ΙΣΤΟΣ & ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ

ΝΕΑ:

ΘΕΜΑΤΑ ΔΙΠΛΩΜΑΤΙΚΩΝ ΕΡΓΑΣΙΩΝ

Εργαστήριο Συστημάτων Βάσεων Γνώσεων & Δεδομένων

Partialist: ΣΥΣΤΗΜΑ ΕΠΕΞΕΡΓΑΣΙΑΣ ΕΡΩΤΗΣΕΩΝ ΓΙΑ ΔΕΝΤΡΙΚΑ ΔΕΔΟΜΕΝΑ

ΠΛΗΡΟΦΟΡΙΕΣ: Σ. Σουλδάτος, 210 772 1402, stef@dblab.ntua.gr, Θ. Δαλαμάγκας, 210 772 1402, dalamag@dblab.ntua.gr

ΠΕΡΙΛΗΨΗ: Η διπλωματική εργασία στοχεύει στο σχεδιασμό και την υλοποίηση συστήματος επεξεργασίας ερωτήσεων για δεντρικά δεδομένα XML. Βασικό χαρακτηριστικό των ερωτήσεων είναι ότι επιτρέπουν μερικό προσδιορισμό δομής.

ΑΤΟΜΑ: 1-2

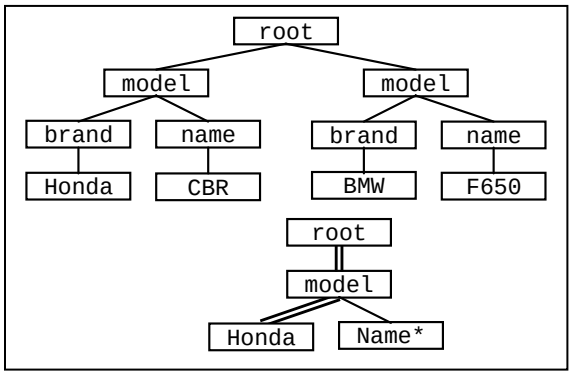
ΠΛΑΤΦΟΡΜΑ ΕΡΓΑΣΙΑΣ: C++/Linux

ΣΥΝΤΟΜΗ ΠΕΡΙΓΡΑΦΗ:

Η **γλώσσα XML** είναι πλέον πρότυπο ανταλλαγής δεδομένων σε εφαρμογές του Παγκόσμιου Ιστού. Η χρήση **ετικετών** (tags/elements) για το μαρκάρισμα και τον χαρακτηρισμό των δεδομένων σε ένα XML αρχείο διευκολύνει την επεξεργασία του από προγράμματα. Δείτε ένα πολύ απλό παράδειγμα XML αρχείου που αποθηκεύει πληροφορίες για μοντέλα μηχανών.

```
<?xml version="1.0">
<root>
    <model>
        <brand>Honda</brand>
        <name>CBR</name>
    </model>
    <model>
        <brand> BMW </brand>
        <name>F650</name>
    </model>
<root>
```

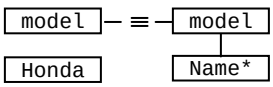
Το XML αρχείο του παραδείγματος αντιστοιχεί σε μια **δενδρική δομή**, αυτή που φαίνεται στο διπλανό σχήμα.



Γλώσσες ερωτήσεων, όπως η XPath, χρησιμοποιούνται ευρέως για την ανάκτηση δεδομένων από τέτοιες δεντρικές δομές. Για παράδειγμα, η ερώτηση //model[//Honda]/Name, ζητά τα ονόματα των μοντέλων της εταιρίας Honda. Η απάντηση της ερώτησης είναι η εικόνα της (mapping) πάνω στο δέντρο, εδώ το μοντέλο CBR.

Οι γλώσσες αυτές απαιτούν από το χρήστη να ξέρει τη δομή του XML αρχείου. Για παράδειγμα, στην προηγούμενη ερώτηση, θα πρέπει ο χρήστης να ξέρει ότι το Honda βρίσκεται κάτω από το model. Αν ο χρήστης δε γνωρίζει τις σχετικές θέσεις τους, θα πρέπει να φτιάξει ερωτήσεις για όλες τις πιθανές θέσεις των κόμβων, δηλ. μια ερώτηση που θα έχει το Honda πάνω από το model και μια που θα τα έχει ανάποδα.

