ΠΕΡΙΓΡΑΦΗ

Μέθοδος και σύστημα για τη συγκέντρωση, περιγραφή, οργάνωση και διαχείριση ψηφιακών/διαδικτυακών εγγράφων με βάση σημασιολογικά χαρακτηριστικά, που πηγάζουν από μια θεματική οντολογία.

Σύντομη περιγραφή της εφεύρεσης

Η εφεύρεση αναφέρεται σε ένα σύστημα (υπολογιστής συνδεδεμένος στο διαδίκτυο με κατάλληλο λογισμικό) που αναλαμβάνει την αυτόματη επιλεκτική συγκέντρωση εγγράφων από τον παγκόσμιο ιστό, την εξαγωγή λέξεων από τους υπερσυνδέσμους και τα περιεχόμενα των ιστοσελίδων και την απεικόνιση αυτών σε έννοιες μιας οντολογίας με αυτόματο τρόπο. Οι λέξεις που εξάγονται και οι αντίστοιχες έννοιες χρησιμοποιούνται για την περιγραφή των εγγράφων. Το σύστημα οργανώνει τα έγγραφα σε ομάδες με κοινά σημασιολογικά χαρακτηριστικά, βασιζόμενο στη σημασιολογική εγγύτητα των εγγράφων, ενώ παράλληλα χαρακτηρίζει τις ομάδες που προκύπτουν με έννοιες της οντολογίας. Το σύστημα απαντά στις λεκτικές ερωτήσεις των χρηστών, βασιζόμενο στη σημασιολογική εγγύτητα των ερωτημάτων με τα έγγραφα της συλλογής.

Παράλληλα η εφεύρεση εισάγει μια νέα μέθοδο για τη συγκέντρωση και οργάνωση εγγράφων με βάση τα σημασιολογικά χαρακτηριστικά που προκύπτουν για τα έγγραφα από μια θεματική οντολογία.

Στη συνέχεια της περιγραφής με τον όρο λεκτική πληροφορία αναφερόμαστε στις λέξεις που περιγράφουν ένα κείμενο, μια ερώτηση, έναν υπερσύνδεσμο ενώ με τον όρο σημασιολογική πληροφορία αναφερόμαστε στις έννοιες που αυτές οι λέξεις αντιπροσωπεύουν. Ενώ οι λέξεις μπορεί να είναι οποιεσδήποτε, οι έννοιες στις οποίες αντιστοιχούν είναι περιορισμένες και μάλιστα οργανωμένες σε μια ιεραρχική δομή (από τις γενικότερες προς τις πιο ειδικές) σε μια οντολογία που ορίζει ο χρήστης του συστήματος.

Για την λειτουργία του συστήματος απαιτείται η κατασκευή μιας θεματικής οντολογίας από το χρήστη (παραδείγματα οντολογιών που υποστηρίζονται υπάρχουν στη διεύθυνση http://www.daml.org/ontologies/). Το σύστημα χρησιμοποιεί ένα

εννοιολογικό θησαυρό για να αντιστοιχίσει την λεκτική πληροφορία που εξάγεται από τα έγγραφα σε σημασιολογική πληροφορία (έννοιες της θεματικής οντολογίας). Η οργάνωση των εγγράφων σε ομάδες γίνεται με βάση την εγγύτητα των σημασιολογικών περιγραφών τους η οποία υπολογίζεται επί της οντολογίας με χρήση ενός νέου μέτρου ομοιότητας μεταξύ εγγράφων. Το νέο μέτρο βασίζεται στο μέτρο των Wu και Palmer (Z. Wu and M. Palmer *"Verb Semantics and Lexical Selection"*, Proceedings of the 32nd Annual Meetings of the Associations for Computational Linguistics, pages 133-138), που υπολογίζει την ομοιότητα δύο όρων μιας ιεραρχίας. Με το μέτρο αυτό υπολογίζουμε άμεσα την ομοιότητα δύο συνόλων όρων μιας ιεραρχίας και το χρησιμοποιούμε για να υπολογίσουμε την ομοιότητα μεταξύ δύο εγγράφων.

<u>Σχετικές τεχνολογίες και συστήματα</u>

Τα υπάρχοντα συστήματα υλοποιούν μέρος της λειτουργικότητας του προσφερόμενου συστήματος και με λιγότερο ευφυή τρόπο. Οι βασικές κατηγορίες συστημάτων σχετικών με την εφεύρεση είναι:

- Συστήματα που αναλαμβάνουν την αυτόματη συγκέντρωση ιστοσελίδων, η οποία σε ορισμένες περιπτώσεις είναι και επιλεκτική (S. Chakrabarti, M. van den Berg, B. Dom, "Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery", 8th WWW Conference). Τα συστήματα αυτά ξεκινούν από μια μικρή ομάδα ιστοσελίδων και επισκέπτονται διαδοχικά τους υπερσυνδέσμους αυτών, είτε όλους είτε μόνον αυτούς που ικανοποιούν συγκεκριμένα κριτήρια. Τα κριτήρια αυτά αφορούν είτε τη δικτυακή διεύθυνση της σελίδας (URL), π.χ. εύρεση σελίδων του ελληνικού δικτυακού χώρου (.gr), είτε την ύπαρξη συγκεκριμένων λέξεων εντός των υπερσυνδέσμων.
- Συστήματα που εξάγουν πληροφορία από το περιεχόμενο ιστοσελίδων και εγγράφων γενικότερα(Web Data Extractor, http://www.webextractor.com/). Τα συστήματα αυτά εξάγουν τις σημαντικότερες λέξεις που περιέχονται σε ένα έγγραφο δίχως να λαμβάνουν υπόψη τους τις έννοιες που αυτές οι λέξεις αντιπροσωπεύουν. Τα συστήματα αυτά μπορούν πολύ εύκολα να χειραγωγηθούν από τους δημιουργούς των εγγράφων (και αυτό είναι πολύ συχνότερο στην περίπτωση των σελίδων του Παγκόσμιου Ιστού), με την προσθήκη περιεχομένου που δεν είναι ορατό στους χρήστες, και το οποίο αλλοιώνει τους εξαγόμενους χαρακτηρισμούς. Η πληροφορία που εξάγουν τα

συστήματα αυτά είναι λοιπόν φτωχή και σε πολλές περιπτώσεις περιορισμένης αξιοπιστίας.

- Συστήματα ταξινόμησης εγγράφων (Yin-Hung Kuo, Man-Hon Wong, "Web Document Classification based on Hyperlinks and Document Semantics", PRICAI Workshop on Text and Web Mining). Τα συστήματα αυτά στηρίζονται σε μια προϋπάρχουσα ιεραρχική ταξινόμηση και εκπαιδεύονται με μια περιορισμένη συλλογή εγγράφων ώστε να αναπτύξουν κανόνες ταξινόμησης. Τα συστήματα αυτά απαιτούν ανθρώπινη παρέμβαση κατά την δημιουργία της ταξινόμησης αλλά και της εκπαίδευσης του συστήματος. Επιπλέον, περιορίζονται στη χρήση των κατηγοριών που ήδη υπάρχουν και δεν προβλέπουν τη δημιουργία νέων (Vivisimo, http://vivisimo.com/)
- Συστήματα ερωτήσεων και συστήματα επέκτασης ερωτήσεων προς συλλογές εγγράφων (S. Brin, L.Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine", 7th WWW Conference). Τα συστήματα αυτά δέχονται τις ερωτήσεις των χρηστών με τη μορφή μιας λίστας λέξεων την οποία επεκτείνουν προσθέτοντας συνώνυμες ή συχνά συμφραζόμενες λέξεις που έχουν προκύψει από ανάλυση προηγούμενων ερωτήσεων. Ο υποβάλλων την ερώτηση έχει τη δυνατότητα να απορρίψει ή να προσθέσει επιπλέον λέξεις. Το αποτέλεσμα αυτών των συστημάτων είναι έγγραφα που περιέχουν ακριβώς κάποιες ή όλες τις λέξεις της ερώτησης. Τα συστήματα αυτά δεν λαμβάνουν υπόψη τους τη σημασιολογική εγγύτητα των λέξεων και συνεπώς δεν φέρνουν ποτέ ως απάντηση έγγραφα που δεν περιέχουν ακριβώς τις λέξεις της ερώτησης αλλά λέξεις με παραπλήσια σημασία. Οι απαντήσεις αυτών των συστημάτων στις ερωτήσεις, αποτελούνται από μια μακροσκελή λίστα συνδέσμων προς τα σχετικά έγγραφα, στην οποία τα έγγραφα ταξινομούνται με φθίνουσα σειρά σημαντικότητας ή σχετικότητας με την ερώτηση.

