Semantic Search – Σημασιολογική έρευνα

ΝΙΚΟΣ ΓΙΑΝΝΟΥΛΗΣ

Η εισαγωγή της σημασιολογίας στον παγκόσμιο ιστό θα οδηγήσει σε μια νέα γενιά υπηρεσιών οι οποίες θα βασίζονται περισσότερο στο περιεχόμενο, παρά στη σύνταξη. Οι μηχανές αναζήτησης θα παρέχουν αναζητήσεις, ανακτώντας τους πόρους οι οποίοι συνδέονται εννοιολογικά με τις πληροφοριακές ανάγκες του χρήστη. Τα ερωτήματα θα είναι δυνατό να διατυπώνονται με διάφορους τρόπους και να εκφραστούν στο σημασιολογικό επίπεδο ορίζοντας θέματα τα οποία θα πρέπει να ανακτηθούν από τον παγκόσμιο ιστό. Δραστηριότητες όπως οι υπηρεσίες στον ιστό και ο σημασιολογικός ιστός εργάζονται για τη δημιουργία ενός ιστού διανεμημένων δεδομένων τα οποία είναι δυνατό να γίνουν κατανοητά από μηχανή. Η παρούσα εργασία αφορά στην παρουσίαση της εφαρμογής που ονομάζεται "σημασιολογική έρευνα – semantic search", η οποία στηρίζεται στις πιο πάνω τεχνολογίες και έχει σχεδιαστεί προκειμένου να βελτιώσει την παραδοσιακή έρευνα του ιστού. Παρέχεται επίσης μια επισκόπηση του TAP, του πλαισίου εργασίας της εφαρμογής πάνω στο οποίο έχει χτιστεί η σημασιολογική έρευνα. Επίσης, παρουσιάζονται δύο εφαρμοσμένα συστήματα σημασιολογικής έρευνας τα οποία στηρίζονται στον προσδιορισμό της ερώτηση αναζήτησης, αυξάνοντας τα παραδοσιακά αποτελέσματα αναζήτησης με σχετικά δεδομένα που συγκεντρώνονται από διανεμημένους πόρους. Τέλος, αναφέρονται κάποια γενικά θέματα σχετικά με την έρευνα και τον σημασιολογικό ιστό (semantic web), καθώς και το πώς η κατανόηση της σημασιολογίας των όρων αναζήτησης μπορεί να χρησιμοποιηθεί για την παροχή βέλτιστων αποτελεσμάτων.

Semantic Web, semantic search, semantic retrieval

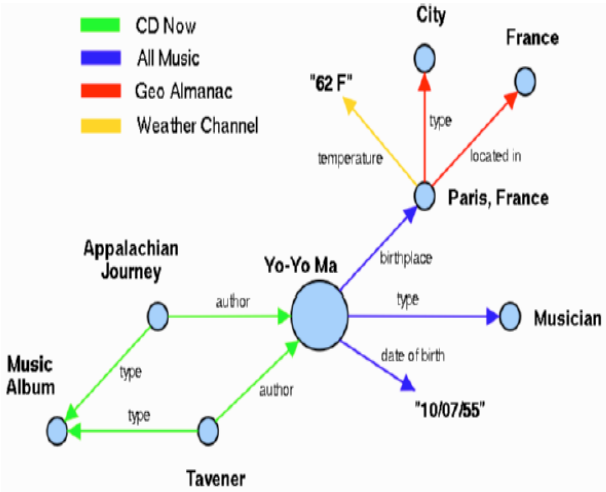
Εισαγωγή

Το διαδίκτυο, όπως το γνωρίζουμε σήμερα, είναι ένα διαρκώς αυξανόμενο μέσο, αφού ο αριθμός των διαδικτυακών τόπων μεγαλώνει ακόμα εκθετικά χάρη στην ευκολία και την ελευθερία που παρέχεται, σχεδόν στον καθένα, να προσθέσει νέα δεδομένα και ιστοσελίδες. Αυτή η ελευθερία παρουσιάζει σίγουρα αρκετά πλεονεκτήματα, ωστόσο συνεπάγεται και έναν αριθμό μειονεκτημάτων, αφού υπάρχει ελάχιστος ή καθόλου έλεγχος του περιεχομένου των ιστοσελίδων. Το γεγονός αυτό οδηγεί σε προβλήματα όπως η ακρίβεια των δεδομένων, η νομιμότητα (ζητήματα πνευματικών δικαιωμάτων) και η αισθητική. Σε ένα περισσότερο υψηλό επίπεδο σημειώνεται η έλλειψη ενός συστήματος δεικτοδότησης στον ιστό, η οποία καθιστά, μερικές φορές, αδύνατη την εύρεση της επιθυμητής πληροφορίας, ακόμα και αν αυτή είναι δημόσια διαθέσιμη στον ιστό. Κατά τη διάρκεια των ετών, οι μηχανικοί λογισμικού έχουν αναπτύξει διάφορα συστήματα ανάκτησης πληροφοριών μέσα στον ιστό. Μεταξύ αυτών, μηχανές αναζήτησης όπως το Google, το AltaVista και κατάλογοι όπως το Yahoo, είναι τα περισσότερο γνωστά. Αν και τα συγκεκριμένα εργαλεία είναι εξαιρετικά χρήσιμα, δε θεωρούνται πλήρη, αφού ο κύριος στόχος τους είναι η προσπάθεια παροχής των πιο ενδιαφερόντων αποτελεσμάτων αναζήτησης στην πρώτη σελίδα. Ένα ακόμα πρόβλημα των μηχανών αναζήτησης όπως λειτουργούν σήμερα είναι και το γεγονός ότι είναι αδύνατο να συνδέσουν μεταξύ τους συσχετισμένα γεγονότα τα οποία εμφανίζονται σε διαφορετικές ιστοσελίδες. Αν και υπάρχει ένας τεράστιος όγκος διαθέσιμων δεδομένων και στοιχείων, οι μηχανές αναζήτησης δε διαθέτουν καμία γνώση σχετικά με το περιεχόμενο μιας ιστοσελίδας με αποτέλεσμα να μην μπορούν να μετατρέψουν τα δεδομένα αυτά σε πληροφορία χρήσιμη για το χρήστη. Το διαδίκτυο, όπως το γνωρίζουμε σήμερα, έχει σχεδιαστεί για ανθρώπους και όχι για μηχανές. Ο παγκόσμιος ιστός ενσωματώνεται στις μέρες μας όλο και περισσότερο στην καθημερινή ζωή του ανθρώπου, γεγονός το οποίο παράλληλα αυξάνει τις απαιτήσεις. Το σημασιολογικό δίκτυο αποτελεί μια απάντηση σε αυτή την αύξηση των απαιτήσεων του χρήστη, αφού ο κύριος σκοπός του είναι η δημιουργία ενός περιβάλλοντος όπου όλες οι πληροφορίες θα είναι δυνατό να γίνουν κατανοητές από τα εργαλεία που στέλνονται από το χρήστη. Από τη στιγμή που τα δεδομένα γίνονται διαθέσιμα με τέτοιο τρόπο έτσι ώστε να γίνονται κατανοητά από τις μηχανές, τότε μπορούν αν θεωρηθούν ως πληροφορία χρήσιμη για το χρήστη.

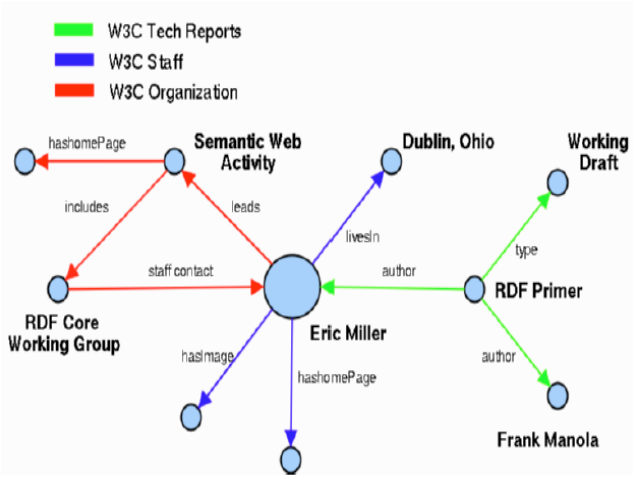
I. Ο ΣΗΜΑΣΙΟΛΟΓΙΚΟΣ ΙΣΤΟΣ – SEMANTIC WEB