Για το λόγο αυτό, έχουμε ορίσει μια **γλώσσα μερικώς ορισμένων δενδρικών ερωτήσεων** (**PTPQ**), παρόμοια με την XPath, η οποία όμως επιτρέπει σε κόμβους να μείνουν ξεκρέμαστοι, όπως ο κόμβος Honda στο παράδειγμα.



Επίσης, έχουμε ξεκινήσει την υλοποίηση αρκετών μηχανισμών, όπως αποτίμηση ερωτήσεων, έλεγχος ικανοποιησιμότητας, εξαγωγή πλήρους μορφής, έλεγχος περιεκτικότητας, κ.λ.π. Βασικό κομμάτι των μηχανισμών αυτών αποτελεί ο Γράφος Διαστάσεων που αποτελεί περίληψη της δομής του XML αρχείου. Το ζητούμενο σύστημα θα εκτελεί τις παρακάτω λειτουργίες:

- *Επεξεργασία Δέντρου*: θα υπάρχει η δυνατότητα γραφικής αποτύπωσης, επεξεργασίας και εκ νέου αποθήκευσης δενδρικών δεδομένων. Επιπλέον, θα υποστηρίζεται η ημιαυτόματη εξαγωγή Γράφου Διαστάσεων.

- *Επεξεργασία Γράφου*: θα υπάρχει η δυνατότητα γραφικής αποτύπωσης, επεξεργασίας και αποθήκευσης γράφων.

- *Επεξεργασία Ερωτήσεων*: θα υπάρχει η δυνατότητα γραφικής αποτύπωσης, επεξεργασίας και αποθήκευσης ερωτήσεων. Επιπλέον, θα ενσωματωθούν υπάρχουσες λειτουργίες (υλοποιημένες), όπως η εξαγωγή ερώτησης πλήρους μορφής, έλεγχος ικανοποιησιμότητας, έλεγχος περιεκτικότητας. Τέλος, θα ενσωματωθούν υπάρχοντες αλγόριθμοι επεξεργασίας και αποτίμησης ερωτήσεων, και θα υλοποιηθούν και νέοι.

Η διπλωματική αυτή επομένως, θα ασχοληθεί με τα παρακάτω θέματα:
1. Μελέτη μερικώς ορισμένων δενδρικών ερωτήσεων.
2. Ανάλυση απαιτήσεων συστήματος.
3. Σχεδίαση συστήματος.
4. Υλοποίηση συστήματος και σύνδεση με τις υλοποιημένες λειτουργίες
5. Υλοποίηση νέων αλγορίθμων επεξεργασίας και αποτίμησης ερωτήσεων.

ΘΕΜΑΤΑ ΔΙΠΛΩΜΑΤΙΚΩΝ ΕΡΓΑΣΙΩΝ

Εργαστήριο Συστημάτων Βάσεων Γνώσεων & Δεδομένων

ΔΙΑΧΕΙΡΙΣΗ MXML ΚΑΙ MXPATH ΜΕΣΩ ΣΧΕΣΙΑΚΩΝ ΒΔ

ΠΛΗΡΟΦΟΡΙΕΣ: Γιάννης Σταύρακας, 210 772 1446, ys@dblab.ntua.gr

ΠΕΡΙΛΗΨΗ: Υλοποίηση συστήματος αποθήκευσης Multidimensional XML (MXML) και απάντησης ερωτημάτων Multidimensional XPath (MXPath).

ΑΤΟΜΑ: 1 εως 2. Συνίσταται προηγούμενη εξοικείωση με Java.

ΠΛΑΤΦΟΡΜΑ ΕΡΓΑΣΙΑΣ: Java, XML, XPath, SQL.

ΣΥΝΤΟΜΗ ΠΕΡΙΓΡΑΦΗ: Θυμίζω ότι η XML είναι μια γλώσσα αναπαράστασης (κωδικοποίησης) πληροφορίας για το Web, και το XPath είναι μια γλώσσα αναζήτησης δεδομένων μέσα σε XML κείμενα. Η MXML από την άλλη, είναι μια επέκταση της XML που αναπαριστά πληροφορία η οποία μπορεί να παρουσιάζει διαφορετικές εναλλακτικές μορφές ανάλογα με το **context**, όπου σαν context θεωρούμε μια σειρά «εξωτερικών» συνθηκών. Έτσι η MXML είναι κατάλληλη για τον χειρισμό δεδομένων των οποίων η δομή και η τιμή αλλάζουν ανάλογα με τις συνθήκες (context). Αντίστοιχα, το MXPath είναι μια επέκταση του XPath που ενσωματώνει context, και είναι κατάλληλο για αναζήτηση πληροφορίας σε MXML κείμενα. Προηγούμενες εργασίες έχουν χρησιμοποιήσει σχεσιακές ΒΔ για την αποθήκευση XML εγγράφων, και για την απάντηση XPath ερωτημάτων (μέσω μετάφρασής τους σε SQL). Η διπλωματική αυτή θα υλοποιήσει ένα αντίστοιχο σύστημα για MXML και MXPath, το οποίο:

- Θα αποθηκεύει σε σχεσιακούς πίνακες με συγκεκριμένο τρόπο οποιοδήποτε κείμενο MXML
- Θα απαντάει σε context-aware ερωτήματα MXPath πάνω στα αποθηκευμένα MXML κείμενα, μεταφράζοντας τα ερωτήματα αυτά σε ερωτήματα SQL πάνω στους σχετικούς πίνακες.

Ο τρόπος που προτείνεται να υλοποιηθούν τα παραπάνω έχει ως εξής:

- Σχεδιασμός και υλοποίηση αλγορίθμου που μετατρέπει και αποθηκεύει MXML σε συγκεκριμένη σχεσιακή δομή.
- Σχεδιασμός και υλοποίηση αλγορίθμου που μεταφράζει MXPath σε «ισοδύναμη» ερώτηση SQL.
- Σχεδιασμός και υλοποίηση αλγορίθμου που μετατρέπει περιεχόμενα των σχεσιακών πινάκων (που έρχονται σαν απάντηση στην ερώτηση SQL) πίσω σε MXML (που είναι και η τελική απάντηση στο αρχικό MXPath).
- Σχεδιασμός και υλοποίηση εφαρμογής που θα ενσωματώνει τους παραπάνω αλγορίθμους κάτω από ένα ενιαίο user interface, και που θα μπορεί να εισάγει και να διαγράφει MXML κείμενα, καθώς και να θέτει MXPath ερωτήματα και να παρουσιάζει τα αποτελέσματα.

Μια εναλλακτική υλοποίηση θα μπορούσε να αποτελείται από αλγόριθμους μετατροπής της MXML σε XML και του MXPath σε XPath, μαζί με χρήση κάποιου συστήματος διαχείρισης XML μέσω σχεσιακής βάσης. Η τελική προσέγγιση θα αποφασιστεί στην αρχή της διπλωματικής.

Η διπλωματική είναι προγραμματιστική και θα δώσει μια καλή πρακτική εξάσκηση σε Java, έχει όμως και αρκετό ερευνητικό υπόβαθρο, και είναι πιθανό να οδηγήσει σε κάποια δημοσίευση.

ΘΕΜΑΤΑ ΔΙΠΛΩΜΑΤΙΚΩΝ ΕΡΓΑΣΙΩΝ

Εργαστήριο Συστημάτων Βάσεων Γνώσεων & Δεδομένων

ΠΡΟΗΓΜΕΝΗ ΑΝΑΖΗΤΗΣΗ ΠΛΗΡΟΦΟΡΙΩΝ ΣΤΟΝ ΙΣΤΟ: Εισαγωγικό σημείωμα

ΠΛΗΡΟΦΟΡΙΕΣ: Θοδωρής Δαλαμάγκας, 210 772 1402, dalamag@dblab.ntua.gr

Η ΤΡΕΧΟΥΣΑ ΚΑΤΑΣΤΑΣΗ

Οι μηχανές αναζήτησης, όπως το Google, είναι το βασικό εργαλείο αναζήτησης πληροφορίας στο Web. Η δημοφιλία στη χρήση τους οφείλεται σε δύο παράγοντες:

1. **Απλή μορφή γλώσσας ερώτησης** (**keyword-based search**): ο χρήστης δεν χρειάζεται να γνωρίζει κάποια γλώσσα ερωτήσεων (π.χ. SQL) με σύνταξη και σημασιολογία για να διατυπώνει την ερώτησή του. Απλά και μόνο πληκτρολογεί ένα σύνολο από **λέξεις-κλειδιά** (**keywords**) που θεωρεί ότι περιγράφουν καλύτερα το θέμα προς αναζήτηση. Στη συνέχεια, η μηχανή επιστρέφει ιστοσελίδες με περιεχόμενο σχετικό ως προς αυτό το θέμα. Μάλιστα οι ιστοσελίδες **ταξινομούνται** (**relevance ranking**) ως προς το **βαθμό ομοιότητάς** (**similarity value**) τους με τις λέξεις-κλειδιά.
2. **Ώριμη τεχνολογία αναζήτησης κειμένου** (**text information retrieval**): οι τεχνολογίες αναζήτησης κειμένων με περιεχόμενο σχετικό ως προς κάποιες λέξεις-κλειδιά έχουν ήδη συμπληρώσει πάνω από 25 χρόνια ζωής[1]. Η προσθήκη μηχανισμών που εκμεταλλεύονται την ύπαρξη συνδέσμων μεταξύ κειμένων στον Ιστό για να επιβεβαιώσουν την ομοιότητά τους και τη σχέση τους, και να αναπροσαρμόσουν την ταξινόμηση των αποτελεσμάτων (το γνωστό **PageRank**[2] του Google), έχει βελτιώσει σημαντικά την ποιότητα των αποτελεσμάτων της αναζήτησης.

Το απλό μοντέλο ερώτησης είναι σημαντικό πλεονέκτημα, τουλάχιστον για αναζητήσεις σε θέμα καλά ορισμένο εκ των προτέρων. Αν για παράδειγμα θέλετε να βρείτε reviews για συσκευές mp3, τότε πληκτρολογώντας απλά τις λέξεις-κλειδιά "reviews mp3 player" όλα τα πρώτα αποτελέσματα ικανοποιούν πλήρως τις ανάγκες σας, όπως φαίνεται και στη σχετική εικόνα.

ΤΟ ΠΡΟΒΛΗΜΑ

Συχνά όμως οι ανάγκες αναζήτησης πληροφορίας είναι πιο σύνθετες. Σκεφτείτε ένα μεταπτυχιακό φοιτητή που ψάχνει πληροφορίες για τις τρέχουσες τεχνολογίες, τις δημοσιεύσεις, τις ερευνητικές ομάδες, κ.λ.π. για μια ερευνητική θεματική περιοχή. Ο φοιτητής έχει στο μυαλό του μια **αφαιρετική περιγραφή του πεδίου γνώσης που θέλει να εξερευνήσει**.

✓ Για παράδειγμα, γνωρίζει ότι τον ενδιαφέρει το θέμα *κατευθυνόμενοι γράφοι* σε σχέση με το θέμα *βάσεις δεδομένων*. Η πληροφορία που αναζητά είναι δημοσιεύσεις τύπου *survey/review papers*,

1 http://www.cs.mu.oz.au/mg/ (υπάρχει στη βιβλιοθήκη του εργαστηρίου)

2 http://www.webworkshop.net/pagerank.html, http://infolab.stanford.edu/~backrub/google.html

δηλαδή σχετικά άρθρα που συνοψίζουν την κατάσταση προόδου μιας συγκεκριμένης ερευνητικής περιοχής, καθώς και *σχετικά άρθρα από blogs*.

- ✓ Εκτός από αυτό, τον ενδιαφέρουν οι *αλγόριθμοι αποτίμησης ερωτήσεων σε γράφους*, και πιο συγκεκριμένα οι ερωτήσεις τύπου *reachability* που απαντούν αν δύο κόμβοι είναι στο ίδιο μονοπάτι. Γνωρίζει ότι οι αλγόριθμοι αυτοί είναι δύο κατηγοριών: αυτοί που χρησιμοποιούν κάποια μορφή αριθμητικής κωδικοποίησης (*labelling scheme*) για τους κόμβους και αυτοί που δεν τη χρησιμοποιούν. Και για τις δύο περιοχές θα ήθελε να βρει *σχετικά άρθρα από συνέδρια και επιστημονικά περιοδικά*.
- ✓ Γνωρίζει επίσης, ότι οι αλγόριθμοι της δεύτερης κατηγορίας, που είναι και οι πιο παλιοί, γνώρισαν μια δεύτερη άνθιση, γιατί πολλοί ερευνητές προσπάθησαν να αντιμετωπίσουν τους "ανταγωνιστές" αλγόριθμους της πρώτης κατηγορίας που επίσης εμφανίστηκαν λίγο πριν. Θέλει επομένως, τα άρθρα της δεύτερης κατηγορίας που θα αναζητηθούν να είναι *πιο νέα από τα άρθρα της πρώτης*.

