουσιάζεται η γλώσσα του THESUS, που επιτρέπει την επιλογή υποσυνόλων του ΠΙ με κριτήρια τον τρόπο σύνδεσης και τους σημασιολογικούς χαρακτηρισμούς συνδέσμων και εγγράφων. Επιτρέπει επίσης τον εμπλουτισμό εγγράφων ή συνόλων με εξαγωγή σημασιολογικής πληροφορίας από τους εισερχόμενους συνδέσμους τους. Η γλώσσα THESUS επίσης παρέχει τις μεθόδους για την υποβολή ερωτήσεων που βασίζονται στα χαρακτηριστικά σύνδεσης και στα σχετικά με αυτά σημασιολογικά χαρακτηριστικά, παράγοντας έτσι πολύτιμα αποτελέσματα που δεν προσφέρουν οι υπάρχοντες μηχανισμοί αναζήτησης.

Στη συνέχεια ακολουθεί ο τυπικός ορισμός των βασικών τελεστών και των απαραίτητων λειτουργιών για τον καθορισμό ενός θεματικού υποσυνόλου (THESU). Οι τύποι δεδομένων, οι τελεστές και οι απαραίτητες επεκτάσεις και βοηθητικές συναρτήσεις συνοψίζονται στο Παράρτημα Β.

Θέματα σχεδίασης της γλώσσας

Κατά το σχεδιασμό μιας γλώσσας ή ενός συνόλου κανόνων πρέπει να εξισορροπούνται δύο σημαντικοί παράγοντες: η λιτότητα και η εκφραστικότητα.

- Στην περίπτωση της ***λιτότητας*** αναζητούμε *το* ***ελάχιστο δυνατό σύνολο τελεστών που δεν επικαλύπτονται*** σε λειτουργικότητα, που είναι απλοί στη σχεδίασή τους και που μπορούν να συνδυαστούν στη δημιουργία πιο σύνθετων λειτουργιών.
- Στην περίπτωση της ***εκφραστικότητας*** επιδιώκουμε ***ευκολία στη χρήση των τελεστών***, οι οποίοι ορίζονται σε ένα υψηλό επίπεδο, γίνονται εύκολα κατανοητοί από τους χρήστες και έχουν ***σαφή σημασία και λειτουργικότητα***.

Το μοντέλο του THESUS εξισορροπεί τους δύο παράγοντες ορίζοντας ένα σύνολο από τελεστές, οργανωμένους σε διακριτές ομάδες (τελεστές περιήγησης, τελεστές εξαγωγής σημασιολογίας από τους συνδέσμους, τελεστές ανάλυσης συνδεσμολογίας), που μπορούν να εφαρμοστούν άμεσα και να εξυπηρετήσουν τις απαιτήσεις του συστήματος (π.χ., διάσχιση του γράφου του ΠΙ, εξαγωγή σημασιολογικών χαρακτηριστικών κτλ). Παρ' όλα αυτά, υπάρχει επικάλυψη μεταξύ των τελεστών, καθώς ορισμένοι τελεστές μπορούν να εκφραστούν ως συνδυασμός άλλων.

Σημαντικό θέμα αποτελεί και η ***κλειστότητα*** του συνόλου των τελεστών έτσι ώστε οι ***παράμετροι και τα αποτελέσματα των τελεστών να εκφράζονται με τύπους που έχουν οριστεί στο μοντέλο***. Το σύνολο των τελεστών του THESUS είναι κλειστό ως προς τους τύπους `{URL}` και `{keyword}`.

Επεκτάσεις της γλώσσας

Εκτός από το βασικό σύνολο τελεστών, ορίζουμε ένα σύνολο ειδικών τελεστών μεγάλης προστιθέμενης αξίας. Ένα τέτοιο παράδειγμα είναι οι τελεστές που ασχολούνται με την ανάλυση της συνδεσμολογίας των σελίδων. Τέτοιοι τελεστές επιτρέπουν την ολοκλήρωση πληροφορίας που αφορά τον τρόπο σύνδεσης των εγγράφων και προσφέρουν πολύτιμη πληροφορία γι' αυτές. Επιπλέον οι χρήστες του συστήματος μπορούν να ορίσουν νέους τελεστές με συνδυασμό των βασικών τελεστών.

Θεωρούμε τα ακόλουθα χαρακτηριστικά: α) τον αριθμό των εξερχόμενων συνδέσμων μιας σελίδας (μια σελίδα θεωρείται *σημαντικός* ***κόμβος παραπομπών*** (**hub**) [K99] αν έχει πολλούς εξερχόμενους συνδέσμους), β) τον αριθμό των εισερχόμενων συνδέσμων μιας σελίδας (μια σελίδα θεωρείται *σημαντικός* ***κόμβος-αυθεντία*** (**authority**) [K99] αν τη "δείχνουν" πολλές άλλες σελίδες) και γ) πρότυπα σύνδεσης όπως **συν-*αναφορές*** (**co-citations** [Sm73]) και ***βιβλιογραφικές συζεύξεις*** (**couplings** [K63]) για δύο ή περισσότερες σελίδες. Τα δύο τελευταία χαρακτηριστικά θεωρούνται ενδείξεις ομοιότητας δύο ή περισσοτέρων σελίδων. Ειδικότερα, οι συν-αναφορές αντιστοιχούν στο πλήθος των κοινών εισερχόμενων συνδέσμων για τις σελίδες ενός συνόλου, ενώ οι βιβλιογραφικές συζεύξεις αντιστοιχούν στο πλήθος των κοινών εξερχόμενων συνδέσμων για δύο ή περισσότερες σελίδες [WK01].

Στη συνέχεια, για λόγους συντομίας θα αναφερόμαστε στους κόμβους παραπομπών και στους κόμβους αυθεντίες με τους όρους ***π-κόμβος*** και ***α-κόμβος*** αντίστοιχα.

Ο Kleinberg [K99] παρέχει ένα πιο σύνθετο ορισμό των σημαντικών π-κόμβων και α-κόμβων, σύμφωνα με τον οποίο ο βαθμός σημαντικότητας ενός κόμβου (και στις δύο

περιπτώσεις) υπολογίζεται αναδρομικά. Ένας κόμβος με υψηλό βαθμό σημαντικότητας ως π-κόμβος θα δείχνει σε πολλούς κόμβους, οι οποίοι θα είναι αντίστοιχα σημαντικοί α-κόμβοι, και αντίστροφα. Στο THESUS για απλούστευση χρησιμοποιούμε τον απλό ορισμό. Εκμεταλλευόμαστε και εμπλουτίζουμε τους πιο πάνω ορισμούς με σημασιολογικά χαρακτηριστικά, που μπορούν να εξαχθούν από τους συνδέσμους. Εισάγουμε έτσι τις παρακάτω έννοιες:

Θεματικοί π-κόμβοι: Όπως προαναφέρθηκε, ένας π-κόμβος έχει πολλούς εξερχόμενους συνδέσμους, κάποιοι από τους οποίους έχουν παρόμοια σημασιολογικά χαρακτηριστικά. Τα χαρακτηριστικά αυτά προσδιορίζουν το θέμα του π-κόμβου. Έχουμε έτσι μια ένδειξη της αξίας του συγκεκριμένου κόμβου στα πλαίσια του συγκεκριμένου θέματος.

Γενικεύοντας, μπορούμε να πούμε ότι αν κάποια σημασιολογικά χαρακτηριστικά εμφανίζονται στους περισσότερους από τους εξερχόμενους συνδέσμους, αυτό αποτελεί ένδειξη ότι η σελίδα είναι σημαντική και μάλιστα ότι επικεντρώνεται σε κάποιο θέμα. Αν επισκεφθούμε τον κόμβο, θα βρούμε πολλούς συνδέσμους σχετικούς με το θέμα του κόμβου.

Θεματικοί α-κόμβοι: Ένας α-κόμβος έχει πολλούς εισερχόμενους συνδέσμους. Η σημασιολογία του κόμβου εμπλουτίζεται αν συλλέξουμε τα σημασιολογικά χαρακτηριστικά που μεταφέρουν οι σύνδεσμοι αυτοί. Αν, για παράδειγμα, ένας κόμβος υποδεικνύεται από πολλούς εισερχόμενους συνδέσμους με την έννοια "βάση δεδομένων", είναι πολύ πιθανό ότι η σελίδα περιέχει πληροφορίες για βάσεις δεδομένων.

Θεματικές συν-αναφορές και συζεύξεις: Οι έννοιες των συναναφορών και συζεύξεων μπορούν να εμπλουτιστούν με σημασιολογικά χαρακτηριστικά. Για παράδειγμα, αν δύο κόμβοι Β και Γ δείχνουν προς έναν κόμβο Α, τότε ο Α αποτελεί σύζευξη των Β και Γ. Επιπλέον, αν οι Β και Γ χρησιμοποιούν παρόμοιες έννοιες για να περιγράψουν τον Α τότε η σημασιολογία του Α καθορίζεται καλύτερα και επιπλέον η ένδειξη ότι οι Β και Γ είναι όμοιοι ισχυροποιείται καθώς όχι μόνο υποδεικνύουν την ίδια σελίδα, αλλά τη χαρακτηρίζουν και με παρόμοιο τρόπο.

Παρόμοια, η έννοια των θεματικών συν-αναφορών ενισχύει την ομοιότητα ανάμεσα σε κόμβους που "δείχνονται" από κάποιον ξένο κόμβο με τα ίδια σημασιολογικά χαρακτηριστικά. Για παράδειγμα, αν τα έγγραφα Α και Β δείχνονται από το Γ με συνδέσμους που φέρουν παρόμοιες έννοιες, τότε έχουμε μια ένδειξη ότι τα Α και Β έχουν παρόμοιο περιεχόμενο.

Σημασιολογική περιγραφή εγγράφου: Όπως προαναφέρθηκε, ο χαρακτηρισμός ενός εγγράφου επηρεάζεται σημαντικά από τα σημασιολογικά χαρακτηριστικά που φέρουν οι εισερχόμενοι σύνδεσμοι. Λαμβάνοντας συνολικά υπόψη τη σημασιολογία των εισερχόμενων συνδέσμων, μπορούμε να καθορίσουμε τη σημασιολογική περιγραφή του εγγράφου.

Στην ενότητα που ακολουθεί ορίζονται οι βασικοί τελεστές που περιγράφηκαν προηγουμένως, αλλά και οι επιπλέον τελεστές που μεταφέρουν πλούσια γνώση για το σύνολο των εγγράφων του ΠΙ. Το σύνολο τελεστών που ακολουθεί, δεν είναι

εξαντλητικό, αλλά προσφέρει μια ένδειξη του βασικού συνόλου τελεστών του THESUS.

4.2.1 Τελεστές περιήγησης

Ορίζουμε τέσσερις τελεστές περιήγησης:

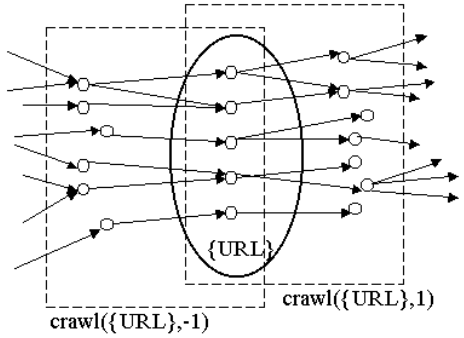
Ορισμός 5: `fetch(URL)` Αν του δώσουμε ένα URL, επιστρέφει είτε το περιεχόμενό του είτε null αν το URL δεν υπάρχει ή δεν αντιστοιχεί σε κάποιο έγγραφο.

Ο επόμενος τελεστής παίρνει ως όρισμα μια πρότυπη διεύθυνση εγγράφου URL και επιστρέφει ένα σύνολο από διευθύνσεις εγγράφων που ταιριάζουν στο αρχικό URL. Χωρίς να εμβαθύνουμε στις λεπτομέρειες δημιουργίας ενός συστήματος περιήγησης του ΠΙ, απαιτούμε από το σύστημα να υλοποιεί τους τελεστές που αναφέρονται στη συνέχεια.

Ορισμός 6: Αν δώσουμε ένα σύνολο από πρότυπες διευθύνσεις ιστού URL, `{URLexpr}` ο τελεστής **`groupMatch({URLexpr})`** επιστρέφει το σύνολο των διευθύνσεων που ταιριάζουν στα αρχικά URLs. Η πρότυπη διεύθυνση είναι μια ημιτελής διεύθυνση και όλες οι διευθύνσεις που επιστρέφονται θα πρέπει να ξεκινούν με αυτή.

Παράδειγμα: Η χρήση του τελεστή ως εξής:
`groupMatch({http://www.nasa.gov/, http://www.ibm.com/})`
επιστρέφει όλα τα έγγραφα κάτω από τους δικτυακούς τόπους http://www.nasa.gov/ or http://www.ibm.com. Τα έγγραφα συλλέγονται ξεκινώντας από τις κεντρικές ιστοσελίδες των δικτυακών τόπων και ακολουθώντας όλους τους συνδέσμους προς έγγραφα των ίδιων τόπων.

Ορισμός 7: Για ένα σύνολο από διευθύνσεις ιστού `({URL})`, και ένα ακέραιο `N` ο τελεστής **`crawl({URL},N)`** επιστρέφει ένα σύνολο από διευθύνσεις, αυτές των εγγράφων, που δείχνονται από οποιοδήποτε έγγραφο του `{URL}` σε οποιοδήποτε επίπεδο, μικρότερο από N, του γράφου του ΠΙ.



Σχήμα 17. Σχηματική αναπαράσταση του τελεστή crawl

Θετικές τιμές του N υποδηλώνουν ότι η περιήγηση του ΠΙ γίνεται "προς τα εμπρός", ακολουθώντας δηλαδή τους εξερχόμενους συνδέσμους των εγγράφων, ενώ αρνητικές τιμές του N υποδηλώνουν περιήγηση "προς τα πίσω", ακολουθώντας τους εισερχόμενους συνδέσμους των εγγράφων στο σύνολο {URL}. Το παράδειγμα του σχήματος 17 δείχνει ότι για N=1 ο τελεστής περιήγησης επιστρέφει τα έγγραφα του {URL} καθώς και όλα τα έγγραφα που δείχνονται από αυτά, ενώ για N= –1 επιστρέφει τα έγγραφα του {URL} και όσα έγγραφα δείχνουν προς αυτά.
Στη συνέχεια δίνεται ο ορισμός του τελεστή σε ψευδοκώδικα:

```
crawl({URL}, N):-
      RES = {}
      if N<=-1
            ∀ d in {URL}
                  if w_links.URLT = d            //βρες τις πλειάδες
                                                 //που έχουν στόχο d
                        RES = RES ∪ w_link.URLS  //πρόσθεσε την αντί-
                                                 //στοιχη πηγή
       return {URL} ∪ crawl(RES, N+1)


       if N=0
        return {URL}
       if N>=1
             ∀ d in {URL}
                  if w_links.URLS = d
                        RES = RES ∪ w_link.URLT
        return {URL} ∪ crawl(RES, N-1)
```

Να σημειωθεί ότι ο αλγόριθμος δεν είναι βέλτιστος και δίνεται στην παρούσα μορφή για λόγους απλότητας.

Παράδειγμα: Στην περίπτωση της `crawl({U1}, 1)`, ο τελεστής επιστρέφει τις διευθύνσεις όλων των εγγράφων που μπορεί κανείς να επισκεφθεί ξεκινώντας από τη U_1 και ακολουθώντας όλους τους συνδέσμους που πηγάζουν από το U_1.

4.2.2 Τελεστές σημασιολογίας των συνδέσμων

Αρχικά προσδιορίζουμε ένα σύνολο από βοηθητικές συναρτήσεις απαραίτητες για τον καθορισμό των τελεστών. Για να εξαγάγουμε σημασιολογική πληροφορία από τους συνδέσμους πρέπει πρώτα να καθορίσουμε τη θέση του συνδέσμου (π.χ. της συμβολοσειράς <A HREF='…'>) μέσα στην ιστοσελίδα και στη συνέχεια να επεξεργαστούμε το γειτονικό υπερκείμενο για να εξαγάγουμε λέξεις.

Για το λόγο αυτό χρησιμοποιούμε τη μέθοδο `getpos(text, URLT)` η οποία με δεδομένο το κείμενο ενός εγγράφου, `text`, και τη συμβολοσειρά που αντιστοιχεί στη διεύθυνση ιστού του εγγράφου στόχου, `URLT`, επιστρέφει ένα σύνολο από θέσεις {`pos`} που αντιστοιχούν στα σημεία εμφάνισης της συμβολοσειράς URLT στο κείμενο του εγγράφου.

Στη συνέχεια επεξεργαζόμαστε τη γειτονική περιοχή υπερκειμένου (π.χ. 100 χαρακτήρες πριν και μετά τα όρια του συνδέσμου). Αυτό γίνεται με κλήση της μεθόδου `process(w_docs.TEXT, pos, 100)` (χρησιμοποιεί την getkeys, βλ.

ενότητα 3.2.1) για κάθε μια θέση εμφάνισης (pos) του URLT. Η μέθοδος αυτή επιστρέφει το σύνολο των λέξεων που εξήχθησαν και αντιπροσωπεύουν τη σημασιολογική πληροφορία του συνδέσμου.

Καθώς θεωρούμε μοναδικούς συνδέσμους από ένα έγγραφο S σε ένα άλλο T, ορίζουμε τον τελεστή linkKeywords που επιστρέφει την ένωση όλων των λέξεων που εμφανίζονται στους συνδέσμους από το έγγραφο S προς το T.

Ορισμός 8: Για δύο έγγραφα URLS και URLT ο τελεστής **linkKeywords (URLS, URLT)** επιστρέφει το σύνολο των λέξεων που εξάγονται απ΄ όλους τους συνδέσμους του URLS προς το URLT.

Ακολουθεί ο ορισμός του τελεστή σε ψευδοκώδικα:

```
linkKeywords (URLS, URLT):-
let Stext =  select w_docs.text where
        w_docs.URL = URLS
KEYS = {}
∀ d in getpos(Stext, URLT)
       KEYS = KEYS ∪ process(w_docs.TEXT, d, 100)
return KEYS
```

Ορισμός 9: Για δύο σύνολα διευθύνσεων εγγράφων {URLS} και {URLT} ο τελεστής **groupKeywords ({URLS}, {URLT})** επιστρέφει τις λέξεις που εξάγονται από όλους τους συνδέσμους που ξεκινούν από έγγραφα του {URLS} και καταλήγουν σε έγγραφα του {URLT}.

Ακολουθεί ο ορισμός του τελεστή σε ψευδοκώδικα:

```
groupKeywords ({URLS}, {URLT}) :-
KEYS = {}
∀ d in {URLS}
      ∀ e in {URLT}
      KEYS = KEYS ∪ linkKeywords(d,e)
return KEYS
```

Τα σύνολα διευθύνσεων εγγράφων {URLS} ή {URLT} στον τελεστή groupKeywords μπορεί να περιέχουν μία μόνο ή και καμία διεύθυνση. Στην πρώτη περίπτωση, αντί για τα {URLS} και {URLT} έχουμε τις διευθύνσεις sourceURL και targetURL αντίστοιχα, έτσι η λειτουργικότητα του τελεστή groupKeywords περιορίζεται σε αυτή του linkKeywords. Στη δεύτερη περίπτωση, αν το σύνολο των πηγαίων διευθύνσεων ιστού {URLS} είναι κενό, ο groupKeywords εξετάζει όλους τους εισερχόμενους συνδέσμους σε έγγραφα του {URLT}, ενώ, αν το σύνολο των τελικών διευθύνσεων ιστού {URLT} είναι κενό, ο τελεστής groupKeywords εξετάζει όλους τους εξερχόμενους συνδέσμους των εγγράφων του {URLS}.

Στη συνέχεια ακολουθεί μια σύνοψη των διαφορετικών ερμηνειών του τελεστή groupKeywords για διαφορετικούς συνδυασμούς των παραμέτρων εισόδου.

groupKeywords({URLS},e) επιστρέφει την ένωση όλων των linkKeywords μεταξύ κάθε εγγράφου d του {URLS} και της e.

groupKeywords({URLS},∅) επιστρέφει την ένωση όλων των `linkKeywords` μεταξύ κάθε εγγράφου d του {URLS} και κάθε εγγράφου e που δείχνεται από κάποιο από τα έγγραφα του {URLS} (e ∈ crawl({URLS},1)-{URLS}). Η πληροφορία αυτή υποδεικνύει τον τρόπο με τον οποίο τα έγγραφα του {URLS} περιγράφουν τα έγγραφα στα οποία δείχνουν, με άλλα λόγια, πώς τα έγγραφα του {URLS} "αντιλαμβάνονται τον κόσμο".

groupKeywords (∅, {URLT}) επιστρέφει την ένωση όλων των `linkKeywords` μεταξύ κάθε εγγράφου d που δείχνει προς ένα έγγραφο e του {URLT} και του e. Αυτή η πληροφορία υποδηλώνει τον τρόπο με τον οποίο περιγράφουν τα έγγραφα του {URLT} όσοι αναφέρονται σε αυτά, δηλαδή "τι πιστεύει ο κόσμος" για τα έγγραφα του {URLT}.

Ορισμός 10: Οι ορισμοί 7 και 8 συνδυάζονται στον ορισμό του τελεστή **`thematicCrawl({URL},{keyword},N)`**, ο οποίος για ένα σύνολο διευθύνσεων εγγράφων (`{URL}`), ένα σύνολο λέξεων (`{keyword}`) και ένα ακέραιο N επιστρέφει ένα σύνολο διευθύνσεων ιστού. Αυτό το σύνολο διευθύνσεων ιστού αποτελείται από έγγραφα που δείχνονται από κάθε έγγραφο στο `{URL}` σε οποιοδήποτε επίπεδο του γράφου του ΠΙ μικρότερο του N, με χρήση μιας τουλάχιστον λέξης από το σύνολο `{keyword}`(link keywords). Σε κάθε διεύθυνση που έχουμε στο τελικό σύνολο, φτάνουμε ακολουθώντας N το πολύ συνδέσμους καθένας από τους οποίους περιέχει μια τουλάχιστον λέξη της οντολογίας.

Ο τελεστής ονομάζεται `thematicCrawl` καθώς, αν χρησιμοποιηθούν λέξεις από κοινό θεματική περιοχή, οι διευθύνσεις που συλλέγονται σχετίζονται θεματικά. Να σημειωθεί ότι προς το παρόν δεν έχει εφαρμοστεί η σημασιολογική επεξεργασία της πληροφορίας.

Ακολουθεί ο ορισμός του τελεστή σε ψευδοκώδικα:

```
thematicCrawl({URL}, {keyword}, N):-
      RES = {}
      if N<=-1
            ∀ d in {URL}
                  if w_links.URLT = d
                  AND linkKeywords(w_link.URLS,d)∩{keyword} ≠ ∅
                        RES = RES ∪ w_link.URLS
       return {URL}∪ thematicCrawl(RES, {keyword}, N+1)
      if N=0
       return {URL}
      if N>=1
            ∀ d in {URL}
                  if w_links.URLS = d
                  AND linkKeywords(d,w_link.URLT)∩{keyword} ≠ ∅
                        RES = RES ∪ w_link.URLT
       return {URL}∪ thematicCrawl(RES, {keyword}, N-1)
```

Να σημειωθεί ότι ο αλγόριθμος δεν είναι βέλτιστος και δίνεται στην παρούσα μορφή για να αναδείξει τη λογική του τελεστή.

Παράδειγμα: Ας θεωρήσουμε τον τελεστή:

`thematicCrawl({U1},{keyword1,keyword2},2)`

ξεκινώντας από ένα έγγραφο στη διεύθυνση `U1` ο τελεστής επιστρέφει όλες τις διευθύνσεις εγγράφων στις οποίες δείχνουν σύνδεσμοι που ξεκινούν από τη `U1`, και μέχρι 2 βήματα και περιέχουν είτε τη λέξη keyword1 είτε τη λέξη keyword2.

4.2.2.1 Σύνθετοι τελεστές σημασιολογίας συνδέσμων

Ο ορισμός του τελεστή `groupKeywords` υποθέτει ότι το τελικό σύνολο από λέξεις είναι η ένωση των λέξεων που εμφανίστηκαν σε κάθε σύνδεσμο, χωρίς να λαμβάνει υπόψη τον αριθμό εμφανίσεων των λέξεων αυτών. Παρ' όλα αυτά, αν λάβουμε υπόψη τον αριθμό εμφανίσεων των λέξεων, ο τροποποιημένος ορισμός της `groupKeywords` αποκτά τρεις διαφορετικές σημασίες ανάλογα με τον τρόπο με τον οποίο αθροίζονται οι εμφανίσεις των λέξεων:

α) αν αθροίσουμε τις εμφανίσεις κάθε λέξης, το αποτέλεσμα της `groupKeywords` θα είναι ένα σύνολο από ζεύγη (KEY,TIMES) όπου TIMES είναι ο αριθμός εμφανίσεων κάθε λέξης KEY στους συνδέσμους από το {URLS} προς το {URLT} (ορισμός 11),

β) αν αθροίσουμε ανά τελικό έγγραφο, τότε το TIMES αντιπροσωπεύει τον αριθμό των διαφορετικών (τελικών) εγγράφων που χαρακτηρίστηκαν με τη λέξη KEY (βλ. ορισμό 12) και

γ) αν αθροίσουμε ανά αρχικό έγγραφο, το TIMES είναι ο αριθμός των διαφορετικών (αρχικών) εγγράφων που χρησιμοποίησαν τη λέξη KEY στους συνδέσμους τους προς τα έγγραφα του {URLT} (βλ. ορισμό 13).

Στους ορισμούς που ακολουθούν χρησιμοποιούμε τρεις βοηθητικές συναρτήσεις:

- ***getkeys(WKEYS)***, η οποία παίρνει ένα σύνολο ζευγών λέξης-αριθμού, το WKEYS = {(KEY,TIMES)}, σαν είσοδο και επιστρέφει το σύνολο των λέξεων μόνο {KEY},
- ***getTimes(WKEYS,K)***, η οποία παίρνει ένα σύνολο ζευγών λέξης-αριθμού, το WKEYS = {(KEY,TIMES)}, και μια λέξη K σαν είσοδο και επιστρέφει τον αριθμό εμφανίσεων (TIMES) της λέξης K, αν το συγκεκριμένο ζεύγος υπάρχει στο WKEYS. Διαφορετικά, επιστρέφει 0.
- ***update(WKEYS,K,N)***, η οποία παίρνει ένα σύνολο ζευγών {(KEY,TIMES)}, μια λέξη K και ένα ακέραιο N σαν είσοδο και αυξάνει το TIMES της λέξης K κατά N (ενημερώνει το ζεύγος (K,T) της WKEYS σε (K,T+N)).

Ορισμός 11: Για δύο σύνολα διευθύνσεων ιστού `{URLS}` και `{URLT}` ο τελεστής **`weightedGroupKeywords ({URLS}, {URLT})`** επιστρέφει ένα σύνολο ζευγών {(KEY, TIMES)} όπου το KEY αντιπροσωπεύει μια λέξη που εμφανίζεται στους συνδέσμους από το `{URLS}` στο `{URLT}` και το TIMES τον αριθμό εμφανίσεων της λέξης KEY.

Ακολουθεί ο ορισμός του τελεστή σε ψευδοκώδικα:

```
weightedGroupKeywords ({URLS}, {URLT}) :-
WKEYS = {}
∀ d in {URLS}
      ∀ e in {URLT}
            ∀ K in linkKeywords(d,e)
            if getkeys(WKEYS) ∩ {K} ≠∅
                  update(WKEYS,K,1)
```

```
            else WKEYS=WKEYS ∪ (K,1)
return WKEYS
```

Ορισμός 12: Για δύο σύνολα διευθύνσεων ιστού {URLS} και {URLT} ο τελεστής **weightedTargetKeywords ({URLS}, {URLT})** επιστρέφει ένα σύνολο από ζεύγη {(KEY,TIMES)} όπου το KEY αντιπροσωπεύει μια λέξη που εμφανίζεται στους συνδέσμους από το {URLS} στο {URLT} και το TIMES τον αριθμό διευθύνσεων εγγράφων του {URLT} που χαρακτηρίζονται με τη λέξη KEY.

Ακολουθεί ο ορισμός του τελεστή σε ψευδοκώδικα:

```
weightedTargetKeywords ({URLS}, {URLT}) :-
WKEYS = {}
      ∀ e in {URLT}
            ∀ K in groupKeywords({URLS},e)
                  if getkeys(WKEYS) ∩ {K} ≠∅
                        update(WKEYS,K,1)
                  else WKEYS = WKEYS ∪ (K,1)
return WKEYS
```

Ορισμός 13: Για δύο σύνολα διευθύνσεων ιστού {URLS} και {URLT} ο τελεστής **weightedSourceKeywords ({URLS}, {URLT})** επιστρέφει ένα σύνολο από ζεύγη {(KEY,TIMES)}, όπου το KEY αντιπροσωπεύει μια λέξη που εμφανίζεται στους συνδέσμους από το {URLS} στο {URLT} και το TIMES τον αριθμό των εγγράφων του {URLS} που χρησιμοποιούν τη λέξη KEY στους συνδέσμους προς τις διευθύνσεις του {URLT}.

Ακολουθεί ο ορισμός του τελεστή σε ψευδοκώδικα:

```
weightedSourceKeywords ({URLS}, {URLT}) :-
WKEYS = {}
      ∀ d in {URLS}
            ∀ K in groupKeywords(d,{URLT})
                  if getkeys(WKEYS) ∩ {K} ≠∅
                        update(WKEYS,K,1)
                  else WKEYS=WKEYS ∪ (K,1)

return WKEYS
```

Τα παραδείγματα της παραγράφου 6.2.2 επιδεικνύουν τη διαφορετική σημασιολογία των τριών υλοποιήσεων της groupKeywords.

4.2.3 Τελεστές ανάλυσης συνδεσμολογίας

Το σύνολο των τελεστών ανάλυσης συνδεσμολογίας εμπλουτίζει με σημασιολογικά χαρακτηριστικά τους δύο βασικούς τελεστές που εμπλέκονται στην ανάλυση των συνδέσμων –τον Hubs και τον Authorities, που εντοπίζουν σημαντικούς π-κόμβους και α-κόμβους αντίστοιχα– και δύο εξίσου σημαντικούς τελεστές –τον Co-citations και τον Couplings, που εξετάζουν την ομοιότητα εγγράφων του ΠΙ με βάση τους κοινούς εισερχόμενους και εξερχόμενους συνδέσμους τους. Οι ορισμοί βασίζονται

στον απλοποιημένο ορισμό των Hubs και Authorities, αν και ο επαναληπτικός ορισμός που παρουσιάζεται στο [K99] μπορεί να επεκταθεί με σημασιολογικά χαρακτηριστικά.

Ορισμός 14: **Θεματικός α-κόμβος - Thematic Authority**. Για ένα σύνολο διευθύνσεων ιστού {URL}, έναν ακέραιο threshold και ένα σύνολο λέξεων {keyword} ο τελεστής **Tauthorities({URL}, threshold, {keyword})** επιστρέφει ένα σύνολο από διευθύνσεις σελίδων που έχουν πάνω από threshold *εισερχόμενους* συνδέσμους, καθένας από τους οποίους περιέχει όλες τις λέξεις του {keyword}.

Ένα σχεδίασμα του αλγορίθμου ακολουθεί:

```
   Tauthorities({URL}, threshold, {keyword}):-
   RES = {}
   ∀ d in {URL}
    WKEYS = weightedSourceKeywords(∅,d)
                              //επιστρέφει τις λέξεις που
                              //εμφανίζονται στους εισερχόμενους
                              //συνδέσμους του d και τον αριθμό των
                              //συνδέσμων ανά λέξη
    if ∀ K in {keyword} getTimes(WKEYS , K) > threshold
                              //αν οι εισερχόμενοι σύνδεσμοι για κάθε
                              //λέξη στο σύνολο  είναι περισσότεροι
                              //από το κατώφλι threshold
         RES = RES ∪ d    //τότε το έγγραφο είναι Θεματικός α-κόμβος
RETURN RES
```

Η κλήση της μεθόδου weightedSourceKeywords(∅,d) με πρώτο όρισμα το κενό σύνολο εντοπίζει λέξεις σε όλους τους εισερχόμενους συνδέσμους του d σε αντιστοιχία με τις κλήσεις της groupKeywords.
Σύμφωνα με τον ορισμό 14, ένα έγγραφο του ΠΙ θεωρείται θεματικός α-κόμβος αν για κάθε λέξη k του {keyword} υπάρχουν πάνω από threshold στο πλήθος σύνδεσμοι που δείχνουν προς το έγγραφο χρησιμοποιώντας τη λέξη k.

Ορισμός 15: Θεματικός π-κόμβος - Thematic Hub: Για ένα σύνολο διευθύνσεων ιστού {URL}, έναν ακέραιο threshold και ένα σύνολο λέξεων {keyword} ο τελεστής **THubs({URL}, threshold, {keyword})** επιστρέφει ένα σύνολο διευθύνσεων ιστού που έχουν πάνω από threshold στο πλήθος *εξερχόμενους* συνδέσμους, καθένας από τους οποίους περιέχει όλες τις λέξεις του {keyword}.

Ένα πρόχειρο σχεδίασμα του αλγορίθμου ακολουθεί:

```
THubs({URL}, threshold, {keyword}):-
   RES = {}
   ∀ d in {URL}
       WKEYS = weightedTargetKeywords(d, ∅)
                            //επιστρέφει τις λέξεις που εμφανίζονται
                            //στους εξερχόμενους συνδέσμους του d
                            //και τον αριθμό των συνδέσμων ανά λέξη
    if ∀ K in {keyword} getTimes(WKEYS , K) > threshold
                            //αν οι εξερχόμενοι σύνδεσμοι για κάθε
                            //λέξη στο σύνολο είναι περισσότεροι από
                            //το κατώφλι threshold
         RES = RES ∪ d    //τότε το έγγραφο είναι Θεματικός π-κόμβος
RETURN RES
```

Σύμφωνα με τον ορισμό αυτό, ένα έγγραφο του ΠΙ θεωρείται θεματικός π-κόμβος αν για κάθε λέξη k του {keyword} το έγγραφο έχει πάνω από threshold στο πλήθος συνδέσμους που χρησιμοποιούν τη λέξη k.

Ορισμός 16: **Θεματική Συν-αναφορά -Thematic Co-citation**: Για ένα σύνολο διευθύνσεων ιστού {URL} και ένα σύνολο λέξεων {keyword} ο τελεστής **TCocitations({URL}, {keyword})** επιστρέφει ένα σύνολο από διευθύνσεις ιστού που *όλες δείχνουν προς κάθε έγγραφο* του {URL} με συνδέσμους που φέρουν τις λέξεις του συνόλου {keyword}.

Ακολουθεί η περιγραφή του αλγορίθμου σε ψευδοκώδικα:

```
TCocitations({URL}, {keyword}):-
   RES = {}
   N= size({URL})            //το πλήθος των εγγράφων του {URL}
   TEMP = crawl({URL},-1)    //έγγραφα που δείχνουν
                             //σε κάποιο έγγραφο του {URL}
      ∀ d in TEMP
       WKEYS = weightedTargetKeywords(d,{URL})
                             //επιστρέφει τις λέξεις που εμφανίζονται
                             //στους συνδέσμους του d προς τα έγγραφα
                             //του {URL} και τον αριθμό των εγγράφων
                             //του {URL} που δείχνονται με κάθε λέξη
       if ∀ K in {keyword} getTimes(WKEYS , K) = N
                             //αν το d δείχνει και στα N έγγραφα του
                             //{URL} με τη λέξη K και αυτό συμβαίνει
                             // για όλες τις λέξεις K του {keyword}
            RES = RES ∪ d    //τότε το d είναι θεματική συναναφορά
RETURN RES
```

Όταν ο τελεστής TCocitations πάρει μεγάλα σύνολα διευθύνσεων ιστού ως είσοδο, επιστρέφει σημαντικούς π-κόμβους που δείχνουν στο συγκεκριμένο σύνολο διευθύνσεων ιστού και μάλιστα με συγκεκριμένα λεξικά χαρακτηριστικά.

Ορισμός 17: Θεματική βιβλιογραφική σύζευξη - Thematic Coupling: Για ένα σύνολο διευθύνσεων ιστού {URL} και ένα σύνολο λέξεων {keyword} ο τελεστής **TCouplings({URL}, {keyword})** επιστρέφει ένα σύνολο διευθύνσεων ιστού όπου *όλα δείχνονται από κάθε έγγραφο του* {URL} με συνδέσμους που φέρουν τις λέξεις του συνόλου {keyword}.

Ένα πρόχειρο σχέδιο του αλγορίθμου που υλοποιεί τον τελεστή ακολουθεί:

```
TCouplings({URL}, {keyword}):-
   RES = {}
   N= size({URL})          //το πλήθος των στοιχείων του {URL}
   TEMP = crawl({URL},1)   //έγγραφα που δείχνονται από κάποιο
                           //έγγραφο του {URL}
      ∀ d in TEMP
       WKEYS = weightedSourceKeywords({URL},d)
                           //επιστρέφει τις λέξεις που εμφανίζονται
                           //στους συνδέσμους από τα έγγραφα του {URL}
                           //προς την d και τον αριθμό των εγγράφων
                           //του {URL} που χρησιμοποιούν κάθε λέξη
       if ∀ K in {keyword} getTimes(WKEYS , K) = N
                           //αν και τα N έγγραφα του {URL} δείχνουν
```

```
                        //στο d με τη λέξη K και αυτό συμβαίνει
                        //για όλες τις λέξεις K του {keyword}
            RES = RES ∪ d     //τότε το d είναι θεματική
                        //βιβλιογραφική σύζευξη
RETURN RES
```

Σύμφωνα με τον ορισμό, ένα έγγραφο θεωρείται θεματική βιβλιογραφική σύζευξη για ένα σύνολο εγγράφων `{URL}` αν κάθε έγγραφο του {URL} δείχνει προς αυτό χρησιμοποιώντας όλες τις λέξεις του `{keyword}`.

5 Μέτρο ομοιότητας εγγράφων με χρήση οντολογίας

Στο κεφάλαιο αυτό περιγράφεται αναλυτικά η διαδικασία απεικόνισης των λέξεων που εξάγονται από τους υπερσυνδέσμους σε έννοιες μιας οντολογίας με χρήση του εννοιολογικού θησαυρού Wordnet. Αυτή η διαδικασία είναι πολύ σημαντική για τη συνέχεια, καθώς οι παραγόμενες σημασιολογικές περιγραφές, για κάθε έγγραφο, θα χρησιμοποιηθούν στη διαδικασία συσταδοποίησης που θα ακολουθήσει. Παρουσιάζεται επίσης το προτεινόμενο μέτρο ομοιότητας μεταξύ συνόλων εννοιών της οντολογίας, με ή χωρίς βάρη, ενώ τέλος μελετάται η πολυπλοκότητα του μέτρου ομοιότητας και κατ' επέκταση της διαδικασίας συσταδοποίησης.

5.1 Απεικόνιση λέξεων σε έννοιες μιας οντολογίας

Wordnet

Το Wordnet είναι ένα σύστημα λεξικογραφικών αναφορών, στο οποίο Αγγλικά ουσιαστικά, ρήματα, επίθετα και επιρρήματα οργανώνονται σε σύνολα συνωνύμων (synsets). Κάθε σύνολο αναπαριστά μια έννοια στο λεξικό. Διάφορες σχέσεις συνδέουν τα σύνολα συνωνύμων, όπως γενίκευση-ειδίκευση για ρήματα και ουσιαστικά, μέρος-όλο για τα ουσιαστικά κτλ. Για ένα ρήμα ή ουσιαστικό το Wordnet παρέχει ένα σύνολο από διαφορετικές έννοιες που μπορεί να έχει το ρήμα ή το ουσιαστικό. Οι έννοιες των ρημάτων και ουσιαστικών οργανώνονται σε θεματικές ιεραρχίες, σχηματίζοντας ένα "δάσος" από 25 δέντρα (IS-A) ουσιαστικών και 15 ρημάτων. Για επίθετα και επιρρήματα το Wordnet παρέχει μόνο συνώνυμα.

Για κάθε λέξη στο Wordnet μπορούμε να έχουμε ένα σύνολο από διαφορετικές έννοιες. Για τα ουσιαστικά και τα ρήματα έχουμε επίσης και ένα σύνολο από μονοπάτια γενικεύσεων από την κάθε έννοια προς την έννοια που βρίσκεται στην κορυφή της ιεραρχίας. Για παράδειγμα η λέξη "wind" έχει 8 έννοιες ως ρήμα και 8 ως ουσιαστικό. Για την πρώτη έννοια της λέξης "wind" ως ουσιαστικό έχουμε το ακόλουθο μονοπάτι γενικεύσεων:

wind => weather, weather condition, atmospheric condition
=> atmospheric phenomenon
=> physical phenomenon
=> natural phenomenon, nature
=> phenomenon

Οντολογία

Για να αντιστοιχίσουμε τις λέξεις, που εξάγονται από τα έγγραφα ή τους συνδέσμους, στις πλησιέστερες έννοιες που μας ενδιαφέρουν, χρειάζεται να αναφερθούμε σε μια οντολογία όρων σχετικών με το πεδίο ενδιαφέροντος μας. Κάθε ιεραρχία όρων μπορεί να χρησιμοποιηθεί αν μπορεί να αναπαρασταθεί ως δέντρο. Παραδείγματα οντολογιών που μπορούν να χρησιμοποιηθούν είναι οι οντολογίες που προτείνονται από το πρόγραμμα DARPA Agent Markup Language Program [DAML] (π.χ. η οντολογία στη μουσική που διατίθεται στη διεύθυνση http://www.daml.org/ontologies/276).

Για τη μετάβαση από λέξεις σε όρους της οντολογίας, χρησιμοποιούμε το Wordnet ως μια γενική οντολογία. Με τη βοήθεια του Wordnet και του μέτρου Wu & Palmer,

εντοπίζουμε για κάθε λέξη την πλησιέστερη έννοια της οντολογίας στα δέντρα του Wordnet.

Για να λειτουργήσει το σύστημα, χρειαζόμαστε μια απεικόνιση κάθε όρου της οντολογίας σε ένα σύνολο από έννοιες στο Wordnet. Αρχικά, όλες οι έννοιες ενός όρου της οντολογίας λαμβάνονται υπόψη. Παρ' όλα αυτά, η εξαγωγή σημασιολογικής πληροφορίας μπορεί να βελτιωθεί σημαντικά αν αγνοηθούν οι έννοιες που δεν είναι σχετικές με το θεματικό πεδίο που περιγράφει η οντολογία. Για παράδειγμα, για μια οντολογία στη μουσική, που περιέχει τον όρο "wind", είναι προτιμότερο να κρατήσουμε μόνο τις έννοιες που σχετίζονται με τη μουσική και να απορρίψουμε έννοιες που σχετίζονται με τη μετεωρολογία. Αυτό το βήμα σύνδεσης της οντολογίας με το Wordnet είναι προαιρετικό και απαιτεί την ανθρώπινη παρέμβαση.

Κατά τη σύνδεση της οντολογίας με το Wordnet, το σύστημα εμφανίζει όλες τις πιθανές έννοιες του όρου που υπάρχουν στο Wordnet, οπότε απορρίπτουμε όσες δε σχετίζονται με το πεδίο ενδιαφέροντος. Για το παράδειγμα της μουσικής, για τον όρο "wind" πρέπει να επιλεγούν μόνο 2 από τις 16 προτεινόμενες έννοιες και τα αντίστοιχα μονοπάτια.

wind instrument, wind => musical instrument => instrument => device
=> instrumentality, instrumentation => artifact, artefact => object,
=> physical object => entity, something
wind, wind up => tighten, fasten => change, alter

Με τη διαδικασία αυτή έχουμε για κάθε όρο της οντολογίας ένα ή περισσότερα σημεία αναφοράς στις ιεραρχίες που ορίζει το Wordnet. Βρίσκοντας, με παρόμοιο τρόπο, σημεία αναφοράς στο Wordnet, για κάθε λέξη που θέλουμε να αντιστοιχίσουμε, και υπολογίζοντας με το μέτρο Wu & Palmer τις αποστάσεις μεταξύ όρων και λέξεων στο Wordnet, είμαστε σε θέση να εντοπίσουμε τον πλησιέστερο όρο της οντολογίας σε κάθε λέξη.

Αντιστοίχιση ενός συνόλου λέξεων που περιγράφουν το έγγραφο σε ένα σύνολο από έννοιες της οντολογίας

Η επεξεργασία των εισερχόμενων συνδέσμων ενός εγγράφου d_i, παράγει ένα σύνολο από ζεύγη λέξεων-βαρών ($\mathbf{k_j}$,$\mathbf{n_j}$), όπου το $\mathbf{n_j}$ αντιστοιχεί στον αριθμό των εισερχόμενων συνδέσμων του d_i που χρησιμοποιούν τη λέξη $\mathbf{k_j}$.

Για να βρούμε τον πλησιέστερο όρο της θεματικής οντολογίας για μια λέξη **k**, υπολογίζουμε την ομοιότητα Wu & Palmer μεταξύ κάθε έννοιας της λέξης **k** και κάθε έννοιας κάθε όρου της οντολογίας $\mathbf{c_j}$. Αν η μέγιστη ομοιότητα υπολογιστεί για το ζεύγος εννοιών $\mathbf{k_x}$ (της λέξης **k**) και $\mathbf{c_{jy}}$ (του όρου της οντολογίας $\mathbf{c_j}$), τότε αντιστοιχίζουμε τη λέξη **k** στον όρο $\mathbf{c_j}$. Η προσέγγιση αυτή ελέγχει εξαντλητικά όλες τις έννοιες όλων των όρων της οντολογίας ακόμη και όσες είναι εκτός θέματος της οντολογίας. Η μέθοδος αυτή είναι αργή και δεν παρέχει τα βέλτιστα δυνατά αποτελέσματα.

Το ***πρώτο βήμα*** για τη βελτίωση της απεικόνισης είναι η ***απόρριψη των άσχετων εννοιών που δίνει το Wordnet για τους όρους της οντολογίας***, όπως περιγράψαμε προηγουμένως. Ο ορισμός των εννοιών για κάθε όρο της οντολογίας είναι

σημαντικός στην ανεύρεση της σημασίας των λέξεων και στην αποφυγή λανθασμένων αντιστοιχίσεων, γι' αυτό και απαιτεί ανθρώπινη παρέμβαση.