Όπως προκύπτει από τα παραπάνω, η χρήση λεκτικών χαρακτηριστικών για τη συγκέντρωση οργάνωση και διαχείριση των σελίδων του παγκόσμιου ιστού δεν είναι επαρκής. Απαιτείται λοιπόν ο εμπλουτισμός της υπάρχουσας πληροφορίας με σημασιολογικά χαρακτηριστικά και η χρήση αυτών για την καλύτερη οργάνωση των εγγράφων του παγκόσμιου ιστού αλλά και άλλων συλλογών γενικότερα. Είναι επίσης προτιμότερη η οργανωμένη παρουσίαση των αποτελεσμάτων μιας αναζήτησης, σε ομάδες εννοιολογικά όμοιων εγγράφων, από την απλή απαρίθμησή τους.

Τα πλεονεκτήματα αυτής της εφεύρεσης προκύπτουν από τη χρήση σημασιολογικής πληροφορίας σε όλα τα επίπεδα του συστήματος. Η πληροφορία αυτή προκύπτει με αυτόματο τρόπο με τη χρήση οντολογίας και τη βοήθεια ενός μηχανισμού απεικόνισης. Η συλλογή των εγγράφων από τον Παγκόσμιο Ιστό, η απάντηση ερωτημάτων και η ομαδοποίηση εγγράφων γίνεται με αυτόματο τρόπο και βασίζεται σε σημασιολογική και όχι λεκτική πληροφορία που εξάγεται από τα έγγραφα και τους μεταξύ τους συνδέσμους. Η αναζήτηση πληροφορίας δεν βασίζεται πλέον σε απόλυτο ταίριασμα λεκτικών κριτηρίων αλλά σε στη σχετική ομοιότητα μεταξύ ερωτημάτων και εγγράφων. Τα έγγραφα που επιστρέφονται ως απάντηση εμφανίζονται ομαδοποιημένα με βάση τα όμοια σημασιολογικά χαρακτηριστικά τους.

Πλεονεκτήματα της εφεύρεσης

Σε κάθε περίπτωση η πληροφορία που σχετίζεται με το κείμενο απεικονίζεται σε έννοιες της οντολογίας. Τα πλεονεκτήματα συνοψίζονται στα εξής:

- Επιλεκτική συγκέντρωση ιστοσελίδων, από τον παγκόσμιο ιστό, που σχετίζονται κάποιο θέμα, όπως αυτό περιγράφεται από την οντολογία, με κριτήριο τη σημασιολογική πληροφορία που φέρουν οι υπερσύνδεσμοι. Δίνοντας σαν είσοδο στο σύστημα μια οντολογία, σχετιζόμενη με κάποιο πεδίο ενδιαφέροντος, έχουμε ως απάντηση μια συλλογή ιστοσελίδων που πραγματεύονται τις έννοιες της οντολογίας.
- Σημασιολογικός χαρακτηρισμός των ιστοσελίδων με εξαγωγή λεκτικής πληροφορίας από το περιεχόμενο τους αλλά και από τους εισερχόμενους προς αυτές υπερσυνδέσμους και απεικόνιση της πληροφορίας αυτής σε σημασιολογικό επίπεδο (σε έννοιες μιας οντολογίας)
- Αυτόματη οργάνωση της συλλογής των ιστοσελίδων με βάση τη σημασιολογική εγγύτητα των χαρακτηρισμών που τους αποδόθηκαν. Οι σελίδες οργανώνονται σε ομάδες που χαρακτηρίζονται με τις αντιπροσωπευτικότερες, για κάθε ομάδα, έννοιες τις οντολογίας.
- Δυνατότητα περιήγησης της οργανωμένης συλλογής και εντοπισμός των ιστοσελίδων μέσα από τις ομάδες στις οποίες ανήκουν
- Δυνατότητα ερωτήσεων προς τη συλλογή, υπό τη μορφή λίστας λέξεων. Οι ερωτήσεις αυτές ανάγονται σε λίστες εννοιών της οντολογίας και τα σχετικά έγγραφα καθορίζονται με χρήση ενός μέτρου ομοιότητας που λαμβάνει υπόψη

του την καθολική εγγύτητα των εννοιών που πρεσβεύει η ερώτηση με τις έννοιες που χαρακτηρίζουν τα έγγραφα.

Αναλυτική περιγραφή της προτεινόμενης εφεύρεσης

Η εφεύρεση περιγράφεται στη συνέχεια με αναφορές στο συνημμένο σχέδιο στα οποία περιγράφεται η γενική αρχιτεκτονική του συστήματος.

Το σύστημα αποτελείται από 4 υποσυστήματα (Thematic Crawler – Σχήμα 1, TC, Information Miner – Σχήμα 1, IM, Document Organizer – Σχήμα 1, DO, Query Manager – Σχήμα 1, QM) μια Βάση δεδομένων (Database – Σχήμα 1, DB) για την αποθήκευση της πληροφορίας, και το λεξιλογικό θησαυρό (WordNet, http://www.cogsci.princeton.edu/~wn/ – Σχήμα 1, WN) εγκατεστημένα στον ίδιο ή σε διαφορετικούς ηλεκτρονικούς υπολογιστές συνδεδεμένους μεταξύ τους σε τοπικό δίκτυο.

Υποσύστημα συλλογής πληροφορίας

Το υποσύστημα συλλογής πληροφορίας (Thematic Crawler – Σχήμα 1, TC) βρίσκεται σε υπολογιστή συνδεδεμένο με το διαδίκτυο (Σχήμα 1, WWW) και προσφέρει τη δυνατότητα αυτόματης συγκέντρωσης εγγράφων από το διαδίκτυο που σχετίζονται με κάποιο θέμα, βασιζόμενο σε ένα περιορισμένων σύνολο διευθύνσεων, που περιλαμβάνει τις πρώτες σελίδες των πιο γνωστών δικτυακών καταλόγων. Το υποσύστημα χρησιμοποιεί την οντολογία και τους υπερσυνδέσμους μεταξύ των δικτυακών εγγράφων για να εντοπίσει σχετικά έγγραφα. Το υποσύστημα δέχεται ως είσοδο μια θεματική οντολογία (Σχήμα 1, O_1) και επιστρέφει στην έξοδο μια συλλογή από έγγραφα του διαδικτύου (Web Document Collection – Σχήμα 1, WDC_1). Ως οντολογία ορίζεται ένα σύνολο εννοιών που συνδέονται μεταξύ τους με μια ή περισσότερες σχέσεις (π.χ. Α είναι Β, Γ περιέχει Δ κλπ.), ενώ κάθε έννοια περιέχει πολλές λέξεις που την περιγράφουν και αναφέρονται ως στιγμιότυπα της έννοιας. Παραδείγματα τέτοιων οντολογιών βρίσκονται στο δικτυακό τόπο www.daml.org .