Πολλές φορές λοιπόν έχουμε στο μυαλό μας κάποιες **έννοιες** ή **θέματα** (θα τα λέμε απλά **έννοιες** – **concepts** – από εδώ και στο εξής**)** που περιγράφουν το γενικότερο **πεδίο γνώσης** που θέλουμε να εξερευνήσουμε και να αναζητήσουμε πληροφορία. Για κάθε μια τέτοια έννοια, μπορεί να αναζητούμε διαφορετικά πράγματα: π.χ. κείμενα σχετικά με την έννοια Α, εικόνες σχετικές με την έννοια Β, blogs και άρθρα εφημερίδων για την έννοια Γ. Επίσης, οι έννοιες μπορεί να σχετίζονται μεταξύ τους (**concept relationships**), και οι σχέσεις αυτές μπορεί να θέλουμε να υπόκειται σε **εξαρτήσεις** (**constraints**).

> Οι αφαιρετικές περιγραφές του πεδίου γνώσης είναι δημοφιλές χαρακτηριστικό των **εργαλείων διαχείρισης της σκέψης** (**mind manager tool)**. Τέτοια εργαλεία είναι γνωστά στο χώρο της εκπαιδευτικής κοινότητας (δείτε σχετικά http://en.wikipedia.org/wiki/Mind_map). Χρησιμοποιούν συνήθως διαγράμματα αναπαράστασης ιδεών και συσχετίσεων μεταξύ τους, ώστε να βοηθήσουν τον εκπαιδευόμενο να κατανοήσει τις βασικές ιδέες τις οποίες αργότερα θέλει να αναλύσει και να εξειδικεύσει. Τα διαγράμματα αυτά είναι σημαντικό βοήθημα για οργάνωση μελέτης, επίλυση προβλημάτων, επιλογή απόφασης, συγγραφή κειμένων, κ.λ.π.

Το τρέχον μοντέλο λειτουργικότητας των μηχανών αναζήτησης αδυνατεί να ικανοποιήσει τις ανάγκες χρηστών όπως ο παραπάνω φοιτητής. Συγκεκριμένα:

1. Το μοντέλο ερώτησης δεν μπορεί να εκφράσει τις παραπάνω ανάγκες σε μια ενιαία διαδικασία αναζήτησης.
2. Η αφαιρετική περιγραφή του πεδίου γνώσης που περιγράψαμε πριν, παρέχει από μόνη της σημασιολογική πληροφορία που μπορεί να χρησιμοποιηθεί για τη βελτίωση των αποτελεσμάτων αναζήτησης. Η βελτίωση αυτή μπορεί να είναι α) σε επίπεδο ποιότητας ταξινόμησης αποτελεσμάτων (**ranking**), αλλά και β) σε επίπεδο παρουσίασης (π.χ. **συσταδοποίηση** – **clustering** - αποτελεσμάτων με κοινό θεματικό περιεχόμενο). Ο μηχανισμός αποτίμησης ερωτήσεων στο τρέχον μοντέλο λειτουργικότητας των μηχανών αναζήτησης αγνοεί τη σημασιολογική αυτή πληροφορία.
3. Η αφαιρετική περιγραφή του πεδίου γνώσης παρέχει επίσης πληροφορία με την οποία μπορεί κάποιος να **συντονίσει την αναζήτηση**, κατευθύνοντας τις ερωτήσεις σε συγκεκριμένες μηχανές (π.χ. technorati για blogs, google images για εικόνες), πάλι έχοντας ως στόχο τη βελτίωση των αποτελεσμάτων αναζήτησης. Και πάλι ο μηχανισμός αποτίμησης ερωτήσεων στο τρέχον μοντέλο λειτουργικότητας των μηχανών αναζήτησης αδυνατεί να πραγματοποιήσει το συντονισμό αυτό.

ΘΕΜΑΤΑ ΔΙΠΛΩΜΑΤΙΚΩΝ ΕΡΓΑΣΙΩΝ

Εργαστήριο Συστημάτων Βάσεων Γνώσεων & Δεδομένων

Ο ΣΤΟΧΟΣ

Φιλοδοξούμε να αναπτύξουμε **προηγμένες τεχνικές αναζήτησης στον Ιστό**. Οι τεχνικές αυτές θα βασίζονται σε μοντέλο ερώτησης που θα χρησιμοποιεί αφαιρετικές περιγραφές και συσχετίσεις των εννοιών προς αναζήτηση. Το μοντέλο αυτό έχει κοινά σημεία με αυτό των εργαλείων διαχείρισης σκέψης. Σκοπός είναι η ενσωμάτωση της λειτουργικότητας τέτοιων εργαλείων στις μηχανές αναζήτησης ώστε να υποστηριχθούν πιο σύνθετες ανάγκες αναζήτησης πληροφορίας.

