ΜΕΛΕΤΗ ΚΑΙ ΕΠΕΚΤΑΣΗ ΑΛΓΟΡΙΘΜΩΝ ΣΥΓΧΩΝΕΥΣΗΣ ΟΝΤΟΤΗΤΩΝ ΣΕ ΣΗΜΑΣΙΟΛΟΓΙΚΑ ΔΕΔΟΜΕΝΑ ΜΕ ΓΕΩΧΩΡΙΚΗ ΠΛΗΡΟΦΟΡΙΑ

ΠΛΗΡΟΦΟΡΙΕΣ: Γιώργος Γιαννόπουλος, giann [at] dblab.ece.ntua.gr
Δημήτρης Σκούτας, dskoutas [at] imis.athena-innovation.gr

ΠΕΡΙΛΗΨΗ: Στόχος της διπλωματικής είναι η μελέτη τεχνικών, αλγορίθμων και εργαλείων για συγχώνευση οντοτήτων σε σημασιολογικά δεδομένα και η επέκτασή τους ή ανάπτυξη νέων τεχνικών για συγχώνευση βασιζόμενη τόσο στη σημασιολογική, όσο και στη γεωχωρική πληροφορία των δεδομένων.

ΑΤΟΜΑ: 1

ΠΛΑΤΦΟΡΜΑ ΕΡΓΑΣΙΑΣ: Java

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].
Από την άλλη πλευρά, η διαχείριση και εκμετάλλευση μεγάλου όγκου γεωχωρικών δεδομένων είναι υψηλής σημασίας τόσο στη βιομηχανία (π.χ. τουριστικά πρακτορεία που αναλύουν τις τάσεις των πελατών τους) όσο και στην κοινωνική ζωή (π.χ. εκμετάλλευση γεωγραφικής πληροφορίας στα κοινωνικά δίκτυα για διασύνδεση χρηστών, διαφημίσεις, κ.α.).
Παρόλο που στις δύο παραπάνω περιοχές (ειδικά στα γεωχωρικά δεδομένα) έχουν γίνει σημαντικά βήματα προόδου, δεν ισχύει το ίδιο για τη διαχείριση δεδομένων που συνδυάζουν ιδιότητες και από τα δύο πεδία, δηλαδή για σημασιολογικά, γεωχωρικά δεδομένα.
Η συγκεκριμένη διπλωματική θα επικεντρωθεί στο πρόβλημα της συγχώνευσης δεδομένων (data fusion), το οποίο συνίσταται στα εξής στάδια: (α) αναζήτηση σημασιολογικών οντοτήτητων οι οποίες ενδέχεται να έχουν κωδικοποιηθεί με διαφορετικό τρόπο, αλλά αντιστοιχούν σε κοινή οντότητα και (β) δημιουργία συσχετίσεων μεταξύ των οντοτήτων και μετασχηματισμός των μεταδεδομένων τους, έτσι ώστε να είναι δυνατή η αυτόματη αντιστοίχισή τους και η εύρεση του συνόλου των μεταδεδομένων τους σε περίπτωση αναζήτησης των οντοτήτων.
Στα πλαίσια της διπλωματικής θα μελετηθούν διάφορες τεχνικές σημασιολογικής αντιστοίχισης και συγχώνευσης δεδομένων [5] καθώς και αντίστοιχα εργαλεία [6], [7] και, στη συνέχεια, θα αναπτυχθούν αλγόριθμοι οι οποίοι θα εκμεταλλεύονται, πέρα από τα σημασιολογικά μεταδεδομένα, και τη γεωχωρική πληροφορία των οντοτήτων για αποδοτική συγχώνευση των δεδομένων. Επιπλέον,

θα σχεδιαστούν benchmarks, πάνω σε κατάλληλα σύνολα σημασιολογικών γεωχωρικών δεδομένων [8], για την αξιολόγηση και σύγκριση των διαφόρων εξεταζόμενων ή προς υλοποίηση μεθόδων.

Η διπλωματική θα πραγματοποιηθεί στα πλαίσια του ερευνητικού έργου GeoKnow το οποίο είναι ένα τριετές, χρηματοδοτούμενο από την ΕΕ ερευνητικό έργο, που αφορά στο γεωχωρικό Σημασιολογικό Ιστό. Συνοπτικά, το GeoKnow καταπιάνεται με τη διασύνδεση, διαχείριση, ποιότητα, συνάθροιση, οπτικοποίηση και δημιουργία γεωχωρικών διαδικτυακών δεδομένων. Τα ερευνητικά μας αποτελέσματα θα εφαρμοστούν στις περιοχές των εφοδιαστικών αλυσίδων και των ταξιδιωτικών εταιρειών.

ΣΧΕΤΙΚΕΣ ΠΛΗΡΟΦΟΡΙΕΣ:

[1] Tim Berners-Lee et al. *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] Bernstein et al. Generic Schema Matching, Ten Years Later, VLDB'11 http://www.sigmod.org/publications/sigmod-record/0906/publications/1003/p41.survey.drosou.pdf
[6] Interlinking tools, http://stack.lod2.eu/
[7] Fusion tools, http://sieve.wbsg.de/
[8] Datasets, http://linkedgeodata.org