Ο σημασιολογικός ιστός (semantic web) αποτελεί μια επέκταση του ήδη υπάρχοντος ιστού, στον οποίο δίνεται στην πληροφορία μια έννοια καθορισμένη με σαφήνεια, επιτρέποντας τη βέλτιστη συνεργασία μεταξύ ανθρώπου και υπολογιστή. Η βασική ιδέα είναι η ύπαρξη δεδομένων στον ιστό, ορισμένων και συνδεδεμένων με έναν τρόπο που να επιτρέπει την αποδοτική ανακάλυψη, αυτοματοποίηση, ενσωμάτωση και επαναχρησιμοποίησή τους μέσα σε διάφορες εφαρμογές. Πιο συγκεκριμένα, ο σημασιολογικός ιστός θα περιέχει πόρους οι οποία θα αντιστοιχούν όχι μόνο σε αντικείμενα πολυμέσων (ιστοσελίδες, εικόνες, ακουστικά αποσπάσματα κτλ.), όπως ο παραδοσιακός ιστός, αλλά και σε αντικείμενα όπως είναι φυσικά πρόσωπα, τοποθεσίες, οργανισμοί και γεγονότα. Επιπρόσθετα, ο σημασιολογικός ιστός δε θα περιέχει μόνο ένα είδος σχέσης (hyperlink) ανάμεσα στους πόρους, αλλά πολλά και διαφορετικά είδη σχέσεων ανάμεσα στα διάφορα είδη πόρων. Η γενική ιδέα θεωρεί ότι τα δεδομένα στο σημασιολογικό ιστό μοντελοποιούνται ως ένα κατευθυνόμενο γράφημα με σήμανση, όπου κάθε κόμβος αντιστοιχεί σε έναν πόρο και κάθε τόξο επισημαίνεται με έναν τύπο ιδιότητας (property type). Αν και υπάρχουν διάφορες προτάσεις για την αναπαράσταση των πόρων και

των αμοιβαίων σχέσεών τους στο σημασιολογικό ιστό βασισμένες στην XML, ένα ιδιαίτερο σύστημα πρέπει να αναλάβει τη δέσμευση για ένα ή περισσότερα σχήματα και πρωτόκολλα ανταλλαγής των πληροφοριών αυτών. Το περιγραφόμενο σύστημα χρησιμοποιεί το πλαίσιο περιγραφής πόρων του W3C με το λεξιλόγιο σχημάτων που παρέχεται από το RDFS [1] ως ένα μέσο για την περιγραφή των πόρων και των μεταξύ τους σχέσεων. Το SOAP [2], χρησιμοποιείται ως πρωτόκολλο για την ερώτηση και την ανταλλαγή αυτών των στιγμιαίων δεδομένων RDF μεταξύ μηχανών.
Οι εικόνες 1 & 2 δείχνουν δύο παραδείγματα του σημασιολογικού ιστού, αναδεικνύοντας εμφανείς πτυχές του, οι οποίες είναι σημαντικές ως προς τη σημασιολογική έρευνα.



Εικόνα 1: ένα τμήμα του σημασιολογικού ιστού που αναφέρεται στον μουσικό Yo-Yo Ma.



Εικόνα 2: ένα τμήμα του σημασιολογικού δικτύου που αναφέρεται στο συγγραφέα του άρθρου Eric Miller.

A. Έγγραφα εναντίον πραγματικών αντικειμένων (Documents vs real-world objects)

Ο σημασιολογικός ιστός δεν αποτελεί έναν ιστό από έγγραφα, αλλά έναν ιστό σχέσεων μεταξύ των πόρων αναδεικνύοντας τα αντικείμενα του πραγματικού κόσμου, αντικείμενα όπως είναι άνθρωποι, τοποθεσίες, και γεγονότα. Στο πρώτο παράδειγμα έχουμε αντικείμενα όπως είναι η πόλη Paris, το μουσικό άλμπουμ με τον τίτλο "Appalachian Journey" κτλ. Στο δεύτερο παράδειγμα, ΄χουμε το φυσικό πρόσωπο Eric Miller, τη σημασιολογική δραστηριότητα ιστού W3C, τον οργανισμό W3C, τις πόλεις Dublin, Ohio κτλ.

B. Ανθρώπινη εναντίον μηχανικής αναγνώσιμης πληροφορίας (Human vs machine readable information).

Στην εικόνα 2, έχουμε μια πηγή της πληροφορίας η οποία αναφέρεται στο φυσικό πρόσωπο Eric Miller. Αυτό δεν είναι μια ακολουθία χαρακτήρων "Eric Miller", αλλά ένας πόρος που αναδεικνύει ένα πρόσωπο. Υπάρχουν αρκετοί άνθρωποι με το συγκεκριμένο όνομα και ο πόρος αυτός συσχετίζει το όνομα με το συγκεκριμένο πρόσωπο. Το εμφανές σημείο σχετικά με το σημασιολογικό δίκτυο είναι το ότι περιέχει πλούσιες μηχανικά αναγνώσιμες πληροφορίες για τους πόρους αυτούς. Αν συγκρίνουμε την εικόνα 2 με την προσωπική ιστοσελίδα του συγκεκριμένου προσώπου, θα διαπιστώσουμε ότι η ιστοσελίδα περιέχει μεγαλύτερο ποσό ανθρώπινα αναγνώσιμης πληροφορίας, αλλά σχεδόν όλα τα μέρη της ιστοσελίδας τα οποία είναι δυνατό να κατανοηθούν από μηχανή αντιστοιχούν στο πώς θα έπρεπε να εμφανίζεται από ένα φυλλομετρητή. Αντίθετα, τα δεδομένα της εικόνας 2 είναι, σχεδόν όλα, μηχανικά αναγνώσιμα. Δηλώνει, σε μια γλώσσα κατανοητή από την μηχανή, ότι ο Eric Miller είναι ένα πρόσωπο, ο οποίος εργάζεται για τον οργανισμό W3C κτλ.

C. Σχέση μεταξύ της γλώσσας HTML και του σημασιολογικού ιστού (Relation between the HTML and the semantic web).

Ο σημασιολογικός ιστός αποτελεί μια επέκταση του ήδη υπάρχοντος ιστού. Όπως δείχνει και η εικόνα 2, υπάρχει ένα πλούσιο σύνολο συνδέσεων μεταξύ των κόμβων στο σημασιολογικό δίκτυο και εγγράφων HTML. Αυτές οι σχέσεις συνδέουν μια έννοια στο σημασιολογικό δίκτυο με τις ιστοσελίδες οι οποίες αναφέρονται περισσότερο σε αυτή. Είναι επίσης δυνατό μερικές από τις ιστοσελίδες στο σύγχρονο ιστό να περιέχουν σημασιολογική σήμανση. Υποθέτουμε την ύπαρξη μηχανισμών για τη συγκέντρωση των σημάνσεων αυτών στο σημασιολογικό ιστό.

D. Διανεμημένη επεκτασιμότητα (Distributed extensibility).

Μια άλλη πτυχή του σημασιολογικού ιστού αποτελεί το ότι διαφορετικοί δικτυακοί τόποι μπορούν να συνεισφέρουν δεδομένα για μια συγκεκριμένη πηγή πληροφορίας. Στο παράδειγμα της εικόνας 1, αρκετές διαφορετικές πηγές διαθέτουν δεδομένα σχετικά με τον μουσικό Yo-Yo Mama, καθώς και σχετικούς πόρους. Οι δικτυακοί τόποι των Amazon και CDNow, διαθέτουν δεδομένα γύρω από τα μουσικά άλμπουμ του συγκεκριμένου μουσικού, ο δικτυακός τόπος e-bay παρουσιάζει δημοπρασίες σχετικές με τα άλμπουμ του καλλιτέχνη κτλ. Καθένας από αυτούς τους δικτυακούς τόπους μπορούν να δημοσιεύσουν

δεδομένα σχετικά με τον συγκεκριμένο μουσικό χωρίς την ύπαρξη άδειας από κάποια κεντρική αρχή, δηλαδή μπορούν να επεκτείνουν τη γνώση σχετικά με το σημασιολογικό ιστό για οποιοδήποτε πόρο με έναν διανεμημένο τρόπο. Η διανεμημένη επεκτασιμότητα αποτελεί μια πολύ σημαντική πτυχή του σημασιολογικού ιστού. Αυτό, ωστόσο οδηγεί στη δημιουργία νέων προβλημάτων, αφού σε έναν κόσμο όπου ο καθένας έχει τη δυνατότητα να δημοσιεύσει ο,τιδήποτε, ένα μεγάλο μέρος της δημοσιευομένης πληροφορίας δεν μπορεί να τύχει εμπιστοσύνης. Στον τρέχοντα ιστό, ως άνθρωποι, χρησιμοποιούμε τη νοημοσύνη μας προκειμένου να αποφασίσουμε τι λέει το περιεχόμενο μιας ιστοσελίδας. Τα υπολογιστικά προγράμματα από την άλλη πλευρά, τα οποία δε διαθέτουν αυτή τη δυνατότητα , δεν έχουν τους πόρους για να εμπιστευτούν τα δεδομένα από μια νέα ιστοσελίδα στο σημασιολογικό ιστό. Αυτό αποτελεί ένα σημαντικό πρόβλημα που θα πρέπει να εξεταστεί.