Οι δύο βασικοί στόχοι μας είναι οι εξής:

- Επέκταση του απλού μοντέλου αναζήτησης με χρήση λέξεων κλειδιών, ενσωματώνοντας ιδέες μοντέλων διαχείρισης σκέψης. Βασικός τρόπος διατύπωσης ερώτησης σχετιζόμενες με έννοιες και ιδέες θα παραμένει η χρήση λέξεων κλειδιών.
- Κατασκευή ενός συνόλου υπηρεσιών σε ένα ενδιάμεσο επίπεδο μεταξύ χρήστη και μηχανών αναζήτησης που θα παρέχουν προηγμένες λειτουργίες αναζήτησης.

Συγκεκριμένα, οι υπηρεσίες αυτές:

1. θα **συντονίζουν την αναζήτηση** σε πολλές μηχανές αναζήτησης και άλλες πηγές πληροφορίας στο Web ανάλογα με την περίπτωση. Το σύστημα θα επικεντρωθεί στην αναζήτηση κειμένων επιστημονικών συνεδρίων, περιοδικών, άρθρων εφημερίδων, κειμένων από blogs και εικόνων.
2. θα παρεμβαίνουν στην **ταξινόμηση των αποτελεσμάτων (ranking)** για να βελτιώσουν την ποιότητα των παρεχόμενων αποτελεσμάτων των μηχανών αναζήτησης. Αυτό μπορεί να γίνει
    a. κοιτώντας το **ιστορικό του χρήστη** (**clickstream data**[3]), αλλά και
    b. λαμβάνοντας υπ' όψη την αφαιρετική περιγραφή του πεδίου γνώσης για αναζήτηση.
3. θα παρεμβαίνουν στην παρουσίαση των αποτελεσμάτων, εφαρμόζοντας δυναμικά **τεχνικές συσταδοποίησης** – **clustering** – με βάση διάφορες ιδιότητες, π.χ. θεματικές περιοχές (π.χ. clustering ανά θέμα, τιμή πεδίου κ.λ.π.).

Στα πλαίσια αυτά μπορούν να ξεκινήσουν άμεσα μια σειρά από διπλωματικές εργασίες. Ενδεικτικά θέματα με τα οποία θα ασχοληθούν οι διπλωματικές αυτές είναι τα εξής:

1. Αναζήτηση με χρήση εννοιών (concept-based search).
2. Τεχνικές μετα-αναζήτησης (metasearch web retrieval).
3. Δυναμική θεματική συσταδοποίηση (clustering) αποτελεσμάτων αναζήτησης.
4. Μηχανές αναζήτησης ειδικευμένες για blogs (blog search engines).
5. Τεχνικές ταξινόμησης αποτελεσμάτων μηχανών αναζήτησης με βάση το ιστορικό αναζητήσεων (clickstream data).

Για πληροφορίες: *dalamag@dblab.ece.ntua.gr*

3http://csdl2.computer.org/persagen/DLAbsToc.jsp?resourcePath=/dl/mags/co/&toc=comp/mags/co/2007/08/r8toc.xml&DOI=10.1109/MC.2007.289

ΘΕΜΑΤΑ ΔΙΠΛΩΜΑΤΙΚΩΝ ΕΡΓΑΣΙΩΝ

Εργαστήριο Συστημάτων Βάσεων Γνώσεων & Δεδομένων

ΤΕΧΝΙΚΕΣ ΤΑΞΙΝΟΜΗΣΗΣ ΑΠΟΤΕΛΕΣΜΑΤΩΝ ΜΗΧΑΝΩΝ ΑΝΑΖΗΤΗΣΗΣ ΜΕ ΒΑΣΗ ΤΗΝ ΙΣΤΟΡΙΑ ΤΟΥ ΧΡΗΣΤΗ

ΠΛΗΡΟΦΟΡΙΕΣ: Θοδωρής Δαλαμάγκας, 210 772 1402, dalamag@dblab.ntua.gr

ΠΕΡΙΛΗΨΗ: Η διπλωματική εργασία στοχεύει στο σχεδιασμό και στην υλοποίηση τεχνικών που θα βελτιώσουν τον τρόπο ταξινόμησης των αποτελεσμάτων μιας μηχανής αναζήτησης.

