ΘΕΜΑΤΑ ΔΙΠΛΩΜΑΤΙΚΩΝ ΕΡΓΑΣΙΩΝ

Εργ. Συστημάτων Βάσεων Γνώσεων & Δεδομένων

ΥΛΟΠΟΙΗΣΗ ΤΕΧΝΙΚΩΝ ΔΙΑΦΟΡΟΠΟΙΗΣΗΣ ΑΠΟΤΕΛΕΣΜΑΤΩΝ ΑΝΑΖΗΤΗΣΗΣ ΜΕ ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ ΣΕ ΣΗΜΑΣΙΟΛΟΓΙΚΑ ΔΕΔΟΜΕΝΑ

ΠΛΗΡΟΦΟΡΙΕΣ: Γιώργος Γιαννόπουλος, 210 772 1402, giann [at] dblab.ece.ntua.gr
Νίκος Μπικάκης, 210 772 1402, bikakis [at] dblab.ece.ntua.gr

ΠΕΡΙΛΗΨΗ: Στόχος της προτεινόμενης διπλωματικής είναι η μελέτη της παρατιθέμενης βιβλιογραφίας και η υλοποίηση τεχνικών διαφοροποίησης (diversification) αποτελεσμάτων αναζήτησης με λέξεις-κλειδιά (keyword queries) σε σημασιολογικά δεδομένα, οργανωμένα σε RDF(S) μορφή.

ΑΤΟΜΑ: 1

ΠΛΑΤΦΟΡΜΑ ΕΡΓΑΣΙΑΣ: Java, Jena Framework, Lucene

ΣΥΝΤΟΜΗ ΠΕΡΙΓΡΑΦΗ: Η ευρεία εξάπλωση και χρήση του διαδικτύου, η διάθεση και διακίνηση μεγάλου όγκου πληροφορίας μέσω αυτού, σε συνδυασμό με την ανάπτυξη πληροφοριακών συστημάτων, τεχνολογιών και προτύπων βασισμένων σε διαφορετικές ανάγκες και ιδιαιτερότητες, έχουν σαν αποτέλεσμα την εμφάνιση *ετερογένειας* (*heterogeneity*) και τον περιορισμό των δυνατοτήτων του σημερινού *παγκόσμιου ιστού* (*WWW*). Τα παραπάνω καλείται να αντιμετωπίσει ο *Σημασιολογικός Ιστός* (*Semantic Web*) [1], ο οποίος αποτελεί τη μεγαλύτερη προσπάθεια αυτόματης ενοποίησης συστημάτων, με σκοπό να συνεργάζονται διαλειτουργικά σε παγκόσμιο επίπεδο. Στον *Σημασιολογικό Ιστό*, τα δεδομένα ακολουθούν το *RDF* (*Resource Description Framework*) [2],[3] μοντέλο, με κυρίαρχη γλώσσα ερωτήσεων, την *SPARQL* (*Simple Protocol and RDF Query Language*) [4]. Η γλώσσα ερωτήσεων SPARQL προσφέρει την δυνατότητα στους χρήστες και στις εφαρμογές του Σημασιολογικού Ιστού να εκφράζουν δομημένες ερωτήσεις (structured queries).

Παρόλο που η SPARQL προσφέρει μεγάλη εκφραστικότητα και δυνατότητες εκτέλεσης πολύπλοκων ερωτημάτων, οι χρήστες, στην πλειοψηφία των περιπτώσεων, προτιμούν να εκφράζουν τα ερωτήματά τους χρησιμοποιώντας απλά λέξεις-κλειδιά (keyword queries). Για αυτό το λόγο, είναι αναγκαία η ανάπτυξη μεθόδων και εργαλείων για την αποτελεσματική εκτέλεση απλών ερωτημάτων με λέξεις-κλειδιά σε σημασιολογικά δεδομένα.

