ΓΕΩΓΡΑΦΙΚΗ ΣΕΛΙΔΟΔΕΙΚΤΟΔΟΤΗΣΗ ΤΟΥ ΙΣΤΟΥ

ΘΕΜΑΤΑ ΔΙΠΛΩΜΑΤΙΚΩΝ ΕΡΓΑΣΙΩΝ

Εργαστήριο Συστημάτων Βάσεων Γνώσεων & Δεδομένων

ΕΙΣΑΓΩΓΙΚΟ ΣΗΜΕΙΩΜΑ ΣΤΗ ΓΕΩΓΡΑΦΙΚΗ ΣΕΛΙΔΟΔΕΙΚΤΟΔΟΤΗΣΗ ΤΟΥ (ΕΛΛΗΝΙΚΟΥ) ΙΣΤΟΥ

Πολλές όψεις μεταδεδομένων μπορούν να χρησιμοποιηθούν για την οργάνωση, ευρετηριοποίηση, αναζήτηση και πλοήγηση σε ιστοσελίδες, όπως λέξεις-κλειδιά, θεματικές κατηγορίες, γεωγραφική θέση, χρόνος. Μια σημαντική, καθότι μονοσήμαντη, τέτοια όψη είναι η γεωγραφική θέση - ένα ζεύγος συντεταγμένων έχει μια και μόνη ερμηνεία. Οι διπλωματικές αυτές εργασίες στοχεύουν στην ανάπτυξη τεχνολογίας για την ημι-αυτόματη δημιουργία γεωγραφικών σελιδοδεικτών στον Ιστού (γεω-σελιδοδεικτοδότηση - geo-bookmarking). Στα πλαίσια των εργασιών θα γίνει χρήση (i) τεχνολογίας γεωκωδικοποίησης, για την αυτόματη αναγνώριση (γεωκωδικοποίηση) λέξεων- και φράσεων- κλειδιών (γεω-αναγνωριστικά) σε ιστοσελίδες, και (ii) ανάδρασης από το χρήστη πάνω στα αποτελέσματα της γεωκωδικοποίησης.

Η προτεινόμενη προσέγγιση εμπίπτει στην ευρύτερη κατηγορία της κοινωνικής σελιδοδεικτοδότησης (βλ. http://del.icio.us) και δημιουργίας (χωρικών) μεταδεδομένων, για την καλύτερη ευρετηριοποίηση του Ιστού. Συνεπώς, η εργασία αυτή μπορεί να αποτελέσει σημαντική συνεισφορά στην θεμελίωση του Web 2.0, δηλαδή την δημιουργία μεταδεδομένων για υπάρχον περιεχόμενο του Ιστού, με αποδοτική χρήση πόρων.

Η γεωγραφική σελιδοδεικτοδότηση του ιστού πρέπει να είναι εξίσου απλή με τη σήμανση ιστοσελίδων στα πλαίσια της κοινωνικής σελιδοδεικτοδότησης. Η απαιτούμενη προσπάθεια για τη σήμανση τμημάτων ιστοσελίδων πρέπει να αντισταθμίζεται από λογισμικό αυτόματης γεωκωδικοποίησης που υποστηρίζει αυτή τη λειτουργία. Η αυτόματη γεωκωδικοποίηση θα παρέχει ένα καλό αρχικό αποτέλεσμα, το οποίο μπορεί να τροποποιείται και να βελτιώνεται εύκολα με την χρήση μιας απλής και διαισθητικής διεπαφής (επέκταση φυλλομετρητή που περιλαμβάνει χάρτες και υπερσυνδέσμους στο κείμενο).

Το όλο έργο υποδιαιρείται στις ακόλουθες δύο εργασίες:

1. Τεχνικές γεωτεχνολόγησης (geoparsing) – αναγνώριση γεωγραφικών "ενδείξεων" σε κείμενα
2. Κοινωνική γεω-σελιδοδεικτοδότηση - αλληλεπίδραση με χρήστες και ολοκλήρωση πληροφορίας από πολλαπλές πηγές

ΘΕΜΑΤΑ ΔΙΠΛΩΜΑΤΙΚΩΝ ΕΡΓΑΣΙΩΝ

Εργαστήριο Συστημάτων Βάσεων Γνώσεων & Δεδομένων

ΕΡΓΑΣΙΑ 1: ΓΕΩΤΕΧΝΟΛΟΓΗΣΗ/ΓΕΩΚΩΔΙΚΟΠΟΙΗΣΗ (GEOPARSING/GEOCODING)

ΠΛΗΡΟΦΟΡΙΕΣ: Dieter Pfoser, 210 6930700, pfoser@cti.gr

ΑΤΟΜΑ: 1

ΠΛΑΤΦΟΡΜΑ ΕΡΓΑΣΙΑΣ: Java, Minor Third, κ.α.