Το ***δεύτερο βήμα***, στη βελτίωση της απεικόνισης, είναι ***να μειώσουμε τις έννοιες που εξετάζονται για κάθε λέξη k, αφαιρώντας τις έννοιες που δε σχετίζονται με το θέμα***. Για το σκοπό αυτό, εξετάζουμε τη σημασία της κάθε λέξης σε συνάρτηση με το γλωσσικό περιβάλλον της (***context***) που το αποτελούν οι υπόλοιπες λέξεις του συνόλου. Έτσι για κάθε λέξη k στο $\{k_i\}$ υπολογίζουμε την ομοιότητα κάθε έννοιάς της με τις έννοιες των υπολοίπων λέξεων στο $\{k_j\}$ και κρατάμε μόνο τις έννοιες που συγκεντρώνουν μεγάλο βαθμό ομοιότητας. Με τον τρόπο αυτό, το σύνολο των εννοιών που απομένει χαρακτηρίζεται από μεγάλη ομοιότητα μεταξύ των εννοιών, απαλείφοντας έτσι τις έννοιες εκείνες που διαφέρουν σημαντικά από το γλωσσικό περιβάλλον τους.

Αν το σύνολο λέξεων περιέχει *n* λέξεις, αρχικά δημιουργούμε όλα τα πιθανά σύνολα εννοιών που περιέχουν μια έννοια από κάθε λέξη (*n*-άδες εννοιών). Στη συνέχεια, καθώς το μέτρο Wu & Palmer χρησιμοποιείται για ζεύγη εννοιών, υπολογίζουμε τη μέση τιμή της ομοιότητας Wu-Palmer για όλα τα διαφορετικά ζεύγη εννοιών σε κάθε *n*-άδα.

Για παράδειγμα, για τις λέξεις *guitar, flute* και *wind*, το Wordnet παρέχει 1, 3 και 8 έννοιες αντίστοιχα. Για τις 24 (1x3x8 συνδυασμοί) τριάδες εννοιών υπολογίζουμε τη μέση τιμή του μέτρου ομοιότητας Wu-Palmer. Αυτός ο υπολογισμός για τις έννοιες του συνόλου (*guitar, flute, wind*) δίνει τιμή 0.8 για την τριάδα εννοιών (*guitar, flute/transverse flute, wind instrument/wind*) και λιγότερο από 0.5 για οποιοδήποτε άλλο συνδυασμό. Παρόμοια για το σύνολο (*storm, cloud, wind*) παίρνουμε μέση ομοιότητα 0.8 για την τριάδα εννοιών (*storm/violent storm, cloud, wind/air moving*) και χαμηλότερες τιμές για τους υπόλοιπους συνδυασμούς. Οι τιμές αυτές δείχνουν ότι η λέξη "*wind*" έχει τη σημασία του "πνευστού μουσικού οργάνου" ("*wind instrument*") όταν εμφανίζεται μαζί με τις λέξεις (*guitar, flute*), και την έννοια του "ανέμου ως φυσικό φαινόμενο ("*wind as weather phenomenon*") όταν εμφανίζεται με τις λέξεις (*storm, cloud*).

Αφού επιλέξουμε τις καταλληλότερες έννοιες μιας λέξης, βρίσκουμε για κάθε λέξη στο σύνολο λέξεων $\{\mathbf{k_j}\}$ τον πλησιέστερο όρο της οντολογίας, στα δένδρα του Wordnet, όπως περιγράφηκε προηγουμένως. Ως αποτέλεσμα κάθε λέξη $\mathbf{k_j}$ αντιστοιχίζεται σε ένα όρο $\mathbf{c_i}$ με βαθμό ομοιότητας $\mathbf{s_j}$. Αναμένεται ότι πάνω από μία λέξεις στο σύνολο λέξεων $\{\mathbf{k_j}\}$ αντιστοιχίζονται στον ίδιο όρο της οντολογίας, συνεπώς το πλήθος στοιχείων του συνόλου $\{\mathbf{c_i}\}$ είναι μικρότερο του $\{\mathbf{k_j}\}$. Το βάρος που συνοδεύει κάθε όρο της οντολογίας $\mathbf{c_i}$ στην περιγραφή του εγγράφου ορίζεται ως η σταθμισμένη μέση τιμή όλων των ομοιοτήτων όλων των λέξεων $\mathbf{k_i}$ που αντιστοιχίζονται στο $\mathbf{c_i}$, λαμβάνοντας υπόψη το αντίστοιχο βάρος $\mathbf{n_j}$ κάθε λέξης (αριθμός εμφανίσεων της λέξης $\mathbf{k_j}$). Η σχετικότητα/βάρος κάθε όρου της περιγραφής υπολογίζεται από την εξίσωση:

$$r_i = \sum_{k_j \to c_i} (n_j \cdot s_j) \Big/ \sum_{k_j \to c_i} n_j \qquad \text{Εξ.26}$$

Έτσι κάθε έγγραφο d_i αναπαρίσταται ως εξής: $(URL_i, \{(c_i, r_i)\})$, όπου το $r_i \in [0,1]$ εφόσον $s_j \in [0,1]$.

Σύμφωνα με την έρευνα που πραγματοποιήθηκε στο πεδίο της αποσαφήνισης εννοιών λέξεων, προέκυψαν εργασίες που χρησιμοποιούν το Wordnet για το σκοπό αυτό, παρόλα αυτά καμία δεν συνδύαζε το γλωσσικό περιβάλλον μιας λέξης και την πληροφορία του Wordnet με τον τρόπο που παρουσιάζουμε. Η παρούσα προσέγγιση είναι πρωτότυπη και εμφανίζει ικανοποιητικά αποτελέσματα.

5.2 Μέτρο ομοιότητας για σύνολα όρων μιας οντολογίας

Για να λειτουργήσει ο αλγόριθμος συσταδοποίησης εγγράφων χρειάζεται ένα μέτρο ομοιότητας μεταξύ εγγράφων όπως αυτά αναπαρίστανται στο THESUS.

Ορισμός ομοιότητας
Τα έγγραφα του ΠΙ στο σύστημα THESUS αναπαρίστανται ως σύνολα από λέξεις και έννοιες μιας οντολογίας με τα αντίστοιχα βάρη. Το πρόβλημα της συσταδοποίησης των εγγράφων στο THESUS ανάγεται λοιπόν στην εύρεση ενός μέτρου ομοιότητας μεταξύ δύο τέτοιων συνόλων και ενός αλγορίθμου συσταδοποίησης, που θα χρησιμοποιήσει το μέτρο αυτό για να εντοπίσει ομάδες σχετικών εγγράφων.

Οι ιδιότητες ενός τέτοιου μέτρου, όπως αναφέρονται στο [L98], είναι:
- **Ιδιότητα 1η:** Η ομοιότητα μεταξύ δύο αντικειμένων Α και Β σχετίζεται με το πλήθος των κοινών χαρακτηριστικών τους. Όσα περισσότερα χαρακτηριστικά μοιράζονται τα Α και Β, τόσο πιο όμοια είναι.
- **Ιδιότητα 2η:** Η ομοιότητα μεταξύ των Α και Β σχετίζεται με το πλήθος των διαφορετικών χαρακτηριστικών τους. Όσο περισσότερες διαφορές έχουν, τόσο λιγότερο όμοια είναι.
- **Ιδιότητα 3η:** Η μέγιστη ομοιότητα των Α και Β επιτυγχάνεται όταν τα Α και Β είναι ταυτόσημα.

Το μέτρο ομοιότητας πρόκειται να χρησιμοποιηθεί κατά τη συσταδοποίηση των εγγράφων αλλά και κατά τη απάντηση ερωτήσεων. Αναμένεται ότι θα χρησιμοποιείται συχνά και γι' αυτό πρέπει να υπολογίζεται γρήγορα, ανεξάρτητα από το συνολικό αριθμό των εγγράφων. Πρέπει λοιπόν να ληφθεί υπόψη η πολυπλοκότητα υπολογισμού της ομοιότητας και μάλιστα να είναι ανεξάρτητη από τον αριθμό των εγγράφων, για λόγους κλιμάκωσης.

Πολλά από τα μέτρα ομοιότητας/απόστασης που βρέθηκαν στη βιβλιογραφία ικανοποιούν τις τρεις αυτές ιδιότητες [EM97], [GH+96]. Το μέτρο ομοιότητας Jaccard (Εξ. 2, ενότητα 2.8.2.), για παράδειγμα, διαθέτει τις πιο πάνω ιδιότητες, εξετάζει όμως μόνο την απόλυτη ομοιότητα των χαρακτηριστικών (exact matching). Το μέτρο ομοιότητας του THESUS έχει τη δυνατότητα να συγκρίνει σύνολα, λαμβάνει όμως υπόψη την σημασιολογική ομοιότητα των όρων σε μια οντολογία και όχι την λεξική ομοιότητα μεταξύ τους. Στα [RS+94], [L98] και [R95] προτείνονται διάφορα μέτρα υπολογισμού της εγγύτητας σε μια οντολογία όπως το Wordnet. Στην περίπτωση του THESUS χρησιμοποιείται το μέτρο Wu και Palmer [WP94] που υπολογίζεται σχετικά γρήγορα, δεν απαιτεί εκπαίδευση για τον καθορισμό του πληροφοριακού περιεχομένου των λέξεων και δίνει εξίσου ικανοποιητικά αποτελέσματα με τα υπόλοιπα μέτρα [L98].

5.2.1 Το μέτρο ομοιότητας Wu και Palmer

Όπως προαναφέρθηκε, οι έννοιες που περιέχονται στο Wordnet είναι οργανωμένες σε ιεραρχίες. Αντίστοιχα, οι έννοιες της θεματικής οντολογίας μας είναι οργανωμένες σε μια ιεραρχία. Για το λόγο αυτό χρησιμοποιούμε ένα μέτρο ομοιότητας σε ιεραρχίες.

Κατά την αναπαράσταση των εγγράφων του THESUS από σύνολα λέξεων, η ιεραρχία που χρησιμοποιείται είναι το Wordnet (κόμβοι της ιεραρχίας είναι τα synsets του Wordnet). Στην περίπτωση της αναπαράστασης με σύνολα εννοιών της οντολογίας, η ιεραρχία είναι, κατ' αντιστοιχία, η ίδια η οντολογία. Για λόγους απλότητας στη συνέχεια θα αναφερόμαστε σε «όρους» της ιεραρχίας και για τις δύο περιπτώσεις.

Θεωρούμε μια ιεραρχία όρων (ένα δένδρο) Ω, και a,b δύο κόμβους αυτής της ιεραρχίας. Έστω c ο πλησιέστερος κοινός πρόγονος των a,b και Depth(x) το βάθος ενός κόμβου x του δέντρου. Θεωρούμε ότι το βάθος της ρίζας είναι 1.

Η ομοιότητα δύο όρων a,b υπολογίζεται σύμφωνα με το μέτρο Wu-Palmer (βλ. Εξ. 3, ενότητα 2.8.3). Το μέτρο επιβεβαιώνει τις ιδιότητες που προαναφέραμε και παίρνει τιμές από 0, όταν τα a και b ανήκουν σε διαφορετικά δένδρα (δεν έχουν κοινό πρόγονο), έως 1, όταν $a \equiv b$. Να σημειωθεί ότι παρόμοια μπορούμε να ορίσουμε την απόσταση δύο όρων a,b ως:

$$D_{W\&P} = 1 - S_{W\&P} \qquad \text{Εξ.27}$$

Εύκολα διαπιστώνουμε ότι, για ένα σύνολο εγγράφων M, η απόσταση επαληθεύει τα εξής:

1. *Για κάθε x, y στο Ω*, $D(x, y) \geq 0$
2. *Για κάθε x στο Ω*, $D(x, x) = 0$
3. *Για κάθε x, y στο Ω*, $D(x, y) = 0 \Leftrightarrow x = y$.
4. *Για κάθε x, y στο Ω*, $D(x, y) = D(y, x)$.

Επίσης, αποδεικνύεται ότι η απόσταση αυτή είναι μετρική, καθώς επαληθεύει και την τριγωνική ανισότητα:

5. *Για κάθε x, y, z στο Ω*, $D(x, z) \leq D(x, y) + D(y, z)$

Συνεπώς, με τη χρήση του συγκεκριμένου μέτρου μπορούμε να ορίσουμε ένα μετρικό χώρο χωρίς διάταξη.

5.2.2 Επεκτείνοντας το μέτρο σε σύνολα όρων μιας ιεραρχίας

Στόχος μας είναι να βασιστούμε σε ένα μέτρο ομοιότητας μεταξύ των στοιχείων δύο συνόλων για να κατασκευάσουμε ένα μέτρο ομοιότητας μεταξύ των συνόλων αυτών. Μπορούμε να χρησιμοποιήσουμε οποιοδήποτε μέτρο ομοιότητας ανάμεσα στα στοιχεία των συνόλων. Μπορούμε επίσης να λάβουμε υπόψη τη σημασιολογική εγγύτητα των όρων, σε αντιστοιχία με τη συχνότητα ταυτόχρονης εμφάνισης [R79] δύο όρων που χρησιμοποιείται στις τεχνικές ανάκτησης πληροφορίας. Στην περίπτωση των εγγράφων του THESUS δε δεσμευόμαστε πλέον στη χρήση της απόλυτης ομοιότητας των όρων.

Στο [EM97] παρουσιάζεται μια πολύ ενδιαφέρουσα μελέτη διαφόρων μέτρων ομοιότητας μεταξύ συνόλων και αξιολογείται η πολυπλοκότητά τους. Η πολυπλοκότητα των αλγορίθμων κυμαίνεται από πολυωνυμική (ως προς το πλήθος στοιχείων του κάθε συνόλου) μέχρι NP. Πολυωνυμική, ως προς το πλήθος όρων που αντιστοιχούν σε κάθε έγγραφο, πρέπει να είναι και η πολυπλοκότητα του αλγόριθμου υπολογισμού του μέτρου ομοιότητας στο THESUS.

Αν και στην περίπτωση των εγγράφων του THESUS έχουμε σύνολαόρων με βάρη, αρχικά για λόγους απλότητας θα θεωρήσουμε σύνολα όρων χωρίς βάρη.

Μέτρο ομοιότητας μεταξύ συνόλων όρων
Συμβολισμοί:
Έστω **Ω** ένα σύνολο όρων.
Έστω $\mathbf{S_{W\&P}(a,b)}$ η ομοιότητα δύο όρων a και b του Ω κατά Wu-Palmer
Έστω **ζ(A,B)** η ομοιότητα μεταξύ δύο εγγράφων A και B (υποσύνολα του Ω).

Σημείωση: Δύο έγγραφα είναι ίδια τότε και μόνο τότε όταν αναπαρίστανται από το ίδιο σύνολο όρων του Ω. Η ομοιότητα των A και B ορίζεται:

$$\zeta(A,B) = \frac{1}{2}\left(\frac{1}{|A|}\sum_{a\in A} \max_{b\in B} \left(S_{W\&P}(a,b)\right) + \frac{1}{|B|}\sum_{b\in B} \max_{a\in A} \left(S_{W\&P}(a,b)\right)\right) \quad \text{Εξ.29}$$

Όπου: |A| και |B| είναι το πλήθος των όρων των A και B αντίστοιχα.

Ιδιότητες του μέτρου ζ(A,B):
1. Εύκολα διαπιστώνουμε ότι το ζ(A,B) είναι μέτρο ομοιότητας για το οποίο:
   - ζ(A,B)=1 ⇔ A=B
   - ζ(A,B)=ζ(B,A)
2. Το μέτρο ομοιότητας για μονοσύνολα (|A|=|B|=1) εκφυλίζεται στο μέτρο Wu-Palmer (π.χ. Αν A={a} και B={b}, τότε ζ(A,B)=SW&P(a,b).)
3. Το μέτρο ομοιότητας υπολογίζεται σε O(|A|x|B|), αν θεωρήσουμε ότι ο χρόνος υπολογισμού του Wu-Palmer είναι της τάξης O(1). Είναι λοιπόν συγκρίσιμο με τα υπόλοιπα μέτρα ομοιότητας που παρουσιάζονται στο [EM97].

Παράδειγμα
Ας υπολογίσουμε την ομοιότητα μεταξύ των εγγράφων A={cat, CD} και B={feline, record, tiger}. Αρχικά υπολογίζουμε την ομοιότητα για όλους τους συνδυασμούς ζευγών μεταξύ των όρων. Αν πρόκειται για λέξεις και δεν υπάρχει οντολογία, ο υπολογισμός μπορεί να γίνει απευθείας στο Wordnet (λέξεις που αντιστοιχούν στο ίδιο synset του Wordnet έχουν ομοιότητα 1). Αν πρόκειται για έννοιες της θεματικής οντολογίας, τότε ο υπολογισμός γίνεται απευθείας στην οντολογία. Έστω:
$S_{W\&P}$(cat, feline) = 0.95; $S_{W\&P}$(cat, record) = 0.59; $S_{W\&P}$(feline, CD) = 0.13; $S_{W\&P}$(CD, record) = 0.83; $S_{W\&P}$(tiger, cat) = 0.95; $S_{W\&P}$(tiger, CD) = 0.2.
Επίσης ισχύει |A|=2 και |B|=3.

Η ομοιότητα των δύο εγγράφων υπολογίζεται:
ζ(A,B) = 0.5 x (1/|A|.($S_{W\&P}$(cat, feline) + $S_{W\&P}$(CD, record)) + 1/|B|.($S_{W\&P}$(feline, cat) + $S_{W\&P}$(record, CD) + $S_{W\&P}$(tiger, cat))
ζ(A,B)=0.5 x (0.5 x(0.95+0.83)+0.33 x(0.95+0.83+0.95)) = 0.89

5.2.3 Μέτρο ομοιότητας μεταξύ συνόλων όρων με βάρη

Για να περιγράψουμε καλύτερα τα έγγραφα, εκτός από τις λέξεις και τις έννοιες που εξάγουμε, χρησιμοποιούμε και βάρη, που δηλώνουν τη σχετικότητα της λέξης ή της έννοιας με το έγγραφο. Για παράδειγμα, ένα έγγραφο που αναφέρεται στη δισκογραφία των Beatles είναι το A={(1, "Beatles"), (1, "Records")}, δηλώνοντας ότι και οι δύο όροι είναι πολύ σημαντικοί για το έγγραφο A. Για ένα δεύτερο έγγραφο B={(1, "Beatles"), (0.2, "Records")} τα βάρη δείχνουν ότι το έγγραφο αφορά περισσότερο τους Beatles και λιγότερο τους δίσκους τους. Όπως αναφέρθηκε στην ενότητα 5.1, το βάρος μιας έννοιας εξαρτάται από τον αριθμό εμφανίσεων της κάθε λέξης που αντιστοιχίζεται στη συγκεκριμένη έννοια, και από την ομοιότητα της συγκεκριμένης λέξης με την αντίστοιχη έννοια.

Σημείωση: Δύο έγγραφα είναι ίδια τότε και μόνο τότε όταν αναπαρίστανται από το τα ***ίδια σύνολα όρων*** με τα ***ίδια ακριβώς βάρη***.

Συμβολισμοί:

- Έστω **Ω** ένα σύνολο όρων.
- Χρησιμοποιούμε καλλιγραφικά κεφαλαία για τα έγγραφα που αναπαρίστανται ως σύνολα όρων με βάρη: $\mathcal{A}=\{(w_i, k_i)\}$, όπου $k_i \in \Omega$ και $w_i \in (0,1]$. Αντίστοιχα $\mathcal{B}=\{(v_j, h_j)$, όπου $h_j \in \Omega$ και $v_j \in (0,1]$.
- Χρησιμοποιούμε τα αντίστοιχα κεφαλαία γράμματα για τα έγγραφα που αναπαρίστανται ως σύνολα όρων χωρίς βάρη όπως πριν.

$$A=\{k_i \mid \exists i, (w_i, k_i) \in \mathcal{A}\}$$

$$B=\{h_j \mid \exists j, (v_j, h_j) \in \mathcal{B}\}$$

Σημείωση: το πλήθος όρων μπορεί να διαφέρει για κάθε έγγραφο($i \neq j$).

<u>Ορισμός 1</u>

Η ομοιότητα μεταξύ δύο εγγράφων $\mathcal{A}$ και $\mathcal{B}$, $\zeta(\mathcal{A},\mathcal{B})$ υπολογίζεται από τη σχέση:

$$\zeta(\mathcal{A},\mathcal{B}) = \frac{1}{2}\left[\frac{\sum_{i=1}^{|\mathcal{A}|}\lambda_{i,j}\ \arg\max_{j\in[1,|\mathcal{B}|]}\left(S_{W\&P}\left(k_i,h_j\right)\right)}{|\mathcal{A}|} + \frac{\sum_{j=1}^{|\mathcal{B}|}\lambda_{i,j}\ \arg\max_{i\in[1,|\mathcal{A}|]}\left(S_{W\&P}\left(k_i,h_j\right)\right)}{|\mathcal{B}|}\right] \quad \text{Εξ.30}$$

όπου $\lambda_{i,j} = \frac{w_i + v_j}{2 \times max\left(w_i, v_j\right)}$ Εξ.31

Κατά τον υπολογισμό της ομοιότητας θα υπολογίσουμε, αρχικά, για καθένα όρο i από τους όρους του πρώτου εγγράφου τη μέγιστη ομοιότητα με κάποιον όρο j από τους όρους του δεύτερου εγγράφου. Για το συνδυασμό αυτό θα λάβουμε υπόψη τους παράγοντες βάρους (στον υπολογισμό του λ_{ij}). Αθροίζοντας και διαιρώντας με το πλήθος των όρων του πρώτου εγγράφου ($|\mathcal{A}|$) προκύπτει η μέση ομοιότητα του εγγράφου $\mathcal{A}$ ως προς το $\mathcal{B}$.

Θυμίζουμε πως το argmax αντιστοιχεί στο όρισμα που αντιστοιχεί στο μέγιστο μιας συνάρτησης (argument of the maximum):

$$\arg\max_{x \in X} \; f(x) = \{x^* \text{ in } X \mid f(x^*) >= f(x), \; \forall x \in X\} \qquad \text{Εξ.32}$$

Για παράδειγμα στην εξίσωση 30 το $\arg\max_{i\in[1,|\mathcal{A}|]}\left(S_{W\&P}\left(k_i, h_j\right)\right)$ επιστρέφει για κάθε j εκείνο το i που μεγιστοποιεί την ομοιότητα του αντίστοιχου ζεύγους όρων.

Ακολούθως, με παρόμοιο τρόπο υπολογίζουμε τη μέση ομοιότητα του εγγράφου $\mathcal{B}$ ως προς το $\mathcal{A}$. Τέλος παίρνουμε το μέσο όρο των δύο ομοιοτήτων.

Εξήγηση των παραγόντων βάρους: Τα βάρη καθορίζουν το ποσοστό με το οποίο ένας όρος θα συμμετέχει στον καθορισμό της ομοιότητας δύο εγγράφων. Για παράδειγμα, αν το ένα έγγραφο περιγράφεται από έναν όρο με χαμηλό βάρος και το άλλο με τον ίδιο όρο και υψηλό βάρος ο όρος αυτός δεν θα επηρεάσει σημαντικά την *ομοιότητα* των εγγράφων. Έτσι τα $\lambda_{i,j}$ μειώνουν τη συνολική επίδραση των όρων με ανόμοια βάρη. Ανεξάρτητα από τις τιμές των βαρών w_i και v_j η μέγιστη τιμή του $\lambda_{i,j}$ ισούται με 1 όταν $w_i = v_j$.

Στον ορισμό αυτό οι παράγοντες βάρους λαμβάνονται υπόψη μόνο κατά τον υπολογισμό των $\lambda_{i,j}$ και δεν επηρεάζουν την επιλογή των συνδυασμών εννοιών. Ο παράγοντας βάρους δεν επηρεάζει την επιλογή των εγγράφων που δίνουν τη μέγιστη ομοιότητα. Έτσι όροι ενός εγγράφου, που δεν είναι πολύ σημαντικοί (μικρό βάρος), λαμβάνονται υπόψη στον υπολογισμό της ομοιότητας, καθώς είναι αρκετά όμοιοι με κάποιον όρο του άλλου εγγράφου. Στον υπολογισμό της τελικής ομοιότητας, για το ζεύγος όρων, λαμβάνεται βέβαια υπόψη το μικρό βάρος, που εξασθενίζει την τελική ομοιότητα.

Αν θέλουμε να λάβουμε υπόψη τον παράγοντα βάρους ενός όρου ενός εγγράφου, κατά την επιλογή του πλησιέστερου όρου του άλλου εγγράφου, προτείνουμε τον επόμενο ορισμό του μέτρου ομοιότητας.

Ορισμός 2

Η ομοιότητα μεταξύ δύο εγγράφων $\mathcal{A}$ και $\mathcal{B}$, $\boldsymbol{\zeta(\mathcal{A},\mathcal{B})}$ υπολογίζεται από τη σχέση:

$$\zeta(\mathcal{A},\mathcal{B}) = \frac{1}{2}\left[\frac{1}{|\mathcal{A}|}\left(\sum_{i=1}^{|\mathcal{A}|}\lambda_{i,x(i)}\cdot S_{W\&P}\left(k_i, h_{x(i)}\right)\right)+\frac{1}{|\mathcal{B}|}\left(\sum_{j=1}^{|\mathcal{B}|}\lambda_{y(j),j}\cdot S_{W\&P}\left(k_{y(j)}, h_j\right)\right)\right] \qquad \text{Εξ.33}$$

όπου τα $\lambda_{i,x(i)}$ και $\lambda_{y(j),j}$, υπολογίζονται όπως πριν σύμφωνα με την εξίσωση 31 και τα *x(i)* και *y(j)* ορίζονται ως εξής:

$$x(i) = \arg\max_{j\in[1,|B|]}\left(\lambda_{i,j}\times S_{W\&P}\left(k_i, h_j\right)\right) \qquad \text{Εξ.34}$$

$$y(j) = \arg\max_{i\in[1,|A|]}\left(\lambda_{i,j}\times S_{W\&P}\left(k_i, h_j\right)\right) \qquad \text{Εξ.35}$$

Σύμφωνα με τον ορισμό αυτό οι παράγοντες βάρους θα επηρεάζουν την επιλογή των καλύτερων ζευγών όρων από τα δύο έγγραφα και κατ' επέκταση τον υπολογισμό της τελικής ομοιότητας.

Ιδιότητες του ζ(*Α*,*Β*)

Και για τους δύο ορισμούς ισχύουν οι ακόλουθες ιδιότητες:

1. Εύκολα διαπιστώνουμε ότι το ζ(A,B) είναι μέτρο ομοιότητας καθώς:
   - $\zeta(\mathcal{A},\mathcal{B}) = 1 \Leftrightarrow A = B$
   - $\zeta(\mathcal{A},\mathcal{B}) = \zeta(\mathcal{B},\mathcal{A})$

2. Το μέτρο ομοιότητας επεκτείνει το προηγούμενο μέτρο ομοιότητας για σύνολα λέξεων χωρίς βάρη. Αν τα βάρη για όλους τους όρους των $\mathcal{A}$ και $\mathcal{B}$ είναι ίδια (π.χ. $\mathcal{A}=\{(\chi, k_i)\}$ και $\mathcal{B}=\{(\chi,h_j)\}$) το μέτρο εκφυλίζεται στο προηγούμενο. Επίσης αν $\mathcal{A}=\{(1,a)\}$ και $\mathcal{B}=\{(1,b)\}$, τότε $\zeta(\mathcal{A},\mathcal{B})=S_{W\&P}(a,b)$.

3. Το μέτρο υπολογίζεται σε $O(|\mathcal{A}|x|\mathcal{B}|)$.

4. Μπορούμε να εφαρμόσουμε στη βάση αυτού του μέτρου, στον υπολογισμό της ομοιότητας μεταξύ δύο όρων, οποιοδήποτε άλλο μέτρο αντί του Wu-Palmer.

1ο Παράδειγμα

Έστω δύο έγγραφα A ={(0.1, cat),(1, magazine)} και B={(1, dog), (1, book) και οι ομοιότητες κατά Wu-Palmer:

$S_{W\&P}$(cat, dog) = 0.80; $S_{W\&P}$(cat, book) = 0.43;

$S_{W\&P}$(dog, magazine) = 0.46; $S_{W\&P}$(book, magazine) = 0.83

Παρατηρούμε ότι οι όροι “cat” και “dog” είναι πλησιέστεροι μεταξύ τους καθώς και οι “magazine” και “book”.

Αν αγνοήσουμε τα βάρη, η ομοιότητα των A και B σύμφωνα με την εξίσωση 29, υπολογίζεται:

$$\zeta_0 = \frac{1}{2} \times \left[\frac{(0.8 + 0.83)}{2} + \frac{(0.83 + 0.8)}{2} \right] = 0.815$$

Οι παράγοντες βάρους υπολογίζονται:

$\lambda_{1,1} = (0.1 + 1)/(2*1) = 0.55$

$\lambda_{2,2} = (1 + 1)/(2*1) = 1$

$\lambda_{1,2} = (0.1 + 1)/(2*1) = 0.55$

$\lambda_{2,1} = (1+1)/(2\cdot 1) = 1$

Αν υπολογίζαμε την ομοιότητα με τον πρώτο ορισμό (εξ. 30) θα είχαμε:

$$\zeta_1 = \frac{1}{2} \times \left[\frac{(0.55 \times 0.8 + 1 \times 0.83)}{2} + \frac{(0.5 \times 0.8 + 1 \times 0.83)}{2} \right] = 0.635$$

Για το δεύτερο ορισμό ελέγχουμε τα “καλύτερα" ζεύγη με βάση τη σημασιολογική ομοιότητα και τους παράγοντες βάρους. Συγκρίνοντας ανά δύο τα:

$\lambda_{1,1} \times S(cat, dog) = 0.44$ και $\lambda_{1,2} \times S(cat, book) = 0.2365$,

$\lambda_{2,1} \times S(magazine, dog) = 0.46$ και $\lambda_{2,2} \times S(magazine, book) = 0.83$,

αποφασίζουμε να χρησιμοποιήσουμε τα ζεύγη όρων με τα μεγαλύτερα γινόμενα.

Καταλήγουμε έτσι στην εξής τιμή ομοιότητας:

$$\zeta_2 = \frac{1}{2} \times \left[\frac{(0.55 \times 0.8 + 1 \times 0.83)}{2} + \frac{(1 \times 0.46 + 1 \times 0.83)}{2} \right] = 0.64$$

Ο δεύτερος ορισμός του μέτρου φροντίζει ώστε να εξασθενεί η επίδραση των όρων με μικρά βάρη, γι'αυτό και κατά τον υπολογισμό του δεύτερου αθροίσματος επιλέγεται το ζευγάρι ((1,dog), (1,magazine)) αντί του ((1,dog), (0,1, cat)).

Στο δεύτερο ορισμό επιλέγονται τα πλησιέστερα ζευγάρια όρων, ανεξάρτητα των βαρών (cat⇔dog, book⇔magazine). Καθώς όμως το βάρος του όρου cat είναι πολύ διαφορετικό από αυτό του όρου dog η ομοιότητα που προκύπτει είναι πολύ μικρή συγκρινόμενη με την ζ_0.

2ο Παράδειγμα

Ας θεωρήσουμε την περίπτωση δύο εγγράφων που αναφέρονται στα ίδια θέματα αλλά με εντελώς διαφορετικά βάρη, έστω A={(1.0, cat) (0.1, dog)} και B={(0.1, cat) (1.0, dog)}.

Αν αγνοήσουμε τα βάρη, τα δύο έγγραφα θεωρούνται όμοια: ζ_0=1

Οι παράγοντες βάρους υπολογίζονται:
$\lambda_{1,1} = (1 + 0.1)/(2*1) = 0.55$
$\lambda_{2,2} = (0.1 + 1)/(2*1) = 0.55$
$\lambda_{1,2} = (1 + 1)/(2 \cdot 1) = 1$
$\lambda_{2,1} = (0.1+0.1)/(2 \cdot 0.1) = 1$

Ο πρώτος ορισμός του μέτρου ομοιότητας για τα δύο έγγραφα θα δώσει:

$$\zeta_1 = \frac{1}{2} \times \left[\frac{(0.55 \times 1 + 0.55 \times 1)}{2} + \frac{(0.55 \times 1 + 0.55 \times 1)}{2} \right] = 0.55$$

Για το δεύτερο ορισμό ελέγχουμε τα "καλύτερα" ζεύγη με βάση τη σημασιολογική ομοιότητα και τους παράγοντες βάρους. Συγκρίνουμε και πάλι ανά δύο τα:
$\lambda_{1,1} \times S(cat, cat) = 0.55$ και $\lambda_{1,2} \times S(cat, dog) = 0.80$,
$\lambda_{2,1} \times S(dog, cat) = 0.80$ και $\lambda_{2,2} \times S(dog, dog) = 0.55$, κτλ.

Ο δεύτερος ορισμός του μέτρου ομοιότητας για τα δύο έγγραφα θα δώσει:

$$\zeta_2 = \frac{1}{2} \times \left[\frac{(1 \times 0.8 + 1 \times 0.8)}{2} + \frac{(1 \times 0.8 + 1 \times 0.8)}{2} \right] = 0.8$$

Στην περίπτωση αυτή βλέπουμε ότι ο πρώτος ορισμός δίνει σημαντικά μικρότερη ομοιότητα από τον δεύτερο και αυτό λόγω της μεγάλης ανομοιότητας βαρών. Αν λάβουμε μάλιστα υπόψην ότι οι όροι cat και dog έχουν σημαντική ομοιότητα τότε το αποτέλεσμα του πρώτου ορισμού φαντάζει πολύ αυστηρό για τα έγγραφα A και B.

Ο δεύτερος ορισμός εμφανίζει μια ομαλότερη συμπεριφορά καθώς εξομαλύνει την επίδραση των όρων με μικρά βάρη στο τελικό αποτέλεσμα πάντοτε όμως σε συνδυασμό με την σημασιολογική ομοιότητα των όρων.

5.3 Μελέτη πολυπλοκότητας

Στην ενότητα αυτή θα υπολογίσουμε την πολυπλοκότητα της διαδικασίας συσταδοποίησης εγγράφων λαμβάνοντας υπόψη και τη διαδικασία υπολογισμού ομοιότητας εγγράφων.

Η πολυπλοκότητα του αλγορίθμου συσταδοποίησης του συστήματος THESUS μπορεί να υπολογιστεί χωριστά για τις δύο βασικές διεργασίες που διακρίνονται στον αλγόριθμο:

- Η πρώτη διαδικασία αφορά τον ***υπολογισμό της ομοιότητας δύο εγγράφων***, με χρήση του μέτρου ομοιότητας. Ο υπολογισμός αυτός θα πρέπει να επαναληφθεί για όλους τους συνδυασμούς εγγράφων του συνόλου.
- Η δεύτερη αφορά στον ***αλγόριθμο συσταδοποίησης***, με δεδομένο το βαθμό ομοιότητας μεταξύ όλων των ζευγών εγγράφων του συνόλου.

Ας θεωρήσουμε ένα σύνολο *n* εγγράφων, τα οποία περιγράφονται από ένα δεδομένο σύνολο εννοιών μιας θεματική οντολογίας. Και έστω *m* το πλήθος εννοιών της οντολογίας O. Έστω ένα έγγραφο $d_i=\{(v_j,c_j)\}$, όπου i=0..n, και $c_j \in O$, v_j το βάρος που αντιστοιχεί στο c_j για το έγγραφο d_i.

Θεωρούμε σταθερό το πλήθος εννοιών της οντολογίας και κατά συνέπεια μπορούμε να προϋπολογίσουμε την ομοιότητα Wu-Palmer για κάθε ζευγάρι εννοιών της οντολογίας O. Ο υπολογισμός αυτός θα γίνει μια φορά.

Η πολυπλοκότητα υπολογισμού του μέτρου ομοιότητας του THESUS, για δύο έγγραφα, με χρήση εννοιολογικών περιγραφών είναι $O(q \cdot p) \cong O(m^2)$, όπου q και p το πλήθος εννοιών κάθε εγγράφου. Τα q και p είναι συνήθως μικρά (~10 έννοιες ανά έγγραφο) και η μέγιστη τιμή τους είναι m. Η αντίστοιχη πολυπλοκότητα υπολογισμού του μέτρου συνημιτόνου είναι O(k), όπου k ο αριθμός των διαφορετικών λέξεων που μπορεί να εμφανιστούν στις περιγραφές των εγγράφων (O(m)-για περιγραφές με έννοιες της οντολογίας).

Ο υπολογισμός θα πρέπει να γίνει για κάθε ζεύγος εγγράφων ($\binom{n}{2}$ ζεύγη). Συνεπώς, οι ομοιότητες μεταξύ των n εγγράφων υπολογίζονται σε: $O(n^2 \times m^2)$, όπου m το πλήθος όρων της οντολογίας, και θεωρούμε ότι η ομοιότητα Wu-Palmer υπολογίζεται σε $O(1)$.

Για την αποθήκευση των αποστάσεων μεταξύ των m όρων της οντολογίας απαιτείται m^2. sizeof(sim). Για μια οντολογία 400 όρων, όπου οι αποστάσεις αποθηκεύονται ως ακέραιοι 1 byte, απαιτούνται περίπου 157KB.

Στο [EK+96] ο DBSCAN χρησιμοποιείται σε χωρικές βάσεις δεδομένων. Ο αλγόριθμος θεωρεί ότι τα στοιχεία που συσταδοποιούνται είναι σημεία ενός μετρικού χώρου δύο ή τριών διαστάσεων. Το μέτρο απόστασης που χρησιμοποιείται είναι η Ευκλείδεια απόσταση. Για την εύρεση των γειτόνων κάθε εγγράφου τα σημεία τοποθετούνται σε ένα R*-Tree [BK+90]. Για τον υπολογισμό της πολυπλοκότητας του αλγορίθμου δε λαμβάνουμε υπόψη το χρόνο δημιουργίας του ευρετηρίου R*-Tree.

Στην περίπτωση των εγγράφων δεν μπορούμε να χρησιμοποιήσουμε R*-Trees, καθώς έχουμε χώρο αποστάσεων, χωρίς διάταξη. Αντίθετα, προϋπολογίζουμε την ομοιότητα μεταξύ των n εγγράφων του συνόλου και αποθηκεύουμε τα αποτελέσματα σε n λίστες μεγέθους n. Όπως είπαμε το κόστος αυτού του υπολογισμού είναι $O(n^2m^2)$.

Ακολούθως, ταξινομούμε κάθε λίστα χρησιμοποιώντας τον αλγόριθμο Quicksort με μέση πολυπλοκότητα $O(n \log n)$ για μια λίστα μήκους n. Για την ταξινόμηση των n λιστών απαιτείται χρόνος $O(n^2 \log n)$.

Μόλις τελειώσει η φάση της προεπεξεργασίας, εφαρμόζουμε τον αλγόριθμο συσταδοποίησης. Η χρονική πολυπλοκότητα της διαδικασίας συσταδοποίησης βασίζεται στη μέση πολυπλοκότητα καθορισμού συνόλων εγγράφων συνδεδεμένων βάσει πυκνότητας, π.χ. στον καθορισμό των γειτόνων ενός εγγράφου. Στο σύστημα THESUS, καθώς η ομοιότητα ενός εγγράφου d_i με όλα τα υπόλοιπα έγγραφα προϋπολογίζεται και τα έγγραφα τοποθετούνται σε μια ταξινομημένη λίστα με σειρά φθίνουσας ομοιότητας προς το d_i, η πολυπλοκότητα καθορισμού της λίστας των πλησιέστερων γειτόνων του d_i (π.χ. $\zeta > MinSim$) είναι της τάξης $O(\log n)$, όπου n ο αριθμός των εγγράφων, χρησιμοποιώντας τη μέθοδο διχοτόμησης. Αυτό γίνεται μια φορά για κάθε έγγραφο της συλλογής, συνεπώς η πολυπλοκότητα γίνεται $O(n \log n)$, και είναι συγκρίσιμη με τη χρήση R*-Trees, σε σύνολα σημείων (επίσης $O(n \log n)$).

Επομένως, η πολυπλοκότητα για την επεξεργασία και τη συσταδοποίηση n εγγράφων είναι:

$$(n^2m^2)+O(n^2 \log n)+ O(n \log n) \cong O(n^2 \log n) \qquad \text{Εξ. 36}$$

Καθώς ο COBWEB είναι αυξητικός, η πολυπλοκότητά του, για λίγα έγγραφα, είναι ανάλογη του αριθμού των εγγράφων O(tN). Καθώς όμως ο αριθμός εγγράφων αυξάνει, η παραγόμενη ιεραρχία μπορεί να κλίνει προς μια κατεύθυνση, και κατά συνέπεια, ο παράγοντας t, που εξαρτάται από τη δομή της παραγόμενης ιεραρχίας, να αυξάνει μη γραμμικά. Προσπάθειες βελτιστοποίησης του αλγορίθμου, όπως αυτή που παρουσιάζεται από τον ίδιο το Fisher [F96], βελτιώνουν τις επιδόσεις του COBWEB καθώς και την ποιότητα του παραγόμενου σχήματος συσταδοποίησης για μεγάλο αριθμό εγγράφων. Η υλοποίηση που ακολουθήσαμε βασίζεται στην αρχική μορφή του αλγορίθμου [F87].

6 Το σύστημα THESUS

Η ενότητα αυτή παρουσιάζει την αρχιτεκτονική του συστήματος THESUS, καθώς και πειραματικά αποτελέσματα που επιβεβαιώνουν την αποτελεσματικότητα του συστήματος στο χαρακτηρισμό και την οργάνωση εγγράφων του ΠΙ. Τα συστατικά στοιχεία του συστήματος, η είσοδος και έξοδός τους, η λειτουργία τους και τα καινοτομικά χαρακτηριστικά τους αναλύονται στην ενότητα αυτή.

Το σύστημα επιδιώκει: α) το χαρακτηρισμό ενός συνόλου ιστοσελίδων κάνοντας χρήση της πληροφορίας των υπερσυνδέσμων, β) τον εμπλουτισμό αυτών των χαρακτηρισμών με σημασιολογικά χαρακτηριστικά και γ) την τμηματοποίηση του συνόλου σε θεματικά συμπαγή υποσύνολα (THESUs). Στόχος είναι ο καθορισμός ενός THESU, δίνοντας έμφαση στη σημασιολογική πληροφορία των συνδέσμων και στα χαρακτηριστικά της συνδεσιμότητας των εγγράφων του ΠΙ. Το σύστημα επιτρέπει την υποβολή ερωτήσεων στο υποσύνολο που δημιουργήθηκε. Οι ερωτήσεις αποσκοπούν να εντοπίσουν σημαντικά έγγραφα (π-κόμβους, α-κόμβους, συν-αναφορές κτλ.), αλλά και να εξετάσουν συλλογικά τη σημασιολογική πληροφορία των υπερσυνδέσμων ενός ή περισσοτέρων εγγράφων. Οι αλγόριθμοι που αναπτύχθηκαν, τμηματοποιούν το THESU σε μικρότερα υποσύνολα σελίδων με παρόμοια σημασία, αναλύοντας περαιτέρω το αρχικό THESU.

6.1 Αρχιτεκτονική – Τεχνολογίες

Το σύστημα THESUS είναι πλήρως υλοποιημένο. Τα υποσυστήματα του THESUS είναι (βλ. Σχήμα 18):

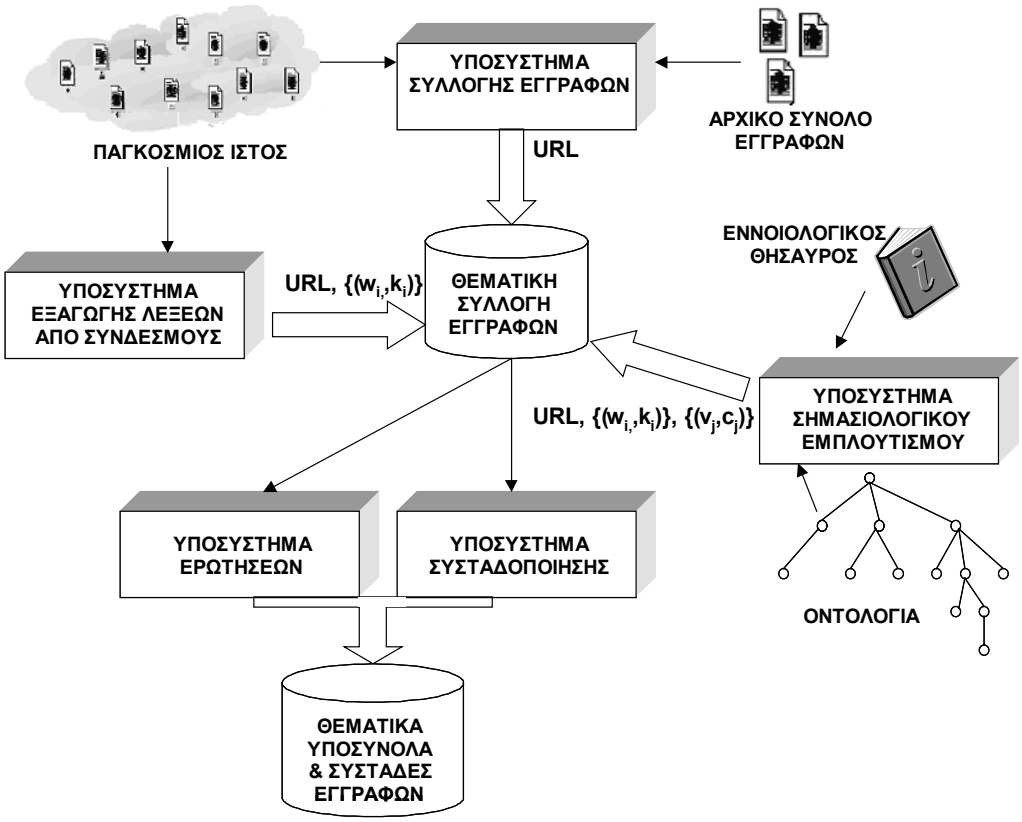
Το υποσύστημα "***συλλογής εγγράφων***": Το συγκεκριμένο υποσύστημα είναι στην ουσία ένας περιηγητής, ο οποίος συλλέγει ένα σύνολο από διευθύνσεις ιστού που φαίνονται σχετικές με κάποιο θέμα.. Αυτό γίνεται επεκτείνοντας είτε το αρχικό σύνολο εγγράφων D, που μπορεί να παραχθεί με μια ερώτηση σε μια μηχανή αναζήτησης, είτε ένα δεδομένο σύνολο εγγράφων σχετικών με κάποιο θέμα. Η επέκταση γίνεται με αλλεπάλληλες επαναλήψεις του αλγορίθμου HITS [K99], οι οποίες περιορίζονται κάθε φορά στους συνδέσμους που σε ένα ευρύτερο παράθυρο υπερκειμένου εμφανίζουν λέξεις σχετικές με το θέμα. Σε αντίθεση με την εστιασμένη συλλογή [CB+99], αυτός ο τύπος περιήγησης επισκέπτεται και συλλέγει απεριόριστο αριθμό εγγράφων. Μετά από Ν επαναλήψεις δημιουργείται ένα επεκταμένο σύνολο διευθύνσεων ιστού. Αυτό το σύνολο (Ε) περιέχει τις διευθύνσεις ιστού των εγγράφων που ασχολούνται με παρόμοια θέματα.