ΑΤΟΜΑ: 1

ΠΛΑΤΦΟΡΜΑ ΕΡΓΑΣΙΑΣ: Συζητήσιμη

ΣΥΝΤΟΜΗ ΠΕΡΙΓΡΑΦΗ: Οι μηχανές αναζήτησης είναι το βασικό εργαλείο αναζήτησης πληροφορίας στο Web. Σε μια τέτοια μηχανή, όπως το Google, ο χρήστης πληκτρολογεί ένα σύνολο από λέξεις-κλειδιά (keywords) που θεωρεί ότι περιγράφουν καλύτερα το θέμα προς αναζήτηση. Στη συνέχεια, η μηχανή επιστρέφει ιστοσελίδες με περιεχόμενο σχετικό ως προς αυτό το θέμα. Μάλιστα, οι ιστοσελίδες ταξινομούνται (relevance ranking) ως προς το βαθμό ομοιότητάς (similarity value) τους με τις λέξεις-κλειδιά. Η προσθήκη μηχανισμών που εκμεταλλεύονται την ύπαρξη συνδέσμων μεταξύ κειμένων στον Ιστό για να αναπροσαρμόσουν την ταξινόμηση των αποτελεσμάτων (το γνωστό PageRank[4] του Google) έχει βελτιώσει σημαντικά την ποιότητα των αποτελεσμάτων της αναζήτησης.

Η διπλωματική αυτή θα αναπτύξει μεθόδους που θα τροποποιούν την ταξινόμηση που κατ' αρχήν προσφέρει μια μηχανή αναζήτησης (Google). Η τροποποίηση θα γίνεται εξάγοντας συμπεράσματα από το ιστορικό χρήσης της μηχανής από τον χρήστη. Έτσι τα αποτελέσματα θα προσαρμόζονται κάθε φορά στις συνήθειες και το υπόβαθρο του χρήστη. Για παράδειγμα, ένας χρήστης από υπολογιστή του Εργαστηρίου Βάσεων Δεδομένων, όταν πληκτρολογεί τη λέξη *delos*, εννοεί κατά πάσα πιθανότητα το ομώνυμο ερευνητικό ευρωπαϊκό πρόγραμμα στο οποίο συμμετέχει το εργαστήριο και όχι το ιερό νησί των Κυκλάδων. Γνωρίζοντας το domain dblab.ece.ntua.gr από το οποίο ήρθε το ερώτημα μπορούμε με κατάλληλο τρόπο να πριμοδοτήσουμε τα αποτελέσματα που αφορούν το πρόγραμμα αυτό.

Το ιστορικό θα περιλαμβάνει μόνο δεδομένα τύπου **clickstream**[5], και δεν θα απαιτείται ο χρήστης να έχει δηλώσει a-priori κάποιες προτιμήσεις ή να έχει ορίσει κάποιο profile. Ως τέτοια δεδομένα θεωρούμε για παράδειγμα τα παρακάτω: α) προηγούμενα clicks προς σελίδες αποτελεσμάτων, β) url διευθύνσεις, γ) συχνές αλληλουχίες από clicks προς σελίδες αποτελεσμάτων, δ) χρονικά διαστήματα μεταξύ click, κ.λ.π.

Η διπλωματική αυτή θα ολοκληρωθεί σε 3 φάσεις: α) ανάλυση απαιτήσεων συστήματος, β) σχεδίαση συστήματος και γ) υλοποίηση συστήματος.

4http://www.webworkshop.net/pagerank.html, http://infolab.stanford.edu/~backrub/google.html

5http://csdl2.computer.org/persagen/DLAbsToc.jsp?resourcePath=/dl/mags/co/&toc=comp/mags/co/2007/08/r8toc.xml&DOI=10.1109/MC.2007.289

ΘΕΜΑΤΑ ΔΙΠΛΩΜΑΤΙΚΩΝ ΕΡΓΑΣΙΩΝ

Εργαστήριο Συστημάτων Βάσεων Γνώσεων & Δεδομένων

P-Miner+: ΕΞΑΤΟΜΙΚΕΥΣΗ ΘΕΜΑΤΙΚΩΝ ΚΑΤΑΛΟΓΩΝ ΜΕ ΥΠΟΣΤΗΡΙΞΗ ΔΙΑΔΙΚΑΣΙΩΝ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ ΧΡΗΣΗΣ ΚΑΙ ΑΝΑΛΥΣΗΣ ΠΡΟΦΙΛ ΧΡΗΣΤΩΝ