ΣΥΝΤΟΜΗ ΠΕΡΙΓΡΑΦΗ: Η εργασία αυτή στοχεύει στον εμπλουτισμό και την επέκταση ήδη υπάρχοντος συστήματος εξαγωγής γεωγραφικής πληροφορίας από ιστοσελίδες, που έχει αναπτυχθεί στο εργαστήριο (Διπλωματική εργασία Α. Άντζελ, 2006, http://www.dblab.ece.ntua.gr/pubs/details.php?id=1427&clang=0).

Θα υλοποιηθεί τεχνολογία αυτόματης γεωκωδικοποίησης για τον εντοπισμό γεω-αναγνωριστικών σε ιστοσελίδες (γεωτεχνολόγηση) και την μετέπειτα τοποθέτησή τους στο χάρτη. Τα γεω-αναγνωριστικά μπορεί να είναι από απλά, όπως τηλέφωνα ή διευθύνσεις, που εμφανίζονται σε σχετικά κανονικές μορφές, έως εξαιρετικά αμφίσημα, όπως τοπωνύμια (π.χ. "Ελ.Βενιζέλος" - αναφέρεται στον πολιτικό, μια οδό ή το αεροδρόμιο;). Το υπάρχον σύστημα επικεντρώνεται στον εντοπισμό σύνθετων γεω-αναγνωριστικών (διευθύνσεων, τηλεφώνων, κ.ο.κ.) - στην παρούσα εργασία έμφαση θα δοθεί στην αποδοτική αναγνώριση τοπωνυμίων, για την οποία και θα εξεταστούν εναλλακτικές μέθοδοι (λ.χ. υπό συνθήκη στοχαστικά πεδία, κανονικές γραμματικές, κ.α.).

Τα γεω-αναγνωριστικά αντιστοιχίζονται στις εγγραφές μιας βάσης δεδομένων (Γεω-Βάση), που περιλαμβάνει πληροφορίες τηλεφωνικού καταλόγου, οδικούς χάρτες, πληροφορίες σχετικά με ταχυδρομικούς κώδικες, διευθύνσεις IP και τοπωνύμια. Η αντιστοίχιση αυτή, για να είναι ανεκτική σε σφάλματα, πρέπει να προβλέπει έναν βαθμό ασάφειας - για παράδειγμα, υπάρχουν εναλλακτικές (π.χ. Θωρικό - Θορικό) ή εσφαλμένες μορφές μιας λέξης, συντμήσεις (π.χ. Αιτ/νίας, Ελ. Βενιζέλου), συνώνυμα (Ελ.Βενιζέλου- Πανεπιστημίου) κ.ο.κ. Η λειτουργικότητα αυτή παρέχεται πλήρως από το υπάρχον σύστημα.

Τέλος, η επιμέρους πληροφορία που συλλέγεται από τα γεω-αναγνωριστικά πρέπει να συνεκτιμάται και να ολοκληρώνεται, για την εξαγωγή πληροφορίας για την ιστοσελίδα συνολικά (ή για τμήματά της). Αυτό είναι απαραίτητο, διότι τα γεω-αναγνωριστικά είναι μια πλούσια αλλά θορυβώδης (δηλ. με αρκετά σφάλματα) πηγή γεωγραφικής πληροφορίας. Κριτήρια γι' αυτήν τη συνεκτίμηση είναι, για παράδειγμα, η επιμέρους σημαντικότητα και βεβαιότητα κάθε γεω-αναγνωριστικού, μια δομική ανάλυση της ιστοσελίδας (π.χ. διάκριση μεταξύ τίτλων/ πινάκων/ κειμένου/ λεζαντών, χωρισμός σε ενότητες, κ.ο.κ.), η απόσταση των γεω-αναγνωριστικών στο κείμενο και στο χάρτη κ.ο.κ.