Το υποσύστημα "***εξαγωγής λέξεων από συνδέσμους***": Εξάγει λέξεις από τους συνδέσμους ενός εγγράφου που επισκεπτόμαστε, και συγκεκριμένα από μια ευρύτερη περιοχή υπερκειμένου γύρω από τον υπερσύνδεσμο.

Το υποσύστημα "***σημασιολογικού εμπλουτισμού***": Εμπλουτίζει την πληροφορία που εξήχθη από τους υπερσυνδέσμους με σημασιολογική πληροφορία αντιστοιχίζοντας τα σύνολα λέξεων σε σύνολα εννοιών από την οντολογία.

Το υποσύστημα "***συσταδοποίησης***": Διαχωρίζει το σύνολο Ε σε σημασιολογικά συμπαγή υποσύνολα βασιζόμενο είτε στις λέξεις που έχουν εξαχθεί για κάθε σελίδα είτε στις έννοιες που αυτά αντιπροσωπεύουν. Η ομοιότητα υπολογίζεται με χρήση ενός νέου μέτρου ομοιότητας, που χρησιμοποιεί την προσεγγιστική ομοιότητα όρων μιας ιεραρχίας και όχι την απόλυτη ομοιότητα λέξεων.

Η "***μηχανή απάντησης ερωτήσεων***": α) Επιτρέπει αναζητήσεις στο σύνολο Ε. Εκμεταλλευόμενη την τμηματοποίηση του Ε εστιάζει την αναζήτηση στα πλησιέστερα υποσύνολα του Ε και ταξινομεί τα αποτελέσματα και β) υλοποιεί τους τελεστές ανάλυσης συνδέσμων, επιτρέποντας αναζητήσεις για θεματικά σημαντικές σελίδες.



Σχήμα 18. Αρχιτεκτονική του συστήματος THESUS

Τα υποσυστήματα αυτά έχουν πρόσβαση σε μια σχεσιακή βάση δεδομένων όπου αποθηκεύεται η πληροφορία για κάθε ιστοσελίδα, ενώ παράλληλα χρησιμοποιούν γνώση από τη βάση πληροφορίας του Wordnet 1.7 [M+93] και την οντολογία. Το σύστημα έχει αξιολογηθεί με τη χρήση διάφορων συνόλων της τάξης των 10^4 εγγράφων, που είναι ένας ικανοποιητικός αριθμός για μια προσωπική θεματική αποθήκη πληροφορίας. Ο τρόπος επικοινωνίας με τη σχεσιακή βάση μπορεί να γίνει είτε απευθείας με χρήση εντολών SQL είτε διαμέσου του υποσυστήματος X-Database με χρήση XML αρχείων δεδομένων και εντολών. Στη δεύτερη περίπτωση, για την οποία θα μιλήσουμε εκτενέστερα στο κεφάλαιο 7, ο τρόπος επικοινωνίας με τη βάση δεδομένων είναι διαφανής για τα υποσυστήματα, διευκολύνοντας έτσι τη μελλοντική προσθήκη νέων υποσυστημάτων εξόρυξης και διαχείρισης γνώσης.

Οι λειτουργίες που προσφέρει το σύστημα είναι:
α) η δημιουργία συλλογών διευθύνσεων ιστού που χαρακτηρίζονται από συγκεκριμένες λέξεις κλειδιά. Αυτό επιτυγχάνεται συνδυάζοντας απλές αναζητήσεις στο δίκτυο (μέσω μηχανών αναζητήσεων, π.χ. αναζήτηση για σελίδες που περιέχουν συγκεκριμένες λέξεις) και ενός επιλεκτικού περιηγητή με θεματικά κριτήρια (θεματική περιήγηση) που επεκτείνει τα αποτελέσματα της μηχανής αναζήτησης,
β) ο χαρακτηρισμός εγγράφων ή συνόλων εγγράφων με λεξική ή σημασιολογική πληροφορία μετά από επεξεργασία των εισερχόμενων και εξερχόμενων συνδέσμων (της περιοχής υπερκειμένου γύρω από κάθε σύνδεσμο),
γ) η τμηματοποίηση ενός χαρακτηρισμένου συνόλου εγγράφων σε υποσύνολα με βάση τα σημασιολογικά χαρακτηριστικά των εγγράφων.

Είναι φυσικό η χρονοβόρα διαδικασία συλλογής και επεξεργασίας ενός μεγάλου όγκου ιστοσελίδων να μην μπορεί να πραγματοποιηθεί μέσα σε λίγα δευτερόλεπτα, που είναι ο μέσος χρόνος αναμονής για μια διεργασία στο διαδίκτυο. Κατά συνέπεια, τα πρώτα τρία υποσυστήματα χρησιμοποιούνται ως ασύγχρονα εργαλεία για τη δημιουργία μιας αποθήκης εγγράφων εμπλουτισμένων με σημασιολογική πληροφορία και το υποσύστημα τμηματοποίησης βοηθά στην οργάνωση της αποθηκευμένης πληροφορίας σε υποσύνολα. Το υποσύστημα ερωτήσεων λειτουργεί ως ένα σύγχρονο εργαλείο που διατίθεται στους χρήστες και τους επιτρέπει να περιηγηθούν και να ψάξουν στην εμπλουτισμένη αποθήκη πληροφορίας.

Ένα σύνολο από υπηρεσίες, που παρουσιάζουν τις δυνατότητες του υποσυστήματος εξαγωγής λέξεων από τους υπερσυνδέσμους και τη λειτουργία των τριών διαφορετικών υλοποιήσεων του τελεστή groupKeywords (βλ. Ενότητα 4.2.2.1), είναι διαθέσιμες στη διεύθυνση www.db-net.aueb.gr/thesus. Οι χρήστες μπορούν να χρησιμοποιούν τους υλοποιημένους τελεστές για αναζητήσεις στο σύνολο του ΠΙ.

Τεχνολογικό πλαίσιο
Το υποσύστημα περιήγησης και συλλογής εγγράφων έχει αναπτυχθεί ως πολυνηματική εφαρμογή Java. Η πληροφορία που εξάγεται από τα έγγραφα και τα σχετικά σημασιολογικά χαρακτηριστικά αποθηκεύονται σε μια σχεσιακή βάση δεδομένων (Microsoft-SQL Server®). Η σύνδεση με τη βάση επιτυγχάνεται μέσω ενός εγγενή JDBC οδηγού. Η εφαρμογή των χρηστών συνδέεται με τη βάση μέσω JDBC και επιτρέπει τη χρήση των τελεστών της γλώσσας του THESUS. Οι δικτυακές υπηρεσίες έχουν υλοποιηθεί ως τάξεις Java που είναι διαθέσιμες μέσω Java Server Pages αλλά και ως Java Web Services μέσω της πλατφόρμας Axis.

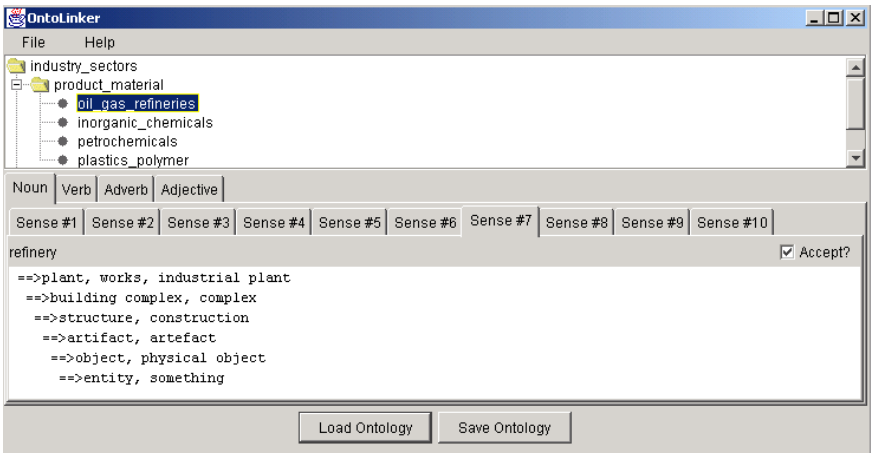
Εκτός από το βασικό σύστημα του THESUS έχουν αναπτυχθεί και ορισμένες βοηθητικές εφαρμογές που εξυπηρετούν συγκεκριμένες διαδικασίες. Στις επόμενες υποενότητες περιγράφουμε την εφαρμογή που χρησιμοποιείται για τη σύνδεση της θεματικής οντολογίας με το λεκτικό θησαυρό Wordnet, καθώς και την εφαρμογή που επιτρέπει τον χαρακτηρισμό ενός εγγράφου του ΠΙ με λεξικά και σημασιολογικά χαρακτηριστικά και αποθηκεύει την πληροφορία αυτή σε ξεχωριστό XML έγγραφο. Στην ενότητα 7 θα αναφερθούμε στη σημασία της εφαρμογής X-Database που χρησιμοποιείται για την αποθήκευση XML εγγράφων σε σχεσιακές βάσεις δεδομένων.

6.1.1 Σύνδεση οντολογίας με το Wordnet

Όπως προαναφέρθηκε, το σύστημα THESUS δεν περιορίζεται στην εξαγωγή λέξεων για ένα έγγραφο, αλλά προσπαθεί να αποσαφηνίσει τις έννοιες που οι λέξεις αυτές αντιπροσωπεύουν. Για να βελτιώσουμε την ποιότητα της απεικόνισης λέξεων σε έννοιες, χρησιμοποιούμε μια θεματική οντολογία. Η αποσαφήνιση των εννοιών γίνεται στο επίπεδο οντολογίας πρώτα και στη συνέχεια σε επίπεδο εγγράφου.

Μάλιστα, κάθε όρος της οντολογίας μπορεί να έχει διαφορετικές σημασίες και αντιστοιχίζεται σε ένα σύνολο εννοιών στο Wordnet. Κάποιες από τις έννοιες ενός του συνόλου είναι σχετικές με το θέμα ενώ κάποιες άλλες όχι. Κατά τη διαδικασία της απεικόνισης των λέξεων σε όρους της οντολογίας εντοπίζουμε για κάθε λέξη τον όρο με τον οποίο έδωσε τη μεγαλύτερη ομοιότητα για κάποιο συνδυασμό εννοιών. Στο στάδιο αυτό είναι πολύ σημαντικό να απορρίψουμε τις έννοιες των όρων της οντολογίας που δεν σχετίζονται με το θεματικό πεδίο. Έτσι θα αποφύγουμε να αντιστοιχίσουμε λέξεις που δε σχετίζονται καθόλου με το θέμα σε κάποιο όρο της οντολογίας, επειδή κάποια έννοια ενός όρου δίνει μεγάλες τιμές ομοιότητας με τις λέξεις αυτές.

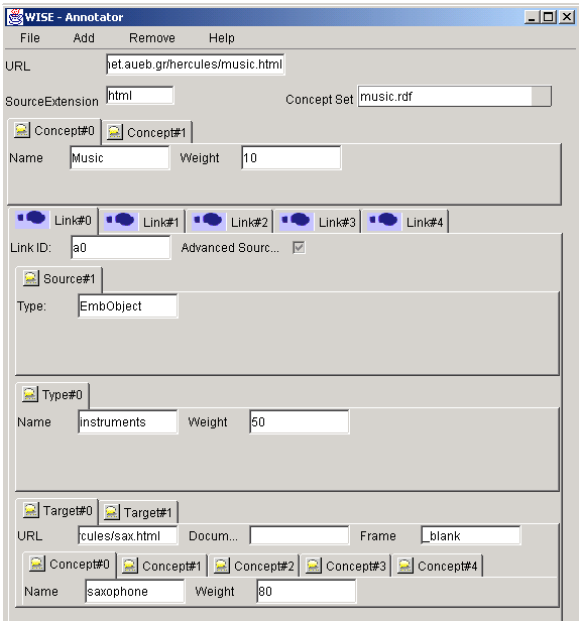
Το βήμα αυτό είναι πολύ σημαντικό για την ποιότητα της τελικής απεικόνισης και γι' αυτό χρειάζεται την ανθρώπινη παρέμβαση για να γίνει σωστά. Για το λόγο αυτό αναπτύχθηκε μια εφαρμογή (βλ. Σχήμα 19) που επιτρέπει στους χρήστες να χτίσουν την οντολογία και ταυτόχρονα να καθορίσουν της έννοιες του κάθε όρου που σχετίζονται με το θέμα. Η εφαρμογή παρουσιάζει όλες τις πιθανές έννοιες ενός όρου όπως αυτές παρέχονται από το Wordnet, οπότε αποφασίζουμε και επιλέγουμε τις σχετικές. Έτσι, για παράδειγμα, για τον όρο "wind" σε μια οντολογία μουσικής κρατούμε μόνο 2 από τις 16 διαφορετικές έννοιες:
wind instrument, wind => musical instrument => instrument => device =>
 => instrumentality, instrumentation => artifact, artefact =>
 => object, physical object => entity, something
wind, wind up => tighten, fasten => change, alter



Σχήμα 19. Εφαρμογή σύνδεσης της οντολογίας με το Wordnet

6.1.2 Εφαρμογή Χαρακτηρισμού Εγγράφων

Για την υποστήριξη της διαδικασίας χαρακτηρισμού εγγράφων του ΠΙ με λεξικά και σημασιολογικά χαρακτηριστικά, αναπτύχθηκε η Εφαρμογή Χαρακτηρισμού Εγγράφων (ΕΧΕ). Η ΕΧΕ επιτρέπει στους χρήστες να επιλέξουν αρχείο θεματικής οντολογίας και στη συνέχεια να δώσουν τη διεύθυνση ενός δικτυακού εγγράφου για χαρακτηρισμό. Η εφαρμογή επεξεργάζεται τους εισερχόμενους και εξερχόμενους συνδέσμους του εγγράφου αλλά και το περιεχόμενό του, εξάγει λεξικές περιγραφές και παράγει τις αντίστοιχες σημασιολογικές. Οι παραγόμενοι χαρακτηρισμοί για κάθε έγγραφο παρουσιάζονται στη διεπαφή, όπου μπορούμε να τις τροποποιήσουμε και να τους αποθηκεύσουμε σε ένα XML έγγραφο. Η διεπαφή της ΕΧΕ παρουσιάζεται στο Σχήμα 20.



Σχήμα 20. Διεπαφή χαρακτηρισμού εγγράφων του ιστού

6.2 Πειραματικά αποτελέσματα

Στην ενότητα αυτή παρουσιάζουμε πληροφορίες σχετικά με την απόδοση του συστήματος THESUS. Αξιολογούμε την ικανότητα των υποσυστημάτων εξαγωγής λέξεων και συσταδοποίησης εγγράφων του δικτύου. Για το πρώτο υποσύστημα επιλέγουμε τυχαίες διευθύνσεις ιστού ή σύνολα διευθύνσεων ιστού, ενώ για το δεύτερο συσταδοποιούμε τα έγγραφα μιας ήδη ταξινομημένης συλλογής, χρησιμοποιώντας τους αλγορίθμους που αναπτύξαμε. Τέλος αξιολογούμε τα αποτελέσματα.

Στο πρώτο σύνολο πειραμάτων εξετάζουμε τη σημαντικότητα της σημασιολογίας, που μεταφέρουν οι εισερχόμενοι σύνδεσμοι, στο χαρακτηρισμό ενός μεμονωμένου εγγράφου ή ενός συνόλου εγγράφων, και παρουσιάζουμε τα αποτελέσματα των τριών διαφορετικών τρόπων με τους οποίους μπορεί να ολοκληρωθεί η πληροφορία των συνδέσμων.

Το δεύτερο πείραμα εξετάζει την ικανότητα του συστήματος να συσταδοποιεί ένα σύνολο εγγράφων σε σύγκριση με τη συσταδοποίηση που γίνεται στο ίδιο σύνολο

από ανθρώπους. Τέλος, παρουσιάζουμε ενδεικτικά αποτελέσματα από τη χρήση των σύνθετων τελεστών ανάλυσης πληροφορίας υπερσυνδέσμων σε ένα μεγάλο σύνολο πληροφορίας, που συγκεντρώθηκε και χαρακτηρίστηκε με τη χρήση του THESUS.

Να σημειώσουμε ότι σε όλα τα πειράματα χρησιμοποιήθηκαν αδέκαστοι χρήστες για την αξιολόγηση των αποτελεσμάτων, ώστε να μην επηρεάζουν τα αποτελέσματα. Οι χρήστες δε διαθέτουν πρότερη γνώση του συστήματος, και επίσης δε γνωρίζουν ποιο από τα συγκρινόμενα συστήματα δίνει τα αποτελέσματα που αξιολογούν.

6.2.1 Αποδοτικότητα του συστήματος

Στην ενότητα αυτή θα παρουσιαστούν ορισμένα αντιπροσωπευτικά στοιχεία της αποδοτικότητας του συστήματος στη συλλογή και επεξεργασία μεγάλου αριθμού εγγράφων του ιστού. Οι επιδόσεις του συστήματος συσταδοποίησης όσο αφορά το χρόνο που απαιτείται για την εκτέλεση του αλγορίθμου DBSCAN στη συλλογή των εγγράφων παρουσιάζονται επίσης στην ενότητα αυτή. Ο αλγόριθμος COBWEB δεν προτιμήθηκε στην περίπτωση αυτή καθώς δεν είναι βελτιστοποιημένος και για μεγάλο αριθμό εγγράφων είναι σημαντικά πιο αργός από τον DBSCAN.

Τα πρώτα τρία υποσυστήματα του THESUS (συλλογής, εξαγωγής λέξεων και εμπλουτισμού πληροφορίας) παίρνουν ως είσοδο μια οντολογία και ένα σύνολο από έγγραφα και παράγουν ένα επεκταμένο σύνολο από έγγραφα και συνδέσμους, που χαρακτηρίζονται από λέξεις, έννοιες και τα αντίστοιχα βάρη. Αυτή η διαδικασία συμβαίνει άπαξ, και δημιουργεί τη δεξαμενή πληροφορίας για τα υποσυστήματα συσταδοποίησης και ερωτήσεων του THESUS.

Η διάρκεια, της διαδικασίας συλλογής πληροφορίας, επηρεάζεται κυρίως από το διαθέσιμο εύρος δικτύου και την αρχιτεκτονική του συστήματος συλλογής, καθώς και από το πλήθος δεδομένων που συλλέγονται. Σε μια συγκεκριμένη υλοποίηση, όπου χρησιμοποιήθηκε μια οντολογία σε μουσικά όργανα (που περιείχε περίπου 70 όρους), η αναζήτηση ξεκίνησε από 3 ιστοσελίδες καταλόγων του ΠΙ (των Yahoo, DMOZ και Google directory) και συγκεκριμένα από την κατηγορία μουσική, και η περιήγηση έγινε για 12 επίπεδα. Η διαδικασία συλλογής έγινε σε έναν υπολογιστή που εκτελούσε μια πολυνηματική εφαρμογή Java, σε μια προσπάθεια να εκμεταλλευθούμε το διαθέσιμο εύρος δικτύου. Η διαδικασία εξαγωγής λέξεων από τους υπερσυνδέσμους έγινε ταυτόχρονα με τη συγκέντρωση των εγγράφων.

Σε λιγότερο από μια ημέρα συλλέξαμε πληροφορία για 190000 έγγραφα, από τα οποία μόνο 4000 είχαν πάνω από 1 εισερχόμενο σύνδεσμο, γεγονός που δείχνει την χαλαρή σύνδεση των εγγράφων του συνόλου μας. Ο πραγματικός αριθμός των συνδέσμων μεταξύ των εγγράφων της συλλογής είναι μεγαλύτερος, όμως κάποιοι από τους συνδέσμους αγνοούνται καθώς δεν περιέχουν κάποια από τις λέξεις της οντολογίας. Συνεπώς η σύνδεση είναι ακόμη πιο χαλαρή από την πραγματική. Η πληροφορία των εγγράφων μπορεί δυνητικά να εμπλουτιστεί αν επισκεφθούμε τα έγγραφα που δείχνουν προς αυτά του συνόλου μας (με χρήση του τελεστή crawl({URL},-1)).

Η επίδοση, του συστήματος εμπλουτισμού της πληροφορίας, επηρεάζεται από το μέγεθος του συνόλου των λέξεων που χαρακτηρίζει το κάθε έγγραφο και από το μέγεθος της οντολογίας, καθώς αυξάνει τον αριθμό των συνδυασμών εννοιών που

πρέπει να ελεγχθούν για να βρεθεί το σύνολο με τις πλησιέστερες έννοιες. Μια αύξηση στον αριθμό των λέξεων, που περιγράφουν ένα έγγραφο, συνεπάγεται αύξηση στον αριθμό των εννοιών που πρέπει να εξεταστούν για το έγγραφο. Μια αύξηση στο μέγεθος της οντολογίας συνεπάγεται αύξηση στο σύνολο των ζευγών (λέξη, όρος οντολογίας) που θα εξεταστούν. Στο παράδειγμα που αναφέρουμε η επεξεργασία για τη συλλογή διήρκεσε λιγότερο από 6 ώρες.

Η διαδικασία συσταδοποίησης δε γίνεται σε πραγματικό χρόνο αλλά επαναλαμβάνεται αρκετές φορές μέχρι να βρεθεί το καλύτερο σχήμα συσταδοποίησης. Η ποιότητα της παραγόμενης συσταδοποίησης ελέγχεται με διάφορα εσωτερικά μέτρα αξιολόγησης. Ο χρόνος που απαιτείται για την συσταδοποίηση ενός συνόλου εγγράφων εξαρτάται από το πλήθος των εγγράφων και τον αλγόριθμο που χρησιμοποιείται. Η διαδικασία συσταδοποίησης μπορεί να επιταχυνθεί με αποθήκευση στη μνήμη της πληροφορίας που υπολογίζεται συχνά. Στην παρούσα υλοποίηση προϋπολογίζουμε τον πίνακα ομοιότητας, που περιέχει τις ομοιότητες μεταξύ όλων των εγγράφων του συνόλου. Για το τρέχον παράδειγμα το μέγεθος του πίνακα στη μνήμη είναι (190000x190000)/2 bytes αν η ομοιότητα δύο εγγράφων παίρνει τιμές από 0 ως 100 και αποθηκευθεί ως byte. Αυτό απαιτεί 16 Gigabytes κύριας μνήμης για να αποθηκευθεί. Λόγω περιορισμών στην υπάρχουσα μνήμη, συσταδοποιούμε μόνο τα 4000 έγγραφα για τα οποία έχουμε πάνω από 1 εισερχόμενο σύνδεσμο. Ο πίνακας ομοιότητας υπολογίζεται σε 38 δευτερόλεπτα και η συσταδοποίηση γίνεται σε 0.8 δευτερόλεπτα με τη χρήση του αλγορίθμου DBSCAN.

Συνοψίζοντας μπορούμε να πούμε ότι οι επιδόσεις του συστήματος συλλογής εγγράφων είναι ικανοποιητικές δεδομένου ότι η διαδικασία αυτή γίνεται μία μόνο φορά και μάλιστα παρασκηνιακά, ενώ παράλληλα εκτελούνται και οι διαδικασίες εξαγωγής λέξεων και εμπλουτισμού με έννοιες της οντολογίας. Η διαδικασία συσταδοποίησης είναι αρκετά γρήγορη στην περίπτωση που ο πίνακας ομοιότητας έχει προϋπολογιστεί.

6.2.2 Αξιολόγηση διαδικασίας εξαγωγής λέξεων από τους εισερχόμενους συνδέσμους

6.2.2.1 Χαρακτηρισμός εγγράφου

Για να δείξουμε τις δυνατότητες που έχει το υποσύστημα εξαγωγής λέξεων να χαρακτηρίζει τα έγγραφα του ιστού, επιλέξαμε 50 τυχαίες διευθύνσεις ιστού. Χαρακτηρίσαμε τα έγγραφα αυτά με βάση την πληροφορία που εξαγάγουμε από τους 100 (το πολύ) πρώτους εισερχόμενους συνδέσμους, όπως αυτοί παρέχονται από τη σχετική υπηρεσία του Google. Ολοκληρώσαμε τα αποτελέσματα με βάση το πηγαίο έγγραφο (`weightedSourceKeywords`) και κρατήσαμε τις 10 λέξεις που χρησιμοποιήθηκαν από τους περισσότερους εισερχόμενους συνδέσμους για να χαρακτηρίσουν το έγγραφο. Στη συνέχεια, παρουσιάσαμε την περιγραφή αυτή, μαζί με τις περιγραφές που δίνουν οι μηχανές αναζήτησης Altavista και Google, σε μια ομάδα κριτών (25 φοιτητές) και τους ζητήσαμε να αξιολογήσουν από 1 ως 5 την ποιότητα του χαρακτηρισμού (1 –κακή, 5 –άριστη).

Τα έγγραφα που επιλέξαμε φροντίσαμε να ανήκουν στους καταλόγους διευθύνσεων είτε του Google είτε του Altavista. Σε περίπτωση που ένας από τους δύο καταλόγους δεν περιείχε μια διεύθυνση τότε κάναμε ερώτηση στην αντίστοιχη μηχανή

αναζήτησης με κριτήριο την διεύθυνση. Αξίζει να αναφέρουμε ότι, για ορισμένες διευθύνσεις ιστού τα συστήματα Google και Altavista έδωσαν περιγραφές που έχουν γραφτεί από τους συντάκτες τους και δεν έχουν εξαχθεί με αυτόματο τρόπο, όπως γίνεται με το THESUS. Σε αντίθετη περίπτωση οι σύντομες περιγραφές που πήραμε έχουν εξαχθεί από το σύνολο των εγγράφων και δεν εστιάζουν σε ένα συγκεκριμένο τμήμα τους (όπως γίνεται για τις συνήθεις ερωτήσεις με λεξικά κριτήρια).

Σε πάνω από το 50% των περιπτώσεων οι περιγραφές που παρήχθησαν από το THESUS θεωρήθηκαν καλύτερες από τις άλλες δύο, ενώ η μέση βαθμολογία ήταν 3.7 στα 5 για το THESUS, 1.9 για το Altavista και 3.5 για το Google.

Για τη μέτρηση του βαθμού συμφωνίας των κριτών χρησιμοποιήσαμε δύο μέτρα: α) το μέσο όρο της μέσης τυπικής απόκλισης (standard deviation) για κάθε διεύθυνση και β) το μέτρο kappa [C96] που αξιολογεί την κατανομή των βαθμολογιών σε κάθε κατηγορία (1 ως 5).
Το μέτρο kappa ορίζεται:

$$K=\frac{P(A)-P(E)}{1-P(E)} \qquad \text{Εξ. 37}$$

όπου P(A) το ποσοστό των περιπτώσεων που οι κριτές συμφωνούν και P(E) το ποσοστό των περιπτώσεων στις οποίες οι κριτές συμφωνούν τυχαία. Είναι προφανές ότι σε απόλυτη συμφωνία των κριτών το K γίνεται μέγιστο και ίσο με 1.
Η τυπική απόκλιση ορίζεται:

$$Stddev=\sqrt{\frac{\sum_{i=1}^{N}\left(x_i-x_{avg}\right)^2}{N}} \qquad \text{Εξ. 38}$$

όπου x_{avg} η μέση βαθμολογία για κάθε διεύθυνση και N το πλήθος των κριτών. Η μέση τυπική απόκλιση ελαχιστοποιείται (Stddev=0) όταν έχουμε απόλυτη συμφωνία των κριτών.
Τελικά λαμβάνουμε τη μέση τιμή της για το σύνολο των διευθύνσεων.

Τα αποτελέσματα για τα αντίστοιχα συστήματα είναι:

	Μέση τυπική απόκλιση	Μέτρο kappa
Altavista	2,1	0,17
Google	2,4	0,06
Thesus	2,4	0,01

Και στις δύο περιπτώσεις βλέπουμε ότι οι διαφορές στην κρίση των χρηστών είναι πολύ μικρές, γεγονός που ενισχύει τα αποτελέσματα της διαδικασίας αξιολόγησης.

Παρατηρούμε λοιπόν ότι οι περιγραφές που παρήχθησαν από το THESUS είναι συγκρίσιμες σε ποιότητα και σε ορισμένες περιπτώσεις πολύ καλύτερες από τις περιγραφές που έδωσαν οι άνθρωποι.

6.2.2.2 Χαρακτηρισμός συνόλου εγγράφων

Στόχος της ενότητας που ακολουθεί είναι να παρουσιάσει τις δυνατότητες των τελεστών του THESUS να παράγουν περιγραφές για σύνολα εγγράφων του ΠΙ.

- Αρχικά παρουσιάζεται, μέσα από ένα παράδειγμα, η χρήση των σύνθετων τελεστών σημασιολογίας συνδέσμων (βλ. ενότητα 4.2.2.1.) και τα διαφορετικά

αποτελέσματα που παράγονται για το ίδιο σύνολο εγγράφων, αν χρησιμοποιηθούν εισερχόμενοι ή εξερχόμενοι σύνδεσμοι.
- Στη συνέχεια αξιολογούμε την ποιότητα των χαρακτηρισμών που παράγονται για γνωστές ομάδες εγγράφων, με χρήση της πληροφορίας των εισερχόμενων συνδέσμων. Επίσης καθορίζουμε το πλήθος των λέξεων που δίνουν μια ικανοποιητική περιγραφή (χωρίς σημαντική μείωση της ποιότητας).
- Τέλος για δεδομένο αριθμό λέξεων στην περιγραφή μιας ομάδας εγγράφων προσπαθούμε να εντοπίσουμε το ελάχιστο πλήθος εισερχόμενων συνδέσμων που απαιτούνται για την περιγραφή του συνόλου.

Τελεστές που παράγουν περιγραφές για ομάδες εγγράφων

Για να δείξουμε τη χρήση και την αξία των τελεστών που εξάγουν σημασιολογική πληροφορία από ομάδες εγγράφων, με βάση είτε τους εισερχόμενους είτε τους εξερχόμενους συνδέσμους, χρησιμοποιούμε τις κεντρικές ιστοσελίδες (homepages) ορισμένων μουσείων του Λονδίνου. Οι διευθύνσεις ιστού των σελίδων παρουσιάζονται στον κατάλογο του Google στο μονοπάτι (http://directory.google.com/Top/Regional/Europe/United_Kingdom/England/London/Arts_and_Entertainment/Museums/). Το έγγραφο αυτό είναι μια σύζευξη για τις κεντρικές ιστοσελίδες των μουσείων. Η λίστα των διευθύνσεων ιστού βρίσκεται στο παράρτημα Γ.

Αν ορίσουμε το σύνολο των διευθύνσεων ιστού των κεντρικών ιστοσελίδων {U}, το σύνολο των εγγράφων που δείχνουν προς αυτές {I} και το σύνολο των εγγράφων που δείχνονται από αυτές {O}, έχουμε τα πιο κάτω αποτελέσματα:

- weightedGroupKeywords({I},{U}) = museum 681, london 264, home 185, freud 102, belfast 100, hms 99, national 96, maritime 92, link 87, soane 84, holmes 81, victoria 80, albert 80, sherlock 77, madame 77, …
- weightedTargetKeywords({I},{U}) = london 17, museum 14, website 13, site 12, visit 11, home 11, web 10, world 10, information 10, history 9, library 9, page 9, art 8, british 8, national 8, …
- weightedGroupKeywords({U},{O}) = museum 56, link 40, war 24, london 17, information 16, collections 16, shop 15, education 15, imperial 14, news 13, exhibitions 13, soane 12, corporate 12, online 12, exhibition 12, …
- weightedSourceKeywords({U},{O}) = museum 11, exhibition 9, collections 9, shop 9, exhibitions 9, online 8, home 7, news 7, events 7, visit 7, information 7, …

Ο πρώτος τελεστής μετρά τις εμφανίσεις των διαφόρων λέξεων στους 100 πρώτους συνδέσμους που δείχνουν προς κάθε έγγραφο του {U}. Ο δεύτερος τελεστής μετρά τους συνδέσμους που δείχνουν προς κάποιο από τα έγγραφα του {U} χρησιμοποιώντας κάποια λέξη. Για να βρούμε τους εισερχόμενους συνδέσμους, για μια ακόμη φορά χρησιμοποιούμε την υπηρεσία του Google (http://www.google.com/advanced_search) που δίνει τους 100 πρώτους εισερχόμενους συνδέσμους ενός εγγράφου. Ο τρίτος τελεστής μετρά τις εμφανίσεις κάποιας λέξης σε όλους τους εξερχόμενους συνδέσμους των εγγράφων του {U}. Ο τελευταίος τελεστής μετρά τον αριθμό των εγγράφων του {U} που χρησιμοποιούν κάθε λέξη στους εξερχόμενους συνδέσμους τους.

Συγκρίνοντας τα δύο πρώτα σύνολα λέξεων, βλέπουμε ότι λέξεις με μεγάλο αριθμό εμφανίσεων στο πρώτο σύνολο δεν εμφανίζονται στο δεύτερο. Αυτό συμβαίνει για

λέξεις που εμφανίζονται πολλές φορές αλλά μόνο σε συνδέσμους προς το ίδιο έγγραφο (για παράδειγμα το όνομα ενός μουσείου εμφανίζεται σε όλους τους συνδέσμους προς την κεντρική ιστοσελίδα του μουσείου). Αν και ο τελεστής `weightedGroupKeywords` θέτει μεγάλες τιμές στις λέξεις αυτές, ο `weightedTargetKeywords` τις μετρά μόνο μία φορά εφόσον χαρακτηρίζουν μόνο ένα από τα έγγραφα του συνόλου.

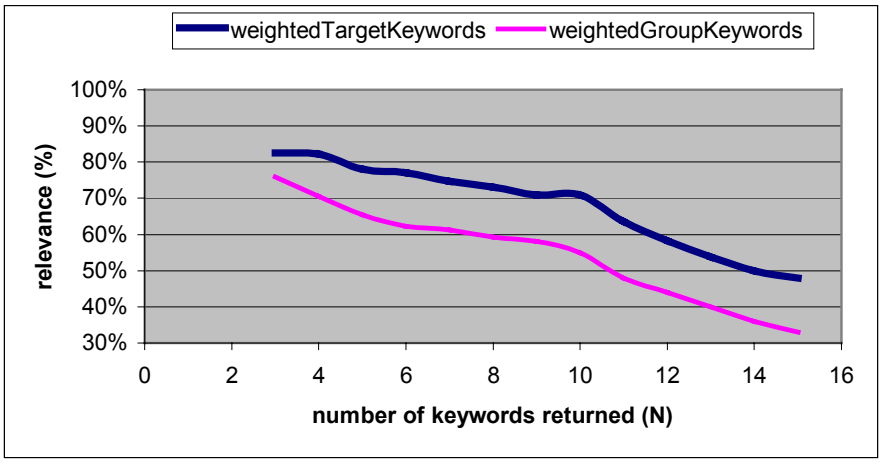
Τα περιεχόμενα του δεύτερου συνόλου χαρακτηρίζουν το σύνολο {U} και όχι τα μεμονωμένα έγγραφά του. Όπως προαναφέραμε, όλα τα έγγραφα του {U} έχουν μια συν-αναφορά, την ιστοσελίδα με τα μουσεία του Google. Μπορούμε λοιπόν να χαρακτηρίσουμε πολύ γρήγορα το σύνολο {U} μέσω του κοινού αναφορέα όλων των εγγράφων του.

Τα δύο τελευταία σύνολα μας ενημερώνουν για τα ενδιαφέροντα των μουσείων, όπως αυτά εκφράζονται στους εξερχόμενους συνδέσμους των πρώτων σελίδων. Στο τελευταίο σύνολο παρατηρούμε τις λέξεις *exhibition, collection, online* και *shop* που αποτελούν συνηθισμένα στοιχεία ενός μουσείου του ΠΙ.

Τα προηγούμενα συμπεράσματα προκύπτουν από ανάλυση της λεξικής πληροφορίας που φέρουν οι σύνδεσμοι. Για να εκμεταλλευτούμε τη σημασιολογική πληροφορία, χρειαζόμαστε μια οντολογία σχετική π.χ. με τον *πολιτισμό*.

Χαρακτηρισμός συνόλων εγγράφων

Στο πείραμα αυτό, εξάγουμε χαρακτηρισμούς για σύνολα σελίδων που βρίσκονται κάτω από μια κατηγορία του καταλόγου διευθύνσεων ιστού του Google, και εκτιμούμε τη σχετικότητα των χαρακτηρισμών με το θέμα της κατηγορίας. Οι δύο πρώτοι τελεστές της προηγούμενης υποενότητας (weightedGroupKeywords και weightedTargetKeywords) χρησιμοποιούνται για το χαρακτηρισμό, και εξάγονται οι N λέξεις με το μεγαλύτερο αριθμό εμφανίσεων (N από 3 ως 15).



Σχήμα 21. Σχετικότητα των λέξεων του χαρακτηρισμού με το θέμα της κατηγορίας

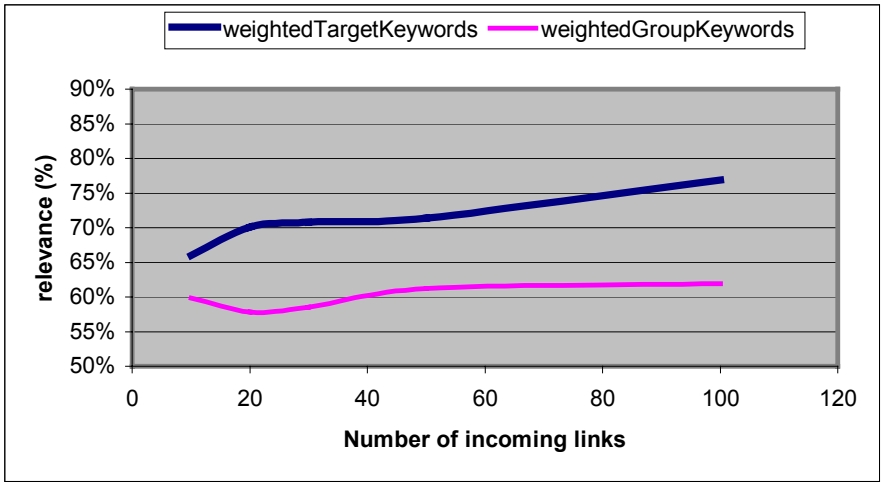
Στη συνέχεια μετρούμε τον αριθμό των λέξεων που είναι σχετικές με το θέμα. Το πείραμα επαναλαμβάνεται για 30 θεματικές κατηγορίες του Google. Για διάφορες τιμές του N υπολογίζουμε το ποσοστό των λέξεων στην περιγραφή που είναι σχετικές με το θέμα.

Θεωρούμε ότι μια λέξη της περιγραφής είναι σχετική με το θέμα αν βρίσκεται στο τίτλο της θεματικής κατηγορίας ή των υπερκατηγοριών της. Το κριτήριο αυτό είναι αρκετά αυστηρό καθώς δεν λαμβάνει υπόψη του συνώνυμες ή συγγενικές λέξεις.
Τα αποτελέσματα απεικονίζονται στο Σχήμα 21.

Τα πιο πάνω παραδείγματα δικαιολογούν την αίσθηση ότι οι τελεστές του THESUS επιτρέπουν το χαρακτηρισμό ενός συνόλου εγγράφων σε ικανοποιητικό επίπεδο, σε σύγκριση με το χαρακτηρισμό που δίνουν άνθρωποι κριτές. Ένα άλλο σημαντικό στοιχείο είναι ότι η ολοκλήρωση των αποτελεσμάτων ανά έγγραφο στόχο (τελεστής weightedTargetKeywords) δίνει καλύτερα αποτελέσματα απ' ό,τι η ολοκλήρωση ανά λέξη (weightedGroupKeywords). Τα πιο πάνω αποτελέσματα δείχνουν ότι, όσο αυξάνει ο αριθμός των λέξεων που χρησιμοποιούνται για το χαρακτηρισμό, τόσο μειώνεται ο βαθμός σχετικότητας με το θέμα. Από το διάγραμμα συμπεραίνουμε ότι αν χρησιμοποιηθούν περισσότερες από 10 μειώνεται σημαντικά η ποιότητα των χαρακτηρισμών, ενώ για λιγότερες από 10 λέξεις τα αποτελέσματα είναι ικανοποιητικά.

Αξίζει να σημειώσουμε ότι τα αποτελέσματα που παρουσιάσαμε χρησιμοποιούν το πολύ 100 εισερχόμενους συνδέσμους για να χαρακτηρίσουν ένα έγγραφο. Παρ' όλα αυτά, το να επισκεφθεί κανείς 100 έγγραφα για να χαρακτηρίσει 1 δεν είναι αρκετά συμφέρον, ακόμη και αν τα έγγραφα βρίσκονται αποθηκευμένα τοπικά και δεν απαιτείται πρόσβαση στο διαδίκτυο. Στο επόμενο πείραμα προσπαθούμε να χαρακτηρίσουμε τα ίδια σύνολα εγγράφων με πριν χρησιμοποιώντας αυτή τη φορά λιγότερους εισερχόμενους συνδέσμους.

Στο Σχήμα 22 παρουσιάζεται η επίδραση του αριθμού των εισερχόμενων συνδέσμων που θα εξεταστούν, στη σχετικότητα των παραγόμενων χαρακτηρισμών. Καθώς από το προηγούμενο διάγραμμα προέκυψε ότι ο αριθμός των λέξεων σε κάθε χαρακτηρισμό δεν πρέπει να ξεπερνά τις 10, και καθώς θεωρούμε ότι λιγότερες από 5 λέξεις δεν είναι αρκετές για να περιγράψουν το σύνολο, επιλέγουμε να χρησιμοποιήσουμε τις 7 πρώτες λέξεις (με το μεγαλύτερο αριθμό εμφανίσεων). Στο σχήμα βλέπουμε ότι ένας περιορισμένος αριθμός συνδέσμων (π.χ. 20 σύνδεσμοι) δίνει συγκρίσιμα αποτελέσματα με αυτά των 100 συνδέσμων (απώλεια σε σχετικότητα λιγότερη από 10% και στους δύο τελεστές).



Σχήμα 22. Πώς επηρεάζει ο αριθμός εισερχόμενων συνδέσμων τη σχετικότητα της περιγραφής

6.2.3 Έλεγχος της ικανότητας του THESUS να ανακαλύπτει θεματικά υποσύνολα

Για να ελέγξουμε την ικανότητα του THESUS να χαρακτηρίζει και να συσταδοποιεί έγγραφα, εξετάζουμε ένα σύνολο από έγγραφα, τα οποία έχουν προηγουμένως ομαδοποιηθεί από ειδικούς και τα συσταδοποιούμε με το THESUS. Στη συνέχεια εκτιμούμε την ποιότητα της συσταδοποίησης που προσφέρουμε, συγκρίνοντάς τη με την υπάρχουσα συσταδοποίηση. Αυτό γίνεται με τη χρήση ενός εξωτερικού μέτρου ελέγχου της ποιότητας της συσταδοποίησης, της ***εντροπίας*** [S48], όπως αυτό ορίζεται στο [KS+00].

Το αρχικό σύνολο περιέχει διευθύνσεις ιστού από ορισμένες κατηγορίες του δικτυακού καταλόγου DMOZ [Dmo], υποκατηγορίες της γενικότερης κατηγορίας Music→Styles. Οι κατηγορίες είναι είδη μουσικής: hip-hop, experimental, world, blues, Indian music, pop, dance, filk, electronic, electronica, country, dance, polka, early music, new age, lounge, rhythm and blues, folk, classical, gamelan, opera, bluegrass, rock, easy listening και περιέχουν από 10 ως 30 διευθύνσεις ιστού η καθεμία. Ο συνολικός αριθμός διευθύνσεων ιστού είναι 481.

Στη συνέχεια υπολογίζουμε την ποιότητα της συσταδοποίησης που παράγεται από τους δύο αλγορίθμους του THESUS, τον DBSCAN και τον COBWEB. Η ελάχιστη τιμή της ***εντροπίας*** είναι 0 και επιτυγχάνεται όταν κάθε έγγραφο τοποθετηθεί σε διαφορετική ομάδα. Η μέγιστη τιμή της ορίζεται όταν κάθε ομάδα περιέχει ένα έγγραφο από κάθε μια διαφορετική κατηγορία. Στην περίπτωση αυτή, η τιμή της εντροπίας γίνεται μέγιστη και ίση με log(N), όπου N είναι ο αριθμός των διαφορετικών κατηγοριών. Υπολογίζουμε τη μέγιστη εντροπία και τη συγκρίνουμε με την εντροπία των συσταδοποιήσεων που προκύπτουν από τη χρήση των DBSCAN και COBWEB για διάφορες τιμές του N.