Κατά τη ***διαδικασία συλλογής διαδικτυακών εγγράφων*** (Document Collection Process – Σχήμα 2, DCP) που ακολουθείται στο υποσύστημα αυτό, από κάθε σελίδα του αρχικού συνόλου εξάγονται όλοι οι υπερσύνδεσμοι που αυτή περιέχει και για κάθε υπερσύνδεσμο, 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 λεκτική περιγραφή που φέρει ο υπερσύνδεσμος. Η λεκτική

περιγραφή ανάγεται σε ένα σύνολο εννοιών της οντολογίας, με χρήση της γνώσης που μεταφέρει η οντολογία καθώς και ενός λεξιλογικού θησαυρού. Το σύνολο των εννοιών αποτελεί την σημασιολογική περιγραφή του υπερσυνδέσμου. Οι στόχοι των συνδέσμων για τους οποίους έχει προκύψει σημασιολογική περιγραφή αποθηκεύονται. Μαζί με αυτούς αποθηκεύεται η σημασιολογική και λεκτική περιγραφή που προκύπτει. Αν η σημασιολογική περιγραφή ενός υπερσυνδέσμου είναι κενή, ο σύνδεσμος αγνοείται.

Οι στόχοι που αποθηκεύθηκαν στο πρώτο επίπεδο και σημειώθηκαν για επίσκεψη, αποτελούν τη βάση για το επόμενο επίπεδο συλλογής πληροφορίας. Το υποσύστημα επισκέπτεται και επεξεργάζεται διαδοχικά τις ιστοσελίδες που αποθηκεύθηκαν, εξάγει τους υπερσυνδέσμους και τις λεκτικές περιγραφές που αυτοί περιέχουν, παράγει τις σημασιολογικές περιγραφές και αποθηκεύει τους στόχους τους επόμενου επιπέδου. Αν ο στόχος ενός συνδέσμου έχει ήδη αποθηκευθεί, η επιπλέον εννοιολογική περιγραφή προστίθεται στην υπάρχουσα, αυξάνοντας τη βαρύτητα της περιγραφής, ενώ δεν σημειώνεται για επίσκεψη.

Η διαδικασία επαναλαμβάνεται μέχρις ότου κανένας από τους στόχους ενός επιπέδου δεν έχει σημειωθεί για επίσκεψη. Το αποτέλεσμα της διαδικασίας είναι η δημιουργία μιας συλλογής ηλεκτρονικών εγγράφων του διαδικτύου (Web Document Collection – Σχήμα 2, WDC).

<u>Υποσύστημα εξαγωγής και εμπλουτισμού πληροφορίας</u>

Το υποσύστημα εξαγωγής και εμπλουτισμού πληροφορίας (Information Miner– Σχήμα 1, IM) εφαρμόζεται είτε στην συλλογή που προήλθε από το υποσύστημα συλλογής πληροφορίας (Σχήμα 1, WDC_1), είτε σε μια υπάρχουσα συλλογή ιστοσελίδων, είτε σε μια υπάρχουσα συλλογή ηλεκτρονικών εγγράφων γενικότερα (Σχήμα 1, DC_2). Το υποσύστημα δέχεται επίσης στην είσοδο μια οντολογία (Σχήμα 1, O_2). Το υποσύστημα εξάγει τις σημαντικότερες λέξεις από τους εισερχόμενους υπερσυνδέσμους μιας ιστοσελίδας στις δύο πρώτες περιπτώσεις και από το περιεχόμενο του ηλεκτρονικού εγγράφου στην τρίτη. Η λεκτική πληροφορία που εξάγεται για κάθε έγγραφο ή ιστοσελίδα ανάγεται σε σημασιολογική (έννοιες της οντολογίας) όπως πριν και αποθηκεύεται στη βάση δεδομένων (Σχήμα 1, DB), εμπλουτίζοντας έτσι τη διαθέσιμη πληροφορία για κάθε έγγραφο.

Κατά τη ***διαδικασία εξαγωγής και εμπλουτισμού της πληροφορίας*** των εγγράφων (Information Enhancement Process – Σχήμα 3, IEP), η συλλογή εγγράφων (Document Collection – Σχήμα 3, DC) αναλύεται ως προς το περιεχόμενο της και το περιεχόμενο των εισερχόμενων υπερσυνδέσμων όταν πρόκειται για ιστοσελίδες. Οι συχνότερα εμφανιζόμενες λέξεις στα περιεχόμενα ενός εγγράφου, ή αντίστοιχα οι λέξεις που εμφανίζονται στους περισσότερους εισερχόμενους συνδέσμους μιας ιστοσελίδας, εξάγονται για κάθε έγγραφο. Οι λέξεις αυτές απεικονίζονται σε έννοιες μιας οντολογίας (Ontology – Σχήμα 3, O) με χρήση του μέτρου Wu & Palmer και του θησαυρού WordNet. Το αποτέλεσμα της διαδικασίας είναι ένα σύνολο λέξεων και αντίστοιχων αριθμών εμφανίσεων, εννοιών της οντολογίας και των αντίστοιχων βαρών σχετικότητας για κάθε έγγραφο της συλλογής (Semantically Characterized Documents, Σχήμα 3, SCD)

<u>Υποσύστημα οργάνωσης εγγράφων</u>

Το υποσύστημα οργάνωσης εγγράφων (Document Organizer – Σχήμα 1, DO), ομαδοποιεί με αυτόματο τρόπο τα έγγραφα, σε μη προκαθορισμένες κατηγορίες. Τα έγγραφα που έχουν μεγάλη ομοιότητα σε σημασιολογικά χαρακτηριστικά (αντιπροσωπεύουν ίδιες ή παρόμοιες έννοιες της οντολογίας) τοποθετούνται στην ίδια ομάδα. Το υποσύστημα ενημερώνει τη Βάση δεδομένων (Σχήμα 1, DB) με τις παραγόμενες πληροφορίες για τις ομάδες που σχηματίζονται.

Η διαδικασία οργάνωσης των εγγράφων (Document Management Process – Σχήμα 3 DMP) οργανώνει τα έγγραφα που έχουν προηγουμένως χαρακτηριστεί με σημασιολογικές περιγραφές (Σχήμα 3, SCD) σε ομάδες εγγράφων που εμφανίζουν ομοιότητα μεταξύ τους (Document Groups - Σχήμα 3, DG).

Η εύρεση της ομοιότητας δύο εγγράφων σε σημασιολογικό επίπεδο γίνεται με τη χρήση ενός νέου μέτρου που λαμβάνει υπόψη του την απόσταση των σημασιολογικών περιγραφών των δύο εγγράφων.

Έστω $A=\{(w_i, k_i)\}$, και $B=\{(v_i, h_i)\}$, δύο έγγραφα - σύνολα όρων οντολογίας με βάρη, όπου k_i, h_i οι όροι της θεματικής οντολογίας και w_i, $v_i \leq 1$ τα αντίστοιχα βάρη.

