ΘΕΜΑΤΑ ΔΙΠΛΩΜΑΤΙΚΩΝ ΕΡΓΑΣΙΩΝ

Εργ. Συστημάτων Βάσεων Γνώσεων & Δεδομένων

ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΩΝ ΑΝΑΖΗΤΗΣΗΣ ΣΕ LINKED DATA ΠΕΡΙΒΑΛΛΟΝ

ΠΛΗΡΟΦΟΡΙΕΣ: Νίκος Μπικάκης, 210 772 1402, bikakis [at] dblab.ece.ntua.gr
Γιάννης Λιαγούρης, 210 772 1436, liagos [at] dblab.ece.ntua.gr

ΠΕΡΙΛΗΨΗ: Στόχος της προτεινόμενης διπλωματικής είναι μελέτη και η ανάπτυξη τεχνικών για την αξιολόγηση αποτελεσμάτων αναζήτησης σε Linked Data περιβάλλον.

ΑΤΟΜΑ: 1

ΠΛΑΤΦΟΡΜΑ ΕΡΓΑΣΙΑΣ: Java

ΣΥΝΤΟΜΗ ΠΕΡΙΓΡΑΦΗ: Ο όρος *Σημασιολογικός Ιστός (Semantic Web)* [1] αναφέρεται στο εγχείρημα εμπλουτισμού του υπάρχοντος (συντακτικού) ιστού (WWW) με σημασιολογική πληροφορία, δηλαδή με πληροφορία που περιγράφει τα ίδια τα δεδομένα που υπάρχουν στο διαδίκτυο και τις σχέσεις μεταξύ τους. Αυτού του είδους η μετα-πληροφορία, ή αλλιώς *μεταδεδομένα (metadata)*, διαφέρει από τις ετικέτες λέξεων-κλειδιών (plain-text tags), που ήδη χρησιμοποιούνται ευρέως, και στοχεύει στο να καταστήσει τους πόρους του διαδικτύου (web resources) εύκολα προσπελάσιμους από αυτοματοποιημένες διαδικασίες οι οποίες θα μπορούν να αξιοποιούν μεγαλύτερο μέρος της πληροφορίας που ενυπάρχει αυτή τη στιγμή στα έγγραφα του ιστού (υπερκείμενα), παρέχοντας έτσι πλουσιότερο και ακριβέστερο υλικό στις αναζητήσεις των τελικών χρηστών.

Καθοριστικό ρόλο στην προσπάθεια μετατροπής του σημερινού παγκόσμιου ιστού από *Ιστό Εγγράφων* (*Web of Documents*) σε *Ιστό Δεδομένων* (*Web of Data*) κατέχει το *Linked Open Data* Project *(LOD)* [3]. Στο LOD, εκατοντάδες σύνολα δεδομένων (datasets) από διάφορους οργανισμούς και υπηρεσίες ανά τον κόσμο δημοσιεύονται και συνδέονται μεταξύ τους ακολουθώντας το μοντέλο *RDF* (*Resource Description Framework*) [2]. Σύμφωνα με το τελευταίο, δεδομένα (και μεταδεδομένα) απεικονίζονται ως τριάδες (RDF triples) της μορφής <*subject, predicate, object*> οι οποίες συνθέτουν έναν κατευθυνόμενο γράφο με κόμβους (subject, object) που παριστάνουν τους πόρους/οντότητες του ιστού και ακμές (predicates) που αντιστοιχούν στις σχέσεις μεταξύ των πόρων. Ενδεικτικό της απήχησης του LOD είναι το γεγονός ότι ο όγκος των δεδομένων που ήδη φιλοξενεί αγγίζει τα 25 εκατ. RDF triples με σύνολα δεδομένων που φέρουν πάνω από 400 εκατ. διασυνδέσεις.

Οι διασυνδέσεις σε ένα περιβάλλον Linked Data μπορούν να θεωρηθούν ως μια επέκταση των διασυνδέσεων που ορίζονται μεταξύ πόρων (ιστοσελίδων) του παγκόσμιου ιστού, τόσο ως προς το είδος τους (στον ιστό αυτή τη στιγμή υπάρχει ένα μόνο είδος σχέσης - η υπερσύνδεση), όσο και ως προς την «λεπτομέρεια» (granularity) του τμήματος της πληροφορίας την οποία συνδέουν (ένας πόρος του Ιστού Δεδομένων μπορεί να αντιστοιχεί σε ένα ή περισσότερα τμήματα μιας ή περισσότερων ιστοσελίδων και όχι αποκλειστικά σε μια ολόκληρη ιστοσελίδα). Υπό αυτή την έννοια, ένα ενδιαφέρον πρόβλημα που προκύπτει σε πρώτη φάση και που αποτελεί αντικείμενο της παρούσας διπλωματικής είναι το πώς οι υπάρχουσες τεχνικές αξιολόγησης της «σημαντικότητας» των πόρων του ιστού (π.χ. PageRank [4], HITS [5], κ.λπ.) - και κατά συνέπεια των αποτελεσμάτων που επιστρέφονται στους χρήστες - μπορούν να προσαρμοστούν στις νέες απαιτήσεις ενός Linked Data περιβάλλοντος. Επομένως, η διπλωματική θα ασχοληθεί με τα εξής θέματα:

- Εξοικείωση με τις γλώσσες περιγραφής (μετα)δεδομένων σε Linked Data περιβάλλον.
- Μελέτη τεχνικών ανάλυσης υπερσυνδέσεων (hyperlink analysis) στον παγκόσμιο ιστό.
- Μελέτη πρόσφατων τεχνικών που εξειδικεύονται στην αξιολόγηση αποτελεσμάτων σε Linked Data περιβάλλον.
- Ορισμός διαδικασίας πειραματικής αξιολόγησης (benchmarking) των διαφόρων τεχνικών που προτείνονται και διεξαγωγή πειραμάτων σε πραγματικά δεδομένα.

ΠΕΡΙΣΣΟΤΕΡΕΣ 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] Linked Data - Connect Distributed Data Across the Web, *http://linkeddata.org/*
[4] PageRank, *http://en.wikipedia.org/wiki/PageRank*
[5] HITS Algorithm, *http://en.wikipedia.org/wiki/HITS_algorithm*