Οι λεπτομέρειες των παραγόμενων συσταδοποιήσεων φαίνονται στους πιο κάτω πίνακες:

Αρ. Εγγράφων	Αρ. Κατηγοριών	MinSim	Αρ. Εγγράφων που συσταδο-ποιούνται	Αρ. Ομάδων	Εντροπία	Μέγ. εντροπία	Μείωση της εντροπίας
400	23	0.55	115	8	0.96	1.36	29%
400	11	0.8	24	3	0.65	1.04	38%
191	22	0.55	74	8	0.76	1.34	43%
191	7	0.8	16	2	0.47	0.85	44%
457	27	0.65	132	6	1.13	1.43	21%
457	5	0.8	10	2	0.47	0.70	33%
182	23	0.6	71	7	0.79	1.36	42%

Πίνακας 1. Αποτελέσματα του DBSCAN για διάφορες παραμέτρους εισόδου (minDocs=3) με χρήση εννοιών

Αρ. Εγγράφων	Αρ. Κατηγοριών	MinSim	Αρ. Εγγράφων που συσταδο-ποιούνται	Αρ. Ομάδων	Εντροπία	Μέγ. εντροπία	Μείωση της εντροπίας
396	8	0.95	18	7	0.16	0.90	82%
396	13	0.85	31	4	0.67	1.11	40%
295	12	0.95	17	4	0.60	1.08	45%
295	20	0.8	48	6	0.90	1.30	31%

Πίνακας 2. Αποτελέσματα του DBSCAN για διάφορες παραμέτρους εισόδου (minDocs=1) με χρήση λέξεων

Αρ. Εγγράφων	Αρ. Κατηγοριών	Αρ. Ομάδων	Εντροπία	Μέγ. εντροπία	Μείωση της εντροπίας
25	2	7	0.11	0.30	63%
37	3	6	0.27	0.48	43%
55	4	12	0.22	0.60	64%
78	5	7	0.45	0.70	35%
100	10	33	0.58	1.00	42%
160	24	56	0.57	1.38	59%
244	26	85	0.59	1.41	58%

Πίνακας 3. Αποτελέσματα του COBWEB (με χρήση εννοιών) για αυξανόμενο αριθμό εγγράφων

Εκτίμηση των αποτελεσμάτων

Χαρακτηρίζουμε τα έγγραφα εξάγοντας λέξεις από τους εισερχόμενους συνδέσμους και εμπλουτίζοντας την πληροφορία με σημασιολογικά χαρακτηριστικά, χρησιμοποιώντας μια οντολογία σχετική με τη μουσική (177 όροι) και το Wordnet. Καθώς δεν έχουν όλα τα έγγραφα της συλλογής εισερχόμενους συνδέσμους, παράγουμε λεξικές περιγραφές μόνο για όσα έχουν ένα τουλάχιστο εισερχόμενο σύνδεσμο. Κατά την απεικόνιση των λέξεων σε έννοιες της οντολογίας, για ορισμένα έγγραφα δεν προκύπτει σημασιολογική περιγραφή. Για τους δύο αυτούς λόγους ο αριθμός των εγγράφων στους τρεις πίνακες δεν μπορεί να είναι ο ίδιος.

Στην περίπτωση του DBSCAN, τα έγγραφα συσταδοποιούνται με βάση τους όρους της οντολογίας που τα χαρακτηρίζουν (Πίνακας 1) και με βάση τις λέξεις που τα χαρακτηρίζουν (Πίνακας 2). Όταν οι λέξεις έχουν αντιστοιχιστεί στους όρους της οντολογίας, η συσταδοποίηση με βάση τους όρους γίνεται πολύ πιο γρήγορα, καθώς οι αποστάσεις ανάμεσα στους όρους της οντολογίας προϋπολογίζονται για τους 177x177 συνδυασμούς (με χρήση του μέτρου Wu και Palmer's στο δέντρο της οντολογίας), ενώ η ομοιότητα ανάμεσα σε λέξεις πρέπει να υπολογιστεί για κάθε διαφορετικό συνδυασμό λέξεων που εμφανίζεται.

Καθώς οι κατηγορίες που χρησιμοποιήσαμε στο παράδειγμα αναφέρονται όλες σε είδη μουσικής, θεωρούμε πως η μεταξύ τους διάκριση δεν είναι πάντοτε σαφής αλλά ούτε και σύμφωνη με την οντολογία. Για το λόγο αυτό σε κάθε περίπτωση μετρούμε την εντροπία σύμφωνα με την αυθεντική κατηγοριοποίηση του DMOZ (27 κατηγορίες συνολικά) και σύμφωνα με μια κατηγοριοποίηση που προκύπτει από

σύμπτυξη παρόμοιων κατά τη γνώμη μας κατηγοριών. Για παράδειγμα ενώ στην πρώτη γραμμή του πίνακα 1, τα 400 έγγραφα του παραδείγματος προέρχονται από 23 κατηγορίες του DMOZ, στη δεύτερη γραμμή θεωρούμε ότι τα ίδια έγγραφα ανήκουν σε 11 συγχωνευμένες κατηγορίες. Στις περιπτώσεις αυτές, όπως αναμέναμε, η μείωση της εντροπίας είναι σημαντικότερη.

Στην περίπτωση του DBSCAN μια αλλαγή στις παραμέτρους MinSim και MinDoc παράγει διαφορετική συσταδοποίηση. Για να βελτιστοποιήσουμε τη συσταδοποίηση χρειαζόμαστε ένα εσωτερικό μέτρο ποιότητας της συσταδοποίησης που θα μας επιτρέψει να βρούμε τις ιδανικές τιμές των παραμέτρων εισόδου. Καθώς το μέτρο αυτό δεν ήταν διαθέσιμο κατά τη διάρκεια των πειραμάτων, προσομοιώσαμε το μέτρο, εκτελώντας τον αλγόριθμο αρκετές φορές με διαφορετικές παραμέτρους και επιλέγοντας το συνδυασμό εκείνο που ελαχιστοποιούσε την εντροπία σε κάθε περίπτωση.

Ο υπολογισμός της απόστασης για δύο λέξεις απαιτεί πρόσβαση στο Wordnet, για την εύρεση των διαφορετικών σημασιών κάθε λέξης και τον υπολογισμό της ελάχιστης απόστασης μεταξύ των σημασιών των δύο λέξεων, κάτι που επιβραδύνει σημαντικά τη διαδικασία. Και στις δύο περιπτώσεις παρατηρείται σημαντική μείωση στην εντροπία του συστήματος, που είναι εντονότερη όταν χρησιμοποιείται ο λεξικός χαρακτηρισμός (49% της μέγιστης εντροπίας, Πίνακας 2) απ' όταν χρησιμοποιείται ο αντίστοιχος σημασιολογικός (36%, Πίνακας 1).

Στην περίπτωση του COBWEB τα έγγραφα συσταδοποιούνται με βάση το σημασιολογικό χαρακτηρισμό και η εντροπία της παραγόμενης συσταδοποίησης είναι μικρότερη από τη μέγιστη εντροπία (μείωση κατά 52%). Ο Πίνακας 3 δείχνει ότι η εντροπία σταθεροποιείται για περισσότερα από 100 περίπου έγγραφα. Αυτό σημαίνει ότι ο αλγόριθμος συσταδοποίησης απαιτεί τη συσταδοποίηση ένα ελάχιστου αριθμού εγγράφων για να δώσει αξιόπιστα αποτελέσματα.

Τα συμπεράσματα από το παραπάνω πείραμα μπορούν να συνοψιστούν στα εξής:
- Τα αποτελέσματα του DBSCAN είναι ικανοποιητικά στις περισσότερες περιπτώσεις και βελτιώνονται όταν ο αριθμός των παραγόμενων συστάδων είναι παρόμοιος με τον αριθμό των κατηγοριών.
- Με τη χρήση σημασιολογικών αντί των λεξικών περιγραφών, χάνουμε σε αποτελεσματικότητα (αυξημένη εντροπία) αλλά κερδίζουμε σε απόδοση.
- Τα αποτελέσματα του COBWEB εμφανίζουν ελαφρώς μικρότερη εντροπία από αυτά του DBSCAN, λόγω όμως της φύσης του αλγορίθμου παράγονται πολλές μικρές συστάδες. Το γεγονός αυτό σε συνδυασμό με τον ορισμό της εντροπίας δείχνει πως τα «καλά» αποτελέσματα του COBWEB είναι πλασματικά καθώς πολλά έγγραφα έχουν τοποθετηθεί σε ξεχωριστές συστάδες μειώνοντας τη συνολική εντροπία.
- Η ταχύτητα του COBWEB μειώνεται σημαντικά μετά από ένα σχετικά μικρό αριθμό εγγράφων. Για το λόγο αυτό στα υπόλοιπα πειράματα χρησιμοποιούμε μόνο τον αλγόριθμο DBSCAN.

6.2.4 Παράγοντες που επηρεάζουν την αποτελεσματικότητα του THESUS

Σε ένα ακόμη μεγαλύτερο πείραμα, προσπαθήσαμε να αξιολογήσουμε τους παράγοντες που επηρεάζουν την αποτελεσματικότητα του συστήματος στη

συσταδοποίηση των εγγράφων. Θεωρήσαμε ότι οι βασικοί παράγοντες που επηρεάζουν την τελική συσταδοποίηση είναι:

- η ποιότητα των παραγόμενων περιγραφών για κάθε έγγραφο,
- η χρήση λεξικών ή σημασιολογικών περιγραφών,
- η εγγύτητα των αρχικών κατηγοριών,
- η επιλογή μέτρου ομοιότητας.

Για το λόγο αυτό επαναλάβαμε την διαδικασία συσταδοποίησης για ένα προκαθορισμένο και προ-κατηγοριοποιημένο σύνολο εγγράφων με διάφορες συνθήκες και συγκρίναμε τις τιμές που μας έδωσαν τα εξωτερικά μέτρα αξιολόγησης των σχημάτων συσταδοποίησης.

Το σενάριο των πειραμάτων έχει ως εξής:
Χρησιμοποιήθηκαν έγγραφα που έχουν περιγραφεί και ταξινομηθεί σε κατηγορίες από ανθρώπους, όπως αυτά του δικτυακού καταλόγου DMOZ. Τα έγγραφα αυτά συσταδοποιήθηκαν από το THESUS χρησιμοποιώντας διαφορετικούς τρόπους για την εξαγωγή χαρακτηρισμών και διαφορετικά μέτρα ομοιότητας, και τα τελικά αποτελέσματα αξιολογήθηκαν. Τα πειράματα έχουν ως σκοπό να συγκρίνουν την επιρροή των προαναφερθέντων παραγόντων.

Η διαδικασία δημιουργίας, χαρακτηρισμού και συσταδοποίησης της συλλογής έγινε στα ακόλουθα στάδια:

1. επιλέχθηκαν έγγραφα από διάφορες κατηγορίες του DMOZ, κατηγορίες σε διάφορα επίπεδα της ιεραρχίας του καταλόγου,
2. τα επιλεγμένα έγγραφα χαρακτηρίστηκαν χρησιμοποιώντας λέξεις που εξήχθησαν: από τις περιγραφές που έδωσαν οι συντάκτες του DMOZ, από τα περιεχόμενα του εγγράφου και από τους εισερχόμενους συνδέσμους των εγγράφων αυτών,
3. οι χαρακτηρισμοί εμπλουτίστηκαν σημασιολογικά αντιστοιχίζοντας τις λέξεις που εξήχθησαν στις πλησιέστερες έννοιες μιας θεματικής ιεραρχίας,
4. η ομοιότητα μεταξύ των εγγράφων υπολογίστηκε με τη χρήση της ομοιότητας συνημιτόνου (cosine similarity) επί των λεξικών περιγραφών και με τη χρήση του μέτρου ομοιότητας του THESUS (THESIM) επί των αντίστοιχων σημασιολογικών περιγραφών

Αξιολόγηση της ποιότητας συσταδοποίησης των εγγράφων

Η ποιότητα των αποτελεσμάτων της συσταδοποίησης μετρήθηκε με δύο μέτρα, το **F-measure** [LA99] (βλ. ενότητα 2.8.6.3.) και το **Rand Statistic** [TK99]. Και τα δύο μέτρα είναι εξωτερικά μέτρα ποιότητας, που σημαίνει ότι μετρούν το βαθμό ομοιότητας ανάμεσα σε μια προκαθορισμένη ομαδοποίηση ενός συνόλου D, και της παραγόμενης συσταδοποίησης από την εφαρμογή του αλγορίθμου συσταδοποίησης στο D. Στην περίπτωση των πειραμάτων μας η προκαθορισμένη ομαδοποίηση, την οποία θεωρούμε και ιδανική, είναι αυτή που παρέχουν οι συντάκτες του DMOZ.

Τα τρία σύνολα εγγράφων

Παρόμοια πειράματα συσταδοποίησης συνόλων εγγράφων του ΠΙ αξιολογούν την ποιότητα της συσταδοποίησης χρησιμοποιώντας συνήθως εξωτερικά μέτρα ποιότητας και συγκρίνονται με μονοεπίπεδες ομαδοποιήσεις εγγράφων, όπως τα έγγραφα της συλλογής TREC για τον ΠΙ [TREC]. Ακόμη και στην περίπτωση που

χρησιμοποιούνται έγγραφα από μια πολυεπίπεδη οργάνωση εγγράφων (όπως το Yahoo ή το DMOZ) τα έγγραφα επιλέγονται από κατηγορίες του ίδιου επιπέδου και μάλιστα από κατηγορίες με κοινό άμεσο πρόγονο. Οι κατηγορίες βρίσκονται είτε στο πρώτο επίπεδο (π.χ. αθλητικά, επιχειρήσεις) ή σε κατώτερα επίπεδα (π.χ. ποδόσφαιρο, βόλευ) [DC00]. Οι κατηγορίες αυτές, συνήθως, περιέχουν τις διευθύνσεις ιστού λίγων εγγράφων και ορισμένες υποκατηγορίες με τις δικές τους διευθύνσεις ιστού. Στις περισσότερες περιπτώσεις, οι κατηγορίες στο κορυφαίο επίπεδο συγχωνεύουν τα περιεχόμενα των υποκατηγοριών τους. Το τελικό σχήμα συσταδοποίησης συγκρίνεται ως προς το συγχωνευμένο αρχικό σχήμα στις κατηγορίες του κορυφαίου επιπέδου.

Την ίδια μέθοδο χρησιμοποιούμε και για την αξιολόγηση της διαδικασίας συσταδοποίησης του THESUS. Η συλλογή εγγράφων αποτελείται από τα έγγραφα στις διευθύνσεις ιστού που βρίσκονται κάτω από την κατηγορία arts/music στο DMOZ. Από τις διευθύνσεις αυτές αφαιρούνται οι κατηγορίες “bands and artists” και “by-letter” καθώς περιέχουν έγγραφα που αναφέρονται σε συγκεκριμένους καλλιτέχνες, εταιρίες κτλ. χωρίς ιδιαίτερο σημασιολογικό περιεχόμενο. Η συλλογή αυτή (FULL-SET) περιέχει περίπου 30000 URLs από 2155 διαφορετικές κατηγορίες.

Η θεματική οντολογία που χρησιμοποιήθηκε στο συγκεκριμένο πείραμα είναι μια ιεραρχία εννοιών σχετικών με τη μουσική. Ακολουθεί σε σημαντικό βαθμό την ταξινομία του DMOZ χωρίς όμως να ταυτίζεται. Καθώς η θεματική οντολογία πρέπει να καθορίζει τα ενδιαφέροντα μιας ομάδας χρηστών σε σημεία που εκτιμήθηκε πως η θεματική ταξινομία του DMOZ δεν εκφράζει τα ενδιαφέροντά μας, διαφοροποιήσαμε τη θεματική οντολογία μας. Για παράδειγμα το DMOZ διακρίνει τις κατηγορίες μουσικής ‘Gamelan’ και ‘Indian’, ενώ στην οντολογία που χρησιμοποιούμε θεωρούμε την έννοια ‘World Music’ και περιμένουμε έγγραφα και των δύο κατηγοριών να χαρακτηριστούν με την έννοια αυτή και να συσταδοποιηθούν. Για το λόγο αυτό είναι αναμενόμενο η τελική συσταδοποίηση να μην είναι ακριβώς ίδια με αυτή του DMOZ.

Πολλές από τις διαφοροποιήσεις της οντολογίας ως προς την κατηγοριοποίηση του DMOZ, έγιναν και για τεχνικούς λόγους και έχουν ως αποτέλεσμα η ποιότητα της τελικής συσταδοποίησης να αξιολογηθεί μικρότερη απ’ ότι πραγματικά είναι.

Παρ' όλα αυτά, αν θεωρήσουμε την επιπεδοποιημένη συσταδοποίηση (όταν όλες οι υποκατηγορίες και οι διευθύνσεις που περιέχονται σε αυτές συγχωνεύονται στη βασική κατηγορία), σε πρώτο ή δεύτερο επίπεδο, τέτοιες διακρίσεις και διαφορές μεταξύ της θεματικής οντολογίας και του DMOZ αναμένεται να μην επηρεάσουν το τελικό αποτέλεσμα. Για το σκοπό θεωρούμε δύο ακόμη συλλογές με επιπεδοποιημένη συσταδοποίηση.

Στην πρώτη συλλογή (LEVEL1) περιέχονται οι πολυπληθέστερες άμεσες υποκατηγορίες του arts/music. Οι 6 αυτές κατηγορίες περιέχουν περίπου 25000 έγγραφα με την ακόλουθη κατανομή (Πίνακας 4).

Sound files	1933
Instruments	6082
Lyrics	853
Vocal	733
Marching	928
Styles	13061

Πίνακας 4. Αριθμός εγγράφων ανά κατηγορία στη συλλογή LEVEL1

Η δεύτερη συλλογή (LEVEL2) περιέχει περίπου 6000 έγγραφα που αφορούν μουσικά όργανα. Τα έγγραφα προέρχονται από 6 κατηγορίες κάτω από την κατηγορία arts/music/instruments. Η κατανομή των εγγράφων ανά επιπεδοποιημένη κατηγορία ακολουθεί (Πίνακας 5).

Electronic	378
Keyboard	784
Percussion	293
Squeezebox	94
Strings	2627
Winds	1735

Πίνακας 5. Αριθμός εγγράφων ανά κατηγορία στη συλλογή LEVEL2

Είναι προφανές από την ιεραρχική δομή του DMOZ ότι:

$$\text{LEVEL2} \subset \text{LEVEL1} \subset \text{FULL-SET}.$$

6.2.4.1 Σύγκριση τεχνικών χαρακτηρισμού

Για να επιλέξουμε τον καλύτερο τρόπο χαρακτηρισμού ενός εγγράφου, παράγουμε τρεις διαφορετικούς χαρακτηρισμούς για κάθε έγγραφο σε κάθε συλλογή. Οι χαρακτηρισμοί αποτελούνται από:

- τις λέξεις που περιέχονται στις ***περιγραφές που δίνουν οι συντάκτες του DMOZ*** (10 με 20 λέξεις) και τον αριθμό εμφανίσεων κάθε λέξης στην περιγραφή (DMOZ)
- από τις 10 συχνότερα χρησιμοποιούμενες ***λέξεις που εξάγονται από τα περιεχόμενα*** του εγγράφου και τον αριθμό εμφανίσεών τους (CONTENT)
- από τις ***λέξεις που εμφανίζονται συχνότερα στους 100 (το πολύ) εισερχόμενους συνδέσμους*** του κάθε εγγράφου, και τον αριθμό των εισερχόμενων συνδέσμων στους οποίους εμφανίζεται κάθε λέξη (INLINKS). Να σημειωθεί ότι ο μέσος αριθμός λέξεων που εξάγεται με τον τρόπο αυτό είναι μικρότερος του 10, γιατί οι εισερχόμενοι σύνδεσμοι χρησιμοποιούν λίγες λέξεις (και συνήθως τις ίδιες) για να χαρακτηρίσουν το έγγραφο.

Διαφορετικές λέξεις στον χαρακτηρισμό ενός εγγράφου μπορεί να αντιστοιχιστούν στην ίδια έννοια της οντολογίας, γι' αυτό οι έννοιες που αποτελούν το σημασιολογικό χαρακτηρισμό των εγγράφων είναι συνήθως λιγότερες από τις αντίστοιχες λέξεις.

6.2.4.2 Σύγκριση του μέτρου ομοιότητας με το μέτρο συνημιτόνου

Για να συγκρίνουμε την αποτελεσματικότητα του μέτρου ομοιότητας εγγράφων του THESUS, αντιστοιχίζουμε τους λεξικούς χαρακτηρισμούς σε σημασιολογικούς για κάθε έγγραφο. Οι έννοιες που χρησιμοποιούνται για το σημασιολογικό χαρακτηρισμό των εγγράφων της συλλογής είναι έννοιες της θεματικής οντολογίας και βρίσκονται οργανωμένες σε μια ιεραρχία. Μπορούμε συνεπώς για αυτές να χρησιμοποιήσουμε το μέτρο Wu-Palmer για να υπολογίσουμε την μεταξύ τους ομοιότητα. Για τα έγγραφα της συλλογής μπορούμε να υπολογίσουμε το μέτρο ομοιότητας του THESUS (THESIM) και να συσταδοποιήσουμε τα έγγραφα. Επαναλαμβάνουμε την ίδια διαδικασία υπολογίζοντας αυτή τη φορά την ομοιότητα με το μέτρο συνημιτόνου (COSINE – βλ. ενότητα 2.8.2, εξ. 1).

6.2.4.3 Αξιολόγηση της ποιότητας συσταδοποίησης με χρήση εξωτερικών μέτρων ποιότητας

Στα πειράματα της ενότητας αυτής εφαρμόζουμε διαφορετικές αρχικές τιμές παραμέτρων για τον αλγόριθμο DBSCAN. Καθώς ο αλγόριθμος δεν έχει τη δυνατότητα να εντοπίσει μόνος του το βέλτιστο σχήμα συσταδοποίησης, επαναλαμβάνουμε τη διαδικασία πολλές φορές και κρατάμε το σχήμα εκείνο για το οποίο έχουμε την καλύτερη ποιότητα συσταδοποίησης. Η ποιότητα της συσταδοποίησης υπολογίζεται με χρήση δύο εξωτερικών μέτρων ποιότητας, των Rand Statistic και F-measure. Και τα δύο μέτρα συγκρίνουν την επικάλυψη του παραγόμενου σχήματος συσταδοποίησης με το αρχικό σχήμα κατηγοριοποίησης (LEVEL1, LEVEL2, FULL-SET)

Ο αριθμός των συστάδων που προκύπτουν σε κάθε περίπτωση εξαρτάται από τις τιμές των παραμέτρων εισόδου του αλγορίθμου συσταδοποίησης. Για κάθε συνδυασμό συλλογής εγγράφων, τρόπου χαρακτηρισμού και μέτρου ομοιότητας, η διαδικασία συσταδοποίησης επαναλαμβάνεται με διαφορετικές τιμές παραμέτρων εισόδου και κάθε φορά το τελικό αποτέλεσμα συγκρίνεται με την αρχική κατηγοριοποίηση (2155 κατηγορίες στο FULL-SET, 6 κατηγορίες στα LEVEL 1 και LEVEL 2). Επιλέγεται το σχήμα με τη μέγιστη ομοιότητα στην αρχική κατηγοριοποίηση και συνεπώς με τη μέγιστη τιμή F-measure και Rand Statistic.

Τα δύο μέτρα χρησιμοποιούν ως μέτρο σύγκρισης την ομαδοποίηση που δίνουν οι συντάκτες του DMOZ. Γενικότερα, τα εξωτερικά μέτρα ποιότητας, όπως το F-measure και το Rand Statistic, εξετάζουν όλα τα δυνατά ζεύγη που παράγονται από το σύνολο των εγγράφων. Η ποιότητα της συσταδοποίησης αυξάνει όταν έγγραφα της ίδιας κατηγορίας τοποθετούνται μαζί και μειώνεται όταν έγγραφα της ίδιας κατηγορίας καταλήγουν σε διαφορετικές συστάδες.

Πιο συγκεκριμένα, τα μέτρα αξιολογούν την ικανότητα του αλγορίθμου συσταδοποίησης μετρώντας το βαθμό ανάκλησης (recall) και ακρίβειας (precision) (ενότητα 2.8.6.3., εξ. 6 & 7) στις παραγόμενες ομάδες σε σχέση με τις αρχικές. Η *ανάκληση* υπολογίζει την αναλογία των εγγράφων μιας κατηγορίας που τοποθετήθηκαν στην ίδια συστάδα, ενώ η *ακρίβεια* την αναλογία των εγγράφων της ίδιας συστάδας που προέρχονται από την ίδια αρχική κατηγορία. Το μέτρο *Rand statistic* μετρά το ποσοστό των ζευγών εγγράφων που προέρχονται από την ίδια κατηγορία και τοποθετήθηκαν στην ίδια συστάδα, και των ζευγών που προέρχονται από διαφορετικές κατηγορίες και τοποθετήθηκαν σε διαφορετικές συστάδες.

Όπως φαίνεται και στα αποτελέσματα:

- Η χρήση του μέτρου ομοιότητας του THESUS σε σύνολα εννοιών δίνει καλύτερα αποτελέσματα από το μέτρο συνημιτόνου για σύνολα λέξεων.
- Τα αποτελέσματα που προκύπτουν, αν χρησιμοποιήσουμε τις περιγραφές των εισερχόμενων συνδέσμων, είναι πολύ καλύτερα από αυτά που δίνουν οι περιγραφές που εξάγονται από τα περιεχόμενα των εγγράφων. Είναι επίσης συγκρίσιμα με τα αποτελέσματα που προκύπτουν αν χρησιμοποιηθούν οι περιγραφές των συντακτών του DMOZ.
- Είναι προφανές ότι ο υπολογισμός του μέτρου ομοιότητας THESIM δεν είναι χρονοβόρος. Η μετάβαση από λέξεις σε έννοιες μιας θεματικής οντολογίας συνεπάγεται πολύ μικρότερα διανύσματα χαρακτηρισμών, καθώς περισσότερες από μια λέξεις αντιστοιχίζονται στις ίδιες έννοιες.

		FULLSET		LEVEL1		LEVEL2	
		THESIM	COSINE	THESIM	COSINE	THESIM	COSINE
DMOZ	Είσοδος						
	MinDoc	1	1	5	5	5	5
	MinSim	1	0.4	0.7	0.2	0.65	0.2
	Έξοδος						
	Clusters	1120	2002	7	17	6	19
	Clust.doc.	16436	17921	14315	22967	2853	5577
	Tot. doc.	22575	29115	18252	23590	4615	5911
	F-measure	0.182	0.196	0.724	0.533	0.694	0.589
	Rand Stat	0.913	0.849	0.502	0.482	0.513	0.378
	Cl. time	80	90	40	60	0.5	0.5
	Init. Time	689	850	499	2285	25	78
CONTENT	Είσοδος						
	MinDoc	1	1	1	1	10	25
	MinSim	1	0.45	0.7	0.15	0.7	0.45
	Έξοδος						
	Clusters	598	1258	6	8	8	6
	Clust.doc.	9891	13507	10006	17545	1679	1663
	Tot. doc.	12774	21904	10242	17743	3011	4636
	F-measure	0.162	0.160	0.720	0.492	0.623	0.483
	Rand Stat	0.917	0.849	0.513	0.501	0.518	0.485
	Cl. time	52	60	40	60	0.60	0.6
	Init. Time	309	650	131	562	12	38
INLINKS	Είσοδος						
	MinDoc	1	1	5	5	5	5
	MinSim	1	0.4	0.65	0.1	0.65	0.1
	Έξοδος						
	Clusters	879	928	4	23	5	5
	Clust.doc.	12080	14397	11312	13534	2737	4062
	Tot. doc.	17913	18944	14248	15048	3473	4147
	F-measure	0.200	0.130	0.730	0.592	0.639	0.511
	Rand Stat	0.751	0.926	0.500	0.441	0.503	0.330
	Cl. time	88	93	52	60	2.7	3.1
	Init. time	365	565	580	598	34	385

Πίνακας 6 Αποτελέσματα συσταδοποίησης για τρία σύνολα δεδομένων με επιλογή των βέλτιστων τιμών των παραμέτρων εισόδου

Είσοδος: Παράμετροι εισόδου, MinSim η ελάχιστη ομοιότητα μεταξύ δύο εγγράφων της ίδιας συστάδας, MinDoc+1 ο ελάχιστος αριθμός εγγράφων σε μια συστάδα.

Έξοδος: Ο αριθμός συστάδων που παράγονται και ο συνολικός αριθμός εγγράφων που τοποθετούνται σε κάποια συστάδα. Τα υπόλοιπα έγγραφα θεωρούνται θόρυβος.

F-measure, Rand Stat.: Τα δύο μέτρα ποιότητας της συσταδοποίησης.

Init. time: Ο χρόνος αρχικοποίησης (σε δευτερόλεπτα) που απαιτείται για να υπολογιστεί η απόσταση μεταξύ όλων των εγγράφων.

Cl. time: Μέσος χρόνος συσταδοποίησης (σε δευτερόλεπτα).

Τεχνική σημείωση: Η διαδικασία συσταδοποίησης επαναλαμβάνεται πολλές φορές σε κάθε πείραμα και για το λόγο αυτό θα πρέπει να είναι όσο το δυνατόν πιο γρήγορη. Ο υπολογισμός ορισμένων μεγεθών εκ των προτέρων και η αποθήκευσή τους στη μνήμη μπορεί να επιταχύνουν σημαντικά τη διαδικασία. Για να μπορεί το σύστημα να διαχειριστεί μεγάλο αριθμό εγγράφων, πρέπει να αποθηκεύεται στην κύρια μνήμη η ελάχιστη δυνατή πληροφορία για κάθε έγγραφο. Για το λόγο αυτό επιλέγουμε να υπολογίσουμε εκ των προτέρων και να αποθηκεύσουμε για κάθε έγγραφο την απόστασή του από τα υπόλοιπα έγγραφα. Με αυτό τον τρόπο μπορούμε να επιταχύνουμε σημαντικά τη διαδικασία συσταδοποίησης η οποία, εφόσον γίνει ο αρχικός υπολογισμός, μπορεί να επαναληφθεί γρήγορα για μεγάλο αριθμό εγγράφων (0,5 sec για 6000 έγγραφα). Ο πίνακας ομοιότητας των εγγράφων της συλλογής είναι ένας τριγωνικός πίνακας με μέγεθος όσο το πλήθος των εγγράφων της συλλογής. Αν θεωρήσουμε για παράδειγμα τη συλλογή FULL-SET με τα 30000 περίπου έγγραφα, το μέγεθος του πίνακα στη μνήμη είναι (30000x30000)/2 και αν θεωρήσουμε ότι οι ομοιότητες των εγγράφων αποθηκεύονται σε bytes (τιμές από 0 έως 100), το μέγεθος μνήμης που απαιτείται είναι περίπου 400 Mbytes.

Μια προσεκτικότερη ανάλυση των αποτελεσμάτων δείχνει ότι ο αριθμός των εγγράφων της συλλογής για τα οποία έχουμε κάποιο χαρακτηρισμό διαφέρει για κάθε μια από τις τρεις τεχνικές χαρακτηρισμού. Αυτό συμβαίνει γιατί δεν υπάρχει διαθέσιμη η αντίστοιχη πληροφορία για όλα τα έγγραφα. Για παράδειγμα, περιγραφές του DMOZ υπάρχουν για 22575 έγγραφα του FULL-SET. Όταν όμως προσπαθήσαμε να διαβάσουμε τα περιεχόμενά τους, μόνο τα 12774 από αυτά ήταν διαθέσιμα. Αυτό συνέβη είτε γιατί τα έγγραφα δεν ήταν προσβάσιμα τη στιγμή του πειράματος, είτε γιατί τα περιεχόμενά τους δεν είναι σε αναγνώσιμη μορφή (έγγραφα pdf, doc, flash κτλ.). Από την άλλη, πληροφορία εισερχόμενων συνδέσμων είχαμε για πολύ περισσότερα έγγραφα (17913). Η ίδια συμπεριφορά παρατηρήθηκε και για τις άλλες δύο συλλογές εγγράφων (LEVEL 1, LEVEL 2).

Επιπλέον, όταν οι λεξικές περιγραφές αντικαταστάθηκαν από τις σημασιολογικές και το μέτρο THESIM χρησιμοποιήθηκε αντί του μέτρου συνημιτόνου, ο αριθμός εγγράφων της συλλογής μειώθηκε. Ο κύριος λόγος για αυτό είναι ότι πολλές λεξικές περιγραφές δεν ήταν σχετικές με το πεδίο ενδιαφέροντος και, κατά συνέπεια, δεν αντιστοιχίστηκαν σε έννοιες της οντολογίας. Με το τρόπο αυτό φαίνεται η ικανότητα του μηχανισμού απεικόνισης να απορρίπτει έγγραφα που δεν είναι σχετικά, πριν ακόμη περάσουμε στη φάση της συσταδοποίησης των εγγράφων.

6.2.5 Τελεστές ανάλυσης υπερσυνδέσμων

Ένα μεγαλύτερο πείραμα πραγματοποιήθηκε για τη συλλογή 40000 εγγράφων σχετικών με την *τεχνολογία*. Τα έγγραφα αυτά συλλέχθηκαν με υποβολή συζευκτικών ερωτήσεων για τους όρους της οντολογίας στη μηχανή Google, όπως περιγράφεται στη μέθοδο β της ενότητας 3.1.2. Χρησιμοποιώντας μια υπηρεσία εισερχόμενων συνδέσμων, εντοπίστηκαν περίπου 700000 διευθύνσεις ιστού που έδειχναν σε ένα τουλάχιστο έγγραφο της συλλογής. Τα έγγραφα αυτά αναλύθηκαν και εξήχθη πληροφορία από αυτά. Συνολικά αποθηκεύθηκαν 1.7 εκατομμύρια σύνδεσμοι.

Στο πείραμα αυτό περιοριστήκαμε στη χρήση της λεξικής πληροφορίας των συνδέσμων και δε χρησιμοποιήσαμε οντολογία. Στη συνέχεια, παρουσιάζουμε ορισμένα ενδεικτικά αποτελέσματα, που δείχνουν τη χρήση των τελεστών ανάλυσης της πληροφορίας των υπερσυνδέσμων (εύρεση θεματικών π-κόμβων και α-κόμβων κτλ.) και δεν αποτελούν αξιολόγηση των τελεστών.

Θεματικοί α-κόμβοι

Το πρώτο ερώτημα που δοκιμάσαμε στη συλλογή εγγράφων αφορούσε θεματικούς α-κόμβους σε "nuclear physics research". Οι απαντήσεις που πήραμε ταξινομήθηκαν με βάση το πλήθος των εισερχόμενων συνδέσμων. Για το ίδιο ερώτημα τα αποτελέσματα των υπολοίπων μηχανών αναζήτησης ([Goo], [Teo], [Yah]) ταξινομούνται με τους αντίστοιχους μηχανισμούς ταξινόμησης κάθε μηχανής. Τα αποτελέσματα που δίνει το THESUS είναι συγκρίσιμα με αυτά των μηχανών αναζήτησης. Αυτό επιβεβαιώνει την αίσθηση ότι οι μηχανές αναζήτησης επιβραβεύουν σελίδες που είναι σημαντικοί α-κόμβοι ή π-κόμβοι.

THESUS	18 σύνδ.	www.cern.ch/	European Organisation for Nuclear Research
	10 σύνδ.	www.rarf.riken.go.jp/rarf/np/	A hub on nuclear physics pages
	7 σύνδ.	www.er.doe.gov/production/henp/henp.html	The US office of High Energy and Nuclear Physics.
Google	1	www.cern.ch/	European Organisation for Nuclear Research.
	2	www.rarf.riken.go.jp/rarf/np/nplab.html	A hub on nuclear physics pages
	3	www.rcnp.osaka-u.ac.jp/index-e.html	Japanese Research centre for nuclear physics
Teoma	1	www.iop.org/	The Institute of Physics
	2	www.pppl.gov/	Princeton Plasma Physics Laboratory (PPPL)
	3	www.rarf.riken.go.jp/rarf/np/nplab.html	A hub on nuclear physics pages
Yahoo	1	www.er.doe.gov/production/henp/henp.html	The US Office of High Energy and Nuclear physics.
	2	www.jinr.dubna.su/	Russian Joint Institute for Nuclear Research
	3	www.nikhef.nl/	Dutch National Institute for Nuclear Physics and High Energy Physics

Πίνακας 7 Θεματικοί α-κόμβοι στο THESUS

Θεματικοί π-κόμβοι

Υποβάλλοντας την ίδια ερώτηση με πριν, ψάχνοντας όμως αυτή τη φορά για π-κόμβους με πολλούς εξερχόμενους συνδέσμους που να περιέχουν τις λέξεις του ερωτήματος, τα αποτελέσματα που παίρνουμε με φθίνουσα σειρά ως προς το πλήθος των εξερχόμενων συνδέσμων είναι:

- (18 σύνδεσμοι) www.cern.ch/Physics/HEP.html: Πάνω από 200 σύνδεσμοι σε σημαντικά κέντρα πυρηνικής φυσικής. Σπουδαίος κόμβος παραπομπών.
- (9 σύνδεσμοι) www.search-beat.com/Science/Technology/Energy/Nuclear/: Περιέχει 75 συνδέσμους σχετικά με έρευνα στην πυρηνική φυσική. Ένας καλός π-κόμβος.
- (6 σύνδεσμοι) netmation.com/www/phystd.htm: Η σελίδα περιέχει συνδέσμους σε σελίδες σχετικές με φυσική και λιγότερο με πυρηνική φυσική

Αξίζει εδώ να σημειωθεί ότι διάκριση ανάμεσα σε ***π-κόμβους*** και ***α-κόμβους*** εμφανίστηκε πρόσφατα από τη μηχανή αναζήτησης ΤΕΟΜΑ, όπου τα έγγραφα έχουν συγκεντρωθεί και αξιολογηθεί προκαταβολικά ως π-κόμβοι ή α-κόμβοι με ανάλυση της συνδεσμολογίας ολόκληρης της συλλογής. Επιπλέον, σε αυτή την περίπτωση η αξία των εγγράφων ως π-κόμβων ή α-κόμβων είναι προκαθορισμένη και ανεξάρτητη από τις ερωτήσεις των χρηστών, ενώ και η αναζήτηση για τις λέξεις της ερώτησης γίνεται και πάλι στο περιεχόμενο των εγγράφων και όχι στους εισερχόμενους ή εξερχόμενους συνδέσμους αυτών.

Θεματικές βιβλιογραφικές συζεύξεις - couplings

Στη συνέχεια ψάχνουμε για σελίδες που δείχνονται και από τους τρεις π-κόμβους για "nuclear physics research" που βρήκαμε στο προηγούμενο πείραμα, και χρησιμοποιούν τους "research" και "institute" στους υπερσυνδέσμους. Τα 16 αποτελέσματα (Πίνακας 8) περιέχουν: α) 8 κεντρικές σελίδες ερευνητικών ιδρυμάτων σχετικών με την πυρηνική φυσική και τη φυσική υψηλών ενεργειών, β) 4 κεντρικές σελίδες πανεπιστημίων ή ιδρυμάτων σχετικού ενδιαφέροντος, γ) 2 κεντρικές σελίδες ιδρυμάτων που δεν ασχολούνται με την πυρηνική φυσική και δ) 2 σημαντικούς π-κόμβους σχετικούς με πυρηνική φυσική.

http://www.rarf.riken.go.jp/
http://www.ijs.si/
http://www.hepi.edu.ge/
http://theory.tifr.res.in/
http://infohelpway.com/featured.htm?http://www.ieer.org
http://www.rarf.riken.go.jp/rarf/np/
http://www.ictp.trieste.it/
http://www.i-b-r.org/index.htm
http://www.rmki.kfki.hu/
http://sunhe.jinr.dubna.su/
http://www.shef.ac.uk/
http://knuhep2.kyungpook.ac.kr/
http://www.inr.ac.ru/welcome.html
http://www.scri.fsu.edu/
http://www.hep.ph.rhbnc.ac.uk/
http://www.uoi.gr/

Πίνακας 8. Παράδειγμα βιβλιογραφικών συζεύξεων

Το παράδειγμα αποδεικνύει την ικανότητα του συστήματος THESUS να εντοπίζει π-κόμβους σε κάποιο θέμα και να τους χρησιμοποιεί για να βρει σελίδες με παρόμοια σημασιολογία που αναφέρονται από όλους αυτούς τους κόμβους. Αυτό το χαρακτηριστικό δε χρησιμοποιείται ιδιαίτερα από τις μηχανές αναζήτησης.

6.3 Συμπέρασμα

Συγκρίνοντας τα αποτελέσματα σε διάφορα σύνολα σελίδων, από μερικές εκατοντάδες σε μερικές χιλιάδες σελίδες, με τα αποτελέσματα που παρέχουν οι μηχανές αναζήτησης και οι κατάλογοι του διαδικτύου καταλήξαμε στα εξής συμπεράσματα:

1- Οι αναζητήσεις που βασίζονται στην πληροφορία των υπερσυνδέσμων προσφέρουν πληροφορία εξίσου αξιόπιστη, ίσως και καλύτερη, από αυτές που βασίζονται σε ολόκληρο το περιεχόμενο των σελίδων.

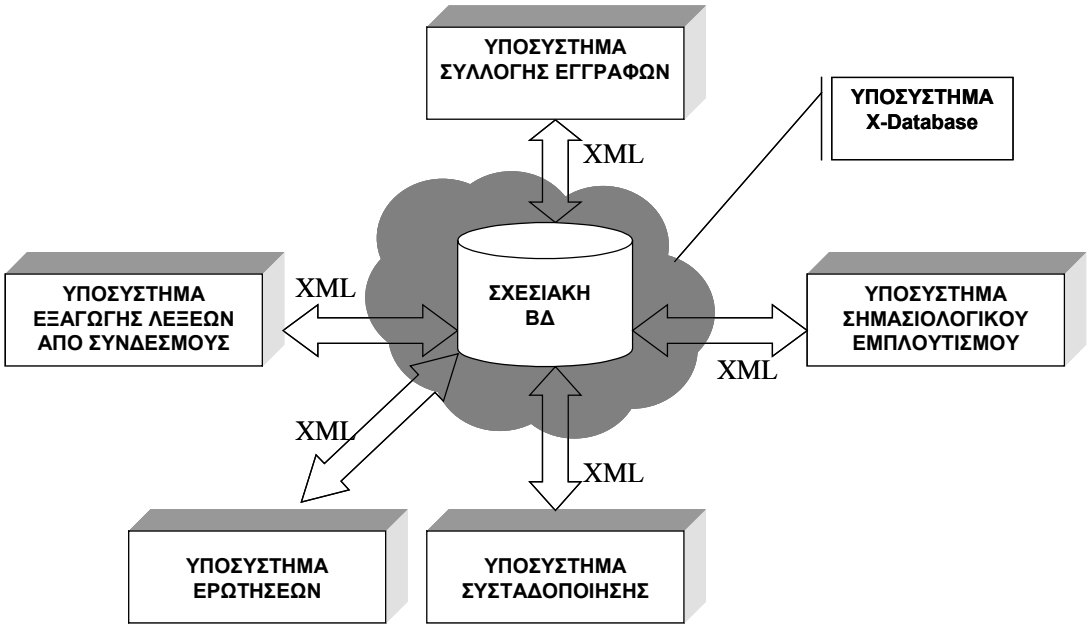
2- Μπορούμε να ταξινομήσουμε σελίδες ως προς τη σημαντικότητά τους με βάση όχι μόνο το πλήθος των εισερχόμενων συνδέσμων τους αλλά και με τη λεξική πληροφορία που αυτοί μεταφέρουν.

3- Ο προσδιορισμός θεματικών π-κόμβων και α-κόμβων φαίνεται να αποτελεί χρήσιμο εργαλείο στις αναζητήσεις στον ΠΙ και η χρήση τους σε σύνθετες ερωτήσεις οδηγεί σε πολύ χρήσιμα συμπεράσματα. Η αξιολόγηση των επιδόσεων του συστήματος στον εντοπισμό τέτοιων κόμβων αποτελεί αντικείμενο μελλοντικής εργασίας.

4- Ο καθορισμός ενός μέτρου ομοιότητας μεταξύ ιστοσελίδων, που βασίζεται στη σημασιολογική πληροφορία που μεταφέρουν οι σύνδεσμοι, μας επιτρέπει να ανακαλύπτουμε ομάδες σελίδων με κοινή σημασιολογία και παρόμοια συνδεσμολογία.

7 Το υποσύστημα X-Database

Το κεφάλαιο αυτό αναφέρεται στο υποσύστημα X-Database, το οποίο αναλαμβάνει την αποθήκευση και διαχείριση των XML εγγράφων που περιέχουν τους λεξικούς και σημασιολογικούς χαρακτηρισμούς των εγγράφων του ιστού και των υπερσυνδέσμων τους, σε μια σχεσιακή βάση δεδομένων. Το σύστημα X-Database λειτουργεί σαν κέλυφος (Σχήμα 23) που περιβάλλει τη σχεσιακή βάση δεδομένων και επιτρέπει στους χρήστες να διαχειρίζονται με διαφάνεια τα περιεχόμενά της. Αν και το σύστημα αναπτύχθηκε αρχικά για να εξυπηρετήσει τους σκοπούς του συστήματος THESUS, στη συνέχεια επεκτάθηκε ώστε να υποστηρίζει ακόμη πιο σύνθετες δομές XML εγγράφων αλλά και πιο σύνθετες λειτουργίες.



Σχήμα 23. Ο ρόλος του X-Database στο σύστημα THESUS

Το σύστημα X-Database ελέγχθηκε συστηματικά με τα XML-Schema αρχεία του μοντέλου μεταπληροφορίας οπτικοακουστικών δεδομένων MPEG-7 [ISOb], [DM01]. Χρησιμοποιήθηκε για την αποθήκευση, ενημέρωση και διαχείριση μεγάλου όγκου αρχείων MPEG-7 σε μια σχεσιακή βάση δεδομένων, την οποία δημιούργησε αυτόματα με βάση τα XML-Schema αρχεία του MPEG-7 μοντέλου πληροφορίας. Μέσα από τη διαδικασία αυτή, αναπτύχθηκε ένα ολοκληρωμένο σύστημα που αναλαμβάνει τη διασύνδεση XML δομών (που καλύπτουν ένα μεγάλο μέρος της γλώσσας XML-Schema) και σχεσιακών βάσεων δεδομένων.

Το σύστημα X-Database εισάγει μια μεθοδολογία για απεικόνιση των δομών της *XML-Schema* στο *σχεσιακό σχήμα*, ενώ ταυτόχρονα επιτρέπει την αυτόματη δημιουργία ενός σχεσιακού σχήματος σε οποιοδήποτε Σχεσιακό Σύστημα ΒΔ από τα XML-Schema αρχεία που περιγράφουν τη δομή των XML εγγράφων. Επιπλέον, αναλαμβάνει την αποθήκευση, διαχείριση και ανάκτηση των XML εγγράφων στη

σχεσιακή βάση. Όλες οι εντολές προς τη βάση γίνονται με τη μορφή XML εγγράφων που ικανοποιούν το ίδιο XML-Schema. Τα αρχεία XML-Schema που διαχειρίζεται το σύστημα περιγράφουν:

- **Τη δομή των XML εγγράφων** που διακινούνται από και προς τη βάση δεδομένων
- **Τη δομή των εντολών διαχείρισης των δεδομένων XML**. Η σύνταξη των εντολών που αφορούν την εισαγωγή, διαχείριση και ερώτηση των XML δεδομένων της βάσης περιγράφεται σε ένα ξεχωριστό αρχείο XML-Schema. Στο ίδιο αρχείο περιγράφονται και οι δομές της βάσης δεδομένων που βελτιστοποιούν τη λειτουργία της, π.χ. ευρετήρια. Το αρχείο αυτό καθορίζει τη διεπαφή του συστήματος X-Database και των υπόλοιπων εφαρμογών.