Η ομοιότητα των δύο εγγράφων δίνεται από το μέτρο:

$$\zeta(\mathcal{A},\mathcal{B}) = \frac{1}{2}\left[\left(\frac{1}{K}\sum_{i=1}^{|\mathcal{A}|}\max_{j\in[1,|\mathcal{B}|]}\left(\lambda_{i,j}\times S_{W\&P}\left(k_i,h_j\right)\right)\right)+\left(\frac{1}{H}\sum_{i=1}^{|\mathcal{B}|}\max_{j\in[1,|\mathcal{A}|]}\left(\mu_{i,j}\times S_{W\&P}\left(h_i,k_j\right)\right)\right)\right]$$

όπου $S_{W\&P}(k_i, h_j)$ η ομοιότητα κατά Wu και Palmer των όρων k_i και h_j,

$$\lambda_{i,j} = \frac{w_i + v_j}{2 \times max(w_i, v_j)}$$ και $$K = \sum_{i=1}^{|A|} \lambda_{i,x(i)}$$ με

$$x(i) = x \mid \lambda_{i,x} \times S_{W\&P}(k_i, h_x) = \max_{j \in [1,|B|]} (\lambda_{i,x} \times S_{W\&P}(k_i, h_j))$$

Τα έγγραφα ομαδοποιούνται με χρήση ενός αλγορίθμου ομαδοποίησης που λειτουργεί χωρίς ανθρώπινη επίβλεψη. Το αποτέλεσμα του υποσυστήματος είναι ο εμπλουτισμός της πληροφορίας για κάθε έγγραφο με ένα αύξοντα αριθμό ομάδας. Επιπλέον στη ΒΔ αποθηκεύεται για κάθε ομάδα εγγράφων, ο αύξων αριθμός και ένα σύνολο εννοιών της οντολογίας που αποτελεί την περιγραφή της ομάδας (7) και εξάγεται από τη σημασιολογική πληροφορία του συνόλου των εγγράφων της ομάδας. Η περιγραφή της κάθε ομάδας προκύπτει από συγχώνευση των περιγραφών των εγγράφων της ομάδας. Οι έννοιες που εμφανίζονται στις λιγότερες σελίδες της ομάδας αντικαθίστανται από τις αμέσως γενικότερες και οι έννοιες με διπλές εμφανίσεις συγχωνεύονται. Η οργάνωση των εγγράφων σε ομάδες γίνεται με τρόπο ιεραρχικό, έτσι ώστε τα έγγραφα να διαμοιράζονται στις ομάδες που βρίσκονται στο ίδιο επίπεδο της ιεραρχίας. Η ιεραρχική οργάνωση σε συνδυασμό με την περιγραφή που εξάγεται για κάθε ομάδα διευκολύνει την περιήγηση στα έγγραφα της συλλογής και τον εντοπισμό των εγγράφων που έχουν συγκεκριμένα σημασιολογικά χαρακτηριστικά.

<u>Υποσύστημα διαχείρισης ερωτήσεων</u>

Το υποσύστημα διαχείρισης ερωτήσεων (Query Management – Σχήμα 1, QM) δέχεται ως είσοδο ερωτήσεις χρηστών υπό τη μορφή λίστας λέξεων (Σχήμα 1, Q_3) και επιστρέφει στην έξοδο έγγραφα της συλλογής που εμφανίζουν μεγάλη σημασιολογική ομοιότητα με την ερώτηση. Η απάντηση (Σχήμα 1, R_3) περιλαμβάνει τα έγγραφα ομαδοποιημένα (στις ομάδες που δημιουργήθηκαν από το υποσύστημα οργάνωσης εγγράφων) και ταξινομημένα με βάση την ομοιότητα προς την ερώτηση.

Η ***διαδικασία διαχείρισης ερωτήσεων*** των χρηστών (Query Management Process - Σχήμα 4, QMP) αρχικά μετατρέπει την λίστα λέξεων του ερωτήματος (Query - Σχήμα 4, Q) σε λίστα εννοιών της οντολογίας (με τον ίδιο τρόπο που περιγράφηκε στα προηγούμενα υποσυστήματα) και κατά συνέπεια την λεκτική ερώτηση σε σημασιολογική ερώτηση. Για την απάντηση της σημασιολογικής ερώτησης

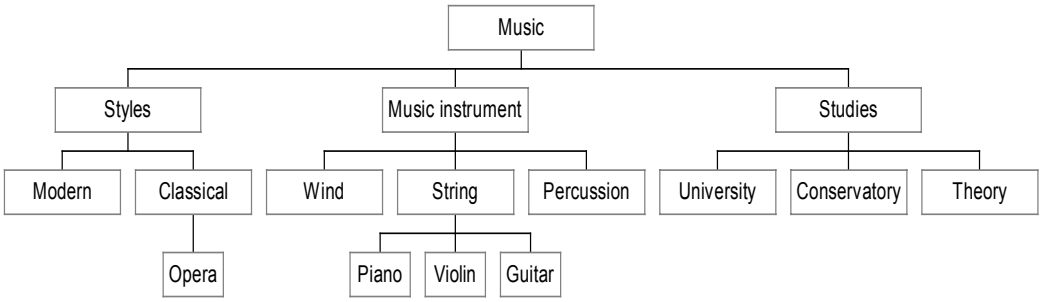
υπολογίζεται αρχικά η ομοιότητα της ερώτησης με τις σημασιολογικές περιγραφές των ομάδων που έχουν δημιουργηθεί. Για την απάντηση επιλέγονται οι ομάδες με τις πλησιέστερες σημασιολογικές περιγραφές και τα έγγραφά τους εμφανίζονται ταξινομημένα με φθίνουσα σειρά εγγύτητας προς την σημασιολογική ερώτηση (Sorted Answer Groups – Σχήμα 4, SAG).

Παράδειγμα

Στη συνέχεια παραθέτουμε ένα ολοκληρωμένο παράδειγμα χρήσης του συστήματος που επιδεικνύει τα καινοτομικά του στοιχεία, σε σύγκριση με υπάρχοντα συστήματα.

Οντολογία

Αρχικά ορίζουμε μια θεματική οντολογία σχετική με τη μουσική. Μια απλοποιημένη μορφή της οντολογίας (περιορίζεται σε μια ιεραρχία εννοιών) παρουσιάζεται στο ακόλουθο σχήμα.



Εξαγωγή λέξεων

Στο ακόλουθο παράδειγμα εξηγείται η διαδικασία εξαγωγής λέξεων από έναν υπερσύνδεσμο προς τη σελίδα της Μουσικής Βιβλιοθήκης του Πανεπιστημίου του Β. Τέξας (http://www.library.unt.edu/music/default.htm).

```
<p><a href="http://www.lib.unc.edu/music/">University of North Carolina, Chapel Hill - Music Library</a></p><p><a href="http://www.library.unt.edu/music/default.htm"> University of North Texas - Music Library</a> <br> &amp;quot;...one of the largest university music collections in the United States, has over 250,000 volumes of books, periodicals,
```

Οι λέξεις που εξάγονται από το υπερκείμενο (hypertext) για τον πιο πάνω υπερσύνδεσμο είναι:

- Από την περιοχή εντός του συνδέσμου (σκούρο φόντο λευκά γράμματα) εξάγονται οι λέξεις: University, North, Texas, Music, Library.

- Από την περιοχή 100 χαρακτήρων πριν το σύνδεσμο (γκρι περιοχή) δεν εξάγεται καμία λέξη καθώς η περιοχή ψαλιδίζεται στο σημείο εμφάνισης της ετικέτας (tag) που υποδηλώνει τα όριο άλλου υπερσυνδέσμου.
- Από τους 100 χαρακτήρες μετά το σύνδεσμο (γκρι περιοχή) αφού αφαιρέσουμε λέξεις που αποτελούν κοινές λέξεις (stop-words) όπως "one", "of", "the", "in", "has", "over" απομένουν οι λέξεις: largest, university, music, collection, United, States.

Οι λέξεις συνεπώς που έχουν εξαχθεί για το συγκεκριμένο σύνδεσμο, μαζί με τον αριθμό εμφανίσεων είναι: {(university,2), (north,1), (texas,1), (music,2), (library,1), (largest,1), (collection,1), (united,1), (states,1)} }.

Απεικόνιση σε έννοιες της οντολογίας

Ακολούθως αυτές οι λέξεις απεικονίζονται στον πλησιέστερο όρο της θεματικής οντολογίας. Δίνουμε ένα απλουστευμένο παράδειγμα του τρόπου με τον οποίο απεικονίζεται σε κάποιο όρο της οντολογίας η λέξη library.