II. ΕΙΣΑΓΩΓΗ ΣΤΗ ΣΗΜΑΣΙΟΛΟΓΙΚΗ ΕΡΕΥΝΑ – SEMANTIC SEARCH INTRODUCTION

Όπως συνέβη με τον παγκόσμιο ιστό, η ανάπτυξη του σημασιολογικού ιστού θα πραγματοποιηθεί μέσα από τις εφαρμογές που θα το χρησιμοποιήσουν. Η σημασιολογική έρευνα (semantic search) αποτελεί μια εφαρμογή έρευνας του σημασιολογικού ιστού (semantic web). Η έρευνα μέσα στο διαδίκτυο είναι μια από τις πιο δημοφιλείς εφαρμογές με μεγάλες προοπτικές βελτίωσης. Η προσθήκη ρητής σημασιολογίας είναι δυνατό να βελτιώσει τα αποτελέσματα της τρέχουσας έρευνας στο παραδοσιακό παγκόσμιο ιστό, μέσω της χρησιμοποίησης δεδομένων από το σημασιολογικό ιστό. Η παραδοσιακή τεχνολογία ανάκτησης της πληροφορίας (Information Retrieval technology) βασίζεται σχεδόν αποκλειστικά στη συχνότητα εμφάνισης συγκεκριμένων λέξεων μέσα σε έγγραφα. Οι μηχανές αναζήτησης, όπως το Google, αυξάνουν αυτή τη δυνατότητα στα πλαίσια του παγκόσμιου ιστού παρέχοντας πληροφορίες σχετικές με τη δομή των υπερσυνδέσμων (hyperlinks) του παγκόσμιου ιστού. Η διαθεσιμότητα μεγάλων ποσών δομημένης, μηχανικά αναγνώσιμης πληροφορίας γύρω από ένα ευρύ φάσμα αντικειμένων του σημασιολογικού ιστού προσφέρει δυνατότητες βελτίωσης της παραδοσιακής έρευνας. Ωστόσο, θα πρέπει να γίνει ένας διαχωρισμός όσον αφορά στα δύο βασικά είδη έρευνας και αναζήτησης:

A. Αναζητήσεις πλοήγησης (navigational searches)

Σε αυτή την τάξη αναζήτησης, ο χρήστης παρέχει στην μηχανή αναζήτησης μια φράση, ή ένα συνδυασμό λέξεων που περιμένει να συναντήσει μέσα στα έγγραφα. Δεν υπάρχει καμία ερμηνεία αυτών των λέξεων για την ανάδειξη μιας έννοιας. Σε αυτές τις περιπτώσεις, ο χρήστης χρησιμοποιεί την μηχανή αναζήτησης ως ένα εργαλείο πλοήγησης προκειμένου να οδηγηθεί στο επιθυμητό έγγραφο. Η αναζήτηση στο σημασιολογικό ιστό δεν ασχολείται με αυτού του είδους την αναζήτηση.

B. Αναζητήσεις διερεύνησης (research searches)

Σε πλήθος άλλων περιπτώσεων, ο χρήστης παρέχει στην μηχανή αναζήτησης μια φράση η οποία προορίζεται να επισημάνει ένα αντικείμενο σχετικά με το οποίο ο χρήστης προσπαθεί να διερευνήσει και να συγκεντρώσει πληροφορίες. Δεν υπάρχει κάποιο συγκεκριμένο έγγραφο για το οποίο ο χρήστης γνωρίζει και προσπαθεί να πλοηγηθεί σε αυτό. Ο χρήστης, μάλλον, προσπαθεί να εντοπίσει ένα σύνολο εγγράφων από το οποίο θα μπορέσει να βρει και να συγκεντρώσει τις πληροφορίες της επιθυμίας του. Αυτή είναι και η κατηγορία των αναζητήσεων που ενδιαφέρουν το σημασιολογικό ιστό.
Αν πάρουμε για παράδειγμα ένα ερώτημα αναζήτησης (search query) όπως είναι το ακόλουθο , “W3C track 2pm Panel”, θα διαπιστώσουμε ότι το συγκεκριμένο ερώτημα δεν αναδεικνύει καμία έννοια. Το πιο πιθανό είναι ότι ο χρήστης προσπαθεί να εντοπίσει την ιστοσελίδα εκείνη που θα περιέχει όλες τις λέξεις του ερωτήματος. Από την άλλη πλευρά, ερωτήσεις αναζήτησης όπως το όνομα ενός ανθρώπου ή μιας τοποθεσίας αναδεικνύουν συγκεκριμένες έννοιες. Το πιο πιθανό είναι ότι ο χρήστης προσπαθεί να κάνει μια διερευνητική αναζήτηση πάνω στο πρόσωπο ή την τοποθεσία που περιγράφεται στο ερώτημα. Η σημασιολογική αναζήτηση έχει σκοπό να βελτιώσει τα αποτελέσματα της έρευνας και το πραγματοποιεί με δύο τρόπους:

- Τα αποτελέσματα της παραδοσιακής αναζήτησης παίρνουν την μορφή μιας λίστας ιστοσελίδων ή εγγράφων. Η λίστα αυτή των εγγράφων εμπλουτίζεται με σχετικά στοιχεία τα οποία εξάγονται από το σημασιολογικό ιστό. Τα αποτελέσματα τα οποία βασίζονται στο σημασιολογικό ιστό είναι ανεξάρτητα των αποτελεσμάτων τα που λαμβάνονται μέσω παραδοσιακών τεχνικών ανάκτησης της πληροφορίας.
- Η αναζήτηση της φράσης στις διερευνητικές αναζητήσεις επισημαίνει μία ή περιστασιακά δύο έννοιες του πραγματικού κόσμου. Πιστεύεται ότι είναι πιο χρήσιμο για το κομμάτι της ανάκτησης κειμένου της μηχανής αναζήτησης να υπάρχει μια κατανόηση των εννοιών αυτών που προσδιορίζονται από τη φράση της αναζήτησης. Η κατανόηση αυτών των προσδιορισμών μπορεί να συμβάλλει στην κατανόηση του περιεχομένου της αναζήτησης, βελτιστοποιώντας τα αποτελέσματα αυτής.

Το μεγαλύτερο κομμάτι του παρόντος άρθρου σχετίζεται με τον εμπλουτισμό των αποτελεσμάτων της αναζήτησης με δεδομένα τα οποία προέρχονται από το σημασιολογικό ιστό. Από τη στιγμή που ο σημασιολογικός ιστός δεν περιέχει ακόμα μεγάλο ποσό πληροφορίας, εκτός από την εφαρμογή της σημασιολογικής αναζήτησης θεωρείται χρήσιμη και η κατασκευή των απαραίτητων μονάδων του σημασιολογικού ιστού για την παροχή δεδομένων για τις εφαρμογές σημασιολογικής αναζήτησης. Τόσο η εφαρμογή σημασιολογικής αναζήτησης, όσο και οι μονάδες του

σημασιολογικού ιστού έχουν χτιστεί πάνω από την υποδομή TAP (TAP infrastructure).

III. Η ΥΠΟΔΟΜΗ TAP (TAP INFRASTRUCTURE)

Η υποδομή TAP [3] προορίζεται ως μια υποδομή για εφαρμογές στο σημασιολογικό ιστό. Παρέχει ένα σύνολο απλών μηχανισμών για δικτυακούς τόπους προκειμένου να δημοσιεύσουν δεδομένα μέσα στο σημασιολογικό ιστό, καθώς και για εφαρμογές προκειμένου να καταναλώσουν τα δεδομένα αυτά μέσω μιας μινιμαλιστικής ερώτησης διασύνδεσης που ονομάζεται GetData.

A. Η ερώτηση διασύνδεσης GetData (The GetData query interface)

Ένας αριθμός γλωσσών ερωτημάτων (query languages) έχει αναπτυχθεί για το πλαίσιο εργασίας περιγραφής πόρων (RDF [4]– Resource Description Framework), τη DAML και πιο γενικά για ημι-δομημένα δεδομένα. Ωστόσο, οι γλώσσες αυτές παρέχουν πολύ εκφραστικούς μηχανισμούς οι οποίοι έχουν σκοπό να καταστήσουν εύκολη τη διατύπωση πολύπλοκων ερωτημάτων. Η ύπαρξη των συγκεκριμένων μηχανισμών έκφρασης κάνει, ωστόσο, εύκολη την κατασκευή ερωτημάτων τα οποία απαιτούν την επεξεργασία ενός μεγάλου πλήθους υπολογιστικών πόρων. Κατά συνέπεια, αφού καμία σημαντική ιστοσελίδα δεν παρέχει μια διασύνδεση SQL στη σχεσιακή βάση δεδομένων, δεν περιμένει κανείς από τους δικτυακούς τόπους, ειδικά τους μεγάλους, να χρησιμοποιούν γλώσσες ερωτημάτων ως εξωτερική διασύνδεση στα δεδομένα τους. Αυτό το οποίο χρειάζεται είναι μια περισσότερο ελαφριά διασύνδεση που θα είναι πιο εύκολο να τυγχάνει υποστήριξης και το πιο σημαντικό, να παρουσιάζει προβλέψιμη συμπεριφορά. Αυτή η προβλέψιμη συμπεριφορά είναι σημαντική όχι μόνο για τον παροχέα υπηρεσιών, αλλά επίσης και για τον πελάτη των υπηρεσιών. Ένα τέτοιο ελαφρύ σύστημα ερωτημάτων είναι δυνατό να είναι συμπληρωματικό σε πιο πλήρεις γλώσσες ερωτημάτων που αναφέρθηκαν πιο πάνω και δεν αποκλείει συγκεκριμένους δικτυακούς τόπους από το να συναθροίζουν δεδομένα από πολλαπλές ιστοσελίδες και να παρέχουν πλουσιότερες διασυνδέσεις ερωτημάτων σε αυτές τις συναθροίσεις.