Οι συνεισφορές του X-Database συνοψίζονται στα εξής:

1. Μια μεθοδολογία για την αυτόματη απεικόνιση αρχείων XML-Schema σε σχεσιακές βάσεις δεδομένων. Τα πλεονεκτήματα της προσέγγισης είναι:
   - ***αυτόματη δημιουργία του σχήματος της βάσης***: το σχεσιακό σχήμα δημιουργείται χωρίς ανθρώπινη παρέμβαση ακολουθώντας συγκεκριμένους κανόνες απεικόνισης,
   - ***αυτόματη απεικόνιση των XML δεδομένων στο σχεσιακό σχήμα***: οι κανόνες απεικόνισης προκύπτουν απευθείας από τα XML-Schema αρχεία που χρησιμοποιούνται για την επικύρωση της δομής των XML εγγράφων,
   - ***έλεγχος εγκυρότητας των δεδομένων της βάσης***: κατά τη δημιουργία του σχεσιακού σχήματος ορίζονται αυτόματα και οι διαδικασίες επικύρωσης της δομής των εγγράφων. Οι διαδικασίες αυτές εγγυώνται την εγκυρότητα των στοιχείων της σχεσιακής βάσης μετά από εισαγωγές, διαγραφές και ενημερώσεις, εγγυώνται δηλαδή ότι τα στοιχεία που απομένουν στη βάση, μετά από μια διαδικασία ενημέρωσης, επαρκούν για να δώσουν έγκυρα (σύμφωνα με το XML-Schema) XML έγγραφα,
   - ***δυνατότητα καθορισμού δομών της βάσης δεδομένων εντός των XML-Schema αρχείων***: ο διαχειριστής της βάσης δεδομένων μπορεί να βελτιστοποιήσει τη λειτουργία της βάσης καθορίζοντας ευρετήρια, απόψεις, ρόλους κτλ. στο XML-Schema αρχείο, κάτι που δεν επιτρέπεται από παρόμοια συστήματα (π.χ. μπορούμε να δηλώσουμε τη δημιουργία ευρετηρίου σε ένα χαρακτηριστικό των XML εγγράφων χωρίς να γνωρίζουμε τον πίνακα στον οποίο έχει αποθηκευθεί).
2. Ένα πλήρως υλοποιημένο σύστημα για την αυτόματη δημιουργία ενός σχεσιακού σχήματος από το XML-Schema την αποθήκευση, ενημέρωση και διαχείριση ερωτήσεων στα XML δεδομένα της βάσης. Το σύστημα:
   - ***παράγει τη σχεσιακή βάση*** αναλύοντας τα αρχεία XML-Schema,
   - ***αποσυνθέτει τα XML έγγραφα*** σε πλειάδες και τα αποθηκεύει στη σχεσιακή βάση, και συνθέτει ***δομημένα και έγκυρα XML έγγραφα*** από τα περιεχόμενα της βάσης,
   - επιτρέπει τη διαχείριση των δεδομένων της βάσης ***αποκλειστικά με χρήση XML αρχείων.*** Είναι ευκολότερο για τους χρήστες του συστήματος να αναφερθούν στα περιεχόμενα της βάσης χρησιμοποιώντας τις δομές και τα ονόματα της XML, παρά αυτά της σχεσιακής βάσης,
   - δημιουργεί στη βάση δεδομένων τις απαραίτητες διαδικασίες που ***εγγυώνται την αξιοπιστία των περιεχομένων της βάσης***. Ένα XML

έγγραφο εισάγεται στη βάση και στη συνέχεια μπορεί να ερωτηθεί. Επιπλέον όμως, μπορεί να ενημερωθεί, π.χ. να αλλάξουν οι τιμές των χαρακτηριστικών του, να προστεθούν ή να αφαιρεθούν περιεχόμενα Στις περιπτώσεις αυτές το σύστημα εγγυάται την εγκυρότητα των ενημερωμένων περιεχομένων. Για παράδειγμα, όταν ένα στοιχείο ενός εγγράφου διαγράφεται, ένας έλεγχος επιτελείται για να εγγυηθεί ότι τα δεδομένα που απομένουν στη βάση και αφορούν το έγγραφο αρκούν για να παράγουν ένα έγκυρο έγγραφο.

7.1 Απεικόνιση σε σχεσιακό μοντέλο

Το σύστημα X-Database αντιμετωπίζει τα συνήθη προβλήματα των συστημάτων απεικόνισης XML δομών σε σχεσιακές βάσεις υιοθετώντας τεχνικές που έχουν προταθεί στη βιβλιογραφία. Παρ' όλα αυτά, προτείνει λύσεις σε όχι και τόσο συνηθισμένα προβλήματα που προκύπτουν από τις ιδιαιτερότητες της XML-Schema, όπως η περιγραφή του σχήματος σε περισσότερα από ένα αρχεία, η κληρονομικότητα μεταξύ τύπων και ο πολυμορφισμός, ο ορισμός απλών και σύνθετων τύπων, ο χειρισμός ενώσεων και λιστών (unions και lists) τύπων. Φροντίζει, επίσης, για τη συνεπή διαχείριση των ερωτήσεων και των ενημερώσεων των δεδομένων που είναι αποθηκευμένα στη βάση εισάγοντας νέες πρακτικές απεικόνισης και τακτικές ελέγχου.

Τα XML-Schema αρχεία περιέχουν πέντε διαφορετικούς τύπους στοιχείων. Χρησιμοποιούμε το πρόθεμα ***xsd:*** για να τους διαχωρίσουμε από αντίστοιχα ονόματα που χρησιμοποιούνται στην XML.

– ***xsd:element***
Καθορίζει το όνομα και τον τύπο κάθε στοιχείου στην XML (XML element).

– ***xsd:attribute***
Καθορίζει το όνομα και τον τύπο ενός χαρακτηριστικού (XML attribute) κάποιου στοιχείου της XML (XML element). Μπορεί να περιέχει ένα χαρακτηριστικό ***use*** με τιμή ***"required"***, που δηλώνει ότι το συγκεκριμένο χαρακτηριστικό είναι υποχρεωτικό για το στοιχείο που ορίζεται. Μπορεί να είναι απλού τύπου (integer, float, string, κτλ.) ή σύνθετου (π.χ. απαρίθμηση, εύρος τιμών κτλ.).

– ***xsd:simpleType***
Ορίζει ένα νέο τύπο που μπορεί να χρησιμοποιηθεί για ένα ή περισσότερα χαρακτηριστικά XML.

– ***xsd:attributeGroup***
Ομαδοποιεί τους ορισμούς χαρακτηριστικών της XML που χρησιμοποιούνται σε πολλά XML στοιχεία. Για παράδειγμα, τα χαρακτηριστικά *οδός*, *αριθμός*, *πόλη*, *ταχ. κώδικας* μπορεί να εμφανίζονται σε περισσότερα από ένα στοιχεία, π.χ. σε έναν *άνθρωπο*, σε μια *εταιρία* κτλ. Για να αποφύγουμε επαναλήψεις ορισμών, ορίζουμε ένα attributeGroup που το ονομάζουμε *διεύθυνση* και τα περιέχει.

– ***xsd:complexType***
Αντιπροσωπεύουν τις διάφορες σύνθετες οντότητες του μοντέλου δεδομένων. Οι σύνθετοι τύποι αποτελούνται από ορισμούς xsd:attributes, και ακολουθίες

xsd:elements. Μπορούν να οριστούν επώνυμα-καθολικά και να χρησιμοποιηθούν ως τύποι πολλών στοιχείων XML. Μπορούν εναλλακτικά να ορίζονται ανώνυμα-τοπικά κατά τη δήλωση του τύπου ενός στοιχείου XML. Περιέχουν:

- μία ή περισσότερες ετικέτες ***<xsd:attribute>***
- μία ή περισσότερες ετικέτες ***<xsd:element>*** που περιγράφουν τα υπο-στοιχεία του σύνθετου XML στοιχείου.

Κάθε υπο-στοιχείο έχει ένα χαρακτηριστικό ***"type"*** ή ***"ref"*** που καθορίζει τον τύπο του ή το στοιχείο XML στο οποίο αναφέρεται.

Τα ζητήματα τα οποία αντιμετωπίζονται από το X-Database κατά την σύνδεση XML και σχεσιακών βάσεων δεδομένων είναι:

1. Ολοκλήρωση περισσοτέρων του ενός αρχείων XML-Schema που χρησιμοποιούν χώρους ονομάτων και συνθέτουν το συνολικό XML-Schema μιας εφαρμογής.
2. Απεικόνιση βασικών τύπων της XML-Schema στις αντίστοιχες σχεσιακές δομές.
3. Απεικόνιση του εμφωλιασμού των στοιχείων, λαμβάνοντας υπόψη τη σειρά εμφάνισης των υπο-στοιχείων και το μέγιστο και ελάχιστο αριθμό εμφανίσεων.
4. Διαχείριση της κληρονομικότητας απλών και σύνθετων τύπων στην XML-Schema.
5. Διαχείριση του πολυμορφισμού που μπορεί να προκύψει στα XML έγγραφα με τη χρήση του χαρακτηριστικού xsi:type [W3C00b].

Τα παραπάνω ζητήματα αναλύονται στη συνέχεια.

7.1.1 Ολοκλήρωση πολλών αρχείων XML-Schema

Στην περίπτωση του σχήματος των εγγράφων του THESUS είναι πολύ εύκολο να περιγράψει κανείς τη δομή τους σε ένα μόνο XML-Schema αρχείο. Στην περίπτωση μεγαλύτερων σχημάτων όμως, όπως για παράδειγμα το MPEG-7, η δομή των XML εγγράφων περιγράφεται σε πέντε διαφορετικά XML-Schema αρχεία, ενώ ένα επιπλέον αρχείο περιγράφει τη σύνταξη των εντολών προς τη βάση δεδομένων. Είναι, επίσης, πολύ συχνό το φαινόμενο, να χρησιμοποιούνται στα αρχεία του XML-Schema περισσότεροι από ένας χώροι ονομάτων (namespaces), με αποτέλεσμα στοιχεία ή χαρακτηριστικά με ίδια ονόματα, αλλά διαφορετικούς χώρους, να συμβολίζουν διαφορετικά πράγματα και να έχουν διαφορετική δομή.

Το X-Database συγχωνεύει τα διαφορετικά αρχεία λαμβάνοντας υπόψη τους διαφορετικούς χώρους ονομάτων ώστε να αποφύγει συγκρούσεις ονομάτων. Στις περιπτώσεις αυτές, η απεικόνιση μεταξύ ονομάτων της XML-Schema και της σχεσιακής βάσης αποθηκεύεται στην κύρια μνήμη, και περιέχει το πλήρες όνομα (όνομα και πρόθεμα χώρου ονομάτων) για κάθε τύπο, στοιχείο ή χαρακτηριστικό και το όνομα του αντίστοιχου πεδίου ή σχέσης στη βάση.

7.1.2 Απεικόνιση των βασικών στοιχείων του XML-Schema

Βασικοί τύποι: Οι βασικοί τύποι της XML-Schema απεικονίζονται στους αντίστοιχους τύπους δεδομένων της ANSI SQL. Αριθμητικοί, αλφαριθμητικοί και τύποι ημερομηνίας απεικονίζονται σε NUMERIC (precision, scale), VARCHAR (maxsize) και DATE τύπους αντίστοιχα. Οι τύποι ID, IDREF, IDREFS, ENTITIES, NMTOKEN, NMTOKENS, NOTATION, που χρησιμοποιούνται για αναφορές εντός ενός XML εγγράφου, απεικονίζονται σε τύπους VARCHAR καθώς αντιστοιχούν σε

συμβολοσειρές. Ανάλογα με τη βάση δεδομένων ο τύπος BINARY της XML-Schema απεικονίζεται στον αντίστοιχο τύπο της βάσης δεδομένων (π.χ. στον τύπο BLOB στην Oracle). Η απεικόνιση αυτή είναι γενική και στερείται της ακρίβειας με την οποία ορίζονται οι τύποι στο σχεσιακό μοντέλο. Για παράδειγμα, το μέγιστο μέγεθος ενός τύπου VARCHAR ή BLOB ή η ακρίβεια και η κλίμακα ενός αριθμητικού τύπου δεν καθορίζονται επακριβώς στην XML-Schema. Εναλλακτικά η δήλωση των βασικών τύπων μπορεί να αντικατασταθεί από τους απλούς τύπους (simpleTypes).

simpleTypes: Στην XML-Schema ένας νέος τύπος ορίζεται στη βάση ενός υπάρχοντος τύπου καθορίζοντας ένα ή περισσότερα χαρακτηριστικά, όπως μέγεθος, ελάχιστη-μέγιστη τιμή κτλ. Ο ορισμός ενός απλού τύπου στην XML-Schema παράγει ένα περιορισμό στο αντίστοιχο πεδίο στη βάση. Για παράδειγμα, ο ορισμός του μέγιστου μήκους (*length*) για τον απλό τύπο 'DirectorNameType' παράγει ένα περιορισμό στο πεδίο στο οποίο αντιστοιχίζεται (Πίνακας 9).

XML-Schema	SQL
<simpleType name="DirectorNameType"> <restriction base="string"> <maxLength value="8"/> </restriction> </simpleType>	DirectorNameType VARCHAR(8)

Πίνακας 9. Απεικόνιση του χαρακτηριστικού length

Οι ορισμοί των απλών τύπων πολύ συχνά περιέχουν ετικέτες *union* και *list* που συνδυάζουν δύο ή περισσότερα πρότυπα βασικών τύπων. Στην περίπτωση της ένωσης (***union***) δύο τύπων ο γενικότερος από τους δύο, σύμφωνα με μια ιεραρχία τύπων, υπερισχύει. Όταν ένα χαρακτηριστικό έχει τύπο που ορίζει λίστα τιμών (***list***), τότε σε επίπεδο XML εγγράφου το συγκεκριμένο χαρακτηριστικό έχει ως τιμή μια λίστα τιμών που χωρίζονται μεταξύ τους με κενά. Το συγκεκριμένο χαρακτηριστικό απεικονίζεται σε ένα ξεχωριστό πίνακα στο σχεσιακό μοντέλο με δύο πεδία, ένα που περιέχει το id του στοιχείου και ένα που περιέχει τις τιμές του χαρακτηριστικού (Πίνακας 10)

attribute, attributeGroup: Αυτά απεικονίζονται σε ένα πεδίο ή ένα σύνολο πεδίων μιας σχέσης στη βάση. Τα χαρακτηριστικά που δηλώνονται ως απαιτούμενα αντιστοιχούν σε περιορισμούς NOT NULL στα αντίστοιχα πεδία.

complexType: Κάθε σύνθετος τύπος στην XML-Schema απεικονίζεται σε μια ξεχωριστή σχέση στη βάση (βλ. Πίνακας 11). Τα στοιχεία του ίδιου σύνθετου τύπου, αν και στην XML μπορεί να έχουν διαφορετικά ονόματα, μοιράζονται την ίδια δομή. Γι' αυτό και αποθηκεύονται στον ίδιο πίνακα. Η προσέγγιση αυτή μειώνει τον συνολικό αριθμό σχέσεων που παράγονται, δίνοντας έτσι λιγότερες σχέσεις από τις κλασσικές προσεγγίσεις binary και edge approach [FK99].

XML-Schema	<element name="root" type=" unionDemoType"/> <complexType name="unionDemoType"> <attribute name=" unionfield " > <simpleType> <union memberTypes="Type1 Type2"/> </simpleType> … </complexType> <simpleType name="Type1"> <restriction base="NMTOKEN"> <enumeration value="noun"/> <enumeration value="pronoun"/> </restriction> </simpleType> <simpleType name="Type2"> <restriction base="NMTOKEN"> <enumeration value="verb"/> <enumeration value="adverb"/> </restriction> </simpleType>
SQL	create table unionDemoType (unionDemoType_id NUMBER(10), unionfield VARCHAR(50) check (unionfield in ('noun', 'pronoun','verb', 'adverb'))

Πίνακας 10. Χρήση των union σε ένα simpleType

XML-Schema	SQL
<complexType name="DigitalStorageDS"> <attribute name="id" type="ID" use="required"/> <attribute name="annotate" type="string"/> </complexType>	CREATE TABLE DigitalStorageDS (id NUMBER NOT NULL, annotate VARCHAR(50));

Πίνακας 11. Απεικόνιση complexType

element: Ένα "στοιχείο" στην XML-Schema περιέχει είτε απλό περιεχόμενο (απλού ή βασικού τύπου) είτε σύνθετο περιεχόμενο (ένα σύνολο από υπο-στοιχεία και χαρακτηριστικά). Στην πρώτη περίπτωση το στοιχείο απεικονίζεται είτε σε ένα πεδίο με τους αντίστοιχους περιορισμούς που ορίζει ο τύπος, είτε σε μια νέα σχέση, ανάλογα με τον πιθανό αριθμό εμφανίσεων στοιχείων του ίδιου τύπου κάτω από το ίδιο εξωτερικό στοιχείο. Στη δεύτερη περίπτωση το στοιχείο απεικονίζεται σε μια νέα σχέση.

Στην XML-Schema οι σύνθετοι τύποι είναι ανώνυμοι ή επώνυμοι. Και στις δύο περιπτώσεις δημιουργούνται σχέσεις στη βάση. Τα ονόματα των σχέσεων αυτών παράγονται δυναμικά από το XML-Schema και η αντιστοίχισή τους αποθηκεύεται στη μνήμη.

7.1.3 Απεικόνιση του εμφωλιασμού των στοιχείων

Ο εμφωλιασμός των στοιχείων στην XML εκφράζεται ως μια σχέση "περιέχει" μεταξύ σύνθετων τύπων στην XML-Schema. Ένας σύνθετος τύπος μπορεί να περιέχει

ένα ή περισσότερα στοιχεία άλλων τύπων. Κατά την απεικόνιση των σύνθετων τύπων στο σχεσιακό σχήμα δημιουργούνται διάφοροι περιορισμοί και διαδικασίες ελέγχου που εγγυώνται την εγκυρότητα των αναφορών στη βάση.

Εμφωλιασμός

Όπως προαναφέρθηκε, τα φωλιασμένα στοιχεία XML αποθηκεύονται σε ξεχωριστές σχέσεις. Για να εγγυηθούμε την εγκυρότητα των αναφορών, δημιουργούμε ένα ξένο κλειδί στη σχέση που αντιστοιχεί στο εσωτερικό στοιχείο. Το ξένο κλειδί αναφέρεται στο πρωτεύον κλειδί της σχέσης που αντιστοιχεί στο εξωτερικό στοιχείο. Αυτό σημαίνει για την XML ότι το εξωτερικό στοιχείο πρέπει να έχει ένα χαρακτηριστικό με μοναδικές τιμές, που να μπορεί να χρησιμοποιηθεί ως πρωτεύον κλειδί της σχέσης. Για το λόγο αυτό το σύστημα X-Database αναθέτει ένα επιπλέον χαρακτηριστικό το *DB_ID* σε κάθε στοιχείο και παράγει ένα μοναδικό κλειδί γι' αυτό. Το επιπλέον χαρακτηριστικό υπάρχει στη βάση αλλά δεν εμφανίζεται στο XML έγγραφο παρά μόνο αν ζητηθεί.

Όταν ένα στοιχείο στην XML είναι βασικού τύπου και εμφανίζεται μοναδικά εντός του στοιχείου που το περιέχει (στοιχείο `SourceURI` Πίνακας 12), τότε ακολουθείται το μοντέλο εσωτερικής απεικόνισης (inlining) [FK99], και το στοιχείο το διαχειριζόμαστε σαν χαρακτηριστικό. Αυτό σημαίνει ότι στη σχέση που περιέχει τα χαρακτηριστικά του σύνθετου τύπου προστίθεται ένα επιπλέον πεδίο. Το ίδιο ισχύει για όλους τους τύπους που κληρονομούν αυτό το βασικό τύπο.

Παρ' όλα αυτά, όταν το ίδιο υπο-στοιχείο εμφανίζεται πάνω από μια φορά στο ίδιο στοιχείο που το περιέχει (στοιχείο `TargetURI` Πίνακας 12), τότε το μοντέλο εσωτερικής απεικόνισης δεν είναι εφαρμόσιμο, καθώς οδηγεί σε επαναλαμβανόμενα πεδία, στη σχέση που αντιστοιχεί στον τύπο του εξωτερικού στοιχείου, παραβιάζοντας έτσι την πρώτη κανονική μορφή. Στην περίπτωση αυτή δημιουργείται μια νέα σχέση που περιέχει ένα πεδίο για τις τιμές του συγκεκριμένου χαρακτηριστικού και ένα πεδίο για το id του εξωτερικού στοιχείου.

XML-Schema

```
<complexType name="SourceDocument">
 <sequence>
  <element name="SourceURI" type="string"/>
  <element   name="TargetURI"   type="string"   minOccurs="0"
maxOccurs="unbounded"/>
 </sequence>
</complexType>
```

SQL

```
CREATE TABLE SourceDocument (
 SourceDocument_ID NUMBER(10),
 SourceURI VARCHAR2(500),
 PRIMARY KEY (SourceDocument_ID));
CREATE TABLE TargetURI (
 TargetURI_Text VARCHAR2(500),
 SourceDocument_ID NUMBER(10),
 FOREIGN KEY (SourceDocument_ID)
  REFERENCES SourceDocument(SourceDocument_ID)
   ON DELETE CASCADE);
```

Πίνακας 12. Απεικόνιση φωλιασμένων στοιχείων

Αποθήκευση της σειράς εμφάνισης των υποστοιχείων

Τα στοιχεία από τα οποία απαρτίζεται ένας σύνθετος τύπος ορίζονται μέσα σε ετικέτες *sequence* ή *choice,* που δηλώνουν ότι τα στοιχεία θα εμφανίζονται με μια συγκεκριμένη σειρά (sequence), ή ότι κάποιο από όλα τα ορισμένα στοιχεία θα εμφανίζεται κάθε φορά (choice). Τα στοιχεία μπορεί να εμφανίζονται καμία ή περισσότερες φορές (0..*unbounded*) με συγκεκριμένη σειρά ανάλογα με τις τιμές των χαρακτηριστικών ***minOccurs*** και ***maxOccurs*** που ορίζονται για κάθε στοιχείο στην XML-Schema. Παρόμοια, μια σειρά ή μια επιλογή στοιχείων μπορεί να εμφανίζεται καμία ή περισσότερες φορές σε ένα σύνθετο τύπο, περιπλέκοντας τα πράγματα ακόμη περισσότερο. Η σειρά με την οποία τα υπο-στοιχεία ενός στοιχείου εμφανίζονται είναι σημαντικό να αποθηκευθεί για την ορθή αναπαραγωγή του XML εγγράφου [TI+01]. Για το λόγο αυτό η σειρά των στοιχείων εντός ενός στοιχείου στην XML αποθηκεύεται στη βάση σε ένα ξεχωριστό πεδίο στην κάθε σχέση που αντιστοιχεί σε κάποιον από τους τύπους των εσωτερικών στοιχείων.

7.1.4 Κληρονομικότητα τύπων

Ένα εξίσου ενδιαφέρον πρόβλημα είναι η διαχείριση της κληρονομικότητας τύπων της XML-Schema και πώς αυτό επηρεάζει τη δημιουργία της σχεσιακής βάσης Κληρονομικότητα μπορεί να οριστεί για απλούς και σύνθετους τύπους, ενώ οι παραγόμενοι τύποι μπορούν είτε να επεκτείνουν, είτε να περιορίζουν τα χαρακτηριστικά που κληρονομούν από τους βασικούς.

Στην περίπτωση των ***simpleTypes*** η κληρονομικότητα γίνεται μόνο με περιορισμό (***by restriction)*** του τύπου και χρησιμοποιείται για να προσθέτουμε περιορισμούς στο βασικό τύπο. Για παράδειγμα, ο απλός τύπος 'Α' μπορεί να κληρονομεί το βασικό τύπο 'string' και να τον περιορίζει σε ένα συγκεκριμένο σύνολο τιμών. Αυτή η κληρονομικότητα συνεπάγεται τη δημιουργία ενός περιορισμού, στο πεδίο που αντιστοιχεί στον παραγόμενο τύπο.

Στην περίπτωση των ***complexTypes*** έχουμε κληρονομικότητα: α) είτε με περιορισμό (***by restriction***) μιας βασικής δομής σε μια απλούστερη, που περιέχει μόνο τα στοιχεία και τα χαρακτηριστικά που ορίζονται στον παραγόμενο τύπο, β) είτε με επέκταση (***by extension***) της βασικής δομής σε μια πιο σύνθετη δομή, η οποία περιέχει όλα τα υπο-στοιχεία και χαρακτηριστικά που ορίζει ο βασικός τύπος και επιπλέον τα χαρακτηριστικά που ορίζει ο παράγωγος τύπος.

Στην κληρονομικότητα με περιορισμό ο τύπος απεικονίζεται στη βάση σαν να ήταν ένας κανονικός complexType, αγνοώντας τη δομή του βασικού τύπου. Στην κληρονομικότητα με επέκταση, ο παραγόμενος τύπος απεικονίζεται στη βάση σαν να ήταν ένας νέος complexType με όλα τα στοιχεία και χαρακτηριστικά που ορίζονται από αυτό και από το βασικό τύπο. Το παράδειγμα που ακολουθεί (Πίνακας 13) παρουσιάζει την υλοποίηση της κληρονομικότητας σύνθετων τύπων στο X-Database. Στο παράδειγμα αυτό και τα δύο στοιχεία είναι βασικού τύπου με απεριόριστο αριθμό εμφανίσεων (maxOccurs="unbounded") γι' αυτό και δημιουργείται ένας πίνακας για το καθένα από αυτά.

XML-Schema	SQL
<complexType name="parentType"> <sequence> <element name="parent_content" type="string" maxOccurs="unbounded"/> </sequence> </complexType> <complexType name="childType"> <complexContent> <extension base="parentType"> <sequence> <element name="child_content" type="string" maxOccurs="unbounded"/> </sequence> </extension> </complexContent> </complexType>	create table parentType(parentType_id NUMBER(10), PRIMARY KEY (parentType_id)) create table childType(childType_id NUMBER(10), PRIMARY KEY (childType_id)) create table parent_contentText(text VARCHAR2(200), seq NUMBER(10), p_id NUMBER(10), CONSTRAINT p_id references parentType(parentType_id)) create table child_contentText(text VARCHAR2(200), seq NUMBER(10), p_id NUMBER(10), c_id NUMBER(10), CONSTRAINT p_id references parentType(parentType_id), CONSTRAINT c_id references childType(childType_id))

Πίνακας 13. Παράδειγμα κληρονομικότητας σύνθετων τύπων

7.1.5 Πολυμορφισμός στοιχείων στην XML

Η XML υλοποιεί τον πολυμορφισμό, εκμεταλλευόμενη την κληρονομικότητα σύνθετων τύπων της XML-Schema. Ένας σύνθετος τύπος στην XML-Schema μπορεί να αποτελέσει τη βάση για έναν παράγωγο σύνθετο τύπο, όπως περιγράφηκε προηγούμενα. Ένα στοιχείο μπορεί να δηλωθεί στην XML-Schema ότι είναι τύπου complexType *X*, και παρ' όλα αυτά, το ίδιο στοιχείο στην XML μπορεί να εμφανιστεί με τη δομή που περιγράφει οποιοσδήποτε παράγωγος τύπος του *X*. Αυτό γίνεται με τη χρήση του χαρακτηριστικού ***xsi:type*** με το οποίο δηλώνεται ο συγκεκριμένος παράγωγος τύπος που ακολουθεί.

Το κυριότερο πρόβλημα που προκύπτει από τη χρήση του xsi:type είναι ότι το χαρακτηριστικό *type* στον ορισμό ενός στοιχείου στην XML-Schema δεν καθορίζει αποκλειστικά τη δομή του στοιχείου αυτού στην XML. Το στοιχείο μπορεί να έχει τη δομή που περιγράφεται σε οποιοδήποτε σύνθετο τύπο, παράγωγο αυτού που δηλώνεται στην XML-Schema. Ένα παράδειγμα παρουσιάζεται στον Πίνακα 14. Η δήλωση XPath "`/content/child/beta`" είναι πιθανή, ακόμη και αν σύμφωνα με το XML-Schema το στοιχείο "`child`" δεν περιέχει στοιχείο "`beta`".

Το χαρακτηριστικό αυτό επηρεάζει τις εισαγωγές, ενημερώσεις, διαγραφές και επιλογές δεδομένων από τη βάση, όπως θα εξηγήσουμε στη συνέχεια.

Κατά την ανάγνωση του XML εγγράφου, το όνομα κάθε στοιχείου συνδέεται με τον αντίστοιχο σύνθετο τύπο που ορίστηκε στην XML-Schema και συνεπώς με την αντίστοιχη σχέση στη βάση. Όταν αποθηκεύουμε ένα στοιχείο που περιέχει το χαρακτηριστικό *xsi:type*, η τιμή του χαρακτηριστικού καθορίζει το σύνθετο τύπο και

τη σχέση που συνδέεται με το συγκεκριμένο στοιχείο. Ως αποτέλεσμα, τα περιεχόμενα του XML στοιχείου μπορεί να αποθηκευθούν σε οποιαδήποτε από τις σχέσεις που αντιστοιχούν στο σύνθετο τύπο του στοιχείου, ή σε οποιοδήποτε παραγόμενο τύπο. Κατά συνέπεια και όλες οι διεργασίες διαχείρισης δεδομένων (ενημερώσεις, διαγραφές, επιλογές) εμπλέκουν όλους τους πιθανούς πίνακες που περιέχουν δεδομένα του στοιχείου αυτού.

XML-Schema	XML
<!--Base type--> <complexType name="baseType"> <sequence> <element name="alpha" type="typeA"/> </sequence> </complexType> <!--Derived type --> <complexType name="derivedType"> <complexContent> <extension base="baseType"> <sequence> <element name="beta" type="typeB"/> </sequence> </extension> </complexContent> </complexType> <!—Element of base type--> <element name="content"> <complexType> <sequence> <element name="child" type="baseType"/> </sequence> </complexType> </element>	<content> <child xsi: type="dType"> <alpha ...> ... </alpha> <beta ...> ... </beta> </child1> </content>

Πίνακας 14. Παράδειγμα χρήσης του χαρακτηριστικού xsi:type

Η υλοποίηση των διεργασιών αυτών απαιτεί την αποσαφήνιση μονοπατιών που περιλαμβάνουν στοιχεία όπως το προηγούμενο. Διακρίνουμε δύο προσεγγίσεις, που αφορούν την αποσαφήνιση των εκφράσεων μονοπατιών που περιέχουν στοιχεία με πολυμορφισμό:

Η πρώτη προσέγγιση είναι να ***παράγουμε όλους τους δυνατούς συνδυασμούς των παράγωγων complexTypes για κάθε στοιχείο στο μονοπάτι*** και να κάνουμε την ίδια διεργασία (εισαγωγή, διαγραφή, επιλογή) για κάθε διαφορετικό μονοπάτι. Η προσέγγιση αυτή βασίζεται σε πληροφορία που έχουμε από το XML-Schema και είναι προτιμότερη για σχήματα με περιορισμένη κληρονομικότητα τύπων. Ένα μειονέκτημα της προσέγγισης αυτής είναι ότι μπορεί να παραχθούν μονοπάτια που δεν αντιστοιχούν σε πίνακες ή πεδία της βάσης και να οδηγήσουν σε λανθασμένα ερωτήματα προς τη σχεσιακή βάση.

Η δεύτερη προσέγγιση είναι να ***αποθηκεύσουμε πληροφορία στη βάση σχετικά με τα στοιχεία που εισήχθησαν με χρήση του χαρακτηριστικού xsi:type***. Η πληροφορία

αυτή καθορίζει τους πίνακες που πραγματικά περιέχουν τα δεδομένα και κατά συνέπεια περιορίζει τον αριθμό των πινάκων που θα ελέγχονται σε κάθε διεργασία Το σύστημα X-Database υλοποιεί ένα μηχανισμό για την αποσαφήνιση των εκφράσεων μονοπατιών ακολουθώντας τη δεύτερη προσέγγιση.

7.2 Οι επιτρεπόμενες λειτουργίες

Οι εφαρμογές επικοινωνούν με τη βάση δεδομένων αποκλειστικά με XML έγγραφα. Τα έγγραφα περιγράφουν τις ενέργειες που θα γίνουν στη βάση δεδομένων και προσδιορίζουν τις παραμέτρους και τις τιμές για κάθε ενέργεια. Εκτός από τους σύνθετους τύπους (complexTypes), που αντιστοιχούν στις οντότητες του μοντέλου πληροφορίας, ένα επιπλέον αρχείο XML-Schema ορίζει τις διαθέσιμες λειτουργίες διαχείρισης της πληροφορίας. Το αρχείο αυτό περιέχει τους σύνθετους τύπους που προσδιορίζουν τη δομή των εντολών προς τη βάση. Με τον τρόπο αυτό, περιγράφεται με σαφήνεια ο τρόπος με τον οποίο μπορεί να εισαχθεί ή να τροποποιηθεί η πληροφορία που υπάρχει στη βάση. Καθορίζονται επίσης με ακρίβεια τα στοιχεία που μπορούν να εισαχθούν ή να τροποποιηθούν, καθώς αυτά ορίζονται ως υπο-στοιχεία των αντίστοιχων στοιχείων εντολών.

Δημιουργώντας διαφορετικά αρχεία εντολών, για κάθε εφαρμογή, χρήστη ή ομάδα χρηστών, μπορούμε έμμεσα να ορίσουμε τα δικαιώματα των χρηστών στα δεδομένα της βάσης. Οι ρόλοι και τα αντίστοιχα δικαιώματα δεν προσδιορίζονται στο επίπεδο της Βάσης Δεδομένων, αλλά στο επίπεδο της επικύρωσης των εντολών προς τη βάση. Καθώς κάθε έγγραφο XML που περιέχει μια εντολή προς τη βάση, επικυρώνεται με βάση το αρχείο XML-Schema των εντολών, αν έχουμε διαφορετικά αρχεία επικύρωσης για κάθε ομάδα χρηστών, καθορίζουμε και διαφορετικούς τύπους πρόσβασης στα αποθηκευμένα δεδομένα.

Το αρχείο XML-Schema (Παράρτημα Δ) ορίζει τους εξής σύνθετους τύπους:

- ***DBInsert*** που περιέχει τα στοιχεία που θα εισαχθούν και τα υποστοιχεία τους και μια δήλωση XPATH [W3h] που προσδιορίζει το στοιχεία κάτω από το οποίο θα εισαχθούν.
- ***DBUpdate*** που περιέχει ένα η περισσότερα στοιχεία χωρίς τα υποστοιχεία τους και μια δήλωση XPATH που προσδιορίζει τη θέση των στοιχείων αυτών στο έγγραφο. Η ενημέρωση μέσω της εντολής DBUpdate αφορά μόνο τις τιμές των χαρακτηριστικών ενός στοιχείου. Η ενημέρωση των υποστοιχείων ενός στοιχείου (εισαγωγή-διαγραφή) γίνεται μέσω των αντίστοιχων εντολών (DBInsert και DBDelete).
- ***DBDelete*** που περιέχει αναφορές στα στοιχεία που πρέπει να διαγραφούν.
- ***DBSelect***. Το στοιχείο αυτό έχει τρία χαρακτηριστικά: α) το χαρακτηριστικό ”where” που περιέχει τα κριτήρια του ερωτήματος, χρησιμοποιώντας δηλώσεις XPATH για να προσδιορίσει τα στοιχεία ή χαρακτηριστικά στα οποία αναφέρονται τα κριτήρια, β) το χαρακτηριστικό ”return” που περιέχει μια δήλωση XPATH που καθορίζει το στοιχείο που θέλουμε να επιστραφεί και γ) το χαρακτηριστικό “bind” που περιέχει μια δήλωση XPATH που προσδιορίζει ένα στοιχείο της XML. Η χρήση του χαρακτηριστικού “bind” αναλύεται στη συνέχεια, στην ενότητα 7.2.2.
- ***DBIndex*** που καθορίζει α) το όνομα του νέου ευρετηρίου και β) τα στοιχεία και χαρακτηριστικά των XML εγγράφων στα οποία θα δημιουργηθεί το ευρετήριο.

- ***DBCommand*** που περιέχει ένα ή περισσότερα στοιχεία τύπου DBInsert, DBUpdate, DBDelete, DBSelect or DBIndex
- ***DBReply*** που περικλείει τα ανακτώμενα στοιχεία XML.

Οι παραπάνω τύποι χρησιμοποιούνται για τη μεταφορά δεδομένων από και προς τη βάση, ενώ παράλληλα καθορίζουν τις δυνατότητες κάθε εφαρμογής. Ο διαχειριστής της βάσης μπορεί να καθορίσει μέσα από αυτό το XML-Schema έγγραφο: α) τις επιτρεπόμενες λειτουργίες για κάθε εφαρμογή, β) τα στοιχεία και τους σύνθετους τύπους στους οποίος έχει πρόσβαση κάθε εφαρμογή. Για παράδειγμα, για χρήστες που ασχολούνται μόνο με την εισαγωγή στοιχείων, χρησιμοποιείται ένα XML-Schema έγγραφο που περιέχει μόνο στοιχεία τύπου DBInsert και DBReply, ενώ για τους χρήστες που θέλουν μόνο να υποβάλουν ερωτήσεις δημιουργείται ένα αρχείο που περιέχει στοιχεία τύπου DBSelect και DBReply.

7.2.1 Ενημέρωση των δεδομένων

Αρκετά από τα υπάρχοντα συστήματα βάσεων δεδομένων επιτρέπουν τη μαζική εισαγωγή XML εγγράφων στη βάση και μάλιστα σε σύντομο χρόνο, δεν υποστηρίζουν όμως την ενημέρωση των αποθηκευμένων δεδομένων XML. Το σύστημα X-Database παρέχει ένα ευέλικτο μηχανισμό διαχείρισης των αποθηκευμένων δεδομένων που συνδυάζει διεργασίες εισαγωγής, ενημέρωσης και διαγραφής και επιτρέπει την τροποποίησή τους, ενώ ταυτόχρονα εγγυάται την ακεραιότητα των περιεχομένων της βάσης.

Όταν ένα XML στοιχείο εισάγεται στη βάση, το στοιχείο και όλα τα υποστοιχεία του αποθηκεύονται στους αντίστοιχους πίνακες. Το προσδιοριστικό (id) που παράγεται από τη βάση για κάθε στοιχείο επιστρέφεται στην απάντηση εντός του στοιχείου DBReply ώστε να μπορούμε στη συνέχεια να το χρησιμοποιήσουμε σε ενημερώσεις.

Οι ενημερώσεις που επιτρέπονται καλύπτουν όλες τις λειτουργίες διαχείρισης που περιγράφονται στα [TI+01] και [LM00], όπως διαγραφή, εισαγωγή, αντικατάσταση των υποστοιχείων σε οποιαδήποτε σειρά. Επιτρέπει επίσης ενημέρωση των περιεχομένων των υποστοιχείων ενός στοιχείου εφόσον αυτή δεν παραβιάζει την εγκυρότητα του τελικού XML εγγράφου. Το σύστημα δεν υποστηρίζει λειτουργίες μετασχηματισμού των XML εγγράφων (όπως περιγράφονται από την γλώσσα ερωτήσεων XML-QL [DF+98]), καθώς απευθύνεται σε περιπτώσεις που η δομή των εγγράφων είναι δεδομένη.

Η ενημέρωση των περιεχομένων της βάσης μπορεί να γίνει με δύο τρόπους. Ο πρώτος είναι με εισαγωγή ή διαγραφή ενός ολόκληρου XML εγγράφου, ενώ ο δεύτερος με εισαγωγή, διαγραφή ή τροποποίηση συγκεκριμένων στοιχείων ενός εγγράφου που έχει ήδη αποθηκευθεί.

Εισαγωγή Εγγράφου

Στην περίπτωση που ένα XML έγγραφο εισάγεται στη βάση, η δομή του επικυρώνεται με βάση το XML-Schema, ώστε να ελεγχθεί ότι τα χαρακτηριστικά και οι τιμές τους είναι σωστά, ο εμφωλιασμός των στοιχείων και των υποστοιχείων τους είναι επιτρεπτός, κλπ. Το σύστημα επεξεργάζεται τα στοιχεία και τα χαρακτηριστικά του εγγράφου, εξάγει τα δεδομένα του και τα αποθηκεύει στη βάση ενώ ταυτόχρονα επιστρέφει τα διακριτικά των στοιχείων (ids).

Διαγραφή Εγγράφου
Ένα ολόκληρο XML έγγραφο μπορεί να διαγραφεί από τη βάση αν δώσουμε στο σύστημα το διακριτικό του κεντρικού του στοιχείου (root element). Μια σειρά από αλυσιδωτές διαγραφές αφαιρεί όλα τα περιεχόμενα του εγγράφου από τους αντίστοιχους πίνακες στη βάση.

Ενημέρωση εγγράφου
Όταν η ενημέρωση ενός εγγράφου απαιτεί τη διαγραφή, εισαγωγή ή ενημέρωση ορισμένων από τα στοιχεία του, τότε η διαδικασία περιπλέκεται. Ένας απλός έλεγχος ορθότητας του στοιχείου που έρχεται ως είσοδος δεν αρκεί για να εγγυηθεί την ακεραιότητα των στοιχείων της βάσης μετά την ενημέρωση. Οι ενημερώσεις ενός στοιχείου XML, που περιλαμβάνουν εισαγωγές και διαγραφές υποστοιχείων, επηρεάζουν τη δομή του εγγράφου. Κατά συνέπεια επηρεάζουν και την εγκυρότητα, ως προς το XML-Schema, του εγγράφου που απομένει στη βάση μετά την ενημέρωση. Για παράδειγμα, η διαγραφή ενός υποστοιχείου που είναι υποχρεωτικό παραβιάζει τους κανόνες εγκυρότητας που περιγράφει το XML-Schema, καθώς η δομή του στοιχείου μετά τη διαγραφή δεν είναι σύμφωνη με το σχήμα.

Για να επιτρέψουμε την ενημέρωση σε οποιοδήποτε XML στοιχείο και ταυτόχρονα να εγγυηθούμε ότι τα ενημερωμένα περιεχόμενα του στοιχείου συμβαδίζουν με το XML-Schema, με αυτόματο τρόπο δημιουργούμε διαδικασίες ελέγχου και σκανδάλες (triggers). Οι μηχανισμοί αυτοί ενεργοποιούνται σε μια ενημέρωση, διαγραφή ή εισαγωγή, ελέγχουν παραμέτρους όπως ο αριθμός, η σειρά των υποστοιχείων ενός στοιχείου κ.ά. και εξασφαλίζουν ότι μετά την ενημέρωση οι τιμές των παραμέτρων αυτών συμβαδίζουν με αυτά που επιβάλλουν τα αρχεία XML.

Οι εντολές που με κατάλληλο συνδυασμό επιτρέπουν την ενημέρωση μέρους ενός XML εγγράφου στο σύστημα X-Database είναι οι εξής:

DBInsert, που επιτρέπει την εισαγωγή υποστοιχείων σε ένα στοιχείο. Καθορίζουμε το στοιχείο στο οποίο θα γίνει η εισαγωγή (*γονικό στοιχείο*), το *προσδιοριστικό του γονικού στοιχείου* και προαιρετικά τη *σειρά* με την οποία θα εισαχθεί (Πίνακας 15, Πίνακας 16). Όταν καθορίζεται η σειρά, που σημαίνει ότι το στοιχείο παρεμβάλλεται στα υπόλοιπα στοιχεία του ίδιου επιπέδου, τότε ενημερώνεται το αντίστοιχο πεδίο στη βάση για το στοιχείο αυτό καθώς και για όλα τα στοιχεία στο ίδιο επίπεδο (Πίνακας 17).

DBUpdate, που προσδιορίζει το στοιχείο που πρόκειται να ενημερωθεί και περιέχει τις νέες τιμές των χαρακτηριστικών του. Η εντολή αυτή ενημερώνει μόνο τα χαρακτηριστικά ενός στοιχείου, χωρίς να επηρεάζει τα υποστοιχεία του (Πίνακας 18). Είναι προφανές ότι η ενημέρωση των υποστοιχείων ενός στοιχείου πρέπει να γίνει με συνδυασμό εισαγωγών, διαγραφών και ενημερώσεων των ίδιων των υποστοιχείων.

DBDelete, που προσδιορίζει το γονικό στοιχείο σε μια δήλωση XPATH και το προσδιοριστικό της ΒΔ για το στοιχείο που θα διαγραφεί (Πίνακας 19). Όλα τα υποστοιχεία του στοιχείου διαγράφονται αλυσιδωτά.