Ο θησαυρός Wordnet που χρησιμοποιούμε δίνει για τη λέξη library πέντε διαφορετικές έννοιες, και για καθεμία από αυτές μια «αλυσίδα» γενικότερων εννοιών:

Library (a room where books are kept)
=> room
=> area
=> structure, construction
=> artifact, artefact
=> object, physical object
=> entity, something

Το σύστημα διατηρεί τις έννοιες και τις «αλυσίδες» γενικεύσεων για τους όρους της θεματικής οντολογίας. Το σύστημα απεικόνισης εξετάζει όλους τους συνδυασμούς εννοιών των όρων της οντολογίας και της λέξης που θέλουμε να απεικονίσουμε και εντοπίζει το ζεύγος εκείνο (έννοια λέξης, έννοια όρου οντολογίας) που δίνει τη μεγαλύτερη ομοιότητα μεταξύ των αλυσίδων γενικεύσεων. Η ομοιότητα υπολογίζεται με χρήση του μέτρου ομοιότητας Wu και Palmer σύμφωνα με το οποίο η ομοιότητα δύο ιεραρχιών εξαρτάται από το μήκος τους και το μήκος της ιεραρχίας των κοινών προγόνων. Για παράδειγμα η ομοιότητα της έννοιας της λέξης Library που παρουσιάστηκε προηγουμένως και της ακόλουθης έννοιας της λέξης Studies:

Study (a room used for reading and writing and studying)

=> room

=> area

=> structure, construction

=> artifact, artefact

=> object, physical object

=> entity, something

δίνεται ώς Sim=$\frac{2*length(commonpath)}{length(path_1)+lenght(path_2)}=\frac{2*6}{7+7}$=0.86 και είναι η μέγιστη δυνατή ομοιότητα από όλους τους συνδυασμούς.

Εφόσον η μέγιστη ομοιότητα ξεπερνά ένα συγκεκριμένο κατώφλι, η λέξη απεικονίζεται στην αντίστοιχη έννοια. Συνεπώς η λέξη Library απεικονίζεται στην έννοια Studies. Αν η μέγιστη ομοιότητα είναι μικρότερη από το κατώφλι η λέξη δεν απεικονίζεται σε καμία έννοια. Βάζοντας για παράδειγμα το κατώφλι=0.6 οι λέξεις north, texas, united, states, largest και collection δεν απεικονίζονται σε όρο της οντολογίας. Είναι σαφές ότι οι λέξεις music και university θα απεικονιστούν ως έχουν στους όρους της οντολογίας με ομοιότητα 1.

Οι έννοιες της οντολογίας που προκύπτουν για το συγκεκριμένο σύνδεσμο, μαζί με το βαθμό ομοιότητας είναι: {(university,1), (music,1), (studies,0.86)}.

Συλλογή ιστοσελίδων

Το υποσύστημα συλλογής σελίδων ξεκινά τη συλλογή παίρνοντας ως είσοδο τη συγκεκριμένη οντολογία. Η συγκέντρωση αρχίζει από τις πρώτες σελίδες ορισμένων δικτυακών καταλόγων (π.χ. www.yahoo.com, www.dmoz.org) που περιέχουν διάφορες θεματικές κατηγορίες και συνδέσμους προς τις σημαντικότερες σελίδες σε κάθε κατηγορία. Το υποσύστημα εξάγει λέξεις από τους συνδέσμους των σελίδων αυτών και στη συνέχεια απεικονίζει τις λέξεις αυτές σε όρους της πιο πάνω οντολογίας όπως περιγράφηκε προηγούμενα.

Στην περίπτωση του συνδέσμου του προηγούμενου παραδείγματος είναι σαφές ότι το υποσύστημα συλλογής θα τον επιλέξει για ανάλυση στο επόμενο στάδιο συλλογής, καθώς περιέχει λέξεις που απεικονίζονται σε έννοιες της θεματικής οντολογίας και συνεπώς η σελίδα που υποδεικνύεται από αυτόν είναι πιθανώς σχετική με το θέμα.

Μέτρο ομοιότητας εγγράφων

Οι σελίδες που προκύπτουν από το υποσύστημα συλλογής χαρακτηρίζονται με λέξεις και έννοιες της θεματικής οντολογίας μέσα από τους υπερσυνδέσμους με τον τρόπο που περιγράφηκε προηγουμένως. Για παράδειγμα δύο έγγραφα που περιγράφονται από τις έννοιες A ={(0.5, university),(1.0, opera)} and B={(0.8, classical), (0.3, studies)}. Οι ομοιότητες μεταξύ των λέξεων κατά Wu και Palmer είναι:

$S_{W\&P}(\text{university, classical})_{1,1} = 0.33$; $S_{W\&P}(\text{university, studies})_{1,2} = 0.8$; $S_{W\&P}(\text{opera, classical})_{2,1} = 0.86$; $S_{W\&P}(\text{opera, studies})_{2,2} = 0.33$

Παρατηρούμε ότι η έννοια "university" είναι πιο κοντά στην έννοια "studies" και αντίστροφα όπως και οι έννοιες "opera " και " classical". Από τα βάρη υπολογίζουμε:

$$\lambda_{1,1}=\mu_{1,1}=(0,5+ 0,8)/1,6=0,81$$

$$\lambda_{2,2}=\mu_{2,2}=(1,0+0,3)/2=0,65$$

$$\lambda_{1,2}= \mu_{2,1}=(0,5+0,3)/1=0,8$$

$$\lambda_{2,1}= \mu_{1,2}=(1,0+0,8)/2=0,9$$

Τελικά η ομοιότητα ανάμεσα στα δύο έγγραφα είναι:

$$\zeta=\frac{1}{2}\times\left[\frac{1}{2}\times\left(\lambda_{1,2}\times S_{W\&P1,2}+\lambda_{2,1}\times S_{W\&P2,1}\right)+\frac{1}{2}\times\left(\mu_{2,1}\times S_{W\&P1,2}+\mu_{1,2}\times S_{W\&P2,1}\right)\right]$$

$$\zeta=\frac{1}{2}\times\left[\frac{1}{2}\times\left(0.8\times 0.8+0.9\times 0.86\right)+\frac{1}{2}\times\left(0.8\times 0.8+0.9\times 0.86\right)\right]=0.71$$

Αν χρησιμοποιούσαμε ένα μέτρο που εξετάζει το ακριβές ταίριασμα λέξεων ή εννοιών για τα δύο έγγραφα η ομοιότητα που θα υπολογίζαμε θα ήταν 0 αφού η τομή των συνόλων A και B είναι το κενό σύνολο.

Οργάνωση ιστοσελίδων

Το σύστημα χρησιμοποιεί έναν αλγόριθμο ομαδοποίησης (clustering) που τοποθετεί στην ίδια ομάδα έγγραφα που εμφανίζουν μεγάλη ομοιότητα και σε διαφορετικές ομάδες έγγραφα που εμφανίζουν μικρή ομοιότητα μεταξύ τους. Ο αριθμός των ομάδων στις οποίες θα χωριστούν τα έγγραφα της συλλογής δεν είναι προκαθορισμένος. Επιπλέον σύμφωνα με τον αλγόριθμο κάθε έγγραφο ανήκει σε μία το μόνο ομάδα. Έγγραφα που διαφέρουν από όλα τα υπόλοιπα θεωρούνται ότι αποτελούν μόνα τους ομάδα.

