ΘΕΜΑΤΑ ΔΙΠΛΩΜΑΤΙΚΩΝ ΕΡΓΑΣΙΩΝ

Εργ. Συστημάτων Βάσεων Γνώσεων & Δεδομένων

ΥΛΟΠΟΙΗΣΗ ΟΛΟΚΛΗΡΩΜΕΝΟΥ ΣΥΣΤΗΜΑΤΟΣ 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 ενός ολοκληρωμένου συστήματος εκτέλεσης ερωτήσεων με λέξεις-κλειδιά (keyword queries) σε σημασιολογικά δεδομένα, οργανωμένα σε RDF(S) μορφή.

ΑΤΟΜΑ: 1-2

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 αποτελεσμάτικη εκτέλεση απλών ερωτημάτων με λέξεις-κλειδιά σε σημασιολογικά δεδομένα.

Στη βιβλιογραφία έχουν προταθεί κάποιες μέθοδοι ([5], [6], [7], [8], [9]) για εκτέλεση τέτοιου τύπου ερωτημάτων, γενικότερα σε (ημι)δομημένα δεδομένα (βάσεις δεδομένων, γράφους) και σε RDF δεδομένα. Τα RDF δεδομένα μπορούν να χαρακτηριστούν ειδική περίπτωση δεδομένων με μορφή γράφου, οπότε, όλες οι παραπάνω μέθοδοι έχουν εφαρμογή σε αυτά. Το πρόβλημα με τις παραπάνω μεθόδους είναι ότι έχουν προταθεί από διαφορετικές ερευνητικές ομάδες, με σκοπό ερευνητικό και, σε κάποιες περιπτώσεις, με προσανατολισμό σε επιμέρους διαφορετικά υποπροβλήματα. Ως αποτέλεσμα, δεν υπάρχει ένα κοινό πλαίσιο υλοποίησης και σύγκρισης των παραπάνω μεθόδων σε RDF δεδομένα.

Στα πλαίσια της διπλωματικής θα πραγματοποιηθούν οι ακόλουθες εργασίες:

1. Θα μελετηθεί η βιβλιογραφία που δίνεται παρακάτω ([5], [6], [7], [8], [9]) και θα επιλεγούν προς υλοποίηση, σε συνεννόηση με τους υπευθύνους, οι βασικότερες μέθοδοι εκτέλεσης και αποτίμησης keyword queries σε δομημένα/RDF δεδομένα.

2. Θα μελετηθούν τα (α) Jena Framework (http://jena.sourceforge.net/ , http://jena.sourceforge.net/ARQ/), το οποίο είναι ένα σύνολο βιβλιοθηκών για διαχείριση (οργάνωση, εκτέλεση ερωτημάτων) σημασιολογικών δεδομένων και (β) Apache Lucene (http://lucene.apache.org/java/docs/index.html), το οποίο είναι μία βιβλιοθήκη για αναζήτηση κειμενικής πληροφορίας με λέξεις κλειδιά.
3. Θα οριστικοποιηθεί η λειτουργικότητα της προς ανάπτυξη εφαρμογής, με βάση την επιλεγμένη βιβλιογραφία και τις διαθέσιμες βιβλιοθήκες κώδικα.
4. 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] BLINKS: Ranked Keyword Searches on Graphs, *http://www.cs.duke.edu/dbgroup/papers/2007-SIGMOD-hwyy-kwgraph.pdf*
[6] Ontology-Based Interpretation of Keywords for Semantic Search, *http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.67.2742&rep=rep1&type=pdf*
[7] EASE: An Effective 3-in-1 Keyword Search Method for Unstructured, Semi-structured and Structured Data, *http://dbgroup.cs.tsinghua.edu.cn/ligl/papers/SIGMOD2008-EASE.pdf*
[8] Top-k Exploration of Query Candidates for Efficient Keyword Search on Graph-Shaped (RDF) Data, *http://people.aifb.kit.edu/dtr/papers/keywordtopk.pdf*
[9] Effective and efficient keyword query interpretation using a hybrid graph, *http://dl.acm.org/citation.cfm?id=1991358*