XML Command	
<DBCommand> <Insert> <AudioVisual Name="Gladiator" AVType="Movie"> <MediaProfile> <StorageID>100</StorageID> </MediaProfile> </AudioVisual> </Insert> </DBCommand>	
XML Reply	XML data after the operation
<DBReply> <DBResult> <AudioVisual id="150"> <MediaProfile id="151"/> </AudioVisual> </DBResult> </DBReply>	<AudioVisual Name="Gladiator" AVType="Movie"> <MediaProfile> <StorageID>100</StorageID> </MediaProfile> </AudioVisual>

Πίνακας 15. Εισαγωγή ενός νέου εγγράφου

XML Command	
<DBCommand> <Insert path="/AudioVisual" DB_ID="151"> <MediaProfile> <StorageID>120</StorageID> </MediaProfile> </Insert> </DBCommand> <!-- Appends a sub-element under MediaInfo element with id="151".DBReply contains the id of the element-->	
XML Reply	XML data after the operation
<DBReply> <DBResult> <MediaProfile id="152"/> </DBResult> </DBReply>	<AudioVisual Name="Gladiator" AVType="Movie"> <MediaProfile> <StorageID>100</StorageID> </MediaProfile> <MediaProfile> <StorageID>120</StorageID> </MediaProfile> </AudioVisual>

Πίνακας 16. Εισαγωγή ενός νέου στοιχείου

XML Command	
<DBCommand> <Insert path="/AudioVisual" DB_ID="150" order="1"> <MediaProfile> <StorageID>150</StorageID> </MediaProfile> </Insert> </DBCommand> <!--Inserts a new sub-element as first sub-element of MediaInfo element with id="150".-->	
XML Reply	XML data after the operation
<DBReply> <DBResult> <MediaProfile id="153"/> </DBResult> </DBReply>	<AudioVisual Name="Gladiator" AVType="Movie"> <MediaProfile> <StorageID>150</StorageID> </MediaProfile> <MediaProfile> <StorageID>100</StorageID> </MediaProfile> <MediaProfile> <StorageID>120</StorageID> </MediaProfile> </AudioVisual>

Πίνακας 17. Εισαγωγή στοιχείου σε συγκεκριμένη σειρά

XML Command	
<DBCommand> <Update path="/AudioVisual" DB_ID="150"> <AudioVisual Name="Gladiator2" AVType="Movie"> </Update> </DBCommand> <!--Updates the contents of an AudioVisual element with id="150"-->	
XML Reply	XML data after the operation
<DBReply> <DBResult> <!-- Element Updated--> </DBResult> </DBReply>	<AudioVisualname="Gladiator2" AVType="Movie"> <MediaProfile> <StorageID>150</StorageID> </MediaProfile> <MediaProfile> <StorageID>100</StorageID> </MediaProfile> <MediaProfile> <StorageID>120</StorageID> </MediaProfile> </AudioVisual>

Πίνακας 18. Ενημέρωση των χαρακτηριστικών ενός στοιχείου

<table>
<tr><td colspan="2">XML Command</td></tr>
<tr><td colspan="2"><DBCommand>
<Delete path="/AudioVisual /MediaProfile" DB_ID="152"/>
</DBCommand>

<i><!--Deletes the</i> MediaProfile <i>sub-element with id=”152” of an AudioVisual element--></i></td></tr>
<tr><td>XML Reply</td><td>XML data after the operation</td></tr>
<tr><td><DBReply>
<DBResult>
<!--Element Deleted-->
</DBResult>
</DBReply></td><td><AudioVisualname="Gladiator2"
AVType="Movie">
<MediaProfile>
<StorageID>150</StorageID>
</MediaProfile>
<MediaProfile>
<StorageID>100</StorageID>
</MediaProfile>
</AudioVisual></td></tr>
</table>

Πίνακας 19. Διαγραφή ενός στοιχείου

Θέματα ακεραιότητας

Το σύστημα X-Database υλοποιεί ένα σύνολο από περιορισμούς ακεραιότητας στη βάση δεδομένων που εξασφαλίζει την εγκυρότητα ως προς το XML-Schema των δεδομένων που ανακτώνται από ένα ερώτημα επιλογής καθώς και τον δεδομένων που απομένουν στη ΒΔ μετά από μια εισαγωγή, ενημέρωση ή διαγραφή.

Η ακεραιότητα των εγγράφων που ανακτώνται από τη βάση εξασφαλίζεται με τον έλεγχο που γίνεται, κατά τη διαδικασία ανασύνθεσης των XML εγγράφων από τα στοιχεία της βάσης. Η διαδικασία της ανασύνθεσης των εγγράφων γίνεται σταδιακά με διαδοχικές ερωτήσεις, προς τη βάση, που αντλούν δεδομένα για το κεντρικό στοιχείο, τα υποστοιχεία του, τα υποστοιχεία αυτών κ.ο.κ. Η ανασύνθεση γίνεται με βάση τη δομική πληροφορία που μεταφέρουν τα XML-Schema αρχεία που είναι συνεχώς στην κύρια μνήμη.

Στην περίπτωση των διαδικασιών επιλογής, η ***ακεραιότητα*** σε δομή και περιεχόμενο των XML εγγράφων που ανακτώνται καθορίζεται ως προς:
α) ***τα στοιχεία και τα χαρακτηριστικά***: όλα τα στοιχεία ενός εγγράφου και τα χαρακτηριστικά τους πρέπει να ανακτηθούν,
β) ***τις τιμές τους***: πρέπει να ανακτηθούν οι σωστές τιμές για κάθε χαρακτηριστικό καθώς και για το περιεχόμενο κείμενο κάθε στοιχείου,
γ) ***τη σειρά των υποστοιχείων***: όλα τα υποστοιχεία ενός στοιχείου πρέπει να ανακτηθούν με τη σειρά με την οποία εισήχθησαν. Σε περίπτωση που μεσολάβησε ενημέρωση του εγγράφου η σειρά των υποστοιχείων μετά την ανάκτηση πρέπει να συμφωνεί με την ενημέρωση.

7.2.2 Διαχείριση ερωτήσεων

Στο σύστημα X-Database τα δεδομένα των XML εγγράφων αποθηκεύονται σε ένα σύνολο από σχεσιακούς πίνακες, των οποίων τα ονόματα έχουν δημιουργηθεί με

αυτόματο τρόπο και είναι άγνωστα στους χρήστες. Τα ονόματα των πεδίων, στα οποία αποθηκεύονται οι τιμές των χαρακτηριστικών της XML, είναι επίσης άγνωστα στους χρήστες και συνεπώς είναι πολύ δύσκολο να γίνουν ερωτήσεις απευθείας στα περιεχόμενα της βάσης δεδομένων. Ο μηχανισμός ερωτήσεων, που παρέχει το σύστημα, χρησιμοποιεί τα ονόματα των XML στοιχείων και χαρακτηριστικών μέσα σε XPATH δηλώσεις αντί για τα ονόματα των πινάκων και των πεδίων της βάσης. Αρκεί οι χρήστες να είναι γνώστες του συγκεκριμένου XML-Schema για να μπορούν να υποβάλουν τις ερωτήσεις τους.

Η γλώσσα ερωτήσεων
Η γλώσσα ερωτήσεων του συστήματος X-Database υποστηρίζει τις δύο βασικές λειτουργίες της γλώσσας XQuery [W3m], την *προβολή* και την *επιλογή*, με διαφορετική σύνταξη που διακρίνει τα κριτήρια επιλογής από τις οντότητες που θα επιλεγούν ως απάντηση.

Καθώς η γλώσσα XQuery είναι ακόμη υπό εξέλιξη, επιλέξαμε την υλοποίηση ενός νέου συντακτικού ερωτήσεων. Παρ' όλα αυτά, η μετάβαση σε μια γλώσσα ερωτήσεων που είναι ευρύτερα αποδεκτή είναι στα άμεσα σχέδια. Η μετάβαση είναι μια σχετικά εύκολη διαδικασία καθώς η γλώσσα ερωτήσεων του X-Database υποστηρίζει μεγάλο μέρος των λειτουργιών και της σημασιολογίας της γλώσσας XQuery.

Η δομή των ερωτήσεων
Όπως προαναφέρθηκε, μια εντολή επιλογής επικυρώνεται ως προς το σχήμα και αποτελείται από τρία μέρη where, return και bind που ορίζονται ως χαρακτηριστικά τύπου string.

Το χαρακτηριστικό ***"where"*** του στοιχείου <Select> μπορεί να περιέχει τους τελεστές της SQL (>, <, =, AND, OR, IN, BETWEEN, κτλ.) και δηλώσεις XPath που καθορίζουν το χαρακτηριστικό ή το στοιχείο του εγγράφου στο οποίο αναφέρεται το κριτήριο. Με επεξεργασία των περιεχομένων του χαρακτηριστικού "where" παράγονται ερωτήσεις SQL προς τη βάση με τις οποίες εντοπίζονται τα προσδιοριστικά των εγγράφων που ικανοποιούν τα κριτήρια.

Το χαρακτηριστικό ***"return"*** περιέχει μία δήλωση XPath που προσδιορίζει τα στοιχεία που θα επιστραφούν. Όταν υπάρχει ένα χαρακτηριστικό "where", τα στοιχεία επιλέγονται από τα έγγραφα εκείνα που ικανοποιούν τα κριτήρια του "where".

Το χαρακτηριστικό "where" προσδιορίζει τα έγγραφα που ικανοποιούν τα κριτήρια και το χαρακτηριστικό "return" τα στοιχεία των εγγράφων αυτών που πρέπει να ανακτηθούν. Αυτό μπορεί να οδηγήσει σε ασάφειες όταν το κριτήριο του where αναφέρεται σε στοιχεία που εμφανίζονται στο return. Καθώς η διαδικασία αυτή γίνεται σε δύο στάδια –επιλογή εγγράφων και ανάκτηση τμημάτων τους– όλα τα στοιχεία (όπως τα προσδιορίζει το return) που περιέχονται σε έγγραφα που ικανοποιούν τα κριτήρια θα ανακτηθούν είτε ικανοποιούν τα κριτήρια είτε όχι.

Το χαρακτηριστικό ***"bind"*** χρησιμοποιείται σε συνδυασμό με τα δύο προηγούμενα χαρακτηριστικά για να αποτρέψει αυτή την ασάφεια. Περιέχει μια δήλωση XPath που προσδιορίζει ποια μέρη των εγγράφων που ικανοποιούν τα κριτήρια θα πρέπει να επιλεγούν. Ακολούθως, θα ανακτηθούν τα στοιχεία που προσδιορίζει το return και

βρίσκονται στα μέρη που ορίζει το bind. Με τον τρόπο αυτό ανακτώνται μόνο τα στοιχεία που περιέχονται σε συγκεκριμένα τμήματα εγγράφων που ικανοποιούν τα κριτήρια.

```
<DBCommand>
 <Insert>
  <Director>
   <Name> Stanley Kubrick </Name>
   <Movie><Title>Eyes Wide Shut</Title></Movie>
   <Movie><Title>2001: A Space Odyssey</Title></Movie>
   <Movie><Title>Spartacus</Title></Movie>
  </Director>
 </Insert>
</DBCommand>
```

```
<DBCommand>
<Select
where="/Director/Movie/Title[text]#=' Eyes Wide Shut'#
return="/Director/Movie"/>
</DBCommand>
```

```
<DBReply>
 <DBResult>
   <Movie><Title>Eyes Wide Shut</Title></Movie>

   <Movie><Title>2001: A Space Odyssey</Title></Movie>

   <Movie><Title>Spartacus</Title></Movie>

 </DBResult>
</DBReply>
```

```
<DBCommand>
<Select
where="/Director/Movie/Title[text]#=' Eyes Wide Shut'
return="/Director/Movie"/ bind="/Director/Movie">
</DBCommand>
```

```
<DBReply>
 <DBResult>
   <Movie><Title>Eyes Wide Shut</Title></Movie>

 </DBResult>
</DBReply>
```

Πίνακας 20. Παραδείγματα εντολών Select

Ένα παράδειγμα εντολής ***DBSelect*** απεικονίζεται στον Πίνακα 20. Το ερώτημα στη γραμμή 2 επιλέγει τα στοιχεία "Director" που περιέχουν υποστοιχείο "*Title*" με τιμή "Eyes Wide Shut" και ανακτά τα υποστοιχεία "*Movie*" αυτών (γραμμή 3). Το δεύτερο ερώτημα, στη γραμμή 4, εντοπίζει τα στοιχεία "Movie" που περιέχουν υποστοιχείο "Title" με την απαιτούμενη τιμή και τα επιστρέφει εντός ενός στοιχείου DBReply (γραμμή 5). Η προτεινόμενη σύνταξη προσομοιώνει τη λειτουργικότητα που παρέχουν οι εκφράσεις FLWOR στην γλώσσα XQuery [W3m]. Τα XML στοιχεία με πλήρη δομή επιστρέφονται πάντα μέσα σε ένα στοιχείο DBReply.

Το σύστημα επιτρέπει ερωτήσεις επιλογής στις οποίες δεν καθορίζεται επακριβώς η θέση του στοιχείου στο έγγραφο, π.χ. ερωτήσεις ανάκτησης στοιχείου με συγκεκριμένο όνομα και δομή σε οποιοδήποτε βάθος του εγγράφου. Οι ερωτήσεις αυτές δηλώνονται όπως και στην XPATH με το σύμβολο "//" μπροστά από το όνομα του αρχείου.

Επεξεργασία ερωτημάτων - ευρετήρια
Η επεξεργασία ενός ερωτήματος από το σύστημα X-Database γίνεται σε δύο βήματα. Το πρώτο βήμα αφορά την επεξεργασία του στοιχείου <Select>, τη σύνθεση και εκτέλεση ενός ερωτήματος SQL που επιστρέφει τα προσδιοριστικά των στοιχείων που θα ανακτηθούν. Στο δεύτερο βήμα, ανακατασκευάζεται πλήρως η εσωτερική δομή των στοιχείων αυτών, με αναδρομικές ερωτήσεις προς τους πίνακες που περιέχουν τα δεδομένα των υποστοιχείων τους. Η ανασύσταση του εγγράφου απάντησης συνεπάγεται πολλά ερωτήματα προς τη βάση δεδομένων. Είναι μια χρονοβόρα διαδικασία, ειδικά όταν τα στοιχεία που πρόκειται να επιστραφούν έχουν μεγάλο βάθος και πολυπλοκότητα.

Η διαδικασία ανάκτησης μπορεί να επιταχυνθεί αν δημιουργηθούν ευρετήρια σε όλα τα ξένα κλειδιά κάθε πίνακα που συμμετέχει σε μια σχέση εμφωλιασμού. Τα ευρετήρια αυτά μπορούν να δημιουργηθούν αυτόματα με ανάλυση του συγκεκριμένου XML-Schema. Με τον τρόπο αυτό, αν το XML-Schema περιέχει πολλά εμφωλιασμένα στοιχεία, θα δημιουργηθούν πολλά ευρετήρια στη βάση, με συνέπεια να μεγαλώσει το μέγεθος της βάσης και να καθυστερούν οι διαδικασίες ενημέρωσης. Εναλλακτικά, η υλοποίηση μιας υβριδικής βάσης που αποθηκεύει τα συχνά ζητούμενα στοιχεία XML ολόκληρα μαζί με τα σχεσιακά δεδομένα μπορεί να επιταχύνει τις αναζητήσεις. Αν, για παράδειγμα, τα δεδομένα του στοιχείου <Movie> στον Πίνακα 20, αποθηκεύονται ταυτόχρονα σε σχεσιακή και XML μορφή, μπορούμε να ελέγχουμε τα κριτήρια στη σχεσιακή μορφή και να ανακτούμε απευθείας τα κατάλληλα στοιχεία <Movie> από το σύστημα αρχείων χρησιμοποιώντας το προσδιοριστικό της βάσης δεδομένων.

Ο χρόνος εκτέλεσης του πρώτου βήματος, σε ένα ερώτημα, εξαρτάται από τον αριθμό των στοιχείων στη βάση και μπορεί να επιταχυνθεί, αν δημιουργηθούν ευρετήρια στα πεδία που ερωτώνται συχνά. Αυτό μπορεί να γίνει χειροκίνητα, μετά τη δημιουργία της βάσης, κάνοντας χρήση του μηχανισμού ορισμού των ευρετηρίων του συστήματος.

Ένα παράδειγμα χρήσης της εντολής CreateIndex και η αντίστοιχη δήλωση SQL που παράγεται απεικονίζονται στον Πίνακα 21.

XML	`<DBCommand>`
	`<CreateIndex Name="MovieNameIndex">`
	`<field>/Director/Movie[name]</field>`
	`<field>/Director/Movie[AVType]</field>`
	`</CreateIndex>`
	`</DBCommand>`

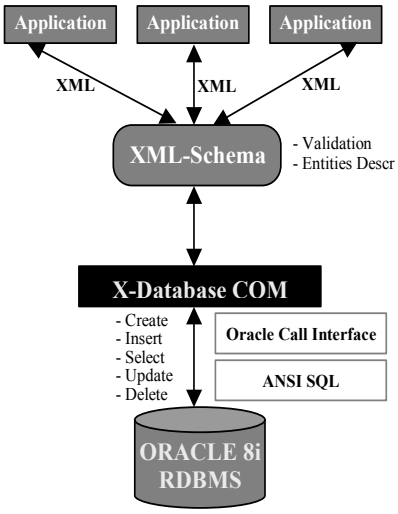
SQL	`CREATE INDEX MovieNameIndex ON Movie(name,AVType)`

Πίνακας 21. Παράδειγμα δημιουργίας ευρετηρίου

7.3 Η αρχιτεκτονική του συστήματος X-Database

Το σύστημα δρα ως διεπαφή μεταξύ της βάσης δεδομένων και διάφορων εφαρμογών που αποθηκεύουν και ανακτούν XML έγγραφα. Το σύστημα αποτελείται από δύο τμήματα: α) τη βάση δεδομένων και β) τη διεπαφή της βάσης, όπως φαίνεται και στο Σχήμα 24. Στο πρότυπο σύστημα που αναπτύχθηκε, επιλέχθηκε η σχεσιακή βάση δεδομένων Oracle 8i και η διεπαφή Oracle Call Interface [OC97] για την επικοινωνία

με αυτή. Καθώς το σύστημα μετατρέπει τις εντολές XML σε εντολές ANSI SQL προς τη βάση, εύκολα το σύστημα μπορεί να μεταφερθεί σε οποιαδήποτε σχεσιακή ΒΔ. Το σχεσιακό σχήμα δημιουργείται αυτόματα έπειτα από επεξεργασία της δομής που περιγράφει το XML-Schema.

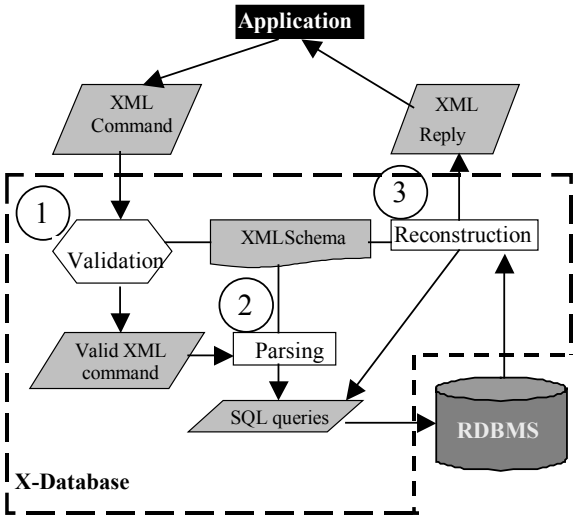


Σχήμα 24. Αρχιτεκτονική του συστήματος X-Database

Τόσο δικτυακές όσο και τοπικές εφαρμογές μπορούν να συνδεθούν με το σύστημα X-Database. Η επικοινωνία των εφαρμογών με το X-Database γίνεται αποκλειστικά με χρήση XML εγγράφων. Η διεπαφή μεταξύ των εφαρμογών και της ΒΔ έχει υλοποιηθεί ως ένα αντικείμενο COM, γραμμένο σε C++. Η αρχιτεκτονική αυτή επιτρέπει σε νέες εφαρμογές να χρησιμοποιούν το σύστημα μέσα από την ίδια διεπαφή, αγνοώντας τη δομή της βάσης.

7.3.1 Ροή δεδομένων

Η επεξεργασία μιας XML εντολής απεικονίζεται στο Σχήμα 25 που ακολουθεί:



Σχήμα 25. Επεξεργασία XML εντολών

1) Μια XML εντολή έρχεται στο σύστημα X-Database από μια εφαρμογή. Η εντολή ***επικυρώνεται*** ως προς το XML-Schema που αντιστοιχεί στη συγκεκριμένη εφαρμογή, ώστε να ελεγχθεί η καταλληλότητα της δομής της εντολής.

2) Το σύστημα ***επεξεργάζεται*** την επικυρωμένη εντολή, με βάση το XML-Schema, και συνθέτει τα κατάλληλα ερωτήματα SQL.
3) Η ΒΔ επεξεργάζεται τα ερωτήματα. Οι εισαγωγές επιστρέφουν ένα XML έγγραφο που περιέχει τα προσδιοριστικά όλων των στοιχείων που εισήχθησαν και των υποστοιχείων τους. Τα ερωτήματα επιλογής ***ανασυνθέτουν*** πλήρως τα XML έγγραφα κάνοντας αναδρομικά ερωτήσεις για τα εμφωλιασμένα στοιχεία των εγγράφων.
4) Το XML έγγραφο απάντησης επιστρέφεται στην εφαρμογή που έστειλε την εντολή.

7.4 Αξιολόγηση του συστήματος

Πέρα από τη χρήση του στα πλαίσια του συστήματος THESUS, το σύστημα X-Database δοκιμάστηκε εκτεταμένα σε μια πραγματική εφαρμογή που αφορούσε τη διαχείριση οπτικοακουστικών μεταδεδομένων [AS00] για ένα σύστημα περιγραφής και ερωτήσεων ενός αρχείου βίντεο. Το πρότυπο MPEG-7 χρησιμοποιήθηκε για την τυποποίηση των μεταδεδομένων.

Για να αξιολογήσουμε την αξιοπιστία και την δυνατότητα κλιμάκωσης του συστήματος εκτελέσαμε μια σειρά διεργασιών που περιελάμβαναν τη δημιουργία της βάσης, τη μαζική εισαγωγή στοιχείων στη βάση και την υποβολή ερωτημάτων. Σε όλες τις λειτουργίες μετρήθηκε ο συνολικός χρόνος απόκρισης του συστήματος. Ο χρόνος αυτός περιλαμβάνει: α) την επεξεργασία των XML εγγράφων που δίνονται στο σύστημα, β) την πρόσβαση στη βάση και γ) τη δημιουργία του XML εγγράφου απάντησης.

Στην περίπτωση των ερωτημάτων επιλογής, ο χρόνος απόκρισης περιλαμβάνει το χρόνο για την ανάκτηση α) των προσδιοριστικών των αρχικών στοιχείων που ικανοποιούν το ερώτημα και β) των υπόλοιπων χαρακτηριστικών και υποστοιχείων που περιέχονται στα αρχικά στοιχεία (πολλές ερωτήσεις SQL σε πολλαπλούς σχετιζόμενους πίνακες).

Η απόδοση του συστήματος μετρήθηκε με χρήση μιας απλής γραφικής διεπαφής όπου τα XML έγγραφα εισάγονται ως κείμενο και τα έγγραφα απάντησης εμφανίζονται σε ξεχωριστό παράθυρο. Το σύστημα X-Database δοκιμάστηκε με διάφορο φόρτο δεδομένων. Τα πειράματα έγιναν σε ένα υπολογιστή με επεξεργαστή Pentium IV (1.4 GHz) και 512Mbytes μνήμη RAM με σκληρό δίσκο IDE. Ο υπολογιστής είχε λειτουργικό σύστημα Windows 2000 professional και τη βάση Oracle v 8.1. Βάση και εφαρμογή εγκαταστάθηκαν στον ίδιο υπολογιστή για να αποφευχθούν καθυστερήσεις λόγω δικτύου.

7.4.1 Δημιουργία σχεσιακού σχήματος

Το σύστημα X-Database αναλύει τη δομή των αρχείων XML-Schema της εφαρμογής και δημιουργεί τους κατάλληλους πίνακες στο σχεσιακό σχήμα, καθώς και τα απαραίτητα πρωτεύοντα κλειδιά, αναφορές και σκανδάλες που εγγυώνται την ακεραιότητα των περιεχομένων της βάσης δεδομένων κατά τις αλυσιδωτές εισαγωγές και διαγραφές.

Οι παράμετροι που εξετάσθηκαν κατά τη διάρκεια των πειραμάτων είναι ενδεικτικοί της πολυπλοκότητας του παραγόμενου σχεσιακού σχήματος: ο αριθμός των πινάκων,

των περιορισμών ορθότητας αναφορών (ξένα κλειδιά), των triggers. Μετρήθηκαν επίσης ο χρόνος που απαιτείται για τη δημιουργία της βάσης αλλά και τη διαγραφή της χωρίς να έχει γίνει εισαγωγή δεδομένων. Καθώς οι πιο πάνω παράμετροι επηρεάζονται από την πολυπλοκότητα του XML-Schema παραθέτουμε και ορισμένα ενδεικτικά στοιχεία της πολυπλοκότητας της δομής του XML-Schema του MPEG-7.

Στον Πίνακα 22 παρουσιάζονται στοιχεία αυτής της απεικόνισης του XML-Schema του προτύπου MPEG-7 σε μια σχεσιακή βάση.

XML-Schema του MPEG-7					
Αρ. αρχείων	Μέγεθος αρχείων	Complex Types	Abstract types	Extensions	simple elements με πολλές εμφανίσεις
6 αρχεία	348 Kb	519 σύνολο 377 καθολικοί 142 τοπικοί	51	387	218
Σχεσιακό σχήμα					
Πίνακες	Ξένα κλειδιά	Δημιουργία βάσης (sec)		Καταστροφή βάσης (sec)	
676	1872	91.6		57.8	

Πίνακας 22. Στατιστικά του XML-Schema και του σχήματος βάσης που δημιουργεί

Στα στοιχεία του πίνακα παρατηρούμε ότι από το συνολικό αριθμό των σύνθετων τύπων (519) ορισμένοι είναι καθολικοί και επώνυμοι, ενώ άλλοι έχουν δηλωθεί τοπικά ως τύποι ενός συγκεκριμένου στοιχείου της XML-Schema και είναι ανώνυμοι. Επίσης, κάποιοι σύνθετοι τύποι (51) έχουν δηλωθεί ως abstract και κατά συνέπεια δεν απαιτείται η δημιουργία πινάκων γι' αυτούς. Για τους εναπομείναντες τύπους δημιουργούνται πίνακες στη βάση. Οι υπόλοιποι πίνακες δημιουργούνται για όσα στοιχεία είναι απλού τύπου αλλά εμφανίζονται ως υπο-στοιχεία άλλων με πολλαπλές εμφανίσεις (218 τέτοια στοιχεία).

Ο αριθμός των ξένων κλειδιών, που δημιουργούνται, οφείλεται στο συνδυασμό φωλιασμένων στοιχείων και περιπτώσεων κληρονομικότητας τύπων που εμφανίζει το σχήμα του MPEG-7. Αν ένας σύνθετος τύπος Α οριστεί να περιέχει κάποιο στοιχείο τύπου Β, τότε στη βάση θα δημιουργηθεί ένα ξένο κλειδί στον πίνακα που αντιστοιχεί στον τύπο Β και θα αναφέρεται στο πεδίο κλειδί του πίνακα που αντιστοιχεί στον τύπο Α. Αν τώρα ο τύπος Α κληρονομείται από άλλους τύπους, το ίδιο θα συμβαίνει και με τη σχέση εμφωλιασμού και συνεπώς στον πίνακα του τύπου Β θα δημιουργηθούν και άλλα ξένα κλειδιά (στα αντίστοιχα πεδία) προς τα κλειδιά των πινάκων κάθε παράγωγου τύπου.

Τα αποτελέσματα αυτά αποδεικνύουν τη δυσκολία της δημιουργίας ενός σχεσιακού σχήματος για το πρότυπο MPEG-7 με μη-αυτόματο τρόπο. Η σύνθετη δομή του MPEG-7 με τα εκατοντάδες στοιχεία και χαρακτηριστικά είναι δύσκολο να αναλυθεί και να απεικονιστεί στο βέλτιστο σχεσιακό σχήμα ακόμη και από έναν έμπειρο διαχειριστή βάσεων δεδομένων. Επιπλέον, είναι σχεδόν αδύνατο να αφομοιώσουμε το σχήμα της βάσης και να παράγουμε τις κατάλληλες ερωτήσεις προς αυτό.

Αντίθετα είναι πολύ πιο εύκολο να αφομοιώσει κανείς την ιεραρχική δομή των XML εγγράφων και να επικοινωνεί με τη βάση δεδομένων είτε για ενημερώσεις, είτε για ερωτήσεις μέσα από XML έγγραφα που ακολουθούν τη δομή του XML-Schema.

7.4.2 Μαζική εισαγωγή δεδομένων

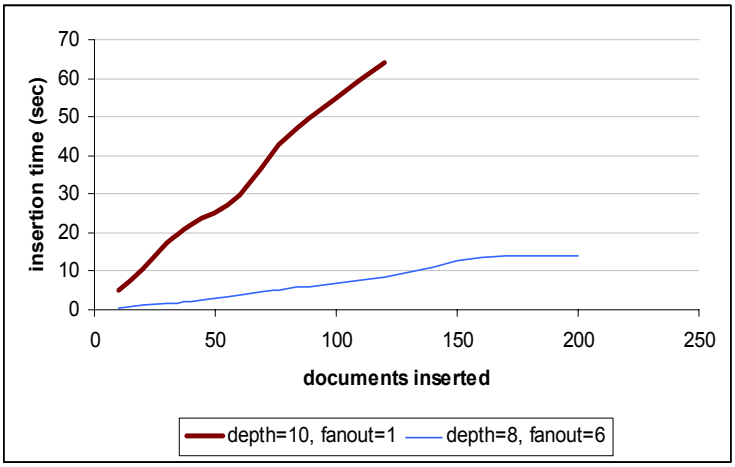
Για να ελέγξουμε την αποδοτικότητα του συστήματος στη μαζική εισαγωγή ολόκληρων XML εγγράφων, τροποποιήσαμε τρεις παραμέτρους των εγγράφων XML και μετρήσαμε την επιρροή τους στο χρόνο που απαιτείται για κάθε εισαγωγή. Οι παράγοντες που ελέγχθηκαν είναι [TI+01]:

- ***Κλιμάκωση - Scaling factor:*** ο αριθμός των εγγράφων που εισάγονται. Χρησιμοποιούμε έγγραφα ίδιου μεγέθους και αναμένουμε ο χρόνος εισαγωγής να είναι ανάλογος του αριθμού των εγγράφων που εισάγονται ακόμη και για πολύ μεγάλο αριθμό εγγράφων.
- ***Βάθος - Depth:*** είναι ο μέγιστος αριθμός επιπέδων εμφωλιασμού και είναι ενδεικτικός της πολυπλοκότητας των εγγράφων. Χρησιμοποιούμε έγγραφα ίδιου μεγέθους με αυξανόμενο βάθος, που συνεπάγεται και διεργασίες εισαγωγής σε περισσότερους πίνακες.
- ***Εύρος - Fan-out:*** προσδιορίζει τον αριθμό των υποστοιχείων στο πρώτο επίπεδο και είναι επίσης ενδεικτικό της πολυπλοκότητας των εγγράφων.

Καθώς οι παράγοντες *εύρος* και *βάθος* αλλάζουν ταυτόχρονα στα διάφορα έγγραφα που εισάγονται, για τις ανάγκες των πειραμάτων κρατάμε σταθερό κάποιον από τους δύο ή και τους δύο. Οι σταθερές τιμές επιλέγονται σύμφωνα με τα δείγματα εγγράφων MPEG-7 που έχουμε στη διάθεσή μας και θεωρούμε ότι η επιλογή τους δεν επηρεάζει τα αποτελέσματα.

Κλιμάκωση

Για να μετρήσουμε την επιρροή της *κλιμάκωσης*, εισάγουμε ένα αυξανόμενο αριθμό εγγράφων με *βάθος=8* και *εύρος=6*. Τα έγγραφα έχουν μέγεθος γύρω στα 1200 bytes. Εισάγουμε από 10 (12Kb) έως 200(240Kb) έγγραφα. Το ίδιο πείραμα επαναλαμβάνεται με έγγραφα που έχουν *βάθος=10* και *εύρος=1* και μέγεθος 8700 bytes. Στην περίπτωση αυτή εισάγουμε από 10 (87Kb) ως 120 (1Mb) έγγραφα. Τα αποτελέσματα παρουσιάζονται στο Σχήμα 26 και δείχνουν ότι ***ο χρόνος εισαγωγής είναι ανάλογος του αριθμού των εγγράφων*** και στις δύο περιπτώσεις (Σχήμα 26). Παρ' όλα αυτά, η κλίση είναι μεγαλύτερη για μεγάλες τιμές *βάθους*, μια ένδειξη ότι η πολυπλοκότητα επηρεάζεται περισσότερο από τον παράγοντα *βάθος* και λιγότερο από τον παράγοντα *εύρος*.

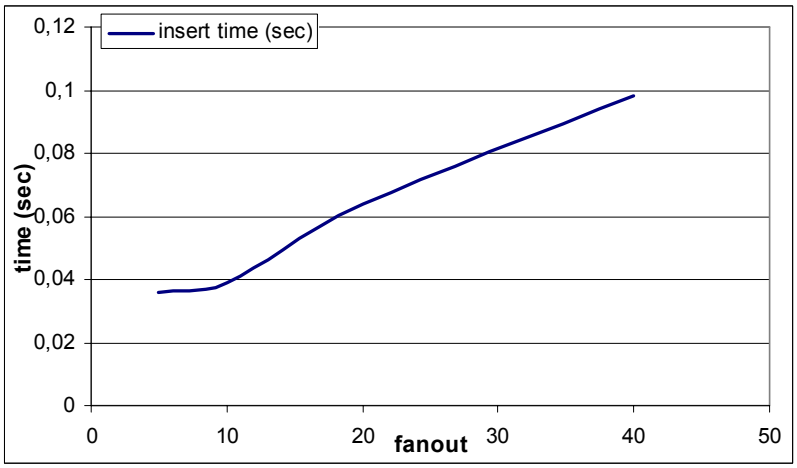


Σχήμα 26. Εξάρτηση του χρόνου εισαγωγής από το πλήθος των εγγράφων

Βάθος

Ένας σταθερός αριθμός εγγράφων (κλιμάκωση=10) με ίδιο εύρος (εύρος=1) και ίδιο μέγεθος σε bytes εισάγονται στη βάση. Όπως φαίνεται και στο Σχήμα 27, ***ο χρόνος εισαγωγής αυξάνει σχεδόν εκθετικά ως προς το βάθος των αρχείων*** καθώς το αυξανόμενο βάθος συνεπάγεται μεγαλύτερο αριθμό εντολών εισαγωγής σε περισσότερους πίνακες στη βάση.

Σχήμα 27. Επίδραση του παράγοντα βάθους στο χρόνο εισαγωγής



Σχήμα 28. Επίδραση του παράγοντα εύρους στο χρόνο εισαγωγής

Εύρος

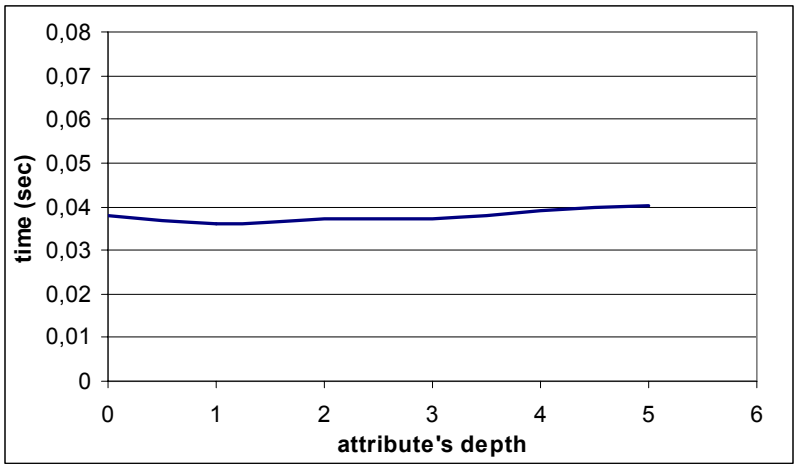
Εισάγουμε ένα σταθερό αριθμό εγγράφων (κλιμάκωση=10), ίδιου μεγέθους και βάθους (βάθος=3), που περιέχουν αυξανόμενο αριθμό υποστοιχείων στο πρώτο επίπεδο (αυξανόμενο εύρος). Από τα αποτελέσματα στο Σχήμα 28 βλέπουμε ότι ***ο χρόνος εισαγωγής είναι περίπου ανάλογος του εύρους των εγγράφων.***

7.4.3 Ενημέρωση εγγράφων

Ένας πολύ σημαντικός παράγοντας, της απόδοσης του συστήματος, είναι ο χρόνος απόκρισης σε ενημερώσεις στοιχείων, για έγγραφα που είναι ήδη αποθηκευμένα στη βάση. Οι απλές ενημερώσεις αφορούν μόνο διαφοροποίηση των τιμών των χαρακτηριστικών ενός στοιχείου. Παρ' όλα αυτά, ενημέρωση ενός εγγράφου μπορεί να έχουμε με εισαγωγή ή διαγραφή στοιχείων του. Η απόδοση του συστήματος στις ενημερώσεις μετρήθηκε ως προς δύο βασικούς παράγοντες: α) το βάθος του στοιχείου που ενημερώνεται, εισάγεται ή διαγράφεται β) τον αριθμό των στοιχείων που ενημερώνονται.

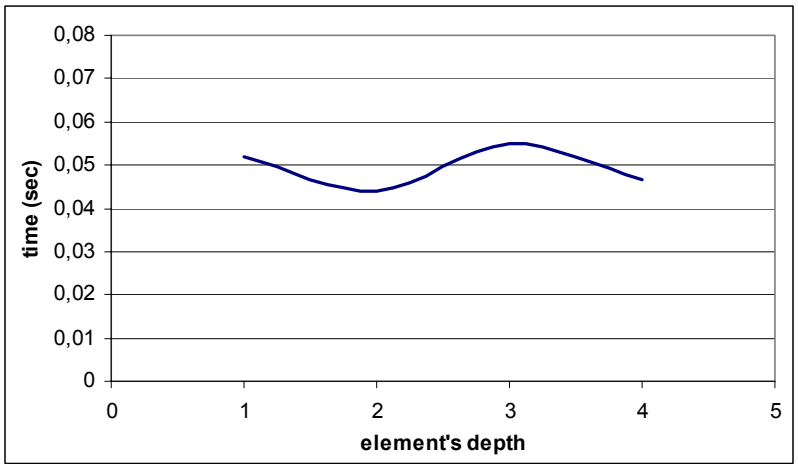
Στα πειράματα που πραγματοποιήσαμε:

- ενημερώσαμε την τιμή ενός χαρακτηριστικού για στοιχείο σε αυξανόμενο βάθος. Όπως αναμέναμε, ο χρόνος που απαιτείται για την ενημέρωση χαρακτηριστικών του στοιχείου είναι σταθερός και ανεξάρτητος του βάθους του στοιχείου, καθώς συνεπάγεται ενημέρωση σε ένα μόνο πίνακα στο σχεσιακό σχήμα (Σχήμα 29),
- εισαγάγαμε ένα στοιχείο XML, χωρίς υποστοιχεία, σε αυξανόμενο βάθος (Σχήμα 30). Στην περίπτωση αυτή έχουμε εισαγωγή μιας νέας εγγραφής σε συγκεκριμένο πίνακα στο σχεσιακό σχήμα και όπως ήταν αναμενόμενο ο χρόνος ενημέρωσης είναι επίσης ανεξάρτητος του βάθους,
- διαγράψαμε ένα στοιχείο, χωρίς υποστοιχεία, σε αυξανόμενο βάθος στο έγγραφο. Παρόμοια η διάρκεια της ενημέρωσης ήταν ανεξάρτητη του βάθους του διαγραφόμενου στοιχείου.



Σχήμα 29. Απόδοση ενημέρωσης ενός χαρακτηριστικού σε αυξανόμενο βάθος (κλιμάκωση=1, βάθος=5, εύρος=1)

Οι διαδικασίες ενημέρωσης επαναλήφθηκαν πολλές φορές, με διαφορετικό φορτίο στη βάση δεδομένων και μετρήθηκε ο μέσος χρόνος ενημέρωσης. Το τελικό συμπέρασμα είναι ότι ***η απόδοση του συστήματος στις ενημερώσεις είναι ανεξάρτητη του βάθους των στοιχείων που ενημερώνονται***.
Ακολούθως, ενημερώσαμε ένα αυξανόμενο αριθμό στοιχείων και καταλήξαμε στο συμπέρασμα ότι ***ο χρόνος ενημέρωσης είναι ανάλογος του αριθμού των στοιχείων που ενημερώνονται***.



Σχήμα 30. Απόδοση ενημέρωσης ενός στοιχείου σε αυξανόμενο βάθος (κλιμάκωση=1, βάθος=2, εύρος=2)

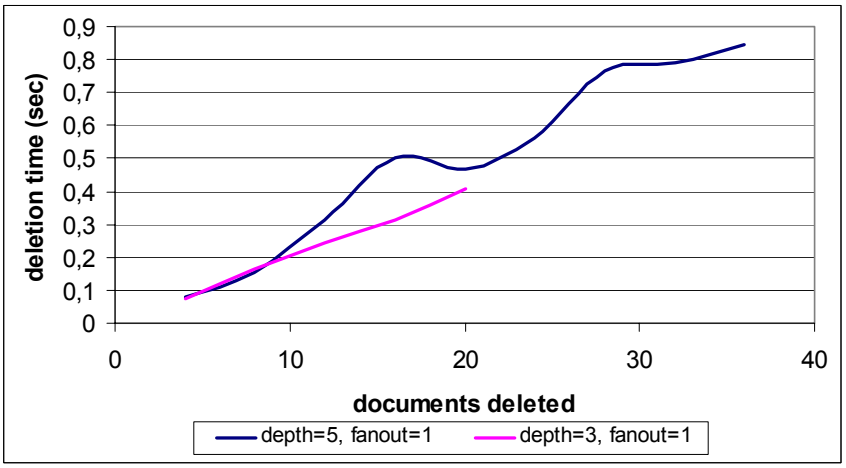
7.4.4 Διαγραφή

Η απόδοση του συστήματος στις διαγραφές μετρήθηκε πρώτα σε επίπεδο εγγράφου και στη συνέχεια σε επίπεδο στοιχείου. Οι διαγραφές έγιναν δίνοντας στο σύστημα το προσδιοριστικό του εγγράφου ή στοιχείου στη βάση. Και στα δύο επίπεδα μετρήθηκαν οι παράγοντες κλιμάκωσης, βάθους και εύρους. Σύμφωνα με τα πειραματικά αποτελέσματα:

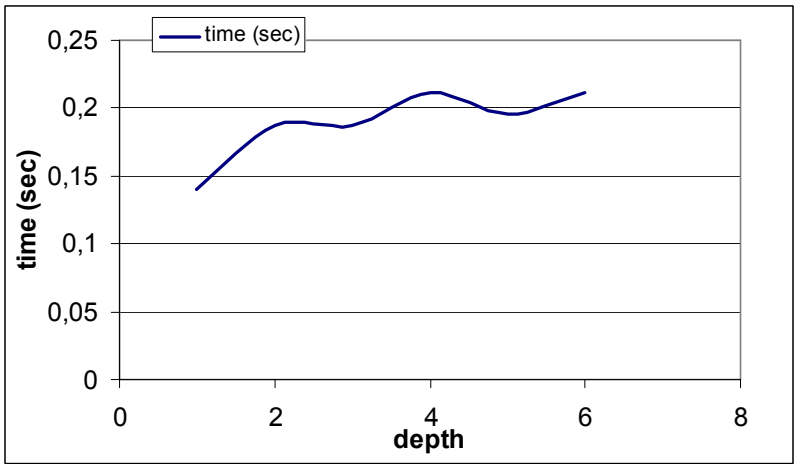
- Για αυξανόμενο αριθμό εγγράφων με σταθερό βάθος και εύρος ο χρόνος διαγραφής αυξάνεται ανάλογα (Σχήμα 31).
- Για σταθερό αριθμό εγγράφων με σταθερό βάθος αλλά αυξανόμενο εύρος ο χρόνος διαγραφής αυξάνεται, καθώς διαγράφονται πολλές εγγραφές στους ίδιους πίνακες (Σχήμα 32).
- Για σταθερό αριθμό εγγράφων με σταθερό εύρος αλλά αυξανόμενο βάθος ο χρόνος διαγραφής αυξάνεται σχεδόν εκθετικά, καθώς διαγράφεται σταθερός αριθμός εγγραφών αλλά από πολλούς πίνακες (Σχήμα 33).

Σε όλες τις περιπτώσεις εισαγωγής, διαγραφής ή ενημέρωσης στοιχείων θεωρούμε ότι τα στοιχεία είναι ίδιου τύπου και κατά συνέπεια αποθηκεύονται στον ίδιο πίνακα στο σχεσιακό σχήμα. Σε διαφορετική περίπτωση είναι πολύ δύσκολο να αξιολογήσουμε τη συμπεριφορά και απόδοση του συστήματος. Πρακτικά η αύξηση του βάθους ή του εύρους ενός εγγράφου ή στοιχείου επηρεάζει σημαντικά τις διαγραφές, ενημερώσεις και εισαγωγές όταν τα στοιχεία XML και υποστοιχεία αυτών είναι διαφορετικών τύπων και κατά συνέπεια πρέπει να γίνουν εισαγωγές, διαγραφές και ενημερώσεις αντίστοιχα σε περισσότερους από έναν πίνακες.

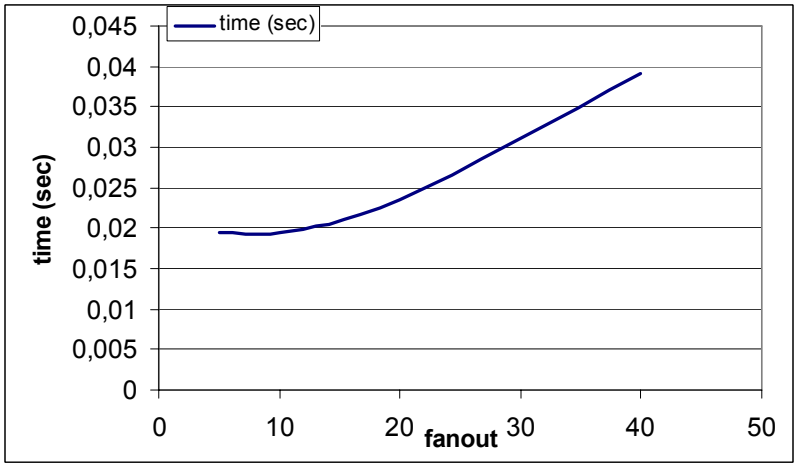
Έστω έγγραφο Α με εύρος=1, βάθος=5 και εύρος=1 για κάθε υποστοιχείο του. Ο συνολικός αριθμός στοιχείων είναι 5 και ο μέγιστος αριθμός πινάκων που εμπλέκονται 5. Για την εισαγωγή/διαγραφή ενός τέτοιου εγγράφου απαιτείται από μια εισαγωγή/διαγραφή εγγραφής σε κάθε ένα από τους 5 πίνακες. Έστω έγγραφο Β με εύρος 5, βάθος 1 και 5 συνολικά στοιχεία. Ας θεωρήσουμε ότι τα 5 στοιχεία του 1^{ου} επιπέδου είναι ίδιου τύπου. Τότε έχουμε εισαγωγή/διαγραφή 5 εγγραφών σε 1 μόνο πίνακα.