Στον ακόλουθο πίνακα απεικονίζονται οι σημασιολογικές περιγραφές για κάποια έγγραφα (E_1, E_2, κλπ) και η ομάδα στην οποία τοποθετούνται (O_1,O_2, κλπ) μετά την

εκτέλεση του αλγορίθμου. Για λόγους απλότητας έχουμε θεωρήσει το βάρος για κάθε έννοια στην περιγραφή ενός εγγράφου ίσο με 1. Αλλάζοντας παραμέτρους του αλγορίθμου παίρνουμε μια «χαλαρότερη» ομαδοποίηση των εγγράφων στην οποία θα έχουμε συγχωνεύσεις των ομάδων που δημιουργήθηκαν στην πρώτη ομαδοποίηση σε μεγαλύτερες ομάδες (G_1, G_2 κλπ.).

Έγγραφο	Περιγραφή	Επίπεδο 1	Επίπεδο 2
E_1	Classical, Violin	O_1 (Classical, String)	G_1 (Styles ,String)
E_2	Opera, Piano		
E_3	Modern, Guitar	O_2 (Modern, String)	
E_4	Modern, String		
E_5	Violin, Guitar, Piano	O_3 (String)	G_2 (String)
E_6	String		
E_7	University, Classical	O_4 (Studies, Classical)	G_3 (Studies, Classical, Music Instrument)
E_8	Studies, Opera		
E_9	Theory, Percussion	O_5 (Studies, Music Instrument)	
E_{10}	Guitar, Conservatory		

Από τα στοιχεία του πίνακα διαπιστώνουμε ότι στην ίδια ομάδα τοποθετούνται έγγραφα που πραγματεύονται παρόμοιες έννοιες. Αυτό δεν θα ήταν εφικτό αν χρησιμοποιούσαμε ένα μέτρο που θα βασιζόταν στο ακριβές ταίριασμα (exact matching) των περιγραφών καθώς τότε πολύ περισσότερα έγγραφα θα τοποθετούνταν σε ξεχωριστές ομάδες.

Με αυτό τον τρόπο οργάνωσης δημιουργούμε μια ιεραρχία στη συλλογή των εγγράφων η οποία περιέχει μια ομάδα με όλα τα έγγραφα στο κορυφαίο επίπεδο και υποομάδες αυτής στα επόμενα επίπεδα. Στις ομάδες του κάθε επιπέδου θα μοιράζονται όλα τα έγγραφα της συλλογής. Επιτρέπουμε έτσι στο χρήστη να εντοπίζει τα έγγραφα που τον ενδιαφέρουν ανάμεσα σε άλλα παρόμοια έγγραφα, επιλέγοντας σε κάθε επίπεδο την ομάδα που τον ενδιαφέρει.

Εξαγωγή περιγραφής ομάδας εγγράφων

Για τη διευκόλυνση του χρήστη το σύστημα παράγει για κάθε ομάδα από τις παραπάνω (O_1, O_2, ..., G_1, G_2, …) μια περιγραφή που αποτελείται από τη συγχώνευση των περιγραφών των εγγράφων της ομάδας. Για παράδειγμα για την ομάδα O_1 οι έννοιες που εμφανίζονται στα έγγραφά της είναι (Classical, Opera, Violin, Piano). Ανάγοντας τη λέξη Opera στην γενικότερή της Classical και τις λέξεις Violin και Piano στην γενικότερή τους String προκύπτει η περιγραφή (Classical, String) για την ομάδα O_1. Οι παραγόμενες περιγραφές για τις ομάδες του πίνακα απεικονίζονται δίπλα στο όνομα της κάθε ομάδας.

Αναζητήσεις – Διαχείριση ερωτήσεων

Στο ακόλουθο παράδειγμα γίνεται σαφής ο τρόπος με τον οποίο υλοποιείται η αναζήτηση στο προτεινόμενο σύστημα. Θεωρούμε ότι ο χρήστης υποβάλει την ερώτηση «courses strings». Μια συνήθης μηχανή αναζήτησης θα επέστρεφε όλα τα έγγραφα που περιέχουν τουλάχιστον μια από τις λέξεις της ερώτησης και θα αγνοήσει όσα έγγραφα περιέχουν παρόμοιες έννοιες, π.χ. όσα έγγραφα αναφέρονται σε μαθήματα κιθάρας. Στο συγκεκριμένο παράδειγμα η μηχανή αναζήτησης δεν θα επέστρεφε κανένα από τα έγγραφα (E_1, ..., E_{10}).

Το προτεινόμενο σύστημα θα ανάγει αρχικά την ερώτηση σε μια λίστα εννοιών της οντολογίας π.χ. «Studies Strings», και στη συνέχεια θα εντοπίσει ομάδες και έγγραφα που χαρακτηρίζονται από συγγενείς με αυτές έννοιες.

Από τα στοιχεία του πίνακα γίνεται σαφές ότι από τις ομάδες στο Επίπεδο 2, η πλησιέστερη είναι η ομάδα G_3 (Studies, Classical, Music Instrument) και από τις υποομάδες αυτής πλησιέστερη η O_5 (Studies, Music Instrument). Τα έγγραφα E_9 και E_{10} θα αποτελέσουν την απάντηση στην ερώτηση.

Η ταξινόμηση των εγγράφων θα γίνει με βάση την σημασιολογική εγγύτητα των περιγραφών τους στην ερώτηση, έτσι το έγγραφο E_{10} θα θεωρηθεί σχετικότερο από το E_9 καθώς το ζευγάρι εννοιών που το περιγράφει είναι πλησιέστερο στο ζευγάρι εννοιών που αντιπροσωπεύει η ερώτηση.

ΑΞΙΩΣΕΙΣ

1. Σύστημα για την ευκολότερη διαχείριση συλλογών ηλεκτρονικών εγγράφων. Το σύστημα περιλαμβάνει τα ακόλουθα υποσυστήματα:

- Το υποσύστημα εξαγωγής και εμπλουτισμού πληροφορίας (Information Miner– Σχήμα 1, IM), που αναλαμβάνει το χαρακτηρισμό των εγγράφων, με ανάλυση του περιεχομένου τους και εξαγωγή λεκτικής πληροφορίας από αυτό. Επίσης αναλαμβάνει το χαρακτηρισμό των εγγράφων με σημασιολογική πληροφορία (όροι της οντολογίας που παρέχεται ως είσοδος στο σύστημα - Σχήμα 1, O_2), η οποία παράγεται από την εξαγόμενη λεκτική πληροφορία με χρήση ενός λεξιλογικού θησαυρού (Σχήμα 1, WN) και ενός μηχανισμού απεικόνισης.
- Το υποσύστημα οργάνωσης εγγράφων (Document Organizer – Σχήμα 1, DO) που αναλαμβάνει την αυτόματη οργάνωση των εγγράφων της συλλογής, κάνοντας χρήση της σημασιολογικής πληροφορίας. Η σημασιολογική οργάνωση των εγγράφων διευκολύνει την περιήγηση των χρηστών στην υπάρχουσα συλλογή πληροφορίας.
- Το υποσύστημα διαχείρισης ερωτήσεων (Query Management – Σχήμα 1, QM) που είναι υπεύθυνο για την επεξεργασία ερωτήσεων, με ανάκτηση των σχετικότερων προς την ερώτηση εγγράφων ή ομάδων εγγράφων.

2. Μέθοδος για την καλύτερη διαχείριση συλλογών ηλεκτρονικών εγγράφων, η οποία περιλαμβάνει:

- το λεκτικό χαρακτηρισμό των εγγράφων, με την εξαγωγή των σημαντικότερων λέξεων από τα περιεχόμενά τους (Information Enhancement Process, Σχήμα 3, IEP). Η αξιολόγηση της σημαντικότητας γίνεται με κριτήρια όπως η συχνότητα εμφάνισης των λέξεων στα έγγραφα κλπ.
- το σημασιολογικό χαρακτηρισμό των εγγράφων, με την απεικόνιση των εξαγομένων λέξεων σε έννοιες μιας θεματικής οντολογίας, την οποία παρέχει ο χρήστης του συστήματος (Information Enhancement Process, Σχήμα 3, IEP). Η απεικόνιση της λεκτικής πληροφορίας σε σημασιολογική γίνεται με τη χρήση ενός εννοιολογικού θησαυρού (Σχήμα 3, WN). Κατά τη διαδικασία αυτή οι λέξεις που

χαρακτηρίζουν το κάθε έγγραφο απεικονίζονται στις πλησιέστερες έννοιες της οντολογίας με χρήση ενός μέτρου ομοιότητας σε ιεραρχίες.