Η GetData προορίζεται να είναι μια απλή διασύνδεση ερωτημάτων σε δεδομένα τα οποία είναι προσβάσιμα στο δίκτυο. Η GetData δεν προορίζεται να είναι μια πλήρης ή εκφραστική γλώσσα ερωτημάτων όπως είναι η SQL, η Xquery, η RQL και η DQL. Σκοπός είναι εύκολη η χρησιμοποίηση, η υποστήριξη και η κατασκευή της, τόσο από την προοπτική των προμηθευτών, όσο και των καταναλωτών των δεδομένων. Αυτό που επιθυμείται είναι η παροχή της δυνατότητας στις μηχανές να πραγματοποιούν ερωτήματα για δεδομένα σε απομακρυσμένους εξυπηρετητές. Η GetData, χτισμένη πάνω στο SOAP, επιτρέπει την πρόσβαση στις τιμές μιας ή περισσότερων ιδιοτήτων ενός πόρου από ένα γράφημα. Κάθε γράφημα που έχει τη δυνατότητα πραγματοποίησης ερωτημάτων έχει μια ενιαία θέση πόρου (URL – uniform resource locator), συσχετισμένη με αυτό. Κάθε ερώτημα τύπου GetData είναι ένα μήνυμα του SOAP το οποίο διευκρινίζει δύο ορίσματα, τον πόρο στις ιδιότητες του οποίου υπάρχει πρόσβαση, και τις ίδιες τις ιδιότητες στις οποίες υπάρχει πρόσβαση. Η απάντηση που επιστρέφεται σε ένα ερώτημα τύπου GetData είναι και αυτό ένα γράφημα το οποίο περιέχει τους πόρους (οι ιδιότητες του οποίου ερωτώνται), μαζί με τα τόξα τα οποία διευκρινίζονται από το ερώτημα και τους αντίστοιχους στόχους /πηγές τους. Οι διασυνδέσεις προγραμματισμού εφαρμογών κρύβουν τα μηνύματα του SOAP και την κωδικοποίηση της XML από τον προγραμματιστή. Έτσι, όσον αφορά μια εφαρμογή χρησιμοποιώντας το σημασιολογικό ιστό, η διασύνδεση προγραμματισμού εφαρμογής (API – Application Programming Interface) έχει την ακόλουθη μορφή:

GetData(<resource>, <property>), => <value>

Πιο κάτω, παρουσιάζονται κάποια παραδείγματα της GetData, με μια αφηρημένη σύνταξη όσον αφορά στα γραφήματα που παρουσιάζονται στις εικόνες 1 και 2.

GetData(<Yo-Yo Ma>, birthplace), => <Paris, France>
GetData(<Paris, France>, temperature), => 57 F
GetData(<Eric Miller>, livesIn), => <Dublin, Ohio>

<Yo-Yo Ma>, <Paris, France>, <Eric Miller> αποτελούν αναφορές στους πόρους που αντιστοιχούν στο συγκεκριμένο μουσικό, στη συγκεκριμένη πόλη και στο συγκεκριμένο πρόσωπο. Τυπικά, οι αναφορές σε αυτούς τους πόρους γίνονται μέσω του ενιαίου αναγνωριστικού πόρων (URI – Uniform Resource Identifier). Αυτά τα URI είναι τα εξής :

http://tap.stanford.edu/data/MusicianMa,_Yo-Yo
http://tap.stanford.edu/data/CityParis,_France
http://tap.stanford.edu/data/W3CPersonMiller,_Eric and
http://tap.stanford.edu/data/CityDublin,_Ohio.

Καθένα από τα παραπάνω ερωτήματα τύπου GetData είναι ένα μήνυμα του SOAP που αντιστοιχεί στο γράφημα με τα δεδομένα. Εκτός από τη διασύνδεση πυρήνα της GetData, υπάρχουν άλλες δύο διασυνδέσεις που παρέχονται από το TAP και βοηθούν στην εξερεύνηση των γραφημάτων. Η πρώτη από αυτές είναι η διασύνδεση της έρευνας η οποία παίρνει μια ακολουθία χαρακτήρων και επιστρέφει όλους τους πόρους, και η δεύτερη είναι η διασύνδεση ανάκλασης, η οποία είναι όμοια με την αντίστοιχη που χρησιμοποιείται στις γλώσσες αντικειμενοστραφούς προγραμματισμού και επιστρέφει μια λίστα από τόξα από και προς τους κόμβους. Αυτό είναι πού χρήσιμο για την εξερεύνηση ενός γραφήματος μέσα στην εγγύτητα ενός κόμβου χωρίς γνώση του τι μπορεί να βρίσκεται γύρω από αυτόν.

B. Τμηματοποίηση του TAP (TAP scraping)

Από τη στιγμή που το σημασιολογικό δίκτυο είναι ακόμα κάπως αραιό, πρέπει να χτίσουμε τα απαραίτητα τμήματα για να επιτραπεί η σημασιολογική έρευνα. Όλα τα δεδομένα για την εφαρμογή της σημασιολογικής έρευνας του W3C, προέρχονται από αρχεία RDF που διατηρούνται από το W3C. Τα δεδομένα αυτά δημοσιεύτηκαν χρησιμοποιώντας τον εξυπηρετητή HTML, TAPache. Ωστόσο, για μεγαλύτερες εφαρμογές που έχουν να κάνουν για παράδειγμα με αθλητές, μουσικούς, τοποθεσίες κτλ, πρέπει να κατασκευάσουμε τμήματα της HTML για να

εξάγουμε τα επιθυμητά δεδομένα από δημοφιλείς δικτυακούς τόπους όπως είναι οι Amazon, All Music, Ticket Master κ.α. που διαθέτουν δεδομένα για τα συγκεκριμένα αντικείμενα. Για να διευκολυνθεί κάτι τέτοιο, το TAP παρέχει την υποδομή για την ερμηνεία ενός αιτήματος τύπου GetData.

C. Δημοσίευση του TAP (TAP publishing)

Από την πλευρά του εξυπηρετητή (server-side), το TAP παρέχει τον εξυπηρετητή TAPache, έναν HTML εξυπηρετητή για την έκθεση των δεδομένων της διασύνδεσης GetData. Ο σκοπός του είναι να κάνει εξαιρετικά απλή τη δημοσίευση δεδομένων στο σημασιολογικό ιστό και δεν προορίζεται να αποτελέσει μια υψηλού επιπέδου λύση για δικτυακούς τόπους με μεγάλα ποσά δεδομένων και κίνησης. Η ευελιξία και η εξελιξιμότητα θεωρούνται περισσότερο σημαντικές από την ευκολία χρήσης για τέτοιου είδους δικτυακούς τόπους. Με τον εξυπηρετητή TAPache, υπάρχει συνήθως ένας κατάλογος (που συνήθως καλείται /html ή /htdocs), στον οποίο κάποιος μπορεί να τοποθετήσει αρχεία (.html,, jpeg, gif κτλ) ή καταλόγους που περιέχουν τα αρχεία αυτά, έτσι ώστε να γίνονται διαθέσιμα μέσω του παγκόσμιου ιστού. Όμοια, υπάρχει άλλος ένας κατάλογος (/data) για την τοποθέτηση αρχείων τύπου RDF. Τα γραφήματα που κωδικοποιούνται μέσα σε αυτά τα αρχεία γίνονται αυτόματα προσβάσιμα μέσω της διασύνδεσης GetData. Η ενιαία θέση πόρου (URL – uniform resource locator) η οποία σχετίζεται με κάθε γράφημα είναι αυτή που ανήκει στο συγκεκριμένο αρχείο. Ο εξυπηρετητής TAPache μεταγλωττίζει κάθε αρχείο με δομές γραφημάτων που μπορούν να χαρτογραφηθούν στην μνήμη, έτσι ώστε τα ερωτήματα που σχετίζονται με το αρχείο αυτό να μπορούν αν απαντηθούν με τη μικρότερη επιβάρυνση ανάλυσης. Τέλος, ο εξυπηρετητής TAPache παρέχει έναν απλό μηχανισμό για τη συγκέντρωση των δεδομένων σε πολλαπλά αρχεία RDF, τα οποία τοποθετούνται σε συγκεκριμένο κατάλογο και είναι διαθέσιμα μέσω της ενιαίας θέσης πόρου που σχετίζεται με τον κατάλογο αυτό.