Συνοψίζοντας, η εργασία αυτή θα ασχοληθεί κυρίως με:

1. Τον αποδοτικό εντοπισμό τοπωνυμίων σε μια ιστοσελίδα, με την παράλληλη επίλυση των σχετικών προβλημάτων αμφισημίας.
2. Την ομαδοποίηση, συνεκτίμηση και συνάθροιση των πληροφοριών από τα επιμέρους γεω-αναγνωριστικά, για την εξαγωγή συνολικότερης πληροφορίας.

ΘΕΜΑΤΑ ΔΙΠΛΩΜΑΤΙΚΩΝ ΕΡΓΑΣΙΩΝ

Εργαστήριο Συστημάτων Βάσεων Γνώσεων & Δεδομένων

ΕΡΓΑΣΙΑ 2: ΚΟΙΝΩΝΙΚΗ ΣΕΛΙΔΟΔΕΙΚΤΟΔΟΤΗΣΗ

ΠΛΗΡΟΦΟΡΙΕΣ: Dieter Pfoser, 210 6930700, pfoser@cti.gr

ΑΤΟΜΑ: 1

ΠΛΑΤΦΟΡΜΑ ΕΡΓΑΣΙΑΣ: Java, MIT PiggyBank, Web Services, Google Maps, κ.α.

ΣΥΝΤΟΜΗ ΠΕΡΙΓΡΑΦΗ: Η γεωκωδικοποίηση παράγει ως αποτέλεσμα χωρικές συντεταγμένες για τμήματα μιας ιστοσελίδας. Πρόκειται, δηλαδή, για μια εργασία σημασιολογικά πλούσια. Επειδή υλοποιείται ως μια αυτόματη διαδικασία, η ποιότητα του αποτελέσματος μπορεί να μην είναι η αναμενόμενη. Για τη βελτίωσή της, θα δημιουργηθεί ένα εργαλείο - επέκταση φυλλομετρητή, με την εξής λειτουργικότητα:

1. Κατά την πλοήγηση σε μια ιστοσελίδα, το εργαλείο καλείται για την αυτόματη γεωκωδικοποίησή της.
2. Χρησιμοποιώντας μια διεπαφή χάρτη με το αρχικό αποτέλεσμα, ο χρήστης μπορεί να το διορθώσει, και
3. Το τελικό αποτέλεσμα, μαζί με ένα βαθμό βεβαιότητας, αποθηκεύεται σε μια δημόσια δεξαμενή γεω-σελιδοδεικτοδοτημένων ιστοσελίδων.

Το εργαλείο θα βασιστεί στην πλατφόρμα MIT Piggybank (http://simile.mit.edu/wiki/Piggy_Bank). Η εκτέλεσή του θα αναλύει/γεωκωδικοποιεί την δεδομένη ιστοσελίδα, κάνοντας χρήση απομακρυσμένης βάσης δεδομένων που περιέχει τα απαραίτητα δεδομένα (ΓεωΒάση). Η πρόσβαση στην ΓεωΒάση θα γίνεται μέσω Web Services.
Το αποτέλεσμα αυτού του αρχικού βήματος θα είναι μια νέα ιστοσελίδα, που θα περιέχει (i) αντίγραφο της αρχικής ιστοσελίδας, με επισημασμένα τα γεω-αναγνωριστικά, με υπερσυνδέσμους στις αντίστοιχες τοποθεσίες σε (ii) ένα χάρτη π.χ. Google Maps.
Στην νέα αυτή ιστοσελίδα (i) νέα γεω-αναγνωριστικά μπορούν να επισημανθούν, και η τοποθεσία τους να επιλεγεί στο χάρτη, (ii) γεω-αναγνωριστικά που έχουν εσφαλμένα ανιχνευθεί μπορούν να διαγραφούν, και (iii) η τοποθεσία των γεω-αναγνωριστικών μπορεί να μεταβληθεί.
Τέλος, οι γεω-σελιδοδείκτες, μαζί με την ιστοσελίδα, μπορούν να αποθηκευτούν σε μια κεντρική δεξαμενή πληροφοριών. Η δεξαμενή αυτή θα παρέχει και απλή λειτουργικότητα (i) χωρικής (χάρτες) και (ii) θεματικής (λέξεις-κλειδιά) αναζήτησης, για την ανάκτηση των αποθηκευμένων γεω-σελιδοδεικτών. Η λειτουργικότητα θεματικής αναζήτησης θα βασιστεί σε υπάρχουσες μηχανές αναζήτησης (π.χ. Apache Lucene).