Σχήμα 31. Εξάρτηση του χρόνου εισαγωγής από το πλήθος των εγγράφων



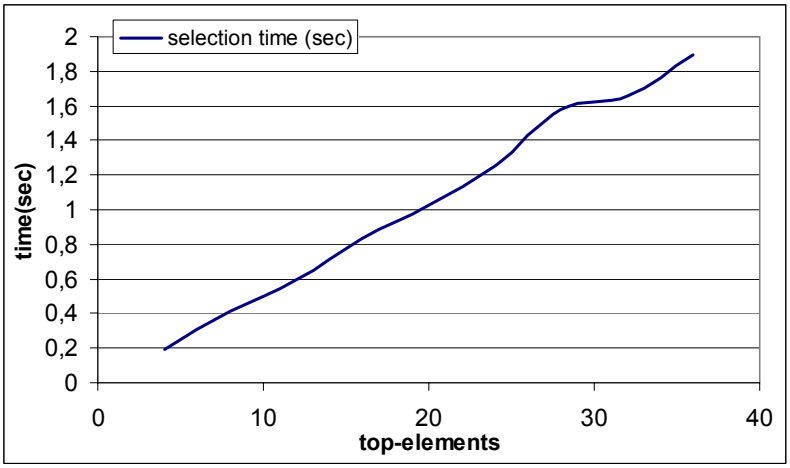
Σχήμα 32. Επίδραση του παράγοντα βάθους στο χρόνο διαγραφής



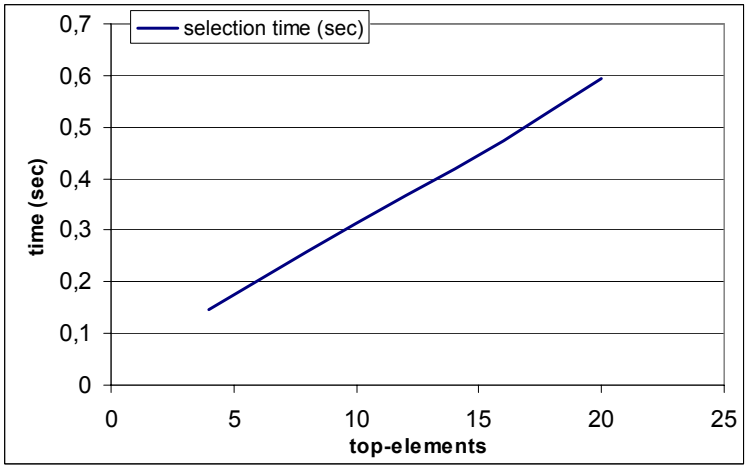
Σχήμα 33. Επίδραση του παράγοντα εύρους στο χρόνο διαγραφής

7.4.5 Επιλογή

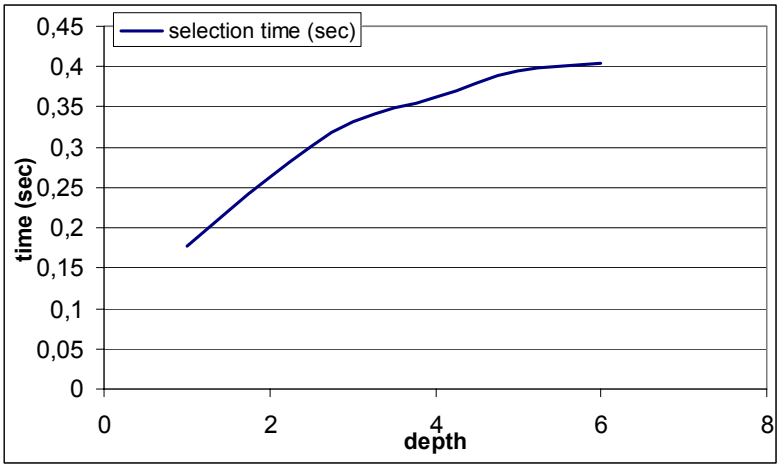
Για να μετρήσουμε την απόδοση του συστήματος στις ερωτήσεις επιλογής πραγματοποιήσαμε επιλογές στοιχείων από έγγραφα που είχαν ήδη αποθηκευθεί στη βάση. Η επιλογή των στοιχείων έγινε με κριτήριο τα προσδιοριστικά τους και αξιολογήθηκαν οι ίδιοι παράγοντες με τα πειράματα εισαγωγής. Τα αποτελέσματα (Σχήματα 34-37) δείχνουν ότι ο χρόνος που απαιτείται για την επιλογή είναι ανάλογος του αριθμού των στοιχείων που επιλέγονται (για στοιχεία ίδιας εσωτερικής δομής), ενώ αυξάνει δυσανάλογα για στοιχεία με αυξανόμενο βάθος και εύρος καθώς αυτές οι διαδικασίες απαιτούν πολλά ερωτήματα επιλογής για την ανακατασκευή των στοιχείων.



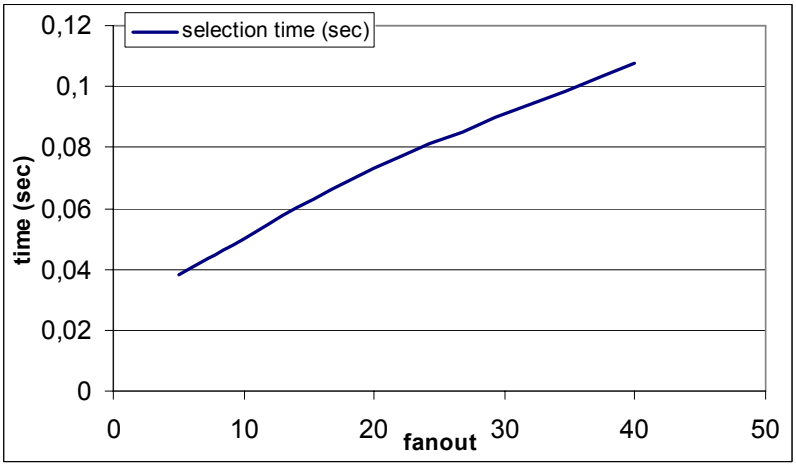
Σχήμα 34. Χρόνος επιλογής για αυξανόμενο αριθμό όμοιων στοιχείων (βάθος=5, εύρος=1)



Σχήμα 35. Χρόνος επιλογής για αυξανόμενο αριθμό διαφορετικών στοιχείων (βάθος=3, εύρος=1)



Σχήμα 36. Χρόνος επιλογής για στοιχεία με αυξανόμενο βάθος (κλιμάκωση=10, εύρος=1)



Σχήμα 37. Χρόνος επιλογής για στοιχεία με αυξανόμενο εύρος (κλιμάκωση=10, βάθος=3)

Σε όλες τις περιπτώσεις, ο χρόνος που απαιτείται, για την κατασκευή του πρώτου ερωτήματος και την ανάκτηση των προσδιοριστικών των στοιχείων που ικανοποιούν τα κριτήρια, είναι της τάξης του χιλιοστού του δευτερολέπτου και είναι κατά πολύ μικρότερος από το χρόνο που απαιτείται για την ανακατασκευή της δομής του εγγράφου. Για το λόγο αυτό τα πειράματα εστιάστηκαν σε επιλογές χωρίς κριτήρια (χαρακτηριστικό where), που καθορίζουν μόνο τα στοιχεία που θα επιστραφούν (χαρακτηριστικό return). Επιλέξαμε τα στοιχεία αυτά να είναι διαφορετικού βάθους και πολυπλοκότητας.

Ενδεικτικά παρουσιάζουμε ορισμένα αποτελέσματα ερωτήσεων προς τη βάση στον πίνακα που ακολουθεί (Πίνακας 23). Στον πίνακα παρουσιάζεται ο αριθμός των στοιχείων που επιστρέφονται, το συνολικό μέγεθος σε Kbytes και ο χρόνος που απαιτείται για την απάντηση σε δευτερόλεπτα.
Στις δύο πρώτες σειρές του πίνακα ανακτώνται στοιχεία σε αυξανόμενο βάθος. Όπως φαίνεται, καθώς το βάθος στο οποίο βρίσκονται τα στοιχεία μειώνεται, αυξάνει η πολυπλοκότητα της υφιστάμενης δομής τους και περιέχονται περισσότερα υποστοιχεία. Αυτό συνεπάγεται αύξηση των ερωτήσεων επιλογής που απαιτούνται για την ανασύνθεση του στοιχείου, και αύξηση στο χρόνο απάντησης και στο μέγεθος του εγγράφου απάντησης.

Οι τρεις τελευταίες σειρές ανακτούν στοιχεία με συγκεκριμένο όνομα ή δομή σε οποιοδήποτε επίπεδο των XML εγγράφων. Για μια ακόμη φορά, όταν η δήλωση XPATH στο χαρακτηριστικό return γίνεται πιο συγκεκριμένη (//Label/Name αντί για //Label) ανακτώνται λιγότερα στοιχεία σε λιγότερο χρόνο.

XPATH δήλωση στο χαρακτηριστικό return	**Πλήθος στοιχείων που επιστρέφονται**	**Μέγεθος (Kb)**	**Χρόνος απόκρισης (sec)**
/Mpeg7/Description /Semantics/Semantic Base/Label/Name	133	9	0.9
/Mpeg7/Description /Semantics/Semantic Base/Label/	133	15	2.8
/Mpeg7/Description /Semantics/Semantic Base	133	63	19.7
/Mpeg7/Description /Semantics	38	96	24.8
//Name	1337	88	16.1
//Label/Name	337	23	2.0
//Label	362	41	10.0

Πίνακας 23. Παραδείγματα εντολών επιλογής και χρόνων απόκρισης

7.4.6 Συμπεράσματα από τη διαδικασία αξιολόγησης

Μια επισκόπηση των πειραματικών αποτελεσμάτων δείχνει ότι η απόδοση του συστήματος δεν επηρεάζεται από τον παράγοντα κλιμάκωσης. Παρ' όλα αυτά, οι χρόνοι εκτέλεσης των εντολών εξαρτώνται σημαντικά από την εσωτερική πολυπλοκότητα των εγγράφων και πιο συγκεκριμένα από τον αριθμό των διαφορετικών τύπων που εμφανίζουν τα υποστοιχεία του εγγράφου.

Στην περίπτωση των ερωτημάτων επιλογής, το κομμάτι εντοπισμού των στοιχείων που ικανοποιούν τα κριτήρια είναι αρκετά γρήγορο, σε αντίθεση με το κομμάτι της ανάκτησης της πλήρους δομής των στοιχείων, που επηρεάζεται από την εσωτερική πολυπλοκότητα των στοιχείων που ανακτώνται.

8 Συμπεράσματα

Όπως έγινε φανερό από τη μελέτη της βιβλιογραφίας και την ανάλυση των υπαρχόντων συστημάτων που σχετίζονται με την οργάνωση και διαχείριση της πληροφορίας του ΠΙ, οι λύσεις που προτείνονται γενικά αγνοούν ορισμένα βασικά χαρακτηριστικά του ΠΙ, όπως:

- Τη διαρκή ανανέωση των περιεχομένων του ΠΙ με έγγραφα:
    - διαφόρων τύπων (κείμενα, εικόνες, ήχοι, παρουσιάσεις, κτλ.) και μορφοποιήσεων,
    - αγνώστου περιεχομένου,
    - χωρίς μεταπληροφορίες,
    - για τα οποία δεν γνωρίζουμε εκ των προτέρων τη θεματική τους ταξινόμηση.
- Την αδυναμία ελέγχου της ορθότητας των περιεχομένων των εγγράφων, τα οποία ενδεχομένως είναι παραπλανητικά.
- Την ύπαρξη συνδέσμων μεταξύ των εγγράφων.

Το τελευταίο είναι και το χρησιμότερο ίσως στοιχείο που διαφοροποιεί τον ΠΙ από μια απλή συλλογή εγγράφων.

Οι μηχανές αναζήτησης απαντούν στις ερωτήσεις χρηστών με σύνολα σελίδων που περιέχουν τις λέξεις των ερωτήσεων. Οι αλγόριθμοι ανάκτησης πληροφορίας ταξινομούν τα αποτελέσματα με βάση την ομοιότητα του ερωτήματος στα περιεχόμενα της ερώτησης. Συχνά, τα περιεχόμενα των σελίδων είναι τέτοια ώστε να επιτυγχάνουν ψηλές θέσεις στην ταξινόμηση. Οι συγγραφείς των σελίδων μπορούν εσκεμμένα να αλλοιώσουν τα αποτελέσματα των αλγορίθμων ταξινόμησης επεμβαίνοντας στο περιεχόμενο των σελίδων.

Η σωστή οργάνωση, διαχείριση και διακομιδή του κατάλληλου περιεχομένου στους χρήστες είναι ένας από τους βασικούς στόχους της «επόμενης μέρας» του ΠΙ. Η εξυπηρέτηση των χρηστών του Σημασιολογικού Ιστού προϋποθέτει την ανάπτυξη γνωστικών λύσεων, που θα στηρίζονται σε πλούσια και αξιόπιστη πληροφορία.

Το πρώτο βήμα, για τη δημιουργία μιας πλούσιας δεξαμενής γνώσης, είναι ο χαρακτηρισμός των εγγράφων του ΠΙ με τρόπο αξιόπιστο, που δε θα επηρεάζεται από τα συμφέροντα των συγγραφέων τους. Με τη χρήση της πληροφορίας των υπερσυνδέσμων η αλλοίωση των αποτελεσμάτων είναι πολύ πιο δύσκολη. Το επόμενο βήμα είναι ο εμπλουτισμός της πληροφορίας που εξάγεται από τους υπερσυνδέσμους με σημασιολογικά χαρακτηριστικά που προκύπτουν από μια οντολογία. Η προσθήκη αυτής της πληροφορίας στα έγγραφα προσθέτει ένα επιπλέον σημασιολογικό επίπεδο αναπαράστασης των εγγράφων, περισσότερο αφηρημένο από το αντίστοιχο λεξικό επίπεδο. Η νέα διάσταση διευκολύνει και επιταχύνει τις διαδικασίες οργάνωσης και ανάκτησης πληροφορίας.

Εμπλουτίζοντας τα περιεχόμενα του ιστού με σημασιολογικά χαρακτηριστικά από τους υπερσυνδέσμους, μπορούμε να αναπτύξουμε νέες μεθόδους συγκέντρωσης και διαχείρισης των εγγράφων του ΠΙ. Οι μηχανισμοί που μπορούν να αναπτυχθούν, στοχεύουν αρχικά στη διευκόλυνση των αναζητήσεων στον Ιστό, μπορούν όμως να επεκταθούν στην εξαγωγή πολύτιμης γνώσης. Τέτοιοι μηχανισμοί είναι:

- Ομαδοποίηση των εγγράφων του ΠΙ σε ομογενή υποσύνολα. Η ομαδοποίηση βασίζεται στη σημασιολογική εγγύτητα των εγγράφων στην οντολογία (πόσο "κοντά" είναι οι έννοιες που αντιπροσωπεύουν) αλλά και στο πλήθος των κοινών εισερχόμενων και εξερχόμενων συνδέσμων. Με τον τρόπο αυτό, επεκτείνονται γνωστές αρχές της θεωρίας Ανάλυσης Συνδέσμων με σημασιολογικά χαρακτηριστικά.
- Αναζήτηση με σημασιολογικά κριτήρια. Η σχετικότητα των εγγράφων σε μια ερώτηση βασίζεται στη σημασιολογική ομοιότητα μεταξύ των όρων της ερώτησης και των όρων που απαρτίζουν την περιγραφή των εγγράφων και όχι στην απόλυτη λεξική ομοιότητα μεταξύ των αντίστοιχων όρων.
- Δομημένη παρουσίαση των αποτελεσμάτων με βάση την οντολογία. Τα έγγραφα του ΠΙ χαρακτηρίζονται με έννοιες της οντολογίας και συνεπώς μπορούν να εμφανίζονται ως στιγμιότυπα (instances) των εννοιών αυτών. Οι χρήστες μπορούν να περιηγούνται στην οντολογία εντοπίζοντας τις έννοιες που τους ενδιαφέρουν και κατ' επέκταση τα έγγραφα που σχετίζονται με αυτές.
- Εξόρυξη γνώσης, από την πληροφορία που υπάρχει στα αρχεία χρήσης (web logs) των εξυπηρετητών Ιστού, με χρήση σημασιολογικών χαρακτηριστικών. Οι υπάρχουσες μέθοδοι εξόρυξης γνώσης από στοιχεία χρήσης του ΠΙ αναλύουν τα αρχεία χρήσης για να ανακαλύψουν πρότυπα πλοήγησης μεταξύ συγκεκριμένων εγγράφων. Η προσθήκη σημασιολογικών χαρακτηρισμών στα έγγραφα του ΠΙ μπορεί να οδηγήσει στην ανακάλυψη σημασιολογικών προτύπων και κανόνων πλοήγησης, π.χ. όσοι επισκέπτονται έγγραφα σχετικά με την έννοια Χ, στη συνέχεια ενδιαφέρονται για την έννοια Υ και επισκέπτονται σχετικά έγγραφα.
- Παροχή προσωποποιημένης πληροφορίας στους χρήστες του ΠΙ. Με βάση τα εξαγόμενα πρότυπα πλοήγησης δημιουργούνται αυτόματα για τους αναγνώστες ενός εγγράφου υποδείξεις προς άλλα έγγραφα που πιθανώς τους ενδιαφέρουν. Τα σημασιολογικά πρότυπα πλοήγησης που εξάγονται από τα αρχεία χρήσης μπορούν να συνδυαστούν με το σημασιολογικό προφίλ των χρηστών (τα ενδιαφέροντά τους), για να διευκολύνουν την περιήγηση των χρηστών σε ένα τόπο του ΠΙ. Αυτοί οι μηχανισμοί μπορούν να εξάγουν ακριβέστερα συμπεράσματα για τη συμπεριφορά των χρηστών και να κάνουν καλύτερες προτάσεις προς τον αναγνώστη ενός εγγράφου, εμπλουτίζοντας τις υπηρεσίες προσωποποίησης των περιεχομένων του ΠΙ.
- Εξέλιξη των οντολογιών. Καθώς τα έγγραφα του ιστού χαρακτηρίζονται από όρους μιας οντολογίας, μπορούν κάλλιστα να αποτελέσουν στιγμιότυπα της οντολογίας. Καθώς νέα έγγραφα προστίθενται, ή παλιότερα έγγραφα αποσύρονται συνεχώς, τα περιεχόμενα της οντολογίας διαρκώς αλλάζουν. Συνεπώς πρέπει να αναπτυχθούν μηχανισμοί που θα προσαρμόζουν τη δομή της οντολογίας στις αλλαγές αυτές (π.χ. θα συγχωνεύουν, θα δημιουργούν και θα διαγράφουν έννοιες) δημιουργώντας έτσι μια συνεχώς αυξανόμενη, ενημερωμένη και δομημένη πηγή γνώσης.

8.1 Συνεισφορά του THESUS

Το THESUS εισάγει ένα μοντέλο και μια γλώσσα, που μπορούν να χρησιμοποιηθούν στην επιλογή και επεξεργασία των περιεχομένων του ΠΙ. Τα ιδιαίτερα χαρακτηριστικά του συστήματος είναι:

α) εξάγει και εκμεταλλεύεται τη σημασιολογία των συνδέσμων,
β) επιτρέπει την οργάνωση των σελίδων με βάση μια συγκεντρωτική θεώρηση της πληροφορίας των συνδέσμων.

Το υλοποιημένο σύστημα THESUS συλλέγει διευθύνσεις ιστού με βάση ένα σύνολο λέξεων (με κοινό θέμα), εξάγει πληροφορία από τους εισερχόμενους και εξερχόμενους συνδέσμους της συλλογής, αφού επεξεργαστεί την περιοχή υπερκειμένου γύρω από τον κάθε σύνδεσμο. Επιπλέον, εμπλουτίζει τη εξαχθείσα πληροφορία με σημασιολογικά χαρακτηριστικά, χρησιμοποιώντας μια οντολογία και ένα λεξιλογικό θησαυρό, και ενημερώνει μια βάση δεδομένων. Η βάση δεδομένων είναι είτε ένα σύνολο XML εγγράφων είτε μια σχεσιακή βάση.

Η προστιθέμενη πληροφορία στα έγγραφα και ο τρόπος με τον οποίο δομείται επιτρέπουν σύνθετες αναζητήσεις που δεν είναι διαθέσιμες σε άλλα συστήματα αναζήτησης στον Ιστό. Ένα υποσύστημα ομαδοποίησης ανακαλύπτει υποσύνολα σελίδων με παρόμοια σημασιολογία.

8.1.1 Για την οργάνωση των εγγράφων

Όσον αφορά την οργάνωση των εγγράφων του ΠΙ, προτείνεται ένα νέο μέτρο ομοιότητας μεταξύ εγγράφων, τα οποία αναπαρίστανται ως σύνολα όρων μιας οντολογίας με τα αντίστοιχα βάρη. Τα βάρη υποδηλώνουν τη σημαντικότητα ενός όρου στην περιγραφή του εγγράφου και τη σχετικότητά του με αυτό. Το νέο μέτρο ομοιότητας διαφέρει από τα συνήθη μέτρα ομοιότητας, που παρουσιάζονται στη βιβλιογραφία, καθώς δεν βασίζεται στην απόλυτη ομοιότητα των όρων που χαρακτηρίζουν τα δύο έγγραφα αλλά στη σχετική ομοιότητά τους.

Η σχετική ομοιότητα δύο όρων υπολογίζεται στη βάση της οντολογίας και λαμβάνει υπόψη της όχι μόνο τη συνωνυμία δύο όρων αλλά και άλλες ιεραρχικές σχέσεις, όπως τη σχέση γενίκευσης-ειδίκευσης ή τη σχέση μέρους-όλου.

Με τον ορισμό ενός μέτρου ομοιότητας μεταξύ των εγγράφων του ιστού επιτύχαμε να προσδιορίσουμε ένα χώρο απόστασης (ένα αυθαίρετο μετρικό χώρο) γι' αυτά. Στο χώρο αυτό που δεν έχει συντεταγμένες και διάταξη εγγράφων χρησιμοποιήσαμε δύο διαφορετικούς αλγορίθμους συσταδοποίησης: έναν ιεραρχικό αυξητικό αλγόριθμο (COBWEB) και έναν αλγόριθμο πυκνότητας (DBSCAN), με τις κατάλληλες τροποποιήσεις και προσαρμογές. Επιτύχαμε να χωρίσουμε το σύνολο των εγγράφων σε ομάδες που περιέχουν έγγραφα που είναι σημασιολογικά παρόμοια.

Το αποτέλεσμα επιβεβαιώθηκε με μια σειρά πειραμάτων με πραγματικά δεδομένα, που έδειξαν ότι το μέτρο ομοιότητας του THESUS:

- ξεπερνά σε απόδοση άλλα μέτρα που βασίζονται στην απόλυτη ομοιότητα όρων,
- δίνει συμπαγείς ομάδες αρκετά όμοιες σε σύσταση με αυτές που δίνουν χειροκίνητα οι «ειδικοί» ενός χώρου.

Στα πλαίσια της οργάνωσης των εγγράφων του ΠΙ με έμφαση στην πληροφορία των υπερσυνδέσμων αναπτύχθηκε η γλώσσα του THESUS που προσφέρει μια σειρά χρήσιμων τελεστών. Με τη χρήση των υλοποιημένων τελεστών της γλώσσας μπορεί

κανείς να εντοπίσει έγγραφα ή ομάδες εγγράφων του ΠΙ με υψηλή συνδεσιμότητα και συγκεκριμένα λεξικά και σημασιολογικά χαρακτηριστικά.

8.1.2 Για την απεικόνιση XML σε σχεσιακή βάση

Στην προσπάθεια δημιουργίας ενός μοντέλου για την αποθήκευση των στοιχείων που εξάγονται από έγγραφα και υπερσυνδέσμους επιλέχθηκε η γλώσσα XML. Η XML είναι η γλώσσα των εγγράφων που παράγει το THESUS με την εξαγόμενη πληροφορία. Τα έγγραφα αυτά έχουν συγκεκριμένη δομή που περιγράφεται με τη γλώσσα XML-Schema.

Το σύνολο των παραγόμενων εγγράφων μπορεί από μόνο του να αποτελέσει μια βάση δεδομένων, όταν όμως πρόκειται να γίνει σύνθετη επεξεργασία των δεδομένων αυτών, είναι προτιμότερη (για λόγους απόδοσης) η χρήση μιας σχεσιακής βάσης δεδομένων. Για το λόγο αυτό αναπτύχθηκε στα πλαίσια του THESUS ένα σύστημα απεικόνισης των δομών της XML σε σχεσιακές βάσεις δεδομένων, το σύστημα X-Database. Το σύστημα επιτρέπει την αποθήκευση, διαχείριση και ανάκτηση XML εγγράφων από τη σχεσιακή βάση μέσα από ένα μηχανισμό εντολών που χρησιμοποιεί παρόμοια σύνταξη με αυτή των εγγράφων. Με τον τρόπο αυτό δεν χρειάζεται να έχουμε κάποια ιδιαίτερη εξοικείωση με το σχεσιακό σχήμα, παρά μόνο με τη δομή των εγγράφων XML που ανταλλάσσουμε με αυτό.

Η εκφραστικότητα της XML-Schema στην περιγραφή της δομής των εγγράφων χρησιμοποιείται για την αυτόματη παραγωγή της σχεσιακής βάσης. Από την ανάλυση του XML-Schema το σύστημα αποκτά σημαντική γνώση τόσο για τη δημιουργία της βάσης και των κατάλληλων μηχανισμών ελέγχου της εγκυρότητας των δεδομένων της, όσο και για την επεξεργασία των XML εγγράφων που διαχειρίζεται το σύστημα.

Η προσφορά του συστήματος X-Database στο ερευνητικό πεδίο της σύνδεσης XML και σχεσιακών Βάσεων Δεδομένων συνοψίζεται στα εξής:

- αυτόματη δημιουργία του σχήματος της βάσης δεδομένων από το XML-Schema,
- αυτόματη απεικόνιση των εγγράφων στη βάση, μέσα από ένα σύνολο κανόνων απεικόνισης. Το σύστημα χρησιμοποιεί κάποιους κανόνες απεικόνισης που ήδη έχουν οριστεί στη βιβλιογραφία για τις απλές περιπτώσεις. Επιπλέον, εντόπισε και προτείνει λύσεις σε κάποια προβλήματα απεικόνισης που μέχρι τώρα δεν είχαν αντιμετωπιστεί.
- καθορισμός ενός μηχανισμού περιγραφής σύνθετων μηχανισμών της σχεσιακής βάσης δεδομένων μέσα από την XML-Schema, όπως ευρετήρια, ρόλους και δικαιώματα κτλ.
- αυτόματη προσθήκη μηχανισμών ορθότητας των δεδομένων που αποθηκεύονται και ανακτώνται από τη βάση, σύμφωνα με το XML-Schema.

8.2 Μελλοντικές βελτιώσεις

8.2.1 Για την οργάνωση των εγγράφων

Η μελλοντική εργασία θα επικεντρωθεί στα εξής:

Βελτίωση της διαδικασίας απεικόνισης λέξεων σε έννοιες της οντολογίας. Η διαδικασία αυτή είναι η σημαντικότερη και ταυτόχρονα δυσκολότερη από όλες τις υπόλοιπες στο σύστημα. Η μόνη γνώση που διαθέτουμε είναι το Wordnet, το οποίο

δεν επαρκεί σε περιπτώσεις πολύ συγκεκριμένων θεματικών περιοχών. Για το λόγο αυτό, μελετάται η χρήση της πληροφορίας που φέρει η ίδια η θεματική οντολογία, αν και προς το παρόν οι υπάρχουσες οντολογίες είναι πολύ φτωχές. Επίσης, για την αποσαφήνιση της έννοιας των λέξεων και για την καλύτερη απεικόνιση στην οντολογία εξετάζεται και η χρήση συντακτικής πληροφορίας, που μπορεί να εξαχθεί από τα ίδια τα έγγραφα ή τους υπερσυνδέσμους. Αυτό βέβαια απαιτεί ποιοτικότερη επεξεργασία των περιεχομένων και αυξάνει την πολυπλοκότητα.

Σημασιολογική περιήγηση του ΠΙ. Το υποσύστημα περιήγησης και συλλογής εγγράφων, προς το παρόν, συλλέγει όσα έγγραφα υποδεικνύονται μέσω συνδέσμων που περιέχουν ακριβώς κάποια από τις λέξεις που μας ενδιαφέρουν. Είναι χρήσιμο να εμπλουτίσουμε τις αρχικές λέξεις με σημασιολογικά χαρακτηριστικά και να ακολουθούμε συνδέσμους που θα χρησιμοποιούν κάποια από αυτά και όχι ακριβώς τις λέξεις που αρχικά δόθηκαν. Ήδη γίνεται προσπάθεια ώστε το υποσύστημα περιήγησης να παρέχει και μια πρώιμη ***ταξινόμηση των σελίδων*** με βάση τη σημαντικότητά τους όπως αυτή υπολογίζεται στη σημασιολογία και στο πλήθος των εισερχόμενων συνδέσμων. Εξετάζεται επίσης η δυνατότητα ενημέρωσης της θεματικής περιοχής με νέες λέξεις και έννοιες που εντοπίζονται κατά τη διάρκεια της συλλογής των εγγράφων.

Χρήση άλλων αλγορίθμων συσταδοποίησης. Τα πειράματα που έγιναν με τον αλγόριθμο DBSCAN έδωσαν χρήσιμα αποτελέσματα, συγκρίσιμα με την ταξινόμηση από ανθρώπους. Παρ' όλα αυτά, λόγω της φύσης του αλγορίθμου πολλά έγγραφα χαρακτηρίζονται θόρυβος και δεν ομαδοποιούνται. Επιπλέον, καθώς ο DBSCAN δεν είναι αυξητικός αλγόριθμος, η διαδικασία ομαδοποίησης πρέπει να επαναληφθεί για όλο το σύνολο των εγγράφων αν προστεθούν νέα έγγραφα. Τα πειράματα που έγιναν με τον αλγόριθμο COBWEB έδωσαν ακόμη καλύτερα αποτελέσματα, αν και η ταχύτητα του αλγορίθμου είναι πολύ μικρή. Ο απώτερος στόχος είναι να μπορούμε να ανακαλύπτουμε συμπαγείς ομάδες σελίδων που είναι σημασιολογικά κοντά και σχηματίζουν ομάδες, με άλλα λόγια να ανακαλύπτουμε Θεματικά Υποσύνολα (THESUs) αντί να τα κατασκευάζουμε με χρήση των τελεστών του THESUS.

Ανάπτυξη ενός νέου αλγορίθμου συσταδοποίησης εγγράφων ΠΙ. Η ανάπτυξη ενός αλγορίθμου που θα προσαρμοστεί στη δομή των εγγράφων του ΠΙ και στις ανάγκες για οργάνωσή τους είναι το επόμενο βήμα για το σύστημα. Οι αλγόριθμοι γράφων δείχνουν να ταιριάζουν πολύ καλά με τη φύση του προβλήματος συσταδοποίησης εγγράφων του ΠΙ, καθώς τα έγγραφα μιας συλλογής αποτελούν από μόνα τους ένα γράφο με ακμές τους μεταξύ τους υπερσυνδέσμους. Η δημιουργία ενός εικονικού γράφου, που θα βασίζεται στη σημασιολογική ομοιότητα των κόμβων και όχι στη φυσική τους σύνδεση, αναμένεται να δώσει μια νέα διάσταση στο πρόβλημα της συσταδοποίησης με χρήση γράφων.

Γραφική διεπαφή αποτελεσμάτων συσταδοποίησης. Ο τρόπος με τον οποίο παρουσιάζεται η γνώση είναι πολλές φορές σημαντικότερος από την ίδια τη γνώση. Αυτό συμβαίνει και στην περίπτωση του ΠΙ, όπου οι χρήστες πρέπει να εντοπίζουν τα έγγραφα που τους ενδιαφέρουν εύκολα και γρήγορα. Η παρουσίαση των αποτελεσμάτων της συσταδοποίησης με μια ιεραρχική μορφή είναι ένα βήμα προς το στόχο αυτό. Ένας διαφορετικός τρόπος αναπαράστασης που θα ταίριαζε περισσότερο στο πνεύμα του Σημασιολογικού Ιστού είναι με τη μορφή ενός γράφου εννοιών (παρόμοιου με αυτόν της οντολογίας) που στους κόμβους του, κάτω από τις έννοιες

της οντολογίας, μπορεί κανείς να βρει τα έγγραφα που τον ενδιαφέρουν. Αυτός και άλλοι εναλλακτικοί τρόποι αναπαράστασης της συσταδοποίησης των εγγράφων εξετάζονται.

Σύνθετη ταξινόμηση των αποτελεσμάτων ερωτήσεων. Ένα από τα επόμενα βήματα περιλαμβάνει την ανάπτυξη ενός μηχανισμού ερωτήσεων που θα αξιοποιεί τα αποτελέσματα της συσταδοποίησης. Η ταξινόμηση των αποτελεσμάτων πρέπει να λαμβάνει υπόψη της δύο διαφορετικά θέματα: α) τη σημαντικότητα του εγγράφου με βάση τους συνδέσμους, β) τη σημασιολογική ομοιότητα του εγγράφου με την ερώτηση.

Η θέση των συνδέσμων στη σελίδα. Έπειτα από πειράματα διαπιστώθηκε ότι η θέση του συνδέσμου στο πηγαίο έγγραφο δίνει αρκετά χρήσιμη πληροφορία για τη σημαντικότητα του συνδέσμου αλλά και για τη σημασιολογία του. Για παράδειγμα, πολλοί σύνδεσμοι με παρόμοια σημασιολογία περιέχονται σε μια λίστα (<LI></LI>). Στην περίπτωση αυτή το «παράθυρο» γύρω από τους συνδέσμους πρέπει να υπερβαίνει τα όρια της λίστας. Επίσης, στις περιπτώσεις κοινών αναφορέων οι ίδιοι σύνδεσμοι βρίσκονται κοντά σε όλους τους αναφορείς.

Μελέτη του συνδυασμού των τελεστών του THESUS. Είναι πολύ σημαντικό να μελετήσουμε τους πιθανούς συνδυασμών των τελεστών, να ερμηνεύσουμε τα αποτελέσματα και να αξιολογήσουμε τη χρησιμότητά τους. Η βελτιστοποίηση των υπαρχόντων τελεστών αποτελεί ένα ακόμη θέμα μελλοντικής έρευνας.

8.2.2 Για την απεικόνιση XML σε σχεσιακή βάση

Η μελλοντική ερευνητική δουλειά θα εστιαστεί σε τρεις τομείς:
Στην απεικόνιση όλων των δομών της XML-Schema στο σχεσιακό σχήμα. Η παρούσα προσπάθεια έχει καλύψει ένα μεγάλο ποσοστό των δυνατοτήτων της γλώσσας XML-Schema (πάνω από 90%) και των υποστηριζόμενων δηλώσεων και δομών. Οι υπόλοιπες δυνατότητες, που δεν καλύφθηκαν, είτε γιατί δεν υπήρξε ανάγκη στο σύστημα THESUS είτε λόγω της μεγάλης πολυπλοκότητας που παρουσιάζουν (π.χ. η ανοικτή δομή της XML-Schema επιτρέπει σε ένα στοιχείο να περιέχει υποστοιχεία οποιουδήποτε τύπου – δήλωση xsd:any), αναμένεται να καλυφθούν στα επόμενα στάδια.

Στη βελτίωση της διαδικασίας ανάκτησης εγγράφων. Η μετάβαση από τη γλώσσα ερωτήσεων του X-Database στη γλώσσα XQuery είναι το πρώτο βήμα, όσον αφορά στις υποστηριζόμενες λειτουργίες. Σε επόμενο στάδιο εξετάζεται η υποστήριξη όλων των δυνατοτήτων της γλώσσας XQuery.

Στην υποστήριξη ενημερώσεων στο XML-Schema. Το μοντέλο απεικόνισης προϋποθέτει ότι το XML-Schema ορίζεται εφάπαξ πριν τη δημιουργία της βάσης και δεν αλλάζει ποτέ (κάτι που συμβαίνει στο XML-Schema του THESUS και του MPEG-7). Σε πολλές περιπτώσεις όμως, το σχήμα των εγγράφων XML μπορεί να αλλάζει με το χρόνο. Με τη χρήση μετασχηματισμών στα XML έγγραφα η μετάβαση στο νέο σχήμα είναι εύκολη. Δε συμβαίνει όμως το ίδιο και στην περίπτωση του σχεσιακού σχήματος. Χρειάζεται επιπλέον προσπάθεια για την επίλυση των προβλημάτων που συνεπάγεται η ενημέρωση του XML-Schema στο σχεσιακό σχήμα.

Βιβλιογραφία

[AA+01] E. Agirre, O. Ansa, D. Martinez, E. Hovy, 'Enriching Wordnet concepts with topic signatures', Proceedings of the NAACL worshop on Wordnet and Other Lexical Resources: Applications, Extensions and Customizations. Pittsburg, USA (2001)

[AB+98] G. Alber, C. Beeri, T.Milo, Y.Sagiv, O.Shmuei, T. Tishbi, D. Konopniki, P. Mogilevski, 'WebSuite - A tool for harnessing web data'. WebDB, (1998).

[AB+99] M. Ankerst, M. M. Breunig, H. P. Kriegel, J. Sander, 'OPTICS: Ordering Points To Identify the Clustering Structure'. SIGMOD Conference, pages 49-60, (1999).

[AC+00] V. Aguilera, S. Cluet, P. Veltri, D. Vodislav, F. Wattez. 'Querying XML documents in Xyleme'. In Proc. of the ACM SIGIR, (2000).

[AF+95] R. Armstrong, D. Freitag, T. Joachims, T. Mitchell, 'Web Watcher: A learning apprentice for the World Wide Web', Proceedings of AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments, pages 6--12. AAAI Press, (1995)

[AG+99] C. Aggarwal, S. Gates, P. Yu, 'On the merits of building categorization systems by supervised clustering', Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Pages 352 - 356, (1999).

[AM+97] P. Atzeni, G. Mecca, P. Merialdo. 'To Weave the Web'. In Proceedings of the 23rd International Conference on Very Large Databases, (1997)

[AM98] G. Arocena, A. Mendelzon, 'WebOQL: Restructuring documents, databases and webs', in ICDE'98, Orlando, Florida, (1998)

[AP+03] S. Abiteboul, M. Preda, G. Cobena, 'Adaptive on-line page importance computation'. WWW Conference, 280-290, (2003)

[AS00] Y. Avrithis, G. Stamou, 'Feathon: unified intelligent access to heterogeneous audiovisual content'. In VLBVO1, International Workshop on Very Low Bitrate Video Coding: Visual Content Representation and Analysis, October 11-12, (2000)

[AV+01] S. Ailleret, P Veltri, L. Mignet, V. Aguilera, 'Xyro : The xyleme robot architecture'. First DIWeb Workshop, (2001).

[B01] R. Bourret, 'XML and Databases', http://www.rpbourret.com/xml/XMLAndDatabases.htm, (2001)

[B93] R.A. Botafogo, 'Cluster analysis for hypertext systems'. Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval,116-125, (1993)

[BC+92] B. T. Bartell, G. W. Cottrell, R. K. Belew, 'Latent semantic indexing is an optimal special case of multidimensional scaling', Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval, 161-167., (1992)

[BE+84] J.C. Bezdek, R. Ehrlich, W. Full, 'FCM: Fuzzy C-Means Algorithm'. Computers and Geosciences (1984)

[BFF02] A. Bidault, B.Safar, Ch. Froidevaux, 'Proximite entre requetes dans un contexte mediateur', RFIA-2002, p653-662, (2002).

[BF+02] P. Bohannon, J. Freire, P. Roy, J. Simeon, 'From XML Schema to Relations: A Cost-Based Approach to XML Storage', Proceedings of ICDE (2002).
[BG+99] D. Boley, M. Gini, R. Gross, E.H. Han, K. Hastings, G. Karypis, V. Kumar, B. Mobasher, J. Moore. 'Partitioning-based clustering for web document categorization'. Decision Support Systems, 27(3),329-341, (1999)
[BH+01] T. Berners-Lee, J. Hendler, O. Lassila, 'The Semantic Web, A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities', Scientific American, May 2001.
[BH+99] P. Bra, G. Houben, H. Wu, 'AHAM: A dexter-based reference model for adaptive hypermedia' In Proc. of ACM Conference on Hypertext and Hypermedia, pages 147-156, (1999).
[BK+00] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. 'Graph structure in the Web'. 9th International World Wide Web Conference, Amsterdam, The Netherlands, May (2000).
[BK+90] N. Beckmann, H.P. Kriegel, R. Schneider, B. Seeger: 'The R*-Tree: An Efficient and Robust Access Method for Points and Rectangles'. SIGMOD Conference (1990).
[BL96] M. Berry, G. Linoff, ‘Data Mining Techniques For marketing, Sales and Customer Support’, John Willey & Sons, Inc, (1996).
[BP+91] P. Brown, S. Della Pietra, V. Della Pietra, R. Mercer, 'Word sense disambiguation using statistical methods', In Proceedings of ACL 264-270, (1991)
[BP98] S. Brin and L. Page. 'The anatomy of a large-scale hypertextual Web search engine'. In 7th International World Wide Web Conference, Brisbane, Australia, (1998)
[BS91] R.A. Botafogo, B. Shneiderman, 'Identifying aggregates in hypertext structures'. Proceedings of the 3rd ACM Conference on Hypertext,63-74, (1991)
[C96] J. Carletta, ‘Assessing agreement on classification tasks: the kappa statistic‘,Computational Linguistics 22(2):249-254, (1996).
[CB+99] S. Chakrabarti, M. van der Berg, B. Dom, 'Focused crawling: A new approach to topic-specific web resource discovery'. Computer Networks, 31(11-16), (1999).
[CC+03] N. Catala, N. Castell, M. Martin, ‘A portable method for acquiring information extraction patterns without annotated corpora’, Natural Language Engineering 9 (2), pp. 151-179, (2003).
[CD+98a] S. Chakrabarti, B. Dom, D. Gibson, J. Keinberg, P. Raghavan, S. Rajagopalan, 'Automatic Resource list Compilation by Analyzing Hyperlink Structure and Associated Text', 7th International WWW Conference (1998)
[CD+98b] S. Chakrabarti, B. Dom, P. Indyk, 'Enhanced hypertext categorization using hyperlinks', In Proceedings of the ACM SIGMOD International Conference on Management of Data, pages 307-318, (1998).
[CD+99] S. Chakrabati, B. Dom, D. Gibson, J. Kleinberg, S. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins. 'Mining the link structure of the world wide web'. IEEE Computer, 32(8):60-67, (1999).
[CG+97] C. Chekuri, M. Goldwasser, P. Raghavan, E. Upfal, 'Web Search Using Automatic Classification', 6th International Conference on the WWW, (1997).