D. Μητρώα και κρυφή μνήμη (Registries and caching)

Οι διάφοροι δικτυακοί τόποι/ γραφήματα έχουν διαφορετικά είδη πληροφορίας (δηλαδή ιδιότητες) για διαφορετικούς τύπους πόρων. Κάθε αίτημα τύπου GetData στοχεύει σε μια συγκεκριμένη θέση πόρου (URL) που αντιστοιχεί σε ένα γράφημα το οποίο υποτίθεται ότι διαθέτει τα δεδομένα. Μόλις έχουμε έναν αριθμό τέτοιων γραφημάτων, η διαδικασία παρακολούθησης του ποιο γράφημα διαθέτει τα δεδομένα μπορεί να γίνει πολύ δύσκολη υπόθεση από την πλευρά του πελάτη. Χρησιμοποιούμε ένα απλό μητρώο, το οποίο είναι διαθέσιμο ως ξεχωριστός εξυπηρετητής, προκειμένου να παρακολουθεί ποια ενιαία θέση πόρου έχει τις τιμές των ιδιοτήτων που σχετίζονται με κάποιο πόρο. Το μητρώο μπορεί να εξαχθεί ως ένας απλός πίνακας αναζήτησης (look up table), οποίος όταν του παρέχεται μια κλάση και μια ιδιότητα επιστρέφει μια λίστα ενιαίων θέσεων πόρων οι οποίες είναι δυνατό να έχουν τις τιμές για της ιδιότητες της κλάσης αυτής. Πιο πολύπλοκα μητρώα, τα οποία στηρίζονται σε περιγραφές των αντικειμένων, είναι επίσης διαθέσιμα μέσω του TAP. Με το μητρώο, ο χρήστης μπορεί να καθοδηγήσει το ερώτημα στο μητρώο, το οποίο με τη σειρά του επανακαθοδηγεί το ερώτημα στους κατάλληλους δικτυακούς τόπους. Η πραγματοποίηση ερωτημάτων σε πολλούς και διαφορετικούς δικτυακούς τόπους με τρόπο δυναμικό μπορεί να έχει ως αποτέλεσμα την αύξηση του χρόνου που απαιτείται προκειμένου το ερώτημα να διανύσει την απόσταση από ένα σημείο του δικτύου σε ένα άλλο (high latency).

Η κρυφή μνήμη είναι μέρος της λειτουργίας του μητρώου. Υπάρχει μεγάλη ομοιότητα με την λειτουργία του συστήματος ονομάτων περιοχών (DNS – Domain Name System) όσον αφορά στα δεδομένα γύρω από κεντρικούς υπολογιστές του διαδικτύου. Το σύστημα ονομάτων περιοχών παρέχει μια ενοποιημένη άποψη των δεδομένων για τους κεντρικούς υπολογιστές που βρίσκονται σε εκατομμύρια δικτυακούς τόπους μέσα στο διαδίκτυο. Τα διαφορετικά είδη εξυπηρετητών αποθηκεύουν τις πληροφορίες που φιλοξενούν με διαφορετικό τρόπο. Αλλά αυτό που κάνει το σύστημα ονομάτων περιοχών είναι η παροχή μιας ενοποιημένης άποψης όλων αυτών των δεδομένων έτσι ώστε ο πελάτης να θεωρεί ότι τα δεδομένα αυτά βρίσκονται τοπικά στον εξυπηρετητή ονομάτων. Προχωρώντας περισσότερο αυτή την αναλογία, η GetData θα μπορούσε να θεωρηθεί ως μια επέκταση της GetHostByName, η οποία είναι η διασύνδεση πυρήνα για το σύστημα ονομάτων περιοχών. Όπως η GetHostByName επιτρέπει στον πελάτη να κάνει ένα ερώτημα για κάποια συγκεκριμένη ιδιότητα (IP address) μιας κλάσης αντικειμένων (Internet hosts), η GetData επιτρέπει στον πελάτη να πραγματοποιήσει ερωτήματα για πολλά διαφορετικά είδη ιδιοτήτων πολλών κλάσεων αντικειμένων. Το TAP επίσης παρέχει δυνατότητες για άλλες λειτουργίες όπως η σημασιολογική διαπραγμάτευση για να βοηθήσει τη συγκέντρωση διαφορετικών ενιαίων αναγνωριστικών πόρων (URIs – Uniform Resource Identifiers) που χρησιμοποιούνται για το ίδιο αντικείμενο.

IV. Πηγες δεδομενων (Data sources)

Στο επόμενο κομμάτι, γίνεται μια περιγραφή των πηγών δεδομένων που χρησιμοποιούνται από δύο συστήματα σημασιολογικής αναζήτησης που έχουν κατασκευαστεί, το ABS και το W3C. Τα δεδομένα για το ABS έρχονται από έναν μεγάλο αριθμό πηγών. Πολλοί διαφορετικοί δικτυακοί τόποι διαθέτουν δεδομένα σχετικά με μουσικούς, αθλητές, τοποθεσίες και προϊόντα. Οι περισσότεροι από αυτούς τους δικτυακούς τόπους δεν έχουν ακόμα διαθέσιμα τα δεδομένα τους σε μια μηχανικά κατανοητή μορφή. Για να αντιμετωπιστεί αυτό, έχουν γραφτεί αρχεία αποκομμάτων HTML (HTML scrapers), δηλαδή αρχεία εφαρμογών όπου αποθηκεύονται δεδομένα για περαιτέρω επεξεργασία, τα οποία εντοπίζουν και μετατρέπουν, με τρόπο δυναμικό, τις σχετικές σελίδες σε αυτούς τους δικτυακούς τόπους σε δεδομένα τα οποία είναι δυνατό να αναγνωριστούν από μηχανή και να τα καταστήσουν διαθέσιμα μέσω της

διασύνδεσης GetData. Μερικούς από τους δικτυακούς τόπους που διαθέτουν τέτοια αρχεία αποκομμάτων περιλαμβάνουν τα All Music, Ebay, Amazon, AOL Shopping, Ticket Master κτλ. Εξαιτίας των αρχείων αυτών, η εφαρμογή της σημασιολογικής έρευνας μπορεί να "προσποιηθεί" ότι τα δεδομένα σε καθένα από αυτούς τους δικτυακούς τόπους είναι διαθέσιμα μέσω της διασύνδεσης GetData. Με δεδομένο τον αριθμό των πηγών δεδομένων και της διανεμημένης φύσης των πηγών αυτών, το σύστημα σημασιολογικής αναζήτησης ABS κάνει χρήση του μητρώου και των μηχανισμών κρυφής μνήμης που παρέχονται από το TAP. Όλες οι πηγές δεδομένων μαζί συγκροτούν ένα σημασιολογικό ιστό. Μια πολύ σημαντική πηγή δεδομένων για το ABS αποτελεί η βάση δεδομένων γνώσης του TAP (TAP Knowledge base), η οποία είναι μια πολύ ευρεία βάση για μια σειρά περιοχών, περιλαμβάνοντας ανθρώπους (μουσικούς, αθλητές, πολιτικούς, ηθοποιούς), οργανισμούς (επιχειρήσεις, αθλητικούς συλλόγους), τοποθεσίες (πόλεις, χώρες, πολιτείες) και προϊόντα. Για κάθε πόρο δεδομένων, παρέχονται ο τύπος και η ετικέτα RDF (rdf:type & rdf:label) για το συγκεκριμένο αντικείμενο. Για το σύστημα ABS, η βάση γνώσης του TAP συνεισφέρει περίπου 65,000 ανθρώπους, οργανισμούς και τοποθεσίες, τα οποία όλα μαζί καλύπτουν το 17% των αναζητήσεων που λαμβάνουν χώρα. Σε αντίθεση με το σύστημα σημασιολογικής έρευνας ABS, τα δεδομένα για το σύστημα W3C προέρχονται από ένα σχετικά μικρό αριθμό πηγών, όλες εσωτερικές στο W3C. Η σημασιολογική έρευνα του W3C έχει πέντε διαφορετικές πηγές δεδομένων , οι οποίες όλες μαζί καλύπτουν τα ακόλουθα είδη αντικειμένων : φυσικά πρόσωπα (αυτό περιλαμβάνει το προσωπικό του W3C και συγγραφείς διαφόρων εγγράφων W3C) , δραστηριότητες του W3C (κάθε δραστηριότητα συνδέεται στους ανθρώπους του W3C), ομάδες εργασίας και άλλες επιτροπές (καθεμία από αυτές συνδέεται με τη δραστηριότητα και το προσωπικό), έγγραφα (αυτό περιλαμβάνει συστάσεις, σχέδια εργασίας, σημειώσεις και καθένα από αυτά συνδέονται με τις ομάδες εργασίας και τις δραστηριότητες που παρήγαγαν), και νέα (η σημασιολογική έρευνα W3C ενσωματώνει δεδομένα από έναν αριθμό τροφοδοτών νέων RSS, δηλαδή της μορφής που συγκροτεί τις ειδήσεις και τι περιεχόμενό τους, για διάφορα νέα και γεγονότα άξια αναφοράς που λαμβάνουν χώρα στο W3C). Επιπρόσθετα, και τα δύο συστήματα σημασιολογικής αναζήτησης ενσωματώνουν μια βασική οντολογία σχετικά με φυσικά πρόσωπα, τοποθεσίες, γεγονότα, οργανισμούς κτλ, η οποία προέρχεται από τη βάση γνώσης του TAP. Αυτή η οντολογία ορίζει έναν μεγάλο αριθμό βασικών όρων λεξιλογίου, οι οποίοι ισχύουν ευρέως σε διάφορες εφαρμογές).