- την ομαδοποίηση των εγγράφων της συλλογής με βάση την μεταξύ τους ομοιότητα (Σχήμα 3, DMP). Η ομοιότητα υπολογίζεται κάνοντας χρήση των σημασιολογικών περιγραφών των εγγράφων, της θεματικής οντολογίας και ενός μέτρου για τον υπολογισμό της εγγύτητας μεταξύ δύο συνόλων όρων της οντολογίας.
- το χαρακτηρισμό των ομάδων που προκύπτουν με έννοιες της οντολογίας.

3. Όπως στην Αξίωση 1, όπου στο υποσύστημα ΙΜ οι λέξεις και έννοιες της οντολογίας που χαρακτηρίζουν τα έγγραφα, τις ομάδες εγγράφων ή τις ερωτήσεις συνοδεύονται από βάρη που υποδηλώνουν την σημαντικότητά και σχετικότητά τους με το περιεχόμενο των εγγράφων. Τα βάρη αυτά λαμβάνονται υπ' όψη στον υπολογισμό της ομοιότητας μεταξύ εγγράφων, μεταξύ ερώτησης και περιγραφής ομάδας, μεταξύ ερώτησης και εγγράφου.

4. Όπως στην Αξίωση 3, όπου στο υποσύστημα ΙΜ:

- Τα βάρη των λέξεων στη λεκτική περιγραφή ενός εγγράφου δηλώνουν τον αριθμό εμφανίσεων της λέξης στο έγγραφο.
- Για τον υπολογισμό των βαρών των εννοιών στη σημασιολογική περιγραφή ενός εγγράφου λαμβάνονται υπόψη:
  - Το βάρος κάθε λέξης που απεικονίζονται στην ίδια έννοια της οντολογίας και
  - η ομοιότητα της λέξης με την έννοια. Η ομοιότητα υπολογίζεται με χρήση του θησαυρού και ενός μέτρου ομοιότητας.
- Για τον υπολογισμό των βαρών των εννοιών στην περιγραφή μιας ομάδας εγγράφων λαμβάνονται υπόψη:
  - Ο αριθμός των εγγράφων της ομάδας που χαρακτηρίζονται από κάθε έννοια
  - Τα αντίστοιχα βάρη της έννοιας σε κάθε ένα από τα έγγραφα.

Παρόμοια υπολογίζονται, στο υποσύστημα QM, τα βάρη των εννοιών που προκύπτουν από την απεικόνιση μιας λεκτικής ερώτησης στην αντίστοιχη σημασιολογική.

5. Όπως στην Αξίωση 4, με εφαρμογή στα έγγραφα του Παγκόσμιου Ιστού, όπου στο υποσύστημα IM:

- Το σύστημα χαρακτηρίζει τα έγγραφα της συλλογής (που είναι ιστοσελίδες) αναλύοντας εκτός από το περιεχόμενό τους και τους εισερχόμενους προς αυτά υπερσυνδέσμους.
- Η λεκτική πληροφορία που εξάγεται για τα έγγραφα περιλαμβάνει τις λέξεις που εμφανίζονται με μεγαλύτερη συχνότητα στους εισερχόμενους συνδέσμους κάθε εγγράφου ξεχωριστά, αγνοώντας κοινότυπες λέξεις (stopwords) όπως άρθρα, μόρια κλπ.
- Τα βάρη των λέξεων στη λεκτική περιγραφή ενός εγγράφου δηλώνουν τον αριθμό διαφορετικών εισερχόμενων συνδέσμων ενός εγγράφου στον οποίο εμφανίζεται κάθε λέξη.

6. Όπως στην Αξίωση 1. Επιπλέον περιλαμβάνει το υποσύστημα συλλογής πληροφορίας (Thematic Crawler – Σχήμα 1, TC) . Το υποσύστημα είναι συνδεδεμένο με το διαδίκτυο και αναλαμβάνει τη συγκέντρωση εγγράφων από τον παγκόσμιο ιστό με μόνη είσοδο μια **θεματική οντολογία**. Τα έγγραφα που συλλέγονται είναι μόνο αυτά που εισερχόμενοι σύνδεσμοι τα χαρακτηρίζουν με λέξεις που απεικονίζονται σε έννοιες της οντολογίας. Το σύστημα ξεκινά από γνωστές δικτυακές πύλες και ακολουθεί τους συνδέσμους που περιέχουν έννοιες από την θεματική οντολογία.

7. Η μέθοδος της αξίωσης 2 όπου:

- Για το χαρακτηρισμό των εγγράφων:
    - Η λεκτική πληροφορία που εξάγεται από τα έγγραφα περιλαμβάνει τις λέξεις που εμφανίζονται με μεγαλύτερη συχνότητα στο κάθε έγγραφο ξεχωριστά, αγνοώντας κοινότυπες λέξεις (stopwords) όπως άρθρα, μόρια κλπ.

- Η σημασιολογική πληροφορία για κάθε έγγραφο αποτελείται από όρους μιας θεματικής οντολογίας,. Η θεματική οντολογία περιέχει περιορισμένο αριθμό όρων (εννοιών) που ενδιαφέρουν τον χρήστη του συστήματος και παρέχεται ως είσοδος στο σύστημα. Η απεικόνιση της λεκτικής πληροφορίας σε σημασιολογική γίνεται με τη χρήση ενός εννοιολογικού θησαυρού.

- Η οργάνωση των εγγράφων της συλλογής γίνεται αυτόματα κάνοντας χρήση της εγγύτητας των σημασιολογικών χαρακτηρισμών των εγγράφων στην οντολογία. Τα έγγραφα ομαδοποιούνται και κάθε ομάδα που προκύπτει περιγράφεται με όρους της οντολογίας

Η επεξεργασία ερωτήσεων που γίνονται με λεκτικά κριτήρια χρησιμοποιεί την προηγούμενη ομαδοποίηση και την σημασιολογική εγγύτητα των ερωτήσεων στις περιγραφές των ομάδων και των εγγράφων τους. Τα αποτελέσματα των αναζητήσεων εμφανίζονται ομαδοποιημένα και ταξινομημένα με φθίνουσα σειρά σχετικότητας στην ερώτηση.

8. Όπως στην Αξίωση 2 και επιπλέον οι λέξεις και οι έννοιες της οντολογίας που χαρακτηρίζουν τα έγγραφα, τις ομάδες εγγράφων ή τις ερωτήσεις συνοδεύονται από βάρη που υποδηλώνουν την σημαντικότητά και σχετικότητά τους με το περιεχόμενο των εγγράφων. Τα βάρη αυτά λαμβάνονται υπ' όψη στον υπολογισμό της ομοιότητας

- μεταξύ εγγράφων,
- μεταξύ ερώτησης και περιγραφής ομάδας,
- μεταξύ ερώτησης και εγγράφου.