[CK+00] M. Carey, J. Kierman, J. Shanmugasundaram, E. Shekita, S. Subramanian. 'XPERANTO: Middleware for Publishing Object Relational Data as XML Documents'. Proceedings of VLDB Conference, (2000).
[CK+92] D. Cutting, D. Karger, J. Pedersen, J. Tukey, 'Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections', SIGIR '92, Pages 318 - 329, (1992).
[CM+98] J. Cho, H. Garcia-Molina, L. Page, 'Efficient crawling through URL ordering'. In Proceedings of the Seventh International WWW Conference, pages 161-172, (1998).
[CM00] J. Cho, H. Garcia-Molina, 'Estimating Frequency of Change'. Technical report, Stanford University, (2000).
[CS96] P. Cheeseman, J. Stutz, 'Bayesian Classification (AutoClass): Theory and Results', Advances in Knowledge Discovery and Data Mining, AAAI/MIT Press, 153-180, (1996)
[CS97] M. Sintichakis, P. Constantopoulos, 'A Method for Monolingual Thesauri Merging', SIGIR (1997).
[D01] I. S. Dhillon. 'Co-clustering documents and words using bipartite spectral graph partitioning'. In Proceedings of The 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining ,(2001).
[D77] D. Defays, 'An efficient algorithm for the complete link method', The Computer Journal 20, 364-366, (1977)
[D94] S.T. Dumais, 'Latent Semantic Indexing (LSI) and TREC-2', In D. K. Harman (Ed.),Proceedings of the 2nd Text REtrieval Conference (TREC-2), 105-115, (1994).
[DAML] DARPA Agent Markup Language Program, www.daml.org
[DBLP] M. Ley, 'DBLP Computer Science Bibliography', http://dblp.uni-trier.de/
[DC00] S. Dumais, H. Chen, 'Hierarchical Classification of Web Content', Proceedings of SIGIR-00, 23rd ACM International Conference on Research and Development in Information Retrieval, (2000)
[DD+90] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, 'Indexing by Latent Semantic Analysis'. Journal of the American Society for Information Science, 41, 391-407, (1990).
[DF+98] A. Deutsch, M. Fernandez, D. Florescu, A. Levy, D. Suciu, 'XML-QL: A Query Language for XML', Submission to the World Wide Web Consortium, http://www.w3.org/TR/NOTE-xml-ql/, (1998)
[DF+99] A. Deutsch, M. Fernandez, D. Suciu. 'Storing Semistructured Data with STORED'. In Proc. of ACM SIGMOD International Conference on Management of Data, pages 431-442, (1999).
[DJ02] E. Desmontils, C. Jacquin. 'Indexing a Web Site with a Terminology Oriented Ontology'. In I.F. Cruz, S. Decker, J. Euzenat et D.L. McGuinness, eds., The Emerging Semantic Web. IOS Press, pages 181-198, (2002).
[DL+77] A.P. Dempster, N.M. Laird, D.B. Rubin, 'Maximum Likelihood from Incomplete Data via theEM algorithm'. Journal of the Royal statistical Society, Series B, 39(1): 1-38, (1977).
[DL01] M. Drake, M. Lehmann, 'Understanding the Oracle9i XMLType', Oracle Magazine November, http://www.oracle.com/oramag/oracle/01-nov/index.html?o61xml.html, (2001),
[DM01] N. Day, J. M. Martinez, 'Introduction to MPEG-7 (v3.0)', http://ipsi.fraunhofer.de/delite/Projects/MPEG7/Documents.html (2001)

[Dmo] ODP - Open Directory Project, http://dmoz.org/
[E99] Eikvil, L.: Information Extraction from World Wide Web - A Survey, Report No. 945, ISBN 82-539-0429-0, July (1999)
[EH81] B.S. Everitt, D.J. Hand, 'Finite Mixture Distributions'. London: Chapman and Hall (1981)
[EK+96] M. Ester, H.P. Kriegel, J. Sander, X. Xu, 'A density based algorithm for discovering clusters in large spatial databases with noise', 2nd International conference on Knowledge Discovery and Data Mining ACM-SIGKDD, (1996).
[EK+98] M. Ester, H.P. Kriegel, J. Sander, M. Wimmer and X. Xu, 'Incremental Clustering for Mining in a Data Warehousing Environment', Proceedings of the 24th VLDB Conference (1998).
[EM97] T. Eiter, H. Mannila, 'Distance measures for point sets and their computation', in Acta Informatica Journal, 34, (1997)
[F87] D. Fisher, 'Knowledge acquisition via incremental conceptual clustering'. Machine learning,2,139-172.(1987)
[F96] D. Fisher, 'Iterative optimization and simplification of hierarchical clustering', Journal of Artificial Intelligence Research, 4, 147-179, (1996).
[FF+97] M. Fernandez, D. Florescu, A. Levy, D. Suciu, 'A query language for a web-site management system', SIGMOD Record, 26(3),(1997)
[FK99] D. Florescu, D. Kossmann. 'A performance evaluation of alternative mapping schemes for storing XML data in a relational database'. Technical Report, INRIA, France, 1999.
[FL+98] D. Florescu, A. Levy, A. Mendelzon. 'Database techniques for the World Wide Web : a survey', ACM SIGMOD Record, 27(3):59-74, (1998).
[FT+00] M. Fernandez, W. Tan, D. Suciu, 'SilkRoute: Trading Between Relations and XML', WWW Conference, (2000).
[G03] Google Inc, Google advanced search, http://www.google.com/advanced_search, (2003)
[G72] E. Garfield. 'Citation analysis as a tool in journal evaluation'. Science, 178, (1972).
[G79] E. Garfield. 'Citation Indexing'. ISI Press, (1979).
[G93] T. R. Gruber. 'A translation approach to portable ontologies'. Knowledge Acquisition, 5(2):199-220, (1993).
[GC+93] W. Gale, K. Church, D. Yarowsky, 'A method for disambiguating word senses in a large corpus', Computers and Humanities 26, 415-439 (1993)
[GG+01] A. Gionis, D. Gunopulos, N. Koudras, 'Efficient and Tunable Similar Set Retrieval', ACM-SIGMOD (2001).
[GH+96] J. Green, N. Horne, E. Orlowska and P. Siemens, 'A Rough Set Model of Information Retrieval', Theoretica Infomaticae 28, pages 273-296 (1996).
[GH+99] R. Goldman, J. McHugh, J. Widom. 'From semistructured data to XML: Migrating the lore data model and query language'. In Proceedings of the 2nd International Workshop on the Web and Databases (WebDB '99), pages 25--30, (1999).
[GR+99] V. Ganti, R. Ramakrishnan, J. Gehrke, A. Powell and J. French. Clustering large datasets in arbitrary metric spaces. In the Proceedings of International Conference on Data Engineering, (1999).

[GT+02] E. J. Glover, K. Tsioutsiouliklis, S. Lawrence, D. M. Pennock, G. W. Flake. 'Using Web Structure for Classifying and Describing Web Pages', WWW Conference (2002).
[Goo] The Google Directory, http://directory.google.com
[H00] M. Henzinger, 'Link Analysis in Web Information Retrieval', Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, (2000).
[Hd01] J. Hendler, 'Agents and the Semantic Web', IEEE Intelligent Systems Journal, March-April (2001)
[Hz01] M. Henzinger, 'Hyperlink Analysis for the Web', IEEE Internet Computing, Jan.Feb.(2001)
[H99] P. Hansen, 'User Guidelines for Dublin Core Creation', http://www.sics.se/~preben/DC/DC_guide.html, (1999)
[HB+98] E.H. Han, D. Boley, M. Gini, R. Gross, K. Hastings, G. Karypis, V. Kumar, B. Mobasher, J. Moore. 'WebACE: a web agent for document categorization and exploration'. In the Proceedings of the 2nd International Conference on Autonomous Agents (Agents'98), May 1998.
[HB+01] M. Halkidi, Y. Batistakis, M. Vazirgiannis, 'On Clustering Validation Techniques', Intelligent Information Systems Journal, Kluwer Pulishers, 2001
[HG+02] T. H. Haveliwala, A. Gionis, D. Klein, P. Indyk. 'Evaluating Strategies for Similarity Search on the Web'. WWW Conference, (2002).
[HG+97] J. Hammer, H. Garcia-Molina, S. Nestorov, R. Yerneni, M. M. Breunig, V. Vassalos.'Template-based wrappers in the TSIMMIS system'. In ACM SIGMOD, (1997).
[HK98] A. Hinneburg, D. Keim, 'An efficient approach to clustering in large multimedia database with noise'. Proc. 4th Int. Con. on Knowledge discovery and data mining, (1998).
[HM+01] A. Hotho, A. Maedche, S. Staab, R. Struder, ' SEAL-II The Soft Spot between Richly Structured and Unstructured Knowledge', Journal of Universal Computer Science, (2001).
[HV01] M. Halkidi, M. Vazirgiannis, 'A Data Set Oriented Approach for Clustering Algorithm Selection'. PKDD (2001).
[HW86] A. El-Hamdouchi, P. Willett, 'Hierarchic document clustering using Ward's method'. Proceedings of the Ninth International Conference on Research and Development in Information Retrieval. ACM, 149-156, (1986)
[HW89] A. El-Hamdouchi, P. Willett, 'Comparison of hierarchic agglomerative clustering methods for document retrieval'. The Computer Journal 32 (1989)
[IBM02] IBM, 'IBM DB2 XML Extender, Add-on Overview', http://www-3.ibm.com/software/data/db2/extenders/xmlext/, (2002)
[IG02] P.G. Ipeirotis, L. Gravano, 'Distributed Search over the Hidden Web: Hierarchical Database Sampling and Selection'. In Proc' of VLDB 02, China, 2002.
[ISOa] ISO/IEC JTC1/SC29/WG11 N4032, 'Introduction to MPEG-7', March 2001, Singapore.
[ISOb] ISO/IEC JTC 1/SC 29/WG 11/N3966, 'Text of 15938-5 FCD Information Technology - Multimedia Content Description Interface - Part 5 Multimedia Description Schemes', March 2001, Singapore.

[ISO 5964:1985] Documentation, Guidelines for the establishment and development of multilingual thesauri (1985)
[ISO 5963:1985] Documentation, Methods for examining documents, determining their subjects, and selecting indexing terms (1985)
[ISO 2788:1986] Documentation, Guidelines for the establishment and development of monolingual thesauri 2nd ed., (1986)
[IV98] N. Ide, J. Veronis. 'Introduction to the special issue on word sense disambiguation: The state of the art', Computational Linguistics, 24(1):1--40, (1998).
[JC97] J. Jiang, D. Conrath, 'Semantic similarity based on corpus statistics and lexical taxonomy'. In Proceedings on International Conference on Research in Computational Linguistics, (1997)
[JM+99] A.K. Jain, M.N. Murty, P.J. Flyn, 'Data Clustering: A Review'. ACM Computing Surveys, Vol. 31, No. 3 (1999)
[K63] M. M. Kessler. 'Bibliographic coupling between scientific papers'. American Documentation, 14, (1963).
[Ko97] G. Kowalski, 'Information Retrieval Systems - Theory and Implementation', Kluwer Academic Publishers, (1997).
[K98] T.Kohonen, 'Self-organization of very large document collections: State of the art'. In L. Niklasson, M. Boden, T. Ziemke editors, Proceedings of ICANN98, the 8th International Conference on Artificial Neural Networks, volume 1, pages 65-74, (1998)
[K99] J. Kleinberg. 'Authoritative sources in a hyperlinked environment'. In Journal of the ACM, 46(5):604-632, (1999).
[Ku97] N. Kushmerick, 'Wrapper Induction for Information Extraction', Intl. Joint Conference on Artificial Intelligence IJCAI, (1997).
[KE+01] C. Kwok, O. Etzioni, D. Weld, 'Scaling Question Answering to the Web', WWW10, Hong Kong, (2001).
[KE+99] R. Kreutz, B. Euler, K Spitzer, 'No longer lost in WWW-based Hyperspaces', ACM Hypertext, Darmstadt, Germany, (1999).
[KH+99] G. Karypis, E.H. Han, V. Kumar, 'CHAMELEON: A Hierarchical Clustering Algorithm Using Dynamic Modelling'. IEEE Computer, Vol. 32, No. 8, 68-75, (1999)
[KL03] B. Katz, J. Lin, 'Selectively Using Relations to Improve Precision in Question Answering', Proceedings of the EACL-2003 Workshop on Natural Language Processing for Question Answering, (2003).
[KM00] C. Kanne, G. Moerkotte, 'Efficient Storage of XML Data'. Proc. Of the 16th Int. Conf. On Data Engineering, ICDE (2000).
[KM99] T. Kopetzky, M. Muhlhauser, 'Visual Preview for Link Traversal on the WWW', 8th International World Wide Web Conference, Toronto, Canada, (1999).
[KR+99] R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, 'Trawling the Web for emerging cybercommunities'. In Proc. of the 8th Intl. WWW Conf., (1999).
[KS+00] G. Karypis, M.G. Steinbach, V. Kumar, 'A Comparison of Document Clustering Techniques'. KDD Workshop on Text Mining (2000)
[KS97] D. Koller, M. Sahami, 'Hierarchically classifying documents using very few words', Proceedings of the 14th International Conference on Machine Learning (ML), Nashville, Tennessee, (1997), Pages 170-178.
[Kar01] The Kartoo System: http://www.kartoo.fr

[L68] J.B. Lovins. 'Development of a Stemming Algorithm'. In Mechanical Translation and Computational Linguistics, 11(1-2), 11-31, (1968)
[L86] M. Lesk, 'Automatic sense disambiguation. How to tell a pine cone from an ice cream cone', In Proceedings of the 1986 SIGDOC Conference, Association for Computing Machinery, (1986)
[L95] K. Lang, 'Newsweeder: Learning to Filter Netnews', Proceedings of the 12th Int. Conference on Machine Learning, pp. 331–339, Stanford, California, (1995).
[L96] R.R. Larson, 'Bibliometrics of the World Wide Web: An Exploratory Analysis of the Intellectual Structure of Cyberspace'. Proceedings of the 1996 American Society for Information Science Annual Meeting (1996)
[L98] D. Lin, 'An Information-Theoretic Definition of Similarity', in proceedings of 15th ICML (1998).
[LA99] B. Larsen, C. Aone, 'Fast and Effective Text Mining Using Linear-time Document Clustering', KDD-99, San Diego, California, (1999).
[LC+98] C. Leacock, M. Chodorow, G. A. Miller. 'Using corpus statistics and Wordnet relations for sense identification'. Computational Linguistics, Volume 24, (1998)
[LG01] O. Lassila, D. McGuinness, 'The Role of Frame-Based Representation on the Semantic Web'. In Electronic Transactions on Artificial Intelligence Volume 5, (2001).
[LM00] A. Laux, L. Martin, XUpdate working draft. http://www.xmldb.org/xupdate/xupdate-wd.html, (2000)
[LS69] M.E. Lesk, G. Salton, 'Relevance assessments and retrieval system evaluation', Information Storage and Retrieval, 4, 343-359 (1969).
[LU+02] G. Li, V. Uren, E. Motta, S. B. Shum, J. Domingue, 'ClaiMaker: Weaving a Semantic Web of Research Papers', ISWC 2002, LNCS 2342, pp. 436-441, (2002).
[M75] M. Minsky, 'A Framework for Representing Knowledge', in Patrick Henry Winston (ed.), The Psychology of Computer Vision, McGraw-Hill, New York, (1975).
[M86] Γ. Μπαμπινιώτης, Θεωρητική Γλωσσολογία, 1986, Εκδ. Δ. Μαυρομάτη.
[M83] F.Murtagh, 'A survey of recent advances in hierarchical clustering algorithms'. The Computer Journal 26, 354-359,(1983)
[M+93] G. Miller et al, 'Five papers on Wordnet', (1993). Available at Wordnet site: http://www.cogsci.princeton.edu/~wn/
[M98] D. Merkl, 'Text Data Mining'. In R. Dale, H. Moisl, H. Somers (eds.), A handbook of natural language processing: techniques and applications for the processing of language as text, Marcel Dekker, New York (1998)
[MA97] G. Mecca, P. Atzeni, 'Cut and Paste', Journal of Computing and System Sciences, Special Issue on PODS'97, (1997)
[MC+91] G.A. Miller, W.G. Charles, 'Contextual correlates of semantic similarity', Language and Cognitive Processes 6(1), 1-28, (1991)
[MM+01] I. Muslea, S. Minton, C. Knoblock, 'Hierarchical Wrapper Induction for Semistructured Information Sources', Journal of Autonomous Agents and Multi-Agent Systems, 4:93-114, (2001) .
[MM+97] A. Mendelzon, G. A. Mihaila, T. Milo, 'Querying the World Wide Web', Journal of Digital Libraries, 1:68-88, (1997).

[MW+72] J. Minker, G.A. Wilson, B.H. Zimmerman, 'An evaluation of query expansion by the addition of clustered terms for a document retrieval system', Information Storage and Retrieval, 8, 329-348 (1972).
[N87] I. Niiniluoto, 'Truthlikeness', D. Reidel Pub. Comp., Dordrecht, Holland 1987
[Nor01] The Northern line search engine: http://www.northernlight.com
[OC97] Oracle Corporation, Oracle Call Interface Programmer's Guide, Volumes 1 & 2 Release 8.0. (1997)
[ODI01] Object Design Inc., 'eXcelon eXtensible Information Server, White Paper', http://support.exln.com/doc/whitepapers/xml/XISwpfinal.pdf, (2001)
[P01] L. Page, 'Method for node ranking in a linked database', United States Patent 6,285,999, (2001)
[PB+98] L. Page, S. Brin, R. Motwani, T. Winograd, 'The PageRank Citation Ranking: Bringing Order to the Web', Technical report, Stanford Digital Library Technologies Project, (1998).
[PM+95] Y. Papakostantinou, H. Garcia-Molina, J. Widom. 'Object Exchange across Heterogeneous Information Sources'. In Proc. of the 11th Intl. Conf. on Data Engineering, (1995).
[PM+97] C. Petrou, D. Martakos, S. Hadjiefthymiades, 'Adding semantics to hypermedia towards Link's enhancement and Dynaming Linking', Hypertext - Information Retrieval - Multimedia, (1997)
[PW00] T. Phelps, R. Wilensky, 'Robust Hyperlinks Cost Just Five Words Each'. UC Berkeley Computer Science Technical Report UCB//CSD-00-1091. Berkeley, CA. (2000)
[Q67] J.B. MacQueen, 'Some Methods for Classification and Analysis of Multivariate Observations'. In Proceedings of 5th Berkley Symposium on Mathematical Statistics and Probability, Volume I: Statistics, pp. 281-297, (1967).
[QF94] Y. Qui and H.P. Frei, 'Improving the retrieval effectiveness by using a similarity thesaurus', (1994).
[R01] U. Reimer, 'Organisational Memories for Capturing, Sharing and Utilizing Knowledge', Tutorial, 3rd International Conference on Enterprise Information Systems, ICEIS (2001)
[R79] C. J. van Rijsbergen, 'Information Retrieval', Buttersworth, London, second edition. (1989)
[R92] E. Rasmussen, 'Clustering Algorithms', In Information Retrieval, W.B. Frakes & R. Baeza-Yates (eds.), Prentice Hall PTR, New Jersey (1992)
[R95] P. Resnik, 'Using Information Content to Evaluate Semantic Similarity in a Taxonomy', IJCAI-95, pages 448-453, (1995).
[R99] P. Resnik, 'Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language', Journal of Artificial Intelligence Research 11, p.95-130 (1999).
[RL94] E. Riloff, W. Lehnert, 'Information Extraction as a Basis for High-Precision Text Classification', ACM Transactions on Information Systems, (1994).
[RS+94] R. Richardson, A. Smeaton, J. Murphy, 'Using Wordnet as a Knowledge Base for Measuring Semantic Similarity between Words'. AICS Conference. Dublin (1994).
[RS97] M. Tork Roth, P. Schwarz. 'Don't Scrap It, Wrap It! A Wrapper Architecture for Legacy Data Sources'. VLDB (1997)

[S48] C. Shannon, 'A mathematical theory of communication', Bell System Technical Journal, vol.27, (1948).
[Sa68] G. Salton. 'Automatic information organization and retrieval', McGraw Hill, New York, (1968).
[Sa73] G. Salton, 'Comment on "an evaluation of query expansion by the addition of clustered terms for a document retrieval system" '. Computing Reviews, 14, 232 (1973).
[Sm73] H. G. Small. 'Co-citation in the scientific literature: A new measure of the relationship between two documents'. Journal of American Society for Information Science, 24(4), (1973).
[Si73] R. Sibson, 'SLINK: an optimally efficient algorithm for the single link cluster method'. The Computer Journal 16, 30-34, (1973)
[S92] H. Schutze, 'Dimensions of meaning', Proceedings of Supercomputing (1992).
[S97] E. Spertus. 'ParaSite: Mining structural information on the Web', Proc. 6th International World Wide Web Conference, (1997)
[SA+89] M. Scholl, S. Abiteboul, F. Bancilhon, N. Bidoit, S. Gamerman, D. Plateau, P. Richard, A. Verroust, 'VERSO: A Database Machine Based On Nested Relations', in Nested Relations and Complex Objects in Databases, (1989).
[SA+03] G. Sakkis, I. Androutsopoulos, G. Paliouras, V. Karkaletsis, C.D. Spyropoulos, P. Stamatopoulos, 'A Memory-Based Approach to Anti-Spam Filtering for Mailing Lists'. Information Retrieval, 6(1):49-73, Kluwer, (2003).
[SA99a] A. Sahuguet, F. Azavant, 'Looking at the Web through XML glasses', Conference on Cooperative Information Systems CoopIS'99, Edinburgh Scotland, (1999)
[SA99b] A. Sahuguet, F. Azavant, 'Web Ecology: Recycling HTML pages as XML documents using W4F', In WebDB, (1999).
[SA99c] A. Sahuguet, F. Azavant, 'WysiWyg Web Wrapper Factory', In WWW8, (1999)
[SB+02] G. Schreiber, I. Blok, D. Carlier, W. van Gent, J. Hokstam, U. Roos, A Mini-experiment in Semantic Annotation,ISWC 2002, LNCS 2342, pp. 404-408, (2002).
[SB+98] R. Studer, V.R. Benjamins, D. Fensel, 'Knowledge engineering: Principles and methods'. Data Knowledge Engineering, 25(1-2), (1998).
[SJ+00] A. Sthehl, G. Joydeep, R. Mooney, 'Impact of Similarity Measures on Web-page Clustering'. Proceedings of the 17th National Conference on Artificial Intelligence: Workshop of Artificial Intelligence for Web Search, 30-31 (2000)
[SJ71] K. Sparck Jones, 'Automatic Keyword Classification for Information Retrieval', Butterworths, London (1971).
[SK+00] A.R. Schmidt, M.L. Kersten, M.A. Windhouwer, F. Waas, 'Efficient Relational Storage and Retrieval of XML Documents'. Workshop on the Web and Databases, WebDB (2000).
[SL+93] K. Shoens, A. Luniewski, P.Schwarz, J. Stamos, J. Thomas'The Rufus system: Information organization for semi-structured data', Proc. of the Int. Conf. On Very Large Data Bases (1993)
[SM83] G. Salton, M. McGill, 'Introduction to Modern Information Retrieval', McGraw-Hill, New-York (1983).

[SQL01] Penton Media Inc., 'XML Perspective. In control with FOR XML Explicit'. SQL Server Magazine, http://msdn.microsoft.com/library/default.asp?url=/library/en-us/dnsqlmag01/html/xmlExplicit.asp, March (2001)
[SS+00] J. Shanmugasundaram, E. Shekita, R. Barr, M. Carey, B. Lindsay, H. Pirahesh, B. Reinwald, 'Efficiently Publishing Relational Data as XML Documents', VLDB Conference, (2000).
[SS98] E. Spetrus, L.A. Stain. 'Just-In-Time Databases and the World-Wide Web'. 7th International ACM Conference on Information & Knowledge Management, (1998).
[ST+99] J. Shanmugasundaram, K. Tufte, G. He, C. Zhang, D. DeWitt, J. Naughton, 'Relational Databases for Querying XML Documents: Limitations and Opportunities, of using semi-structured techniques'. Proceedings of the 25th VLDB Conference, (1999).
[SW+75] G. Salton, A. Wang, C. Yang, 'A vector space model for information retrieval', Journal of the American Society for Information Science, volume 18 (1975)
[SW03] The Semantic Web Community Portal, http://www.semanticweb.org/
[TI+01] I. Tatarinov, Z. Ives, A. Halevy, D. Weld, Updating XML. ACM SIGMOD, (2001).
[TK99] S. Theodoridis, K. Koutroubas, 'Pattern Recognition'. Academic Press. (1999)
[TREC] Web Research Collections - TREC Web Track, http://www.ted.cmis.csiro.au/TRECWeb/
[Teo] Teoma search engine, http://www.teoma.com
[V86] E.M. Voorhees, 'Implementing agglomerative hierarchic clustering algorithms for use in document retrieval'. Information Processing & Management, 22, 465-476, (1986)
[VV01] I. Varlamis, M. Vazirgiannis, 'Web document searching using enhanced hyperlink semantics based on XML', in the proceedings of IDEAS conference, pp 34-43, (2001)
[Viv01] Vivisimo search engine: http://www.vivisimo.com/
[W3a] http://www.w3.org/HTML
[W3b] http://www.w3.org/XML
[W3c] http://www.w3.org/RDF
[W3d] S.Adler, A. Berglund et al, 'Extensible Stylesheet Language (XSL), W3C Recommendation, http://www.w3.org/TR/xsl/, (2001)
[W3e] The Document Object Model. http://www.w3.org/DOM
[W3f] T. Bray, J. Paoli, C. M. Sperberg-McQueen, E. Maler, 'Extensible Markup Language Second Edition', W3C Recommendation, http://www.w3.org/TR/REC-xml (2000)
[W3g] D. Fallside, 'XML Schema Part 0: Primer' W3C Recommendation, http://www.w3.org/TR/xmlschema-0/ (2001)
[W3h] J. Clark, S. DeRose, XML Path Language, W3C Recommendation, www.w3.org/TR/xpath, (1999).
[W3i] W3 Consortium, Semantic Web Activity, http://www.w3.org/2001/sw/, 2001
[W3j] D. Brickley, R.V. Guha, 'RDF Vocabulary Description Language 1.0: RDF Schema, Working Draft', http://www.w3.org/TR/2003/WD-rdf-schema-20030123/, (2003)

[W3k] D. Connolly, F. van Harmelen, I. Horrocks, D. L. McGuinness, P. F. Patel-Schneider, L. A. Stein, 'DAML+OIL (March 2001) Reference Description', http://www.w3.org/TR/daml+oil-reference, (2001)
[W3l] Web-Ontology (WebOnt) Working Group, http://www.w3.org/2001/sw/WebOnt/, (2003)
[W3m] S. Boag, D. Chamberlin, M. F. Fernandez, D. Florescu, J. Robie, J. Simeon, 'XQuery 1.0: An XML Query Language', W3C Working Draft, http://www.w3.org/TR/xquery/, May (2003)
[W88] P. Willett, 'Recent Trends in Hierarchic document Clustering: a critical review'. Information & Management. 24(5) (1988)
[WK01] Y. Wang, M. Kitsuregawa, 'Link Based Clustering of Web Search Results', in the proceedings of WAIM 2001 conference, LNCS 2118, pp. 225-236, (2001)
[WP94] Z. Wu, M. Palmer, 'Verb Semantics and Lexical Selection', Proceedings of the 32nd Annual Meetings of the Associations for Computational Linguistics, pages 133-138.
[XDI] Microsoft Corp. 'XML Data Islands', Microsoft XML 3.0 - XML Developer's Guide
[XGT01] XML Global Technologies, 'GoXML Database', http://www.xmlglobal.com/prod/db/index.jsp, (2001)
[Y94] D. Yarowsky, 'Decision lists for lexical ambiguity resolution. Application to accent restoration in Spanish and French', In Proceedings of ACL 32, (1994)
[Y95] D. Yarowsky, 'Unsupervised word sense disambiguation rivaling supervised methods', In Proceedings of ACL 33, (1995)
[Y00] D. Yarowsky, ‘ Word-Sense Disambiguation’, Handbook of Natural Language Processing, ed. by R. Dale, H. Moisl, H. Somers, Marcel Dekker Inc., New York, 2000, pp. 629-654.
[YA88] E. J. Yannakoudakis, G. Angelidakis, ‘An Insight into the Entropy and Redundancy of the English Dictionary’, IEEE Transactions on Pattern Analysis and Machine Intelligence 10 (6): 960-970 (1988)
[Yah] The Yahoo Directory, http://www.yahoo.com
[ZE+97] O. Zamir, O. Etzioni, O. Madani, R. Karp, 'Fast and Intuitive Clustering of Web Documents', KDD '97, Pages 287-290, (1997).
[ZE98] O.Zamir, O. Etzioni, 'Web document clustering: a feasibility demonstration', in proceedings of ACM-SIGIR (1998)
[ZK02] Y. Zhao, G. Karypis, 'Evaluation of Hierarchical Clustering Algorithms for Document Datasets', in CIKM (2002)

Παράρτημα Α – XML-Schema για τα έγγραφα του THESUS

```
<?xml version="1.0" encoding="UTF-8"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema"
elementFormDefault="qualified" attributeFormDefault="unqualified">
 <xs:element name="SourceDocument">
  <xs:complexType>
   <xs:sequence>
    <xs:element name="SourceURI" type="xs:anyURI"/>
<xs:element name="Keyword" type="KeyValuePairType" minOccurs="0"
maxOccurs="unbounded"/>
<xs:element name="Concept" type="KeyValuePairType" minOccurs="0"
maxOccurs="unbounded"/>
    <xs:element name="Hyperlink" minOccurs="0" maxOccurs="unbounded">
     <xs:complexType>
      <xs:sequence>
       <xs:element name="TargetURI" type="xs:anyURI"/>
<xs:element name="Keyword" type="KeyValuePairType" minOccurs="0"
maxOccurs="unbounded"/>
<xs:element name="Concept" type="KeyValuePairType" minOccurs="0"
maxOccurs="unbounded"/>
      </xs:sequence>
     </xs:complexType>
    </xs:element>
   </xs:sequence>
  </xs:complexType>
 </xs:element>
 <xs:complexType name="KeyValuePairType">
  <xs:attributeGroup ref="KeyValuePair"/>
 </xs:complexType>
 <xs:attributeGroup name="KeyValuePair">
  <xs:attribute name="key" type="xs:NMTOKEN" use="required"/>
  <xs:attribute name="value" type="xs:float" use="required"/>
 </xs:attributeGroup>
</xs:schema>
```

Παράρτημα Β – Σύνοψη κατηγορημάτων του THESUS

Τύποι Δεδομένων		
Doc	(URL, {keyword}, {ontology_term})	Τα έγγραφα του THESUS
Link	(URLS, URLT, {keyword}, {ontology_term})	Οι σύνδεσμοι του THESUS
w_docs	{(URL, text)}	Τα έγγραφα του ΠΙ
w_links	{(URLS, URLT)}	Οι σύνδεσμοι του ΠΙ

Βοηθητικές Συναρτήσεις		
getpos(text, URLT)	returns {pos}	Βρίσκει τη θέση του URLT στη σελίδα
process(w_docs.TEXT, pos, M)	» {keyword}	Εξάγει λέξεις από μια περιοχή Μ χαρακτήρων γύρω από το σύνδεσμο
getkeys({(keyword, times)})	» {keyword}	Επιστρέφει τα keywords στο σύνολο (keyword,times)
getTimes({(keyword, times)},keyword)	» {keyword}	Επιστρέφει τον αριθμό εμφανίσεων times στο σύνολο (keyword,times)
update({(keyword, times)}, keyword, N)	» null	Αυξάνει τον αριθμό times για τη λέξη keyword στο σύνολο (keyword,times) κατά Ν
getSemantics ({keyword})	» {ontology_term}	Επιστρέφει το σύνολο όρων της οντολογίας που αντιστοιχεί στις λέξεις

Τελεστές	
Όνομα	**Επιστρέφει**
fetch(URL)	**Text** Τα περιεχόμενα του URL
GroupMatch({URLexpr})	**{URL}** τα URLs που ταιριάζουν στην έκφραση URLexpr
crawl({SURL},N)	**{URL}** τα URLs όπου μπορούμε να φτάσουμε ξεκινώντας από ένα URL στο {SURL} και ακολουθώντας το πολύ Ν συνδέσμους
thematicCrawl({SURL}, {keyword},N)	**{URL}** τα URLs όπου μπορούμε να φτάσουμε ξεκινώντας από ένα URL στο {SURL} και ακολουθώντας το πολύ Ν συνδέσμους που περιέχουν τουλάχιστο μια λέξη απ' το σύνολο {keyword}
linkKeywords (URLS, URLT)	**{keyword}** οι λέξεις όλων των συνδέσμων από το URLS στο URLT
groupKeywords ({URLS}, {URLT})	**{keyword}** οι λέξεις όλων των συνδέσμων από κάποιο URL στο {URLS} προς οποιοδήποτε URL στο {URLT}
weightedGroupKeywords ({URLS}, {URLT})	**{(keyword,times)}** οι λέξεις και ο αριθμός εμφανίσεων από όλους τους συνδέσμους του {URLS} προς το {URLT}

Παράρτημα Γ – Οι διευθύνσεις ιστού των κεντρικών ιστοσελίδων μουσείων του Λονδίνου

http://www.nhm.ac.uk/
http://www.vam.ac.uk/
http://www.sciencemuseum.org.uk/
http://www.thebritishmuseum.ac.uk/
http://www.nmm.ac.uk/
http://www.iwm.org.uk/
http://www.ri.ac.uk
http://www.museumoflondon.org.uk
http://www.freud.org.uk/
http://www.iwm.org.uk/belfast/index.htm
http://www.the-wallace-collection.org.uk/
http://www.ltmuseum.co.uk
http://www.sherlock-holmes.co.uk/
http://www.madame-tussauds.com/
http://www.commonwealth.org.uk
http://www.soane.org/
http://www.tower-of-london.com/
http://www.cix.co.uk/~museumgh/

Παράρτημα Δ – XML-Schema αρχείο εντολών διαχείρισης

```
<?xml version="1.0" encoding="UTF-8" ?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema"
 elementFormDefault="qualified" attributeFormDefault="unqualified">
 <xs:element name="DBCommand">
  <xs:complexType>
   <xs:sequence>
    <xs:element    name="Insert"    type="DBInsert"    minOccurs="0"
maxOccurs="unbounded" />
    <xs:element    name="Update"    type="DBUpdate"    minOccurs="0"
maxOccurs="unbounded" />
    <xs:element    name="Delete"    type="DBDelete"    minOccurs="0"
maxOccurs="unbounded" />
    <xs:element    name="Select"    type="DBSelect"    minOccurs="0"
maxOccurs="unbounded" />
    <xs:element   name="CreateIndex"   type="DBIndex"   minOccurs="0"
maxOccurs="unbounded" />
   </xs:sequence>
  </xs:complexType>
 </xs:element>
 <xs:complexType name="DBInsert">
  <xs:sequence>
   <xs:any namespace="urn:mpeg:mpeg7:schema:2001" />
  </xs:sequence>
  <xs:attribute name="path" type="xs:string" />
  <xs:attribute name="DB_ID" type="xs:integer" />
  <xs:attribute name="order" type="xs:integer" />
 </xs:complexType>
 <xs:complexType name="DBUpdate">
  <xs:sequence>
   <xs:any />
  </xs:sequence>
  <xs:attribute name="path" type="xs:string" />
  <xs:attribute name="DB_ID" type="xs:integer" />
 </xs:complexType>
 <xs:complexType name="DBDelete">
  <xs:attribute name="path" type="xs:string" />
  <xs:attribute name="DB_ID" type="xs:integer" />
 </xs:complexType>
 <xs:complexType name="DBSelect">
  <xs:attribute name="where" type="xs:string" />
  <xs:attribute name="return" type="xs:string" />
  <xs:attribute name="bind" type="xs:string" />
  <xs:attribute name="noDB_IDs" type="xs:boolean" default="false"/>
  <xs:attribute         name="noSubElements"         type="xs:boolean"
default="false"/>
 </xs:complexType>
 <xs:complexType name="DBIndex">
  <xs:attribute name="name" type="xs:string" />
  <xs:sequence>
   <xs:element name="field" type="xs:string" maxOccurs="unbounded" />
  </xs:sequence>
 </xs:complexType>
 <xs:element name="DBReply">
  <xs:complexType>
   <xs:choice maxOccurs="unbounded">
    <xs:element name="Result" type="DBResult" minOccurs="0" />
    <xs:element name="Error" type="DBError" minOccurs="0" />
```

```
   </xs:choice>
  </xs:complexType>
 </xs:element>
 <xs:complexType name="DBResult">
  <xs:sequence>
   <xs:any    namespace="urn:mpeg:mpeg7:schema:2001"    minOccurs="0"
maxOccurs="unbounded" />
  </xs:sequence>
 </xs:complexType>
 <xs:complexType name="DBError">
  <xs:attribute name="main" type="xs:string" />
  <xs:attribute name="detail" type="xs:string" />
 </xs:complexType>
</xs:schema>
```

ΓΛΩΣΣΑΡΙΟ

Ξένος όρος	**Απόδοση στο κείμενο**	**Επεξήγηση**
Authority	Κόμβος-αυθεντία	Ένα έγγραφο του ΠΙ με πολλούς εισερχόμενους συνδέσμους. Στο κείμενο εμφανίζεται και ως **α-κόμβος** για συντομία.
Centrality	Κεντρικότητα	Προσδιορίζει τη σημαντικότητα ενός εγγράφου για την αναζήτηση στο γράφο του ΠΙ.
Classification	Κατηγοριοποίηση Ταξινόμηση	
Classifiers	Κατηγοριοποιητές	Οι κανόνες με τους οποίους αποφασίζεται η ένταξη νέων αντικειμένων στις υπάρχουσες κατηγορίες
Co-Citation	Συν-αναφορά	Για δύο έγγραφα του ΠΙ: είναι η τομή των εγγράφων που δείχνουν προς τα δύο έγγραφα (με κάποιο σύνδεσμο).
Common words	Κοινές λέξεις.	Λέξεις χωρίς σημασιολογικό περιεχόμενο: άρθρα, σύνδεσμοι, μόρια κτλ.
Connectivity-based ranking	Ταξινόμηση που βασίζεται στη συνδεσμολογία	
Context	Γλωσσικό περιβάλλον	
Cosine measure	Μέτρο συνημιτόνου	
Coupling	Βιβλιογραφική σύζευξη	Για δύο έγγραφα του ΠΙ: είναι η τομή των εγγράφων προς τα οποία δείχνουν τα δύο έγγραφα (με κάποιο σύνδεσμο).
Crawler	Περιηγητής	Ένα πρόγραμμα που επισκέπτεται τα έγγραφα του ΠΙ ακολουθώντας τους μεταξύ τους υπερσυνδέσμους.
Damping factor	Παράγοντας απόσβεσης	Έχει σκοπό να εξαλείψει τη σημαντικότητα που μεταφέρει ένα έγγραφο στα έγγραφα που αναφέρει.
Domain	Περιοχή ιστού	
Domain ontology	Οντολογία περιοχής	Αφορά μια συγκεκριμένη θεματική περιοχή
Density-based clustering	Συσταδοποίηση βασισμένη στην πυκνότητα	
Heuristics	Ευριστικές τεχνικές	
Hub	Πλήμνη, κόμβος παραπομπών	. Ένα έγγραφο του ΠΙ με πολλούς εξερχόμενους συνδέσμους. Στο κείμενο εμφανίζεται και ως **π-κόμβος** για συντομία.
Hypernym	Υπερώνυμα.	Οι γενικότερες μιας έννοιας
Hyponym	Υπώνυμα.	Οι ειδικότερες μιας έννοιας

Generic ontology	Γενική οντολογία	Καλύπτει ένα ευρύ φάσμα θεματικών περιοχών
Latent Semantic Indexing	Ευρετηρίων Λανθάνουσας Σημασιολογίας	Κάθε αντικείμενο αναπαρίσταται ως διάνυσμα χαρακτηριστικών που ανήκουν σε συγκεκριμένο ευρετήριο
Link analysis theory	Θεωρία ανάλυσης των συνδέσμων	
Link semantics	Σημασιολογία συνδέσμου	Το σύνολο των εννοιών που εξάγονται για έναν σύνδεσμο
Mediator	Μεσάζων	Πρόγραμμα που δέχεται ερωτήσεις σε μια κοινή γλώσσα ερωτήσεων και τις προωθεί στις υποκείμενες διαφορετικές δομές δεδομένων, δίνοντας έτσι την αίσθηση μιας ομοιόμορφης δομής.
Object Exchange Model - OEM	Μοντέλο Ανταλλαγής Αντικειμένων	
Polysemy	Πολυσημία.	Η ιδιότητα μιας λέξης να έχει διαφορετικές σημασίες.
Precision	Ποσοστό ακριβείας	το πλήθος των σχετικών εγγράφων που επιστρέφονται ως απάντηση σε μια ερώτηση ως προς το συνολικό αριθμό απαντήσεων
Recall	Ποσοστό ανάκλησης	το πλήθος των σχετικών εγγράφων που επιστρέφονται ως απάντηση σε μια ερώτηση ως προς το συνολικό πλήθος των σχετικών εγγράφων
Semantics	Σημασιολογία.	Ο κλάδος της σημειολογίας που εξετάζει τις σχέσεις σημείου (π.χ λέξης, εγγράφου) και σημασίας
Semantic Web	Σημασιολογικός Ιστός	
Stemming	Αποστελέχωση.	Η αποκοπή των επιθεμάτων και προθεμάτων από τη ρίζα μιας λέξης
Synonymy	Συνωνυμία.	Το φαινόμενο κατά το οποίο δύο διαφορετικές λέξεις έχουν την ίδια σημασία.
Taxonomy	Ταξινομία.	Μια ιεραρχική οργάνωση λέξεων, εννοιών κτλ.
URL (uniform resource locator)	Ενιαίος εντοπιστής πόρου.	Εξαρτάται από τη φυσική θέση του πόρου στο δίκτυο. Στο κείμενο αποδίδεται ως «**διεύθυνση ιστού**» για λόγους απλότητας.
URI (Uniform Resource Identifier)	Ενιαίους προσδιοριστής πόρου	Είναι ανεξάρτητος από τη φυσική θέση του πόρου στο δίκτυο.
Vector Space Model	Μοντέλο του Χώρου Διανυσμάτων	Κάθε αντικείμενο αναπαρίσταται ως διάνυσμα χαρακτηριστικών με μέγεθος όσο το πλήθος των διαφορετικών χαρακτηριστικών των εγγράφων

Web servers	Εξυπηρετητές Παγκόσμιου Ιστού.	
Word sense disambiguation	Αποσαφήνιση της έννοιας των λέξεων	
Wrapper	Περίβλημα (εγγράφου).	Προσαρμοσμένο πρόγραμμα ανάγνωσης εγγράφων
WWW (world wide web)	Παγκόσμιος Ιστός	

ΕΥΡΕΤΗΡΙΟ ΣΧΗΜΑΤΩΝ

Σχήμα 1. Παράδειγμα δέντρου που αντιστοιχεί σε μια ιστοσελίδα HTML 15
Σχήμα 2. Παράδειγμα δομής του OEM 16
Σχήμα 3. Παράδειγμα αναπαράστασης του OEM 17
Σχήμα 4. Παράδειγμα υπερ-δένδρου (http://www.db-net.aueb.gr/pubs.php) 17
Σχήμα 5. Σύνδεση προτύπων εγγράφων στο σύστημα Araneus 18
Σχήμα 6. Ιεραρχίες ομάδων που παράγει το THESUS 44
Σχήμα 7. Η μέθοδος συγκέντρωσης, χαρακτηρισμού και διαχείρισης εγγράφων στο THESUS 56
Σχήμα 8. Εισαγωγή αντικειμένου σε υπάρχουσα υποομάδα 66
Σχήμα 9. Δημιουργία νέας υποομάδας 67
Σχήμα 10. Συγχώνευση των δύο καλύτερων υποομάδων 67
Σχήμα 11. Διάσπαση της καλύτερης υποομάδας 67
Σχήμα 12. Παράδειγμα ταξινομίας 71
Σχήμα 13. Αναπαράσταση του XML-Schema των εγγράφων του THESUS 75
Σχήμα 14. Διαδικασία συλλογής διαδικτυακών εγγράφων 76
Σχήμα 15. Διαδικασίες Εξαγωγής και Εμπλουτισμού πληροφορίας, Οργάνωσης Εγγράφων 77
Σχήμα 16. Διαδικασία διαχείρισης ερωτήσεων 77
Σχήμα 17. Σχηματική αναπαράσταση του τελεστή crawl 83
Σχήμα 18. Αρχιτεκτονική του συστήματος THESUS 106
Σχήμα 19. Εφαρμογή σύνδεσης της οντολογίας με το Wordnet 108
Σχήμα 20. Διεπαφή χαρακτηρισμού εγγράφων του ιστού 109
Σχήμα 21. Σχετικότητα των λέξεων του χαρακτηρισμού με το θέμα της κατηγορίας 114
Σχήμα 22. Πώς επηρεάζει ο αριθμός εισερχόμενων συνδέσμων τη σχετικότητα της περιγραφής 115
Σχήμα 23. Ο ρόλος του X-Database στο σύστημα THESUS 129
Σχήμα 24. Αρχιτεκτονική του συστήματος X-Database 148
Σχήμα 25. Επεξεργασία XML εντολών 148
Σχήμα 26. Εξάρτηση του χρόνου εισαγωγής από το πλήθος των εγγράφων 151
Σχήμα 27. Επίδραση του παράγοντα βάθους στο χρόνο εισαγωγής 152
Σχήμα 28. Επίδραση του παράγοντα εύρους στο χρόνο εισαγωγής 152
Σχήμα 29. Απόδοση ενημέρωσης ενός χαρακτηριστικού σε αυξανόμενο βάθος 153
Σχήμα 30. Απόδοση ενημέρωσης ενός στοιχείου σε αυξανόμενο βάθος 154
Σχήμα 31. Εξάρτηση του χρόνου εισαγωγής από το πλήθος των εγγράφων 155
Σχήμα 32. Επίδραση του παράγοντα βάθους στο χρόνο διαγραφής 155
Σχήμα 33. Επίδραση του παράγοντα εύρους στο χρόνο διαγραφής 155
Σχήμα 34. Χρόνος επιλογής για αυξανόμενο αριθμό όμοιων στοιχείων (βάθος=5, εύρος=1) 156
Σχήμα 35. Χρόνος επιλογής για αυξανόμενο αριθμό διαφορετικών στοιχείων (βάθος=3, εύρος=1) 156
Σχήμα 36. Χρόνος επιλογής για στοιχεία με αυξανόμενο βάθος (κλιμάκωση=10, εύρος=1) 157
Σχήμα 37. Χρόνος επιλογής για στοιχεία με αυξανόμενο εύρος (κλιμάκωση=10, βάθος=3) 157

ΕΥΡΕΤΗΡΙΟ ΠΙΝΑΚΩΝ

Πίνακας 1. Αποτελέσματα του DBSCAN για διάφορες παραμέτρους εισόδου (minDocs=3) με χρήση εννοιών 116
Πίνακας 2. Αποτελέσματα του DBSCAN για διάφορες παραμέτρους εισόδου (minDocs=1) με χρήση λέξεων....117
Πίνακας 3. Αποτελέσματα του COBWEB (με χρήση εννοιών) για αυξανόμενο αριθμό εγγράφων 117
Πίνακας 4. Αριθμός εγγράφων ανά κατηγορία στη συλλογή LEVEL1 120
Πίνακας 5. Αριθμός εγγράφων ανά κατηγορία στη συλλογή LEVEL2 121
Πίνακας 6 Αποτελέσματα συσταδοποίησης για τρία σύνολα δεδομένων με επιλογή των βέλτιστων τιμών των παραμέτρων εισόδου....123
Πίνακας 7 Θεματικοί α-κόμβοι στο THESUS....125
Πίνακας 8. Παράδειγμα βιβλιογραφικών συζεύξεων....126
Πίνακας 9. Απεικόνιση του χαρακτηριστικού length....133
Πίνακας 10. Χρήση των union σε ένα simpleType 134
Πίνακας 11. Απεικόνιση complexType 134
Πίνακας 12. Απεικόνιση φωλιασμένων στοιχείων....135
Πίνακας 13. Παράδειγμα κληρονομικότητας σύνθετων τύπων....137
Πίνακας 14. Παράδειγμα χρήσης του χαρακτηριστικού xsi:type....138
Πίνακας 15. Εισαγωγή ενός νέου εγγράφου....142
Πίνακας 16. Εισαγωγή ενός νέου στοιχείου 142
Πίνακας 17. Εισαγωγή στοιχείου σε συγκεκριμένη σειρά 143
Πίνακας 18. Ενημέρωση των χαρακτηριστικών ενός στοιχείου 143
Πίνακας 19. Διαγραφή ενός στοιχείου....144
Πίνακας 20. Παραδείγματα εντολών Select 146
Πίνακας 21. Παράδειγμα δημιουργίας ευρετηρίου 147
Πίνακας 22. Στατιστικά του XML-Schema και του σχήματος βάσης που δημιουργεί 150
Πίνακας 23. Παραδείγματα εντολών επιλογής και χρόνων απόκρισης....158