V. ΕΜΠΛΟΥΤΙΣΜΟΣ ΤΗΣ ΑΝΑΖΗΤΗΣΗΣ ΜΕ ΔΕΔΟΜΕΝΑ (AUGMENTING SEARCH WITH DATA)

Όπως αναφέρθηκε νωρίτερα, δύο είναι οι βασικοί σκοποί της σημασιολογικής αναζήτησης. Ο πρώτος είναι ο εμπλουτισμός των αποτελεσμάτων της παραδοσιακής αναζήτησης με δεδομένα τα οποία εξάγονται από το σημασιολογικό ιστό. Ο δεύτερος είναι η χρησιμοποίηση της κατανόησης του προσδιορισμού των όρων αναζήτησης, προκειμένου να βελτιστοποιηθεί η παραδοσιακή αναζήτηση. Σε αυτό το κομμάτι περιγράφεται το πώς η σημασιολογική αναζήτηση εμπλουτίζει τα αποτελέσματα της παραδοσιακής έρευνας. Ωστόσο, υπάρχουν τρία προβλήματα που εξετάζονται προκειμένου να πραγματοποιηθεί κάτι τέτοιο:

- Προσδιορισμός : πρέπει να καθοριστεί η έννοια που προσδιορίζεται από την ερώτηση αναζήτησης.
- Περιεχόμενο παρουσίασης : πρέπει να καθοριστεί ποια σχετικά δεδομένα θα εξαχθούν από το σημασιολογικό ιστό.
- Παρουσίαση : τα δεδομένα πρέπει να διαμορφωθούν κατάλληλα για να συμπεριληφθούν στα αποτελέσματα της έρευνας.

Ωστόσο, πριν από αυτό, θα πρέπει να αναφερθούν κάποιες προϋποθέσεις που λαμβάνονται από τη σημασιολογική αναζήτηση σχετικά με τα δεδομένα, προκειμένου αργότερα να εξεταστεί η αρχιτεκτονική του συστήματος.

A. Προϋποθέσεις σχετικά με τα δεδομένα (Assumptions about the data).

Οι διαφορετικοί δικτυακοί τόποι είναι δυνατό να παρέχουν διαφορετικά είδη δεδομένων για ένα αντικείμενο. Γενικά, θα είναι δύσκολος, αν όχι αδύνατος, ο έλεγχος του είδους των δεδομένων που είναι διαθέσιμα για κάποιο αντικείμενο. Κάτι τέτοιο έρχεται σε αντίθεση με τα παραδοσιακά συστήματα σχεσιακών βάσεων δεδομένων τα οποία διαθέτουν καθορισμένα σχήματα και περιορισμούς ενσωμάτωσης που εξασφαλίζουν ότι ένα ομοιόμορφο σύνολο πληροφοριών είναι διαθέσιμο για οποιοδήποτε τύπο δεδομένων. Αυτή η έλλειψη ομοιομορφίας στα διαθέσιμα δεδομένα μπορεί να αποδειχθεί προβληματική για τις εφαρμογές που αφορούν στο σημασιολογικό ιστό. Από την άλλη πλευρά, οι εφαρμογές του σημασιολογικού ιστού θα πρέπει να μπορούν να εκμεταλλευτούν τα νέα είδη δεδομένων που είναι διαθέσιμα μέσα σε αυτόν. Κατά συνέπεια, θα πρέπει να εξισορροπιστούν οι παραλλαγές στη διαθεσιμότητα των δεδομένων, εξασφαλίζοντας ότι ορισμένες ιδιότητες είναι διαθέσιμες για όλα τα αντικείμενα. Πιο συγκεκριμένα, εξασφαλίζεται ότι για κάθε αντικείμενο υπάρχουν δεδομένα σχετικά με αυτό, καθώς και ο τρόπος με τον οποίο αναφέρεται στο αντικείμενο αυτό (δηλαδή αν είναι διαθέσιμα ο τύπος και η ετικέτα, rdf:type & rdf:label). Επιπρόσθετα, εξασφαλίζεται το ότι ένας μικρός αριθμός πόρων δεδομένων διαθέτει αυτό τον πυρήνα πληροφοριών σχετικά με όλα τα αντικείμενα που τυγχάνουν ενδιαφέροντος. Σε μερικές περιπτώσεις, είναι δυνατό να βασιστούμε σε συγκεκριμένες πηγές δεδομένων που παρέχουν συγκεκριμένα στοιχεία δεδομένων.

B. Αρχιτεκτονική του συστήματος (System architecture)

Η εφαρμογή της σημασιολογικής αναζήτησης τρέχει ως πελάτης της διασύνδεσης TAP, παράλληλα με την μηχανή αναζήτησης. Όταν λαμβάνεται η επερώτηση αναζήτησης, η εμπροσθοφυλακή αναζήτησης (search front end), εκτός από την αποστολή της ερώτησης, καλεί επίσης και την εφαρμογή σημασιολογικής αναζήτησης. Η τελευταία, αποκτά πρόσβαση στο σημασιολογικό ιστό μέσω της διασύνδεσης πελάτη του TAP, προκειμένου να πραγματοποιήσει τα τρία βήματα του προσδιορισμού, της επιλογής του περιεχομένου της παρουσίασης και τέλος της ίδιας της παρουσίασης.

C. Επιλογή του προσδιορισμού (Choosing a denotation)

Το πρώτο βήμα είναι η αντιστοίχιση των όρων αναζήτησης σε ένα ή περισσότερους κόμβους του σημασιολογικού ιστού. Αυτό πραγματοποιείται με τη χρησιμοποίηση της διασύνδεσης που παρέχεται από το TAP. Μια επερώτηση αναζήτησης διανέμεται σε γνωστούς δικτυακούς τόπους που περιέχουν τις πληροφορίες για όλα τα αντικείμενα για τα οποία ενδιαφερόμαστε. Υπάρχουν δύο διαφορετικοί τρόποι με τους οποίους ένας όρος αναζήτησης είναι δυνατό να αντιστοιχιστεί σε έναν ή περισσότερους κόμβους του σημασιολογικού ιστού. Ο πρώτος είναι η ασάφεια (ambiguity), όπου ένας ή περισσότεροι όροι μέσα στο σημασιολογικό ιστό έχει τον όρο αναζήτησης ή ένα υποσύνολο του όρου, σαν ιδιότητα που συντάσσεται από τη διασύνδεση αναζήτησης. Για παράδειγμα, στο σύστημα σημασιολογικής αναζήτησης ABS, η επερώτηση αναζήτησης "Paris" θα μπορούσε να αντιστοιχιστεί στη συγκεκριμένη πόλη, ή στην πόλη με το ίδιο όνομα που ανήκει στην πολιτεία του Texas, ή το μουσικό σχήμα με τι ίδιο όνομα κτλ. Σε αυτή την περίπτωση, μπορούμε να επιλέξουμε μία από αυτές τις αντιστοιχίες σαν τον επιθυμητό προσδιορισμό. Η επιλογή μπορεί να γίνει με βάση έναν αριθμό διαφορετικών παραγόντων, όπως είναι η δημοτικότητα το όρου αναζήτησης όπως υπολογίζεται από το ρυθμό εμφάνισής του σε ένα σώμα κειμένου ή από τη διαθεσιμότητα των δεδομένων στο σημασιολογικό ιστό. Για παράδειγμα, "Paris" που ανήκει στη Γαλλία, είναι πιο επιθυμητός και πιο δημοφιλής προσδιορισμός από το "Paris" που ανήκει στις Ηνωμένες Πολιτείες. Επίσης, το προφίλ του χρήστη μπορεί να είναι οδηγός επιλογής του προσδιορισμού, αλλά και το πλαίσιο της αναζήτησης μπορεί να βοηθήσει την έρευνα. Αν ο χρήστης αναζητά πληροφορίες σχετικά με μουσικούς, το ερώτημα "Μπλέ" είναι πιο πιθανό να προσδιορίσει το συγκεκριμένο μουσικό σχήμα, παρά το ίδιο το χρώμα. Άλλοι προσδιορισμοί εκτός των επιθυμητών είναι επίσης δυνατό να προσφερθούν στο χρήστη από το σύστημα σαν μέρος της διασύνδεσης αναζήτησης του χρήστη, έτσι ώστε αυτός να μπορεί να επιλέξει έναν ή περισσότερους, εάν στοχεύει σε κάτι τέτοιο.