Μία άλλη σημαντική πτυχή της διαδικασίας αναζήτησης και ανάκτησης αποτελεσμάτων είναι η δυνατότητα κάλυψης όσο το δυνατόν περισσότερων και ετερογενών αναγκών αναζήτησης, για το ίδιο ερώτημα, έτσι ώστε να ικανοποιείται όσο το δυνατόν μεγαλύτερος αριθμός χρηστών, από τα πρώτα αποτελέσματα αναζήτησης που επιστρέφονται. Αυτό επιτυγχάνεται με τεχνικές διαφοροποίησης αποτελεσμάτων, οι οποίες συνίστανται στην αναταξινόμηση των αποτελεσμάτων και/ή στη συλλογή ενός περιορισμένου αριθμού αποτελεσμάτων, με τέτοιο τρόπο ώστε τα πρώτα *k* αποτελέσματα που συλλέγονται να είναι όσο πιο ετερογενή μεταξύ τους γίνεται.

Σκοπός της εργασίας είναι η επέκταση και προσαρμογή τεχνικών διαφοροποίησης στο σενάριο της αναζήτησης, σύνθεσης και ταξινόμησης αποτελεσμάτων αναζήτησης σε σημασιολογικά δεδομένα, λαμβάνοντας υπόψη ότι τα σημασιολογικά δεδομένα είναι πιο πολύπλοκες δομές (γράφοι, δέντρα, μονοπάτια) από τις απλές ιστοσελίδες/έγγραφα. Για αυτό το λόγο, θα πρέπει να αναπτυχθούν εξειδικευμένα κριτήρια και αλγόριθμοι διαφοροποίησης που θα βρίσκουν εφαρμογή στον

συγκεκριμένο τύπο δεδομένων. Για τα ζητούμενα της διπλωματικής έχει ήδη γίνει προεργασία σε ερευνητικό επίπεδο και σε επίπεδο κώδικα, ο οποίος θα μπορεί να επαναχρησιμοποιηθεί/επεκταθεί.
Στα πλαίσια της διπλωματικής θα πραγματοποιηθούν οι ακόλουθες εργασίες:

1. Θα μελετηθεί η βιβλιογραφία που δίνεται παρακάτω και θα συζητηθούν προσεγγίσεις/ιδέες για διαφοροποίηση αποτελεσμάτων σημασιολογικής αναζήτησης.
2. Θα μελετηθούν τα (α) Jena Framework, το οποίο είναι ένα σύνολο βιβλιοθηκών για διαχείριση (οργάνωση, εκτέλεση ερωτημάτων) σημασιολογικών δεδομένων και (β) Apache Lucene, το οποίο είναι μία βιβλιοθήκη για αναζήτηση κειμενικής πληροφορίας με λέξεις κλειδιά (γ) ο υπάρχων κώδικας που θα αποτελέσει βάση τις διπλωματικής.
3. Θα υλοποιηθούν, σε μία κοινή εφαρμογή, οι επιλεγμένες μέθοδοι αναζήτησης, καθώς και η αντίστοιχη γραφική διεπιφάνεια επαφής χρηστών του συστήματος.

ΣΧΕΤΙΚΕΣ ΠΛΗΡΟΦΟΡΙΕΣ:

[1] Tim Berners-Lee, James Hendler, and Ora Lassila. *The Semantic Web*, Scientific American, May 17, 2001, Available at: *www.dblab.ece.ntua.gr/~bikakis/SW.pdf*
[2] Resource Description Framework (RDF), *http://www.w3schools.com/rdf/rdf_intro.asp*
[3] RDF Primer, *http://www.w3.org/TR/rdf-syntax/*
[4] Simple Protocol and RDF Query Language (SPARQL), *http://www.slideshare.net/olafhartig/an-introduction-to-sparql*
[5] M Drosou, E. Pitoura, Search result diversification, SIGMOD Record, 2010 http://www.sigmod.org/publications/sigmod-record/0906/publications/1003/p41.survey.drosou.pdf
[6] Dbpedia, http://dbpedia.org/About
[7] Lucene, http://lucene.apache.org/
[8] Jena, http://jena.apache.org/