9. Μέθοδος για την απάντηση λεκτικών ερωτήσεων με έγγραφα μιας συλλογής που έχει δημιουργηθεί και οργανωθεί με τη μέθοδο που περιγράφεται στις Αξιώσεις 2 και 8 αντίστοιχα. Οι ερωτήσεις υποβάλλονται ως σύνολα λέξεων (λεκτικά κριτήρια) τα οποία το υποσύστημα διαχείρισης ερωτήσεων (Σχήμα 1, QM) μετατρέπει σε σύνολα εννοιών της θεματικής οντολογίας (σημασιολογικά κριτήρια). Ο εντοπισμός των κατάλληλων εγγράφων γίνεται πρώτα σε επίπεδο σχετικών ομάδων κάνοντας χρήση της ομαδοποίησης που προκύπτει από την Αξίωση 8 και της σημασιολογικής

εγγύτητας των κριτηρίων στις περιγραφές των ομάδων και των εγγράφων τους και στη συνέχεια σε επίπεδο εγγράφου. Τα έγγραφα που επιστρέφονται ως αποτελέσματα των αναζητήσεων εμφανίζονται ομαδοποιημένα και ταξινομημένα με φθίνουσα σειρά σχετικότητας στην ερώτηση.

10. Αποτελεί εξειδίκευση της μεθόδου που περιγράφεται στην Αξίωση 2, για τα έγγραφα του Παγκόσμιου Ιστού. Τα έγγραφα της συλλογής (που είναι ιστοσελίδες) χαρακτηρίζονται με σημασιολογικά χαρακτηριστικά που προκύπτουν από τη χρήση θεματικής οντολογίας.

- Από μια συλλογή των πρώτων σελίδων (homepages) δικτυακών πυλών εξάγεται λεκτική πληροφορία από όλους τους υπερσυνδέσμους
- Η λεκτική πληροφορία απεικονίζεται σε σημασιολογική πληροφορία όπως στην Αξίωση 2.
- Στη συνέχεια επιλέγονται μόνο οι σύνδεσμοι που περιέχουν έννοιες από την θεματική οντολογία και προστίθενται στη συλλογή οι σελίδες στις οποίες αυτοί δείχνουν.

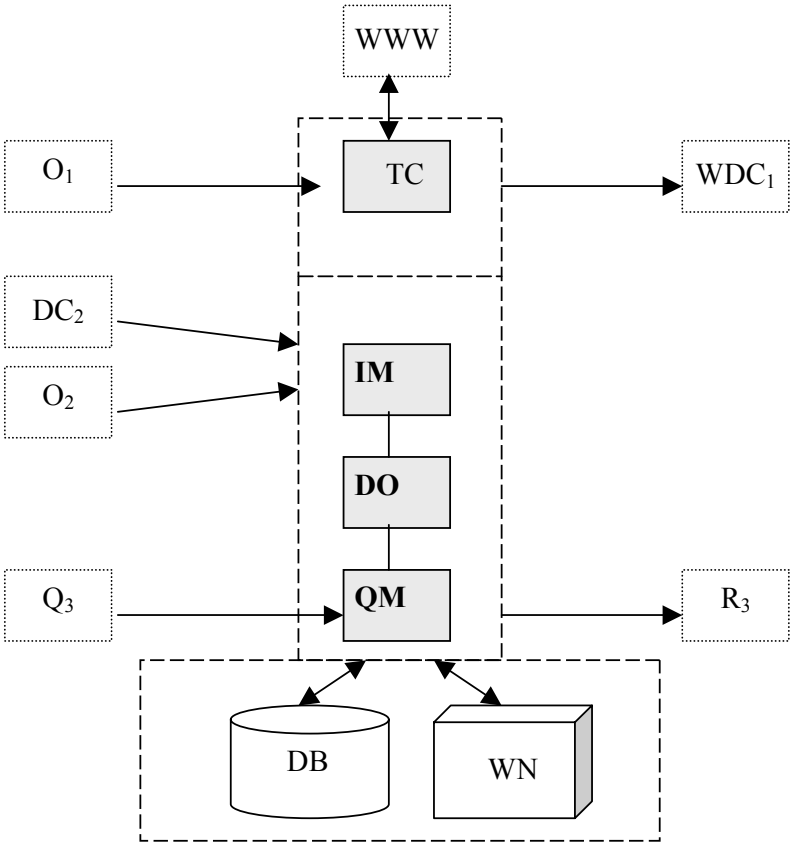
ΠΕΡΙΛΗΨΗ

Σύστημα για τη συγκέντρωση, περιγραφή, οργάνωση και διαχείριση εγγράφων με βάση σημασιολογικά χαρακτηριστικά, που πηγάζουν από μια θεματική οντολογία.

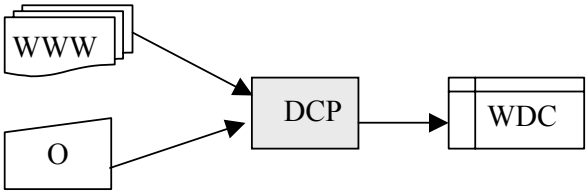
Το σύστημα αποτελείται από 4 υποσυστήματα (Thematic Crawler - TC, Information Miner - IM, Document Organizer - DO, Query Manager - QM) μια Βάση Δεδομένων (Database - DB) για την αποθήκευση της πληροφορίας, και τη βάση γνώσης (WordNet – WN) εγκατεστημένα στον ίδιο ή σε διαφορετικούς ηλεκτρονικούς υπολογιστές συνδεδεμένους μεταξύ τους σε τοπικό δίκτυο.

Το σύστημα βρίσκεται σε υπολογιστή συνδεδεμένο με το διαδίκτυο (WWW) και προσφέρει τη δυνατότητα αυτόματης συγκέντρωσης εγγράφων από το διαδίκτυο που σχετίζονται με κάποιο θέμα, κάνοντας χρήση της οντολογίας και των υπερσυνδέσμων μεταξύ των δικτυακών εγγράφων. Το υποσύστημα συλλογής πληροφορίας δέχεται ως είσοδο μια θεματική οντολογία O_1 και επιστρέφει στην έξοδο μια συλλογή από έγγραφα του διαδικτύου. Ακολούθως το υποσύστημα εξαγωγής και εμπλουτισμού πληροφορίας εξάγει τις σημαντικότερες λέξεις για κάθε έγγραφο και τις αντιστοιχεί σε έννοιες της οντολογίας. Το υποσύστημα οργάνωσης εγγράφων, ομαδοποιεί με αυτόματο τρόπο τα έγγραφα, σε μη προκαθορισμένες κατηγορίες βασιζόμενο στην ομοιότητα των σημασιολογικών χαρακτηριστικών τους. Η πληροφορία που παράγεται από τα δύο υποσυστήματα (IM και DO) αποθηκεύεται στη βάση δεδομένων (DB). Το υποσύστημα απάντησης ερωτήσεων (Query Manager - QM) δέχεται στην είσοδο λεκτικές ερωτήσεις (Q_3) και απαντά με έγγραφα της συλλογής που είναι σημασιολογικά παρόμοια με τις έννοιες που αντιστοιχούν στην ερώτηση. Τα έγγραφα παρουσιάζονται ομαδοποιημένα και ταξινομημένα (R_3).

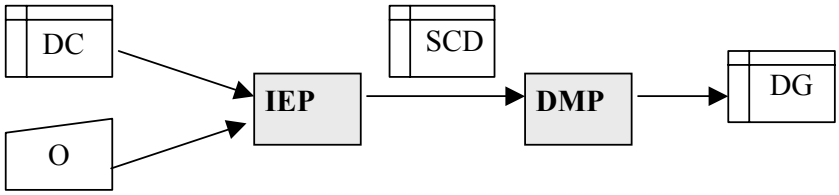
ΣΧΕΔΙΑ



Σχήμα 1 Αρχιτεκτονική του συστήματος (διακεκομμένη γραμμή)



Σχήμα 2 Διαδικασία συλλογής διαδικτυακών εγγράφων



Σχήμα 3 Διαδικασίες Εξαγωγής και Εμπλουτισμού πληροφορίας, Οργάνωσης Εγγράφων

Σχήμα 4 Διαδικασία διαχείρισης ερωτήσεων