Ο δεύτερος τρόπος με το οποίο κάποιος μπορεί να αντιστοιχίσει έναν όρο αναζήτησης σε έναν ή περισσότερους κόμβους στο σημασιολογικό ιστό είναι μέσω της υποβολής πολύπλοκων όρων αναζήτησης, όπου σε αρκετές περιπτώσεις, υποσύνολα του όρου αντιστοιχίζονται σε διαφορετικούς κόμβους. Για παράδειγμα, το ερώτημα "Eric Miller" μπορεί να διασπαστεί σε "Eric Miller + rdf", με το πρώτο μέρος να αντιστοιχεί στον κόμβο με το συγκεκριμένο όνομα και το δεύτερο να αντιστοιχεί στο περιγραφικό πλαίσιο εργασίας πόρων (RDF – Resource Description Framework). Αν ο όρος αναζήτησης δεν προσδιορίζει κάτι γνωστό στο σημασιολογικό ιστό, δεν έχουμε τη δυνατότητα συνεισφοράς του ο,τιδήποτε στα αποτελέσματα της έρευνας. Κατά συνέπεια, θεωρείται απαραίτητο για το σημασιολογικό ιστό να διαθέτει γνώση για μια ευρεία σειρά όρων.

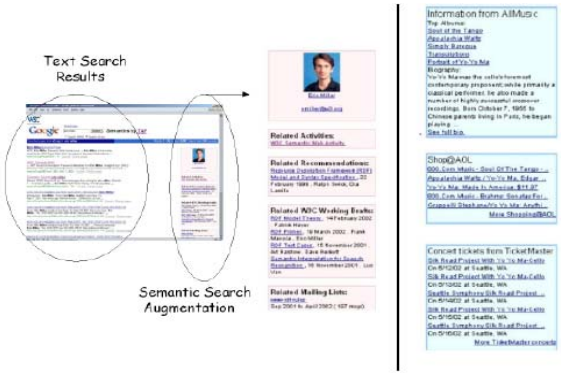
D. Καθορισμός του περιεχομένου παρουσίασης (Determining what to show)

Από τη στιγμή που διαθέτουμε είτε έναν μοναδικό κόμβο, ή ένα ζευγάρι κόμβων, ο επόμενος στόχος είναι ο καθορισμός των δεδομένων που θα πρέπει να ενσωματωθούν στο σημασιολογικό ιστό, και με ποια σειρά. Όπως με την παραδοσιακή αναζήτηση, η απόφαση σχετικά με το περιεχόμενο της παρουσίασης αποτελεί ένα από τα κεντρικά προβλήματα της σημασιολογικής αναζήτησης. Το συγκεκριμένο πρόβλημα μπορεί να απεικονιστεί από την άποψη του γραφήματος σημασιολογικού ιστού. Ο κόμβος ο οποίος αποτελεί τον επιλεγόμενο προσδιορισμό , παρέχει ένα σημείο εκκίνησης. Σε αυτή την περίπτωση, ο συγκεκριμένος κόμβος αναφέρεται ως κόμβος αγκύρωσης (anchor node). Στη συνέχεια, θα πρέπει να γίνει επιλογή ενός υπο-γραφήματος γύρω από το συγκεκριμένο κόμβο, τον οποίο και θέλουμε να παρουσιάσουμε. Στην περίπτωση που ο όρος αναζήτησης προσδιορίζει έναν συνδυασμό όρων, έχουμε δύο κόμβους αγκύρωσης από τους οποίους πρέπει να επιλέξουμε το υπο-γράφημα. έχοντας επιλέξει το υπο-γράφημα, πρέπει να αποφασιστεί η σειρά με την οποία αυτό θα δημοσιευτεί στα αποτελέσματα που παρουσιάζονται στο χρήστη. Ξεκινάμε με μια απλή συντακτική προσέγγιση ο οποία είναι ευρέως εφαρμόσιμη, αλλά παρουσιάζει ορισμένους περιορισμούς. Η απλή προσέγγιση για την επιλογή του υπο-γραφήματος βασίζεται απόλυτα στη δομή του γραφήματος, μεταχειριζόμενη όλους τους τύπους ιδιοτήτων ως εξίσου σχετικούς, αρχίζοντας από τον κόμβο αγκύρωσης, συγκεντρώνοντας τις πρώτες N τριάδες, όπου N είναι κάποιο προκαθορισμένο όριο. Αυτή η βασική προσέγγιση μπορεί να βελτιωθεί με την ενσωμάτωση διαφόρων ειδών ευρετικών, προκειμένου να παραχθεί ένα περισσότερο ισορροπημένο γράφημα. Αυτή η προσέγγιση έχει το πλεονέκτημα ότι δεν απαιτεί γραπτό κώδικα αλλά έχει από την άλλη πλευρά το μειονέκτημα ότι είναι πολύ ευαίσθητη στις επιλογές της αναπαράστασης που γίνονται από τις πηγές του σημασιολογικού ιστού. Αυτή η προσέγγιση έχει επίσης την ιδιότητα της δυνατότητας ενσωμάτωσης νέων στοιχείων πληροφορίας, σχετικά με τον κόμβο αγκύρωσης και τους γειτονικούς, όπως αυτοί εμφανίζονται στο σημασιολογικό ιστό, χωρίς να μεταβάλλει τίποτα στη σημασιολογική μηχανή αναζήτησης. Αυτή η ιδιότητα είναι ταυτόχρονα δυνατότητα και ελάττωμα, αφού από τη μία παρέχει πιο πλούσια αποτελέσματα, αλλά από την άλλη κάνει το σύστημα πιο ευαίσθητο σε μαζικά οχληρά μηνύματα (spam) και τις άσχετες πληροφορίες. Ένα επιπλέον πρόβλημα με αυτή την

προσέγγιση είναι ότι αγνοεί το πλαίσιο αναζήτησης. Για παράδειγμα, η αναζήτηση για το πρόσωπο “Eric Miller” στο δικτυακό τόπο του W3C, θα χρησιμοποιούσε διαφορετικά δεδομένα από το σημασιολογικό ιστό σε σύγκριση με το δικτυακό τόπο της οικογένειας Miller. Δεν υπάρχει κάποιος τρόπος να πραγματοποιηθεί κάτι τέτοιο μέσω της συγκεκριμένης προσέγγισης. Μια διαφορετική προσέγγιση θα μπορούσε να ήταν η χειρονακτική διευκρίνιση, για κάθε κατηγορία αντικειμένων που ενδιαφερόμαστε, του συνόλου των ιδιοτήτων που θα πρέπει να συγκεντρωθούν, Παρουσιάζει το πλεονέκτημα ότι η διευκρίνιση μόνο συγκεκριμένων ιδιοτήτων παρέχει ένα είδος φίλτρου, παράγοντας περισσότερο αξιόπιστα αποτελέσματα. Επίσης, είναι εύκολα εξατομικεύσιμη, έτσι ώστε να καθορίσει ποιες ιδιότητες θα πρέπει να ανακτηθούν. Ωστόσο, απαιτεί μεγαλύτερο κόπο, και δεν είναι σε θέση να ενσωματώσει τα νέα είδη πληροφοριών που παρουσιάζονται στο σημασιολογικό ιστό.

Ε. Διαμόρφωση (Formatting)

Το τελικό πρόβλημα αφορά την προβολή στο χρήστη των δεδομένων που συγκεντρώθηκαν. Η προβολή των εγγραφών στη σελίδα των αποτελεσμάτων της αναζήτησης γίνεται μέσω ενός συνόλου προτύπων (templates). Για κάθε τάξη αντικειμένων για τα οποία ενδιαφερόμαστε, συσχετίζουμε ένα διαταγμένο σύνολο προτύπων. Κάθε πρότυπο ορίζει την τάξη για την οποία ισχύει, τις ιδιότητες που πρέπει να είναι διαθέσιμες προκειμένου να εφαρμοστεί, καθώς και ένα πρότυπο html για την παρουσίαση των αποτελεσμάτων. Παρατηρώντας τη διαταγμένη λίστα των συγκεντρωμένων κόμβων, αναγνωρίζουμε το πρότυπο, δημιουργούμε το πρότυπο html και το ενσωματώνουμε στη διασύνδεση του χρήστη μαζί με τα αποτελέσματα της αναζήτησης. Τα πρότυπα είναι δυνατό να κωδικοποιηθούν με κάποια δηλωτική γλώσσα, ή με μία γλώσσα όπως είναι η Perl. Η εικόνα 3 δείχνει τα αποτελέσματα του εμπλουτισμού της αναζήτησης για τον “Eric Miller” στο σύστημα σημασιολογικής αναζήτησης W3C και τα αντίστοιχα για τον μουσικό “Yo-Yo Ma” στο σύστημα ABS.



Εικόνα 3 : Τα αποτελέσματα του εμπλουτισμού της αναζήτησης για τον “Eric Miller” στο σύστημα σημασιολογικής αναζήτησης W3C και τα αντίστοιχα για τον μουσικό “Yo-Yo Ma” στο σύστημα ABS.

VI. Η ΣΗΜΑΣΙΟΛΟΓΙΑ ΣΤΗΝ ΑΝΑΖΗΤΗΣΗ ΚΕΙΜΕΝΩΝ (SEMANTICS FOR TEXT SEARCH)

Εκτός από τον εμπλουτισμό της παραδοσιακής αναζήτησης με δεδομένα που προέρχονται από το σημασιολογικό ιστό, θα ήταν επιθυμητή η δυνατότητα χρησιμοποίησης του σημασιολογικού ιστού για τη βελτίωση των αποτελεσμάτων της αναζήτησης κειμένου. Η αναζήτηση κειμένου θα πρέπει να μπορεί να εκμεταλλευτεί την κατανόηση της επιθυμητής, από την πλευρά του χρήστη, πληροφορίας. Είναι πιθανό να υπάρχουν αρκετοί διαφορετικοί τρόποι με τους οποίους τα δεδομένα από το σημασιολογικό ιστό είναι δυνατό να χρησιμοποιηθούν προκειμένου να φιλτραριστούν τα αποτελέσματα της αναζήτησης κειμένου. Πιο κάτω περιγράφεται η προσπάθεια επίλυσης ενός συγκεκριμένου προβλήματος. Η μηχανή αναζήτησης του Google, παρουσιάζει περίπου 136,000 αποτελέσματα στην αναζήτηση για τον μουσικό “Yo-Yo Ma”. Μια χειροκίνητη ανάλυση των πρώτων 500 αποτελεσμάτων δείχνει ότι στο σύνολό τους αναφέρονται στο συγκεκριμένο μουσικό. Από την άλλη πλευρά, η αναζήτηση στην ίδια μηχανή αναζήτησης για το πρόσωπο “Eric Miller”παράγει περίπου 1,400,000 αποτελέσματα, από τα οποία τα 20 πρώτα έχουν να κάνουν με 16 διαφορετικά πρόσωπα που διαθέτουν το συγκεκριμένο όνομα. Το πιο πιθανό είναι ότι ο χρήστης επιθυμεί την εύρεση πληροφοριών για κάποιο συγκεκριμένο πρόσωπο με το όνομα αυτό. Δυστυχώς, δεν υπάρχει κάποιος εύκολος τρόπος για να διαβιβαστεί κάτι τέτοιο στο σύστημα.
Ο σκοπός είναι να επιτραπεί στην μηχανή αναζήτησης να κατανοήσει ότι διαφορετικές εκδοχές της ίδιας ακολουθίας χαρακτήρων προσδιορίζουν διαφορετικές έννοιες και περαιτέρω να φιλτράρει, να αξιολογήσει και να προβάλει τα αποτελέσματα των εγγράφων που ανταποκρίνονται στον επιθυμητό προσδιορισμό. Η αρχική εστίαση γίνεται πάνω στα ερωτήματα αναζήτησης φυσικών προσώπων.
Θεωρείται χρήσιμη η παροχή στο χρήστη ενός απλού μηχανισμού για την αναγνώριση του κατάλληλου προσδιορισμού. Σε κάποιες περιπτώσεις (π.χ. στην λέξη “Jaguar”), υπάρχει ένας μικρός αριθμός αρχικών προσδιορισμών (το ζώο, το αυτοκίνητο, ο αθλητικός σύλλογος), έτσι ώστε να μπορούμε να επιλέξουμε έναν και να απαριθμήσουμε τους υπόλοιπους προκειμένου να επιτραπεί στο χρήστη να κάνει μια επιλογή του επιθυμητού μεταξύ αυτών. Στην περίπτωση που υπάρχουν χιλιάδες εν δυνάμει προσδιορισμοί (όπως στην περίπτωση της αναζήτησης στο Google του ανθρώπου “”Eric Miller”), η προσέγγιση της απαρίθμησης όλων των πιθανών προσδιορισμών ξεχωριστά από τα αποτελέσματα αναζήτησης δεν είναι δυνατή. Έτσι, πρέπει να τροποποιήσουμε την παρουσίαση των αποτελεσμάτων της αναζήτησης έτσι ώστε κάθε αποτέλεσμα να έχει μια επιπρόσθετη σύνδεση δίπλα του, χρησιμοποιώντας την οποία ο χρήστης μπορεί να καθορίσει στην μηχανή αναζήτησης ότι αυτός είναι ο επιθυμητός προσδιορισμός. Για παράδειγμα, στην μηχανή αναζήτησης του Google, το πρώτο αποτέλεσμα της αναζήτησης για το πρόσωπο “Eric Miller”, θα μπορούσε να είναι ως εξής :

Eric Miller's Home Page
W3C, Eric Miller, Semantic Web Activity Lead
Lead for the W3C, World Wide Web Consortium
www.w3.org/EM/ - 4k - Cached - Similar Pages- This Eric Miller

Χρησιμοποιείται μια ευρετική προσέγγιση που στηρίζεται στη γνώση, στην οποία αναγνωρίζουμε και κωδικοποιούμε έναν αριθμό διαφορετικών ευρετικών, καθεμία από τις οποίες μπορεί να βρίσκει εφαρμογή σε διαφορετικές περιπτώσεις. Γνωρίζοντας μόνο το ότι ο χρήστης αναζητά πληροφορίες για ένα συγκεκριμένο πρόσωπο, μπορεί να συμβάλλει στην αποφυγή διαφόρων λαθών. Επίσης, ο τύπος του προσώπου (που προσδιορίζεται από τον όρο της αναζήτησης) δημιουργεί προσδοκίες για τις διάφορες κατηγορίες πληροφοριών που μπορεί να είναι διαθέσιμες. Για παράδειγμα, αν το πρόσωπο είναι ένας μουσικός, μπορούμε να αναμένουμε την επιστροφή ιστοσελίδων σχετικά με τα μουσικά του άλμπουμ, τις συναυλίες του κτλ. Αν το αντικείμενο είναι ένας καθηγητής, μπορούμε να αναμένουμε τα αποτελέσματα να αφορούν στις δημοσιεύσεις του, τα μαθήματα ή την έρευνά του. Αυτή η ευρετική μπορεί να χρησιμοποιηθεί όταν ο σημασιολογικός ιστός διαθέτει κάποια γνώση σχετικά με τον επιθυμητό προσδιορισμό. Η συγκεκριμένη προσέγγιση μπορεί να βοηθήσει τους χρήστες να αναγνωρίσουν για ποια αντικείμενα αναζητούν πληροφορίες και να ανακτήσουν έγγραφα με μεγάλη πιθανότητα αυτά να ανταποκρίνονται στις αρχικές τους επιθυμίες.

VII. ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΜΕΛΛΟΝΤΙΚΗ ΕΡΕΥΝΑ

Η ευρέως διαδεδομένη διαθεσιμότητα των μηχανικά αναγνώσιμων πληροφοριών έχει τη δυνατότητα να επηρεάσει πολλές εφαρμογές του παγκόσμιου ιστού, περιλαμβανομένης της αναζήτησης της επιθυμητής πληροφορίας. Εκτιμάται ότι οι προσανατολισμένες προς την έρευνα ερωτήσεις αναζήτησης μπορούν να εκμεταλλευτούν τον αναδυόμενο σημασιολογικό ιστό. Στο παρόν άρθρο, έγινε μια προσπάθεια περιγραφής δύο μηχανισμών για την επίτευξη του συγκεκριμένου στόχου, δίνοντας έμφαση στον εμπλουτισμό των αποτελεσμάτων της αναζήτησης με δεδομένα που προέρχονται από το σημασιολογικό ιστό. Η μελλοντική έρευνα θα αφορά στη βοήθεια της αναζήτησης στοιχείων κειμένου έτσι ώστε ένα τέτοιο σύστημα να μπορέσει να εκμεταλλευτεί μια βαθύτερη κατανόηση του προσδιορισμού των όρων της αναζήτησης.

VIII. ΑΝΑΦΟΡΕΣ

[1] D.Brickley and R.V.Guha, Rdf schema http://www.w3.org/TR/rdf-schema/
[2] D. Box, D. Ehnebuske, G. Kakivaya, A. Layman, N. Mendelsohn, H. F. Nielsen, S. Thatte, and D. Winder. Simple Object Acceess Protocol. http://www.w3.org/TR/SOAP/, May 2000.
[3] TAP: R.Guha and R. McCool. Tap: Towards a web of data. http://tap.stanford.edu/.
[4] B.McBride. Jena: Implementing the rdf model and syntax specification. Hewlett Packard Laboratories. http://wwwuk.hpl.hp.com/people/bwm/papers/20001221-paper/, 2001
[5] R.Guha, Eric Miller and R. McCool : Semantic Search
[6] http://www.w3.org/DesignIssues/Semantic.html