ΠΛΗΡΟΦΟΡΙΕΣ: Π. Μπούρος, 210 772 1402, pbour@dblab.ntua.gr

ΠΕΡΙΛΗΨΗ: Η διπλωματική εργασία στοχεύει στην ανάπτυξη του πρότυπου συστήματος εξατομίκευσης θεματικών καταλόγου *P*-miner+ για συγκεκριμένο πεδίο γνώσης, υποστηριζόμενη από διαδικασίες εξόρυξης δεδομένων χρήσης του καταλόγου.

ΑΤΟΜΑ: 1-2

ΠΛΑΤΦΟΡΜΑ ΕΡΓΑΣΙΑΣ: Java/Apache/PHP/MySQL/OWL

ΣΥΝΤΟΜΗ ΠΕΡΙΓΡΑΦΗ: Τα **Topic Directories Web Sites** είναι **θεματικοί κατάλογοι** στον Ιστό που παρέχουν πληροφορίες σχετικές με μία θεματολογία. Παραδείγματα αποτελούν οι κατάλογοι για μουσεία, μουσική, πώληση Η/Υ, βιβλίων κ.λ.π.. Χαρακτηριστικό των καταλόγων αυτών είναι ότι η πληροφορία που παρέχουν είναι οργανωμένη με μορφή **γράφου** με σχέσεις κατηγορίας/υποκατηγορίας.

Στο εργαστήριο έχει ήδη αναπτυχθεί ένας θεματικός κατάλογος με βάση το περιεχόμενο του www.dmoz.org καθώς και το σύστημα P-miner για την πλήρη διαχείρισή του. Το σύστημα μελετά, με δυνατότητες εξόρυξης γνώσης, τις **πλοηγήσεις χρηστών** στις θεματικές κατηγορίες. Εντοπίζει **δημοφιλείς ή προβληματικές περιοχές** του καταλόγου και προτείνει στο διαχειριστή βελτιώσεις με τη μορφή **συντομεύσεων**. Επίσης δίνεται η δυνατότητα στο διαχειριστή να ομαδοποιεί τους χρήστες ανάλογα με την **ομοιότητα των πλοηγήσεών τους** εφαρμόζοντας τεχνικές συσταδοποίησης (clustering), και να εξατομικεύει τον κατάλογο ανάλογα με τις ομάδες που ανήκει κάθε χρήστης,

Η διπλωματική θα επεκτείνει το υπάρχον σύστημα P-miner, προσφέροντας ισχυρότερο μηχανισμό εξατομίκευσης του καταλόγου πέραν των προτάσεων δημιουργίας συντομεύσεων που προσφέρει το παρόν σύστημα. Θα μελετηθεί η ενσωμάτωση προφίλ χρηστών που θα χρησιμοποιείται από το μηχανισμό πρότασης κατηγοριών/σελίδων που θα επισκέπτεται ο χρήστης. Επίσης θα μελετηθούν σημασιολογικά εμπλουτισμένοι θεματικοί κατάλογοι. Συγκεκριμένα θα διερευνηθεί πως η σημασιολογία των σχέσεων π.χ. IS_A και PART_OF, μπορεί να εμπλουτίσει τις διαδικασίες εξόρυξης δεδομένων χρήσης, αλλά και τις διαδικασίες εξατομίκευσης. Η διπλωματική αυτή επομένως, θα ασχοληθεί με τα παρακάτω θέματα:

1. Μελέτη και εξοικείωση με το σύστημα P-miner που έχει αναπτυχθεί στον εργαστήριο.
2. Ανάλυση απαιτήσεων εξατομίκευσης πυλών καταλόγων.
3. Μελέτη σημασιολογικά εμπλουτισμένων θεματικών καταλόγων
4. Σχεδίαση συστήματος *P*-miner+ για την εξατομίκευση θεματικών καταλόγων με υποστήριξη διαδικασιών εξόρυξης γνώσης από πλοηγήσεις χρηστών και επισκέψεις σελίδων.
5. Υλοποίηση συστήματος *P*-miner+.

Για περισσότερες πληροφορίες δείτε τις αναφορές http://www.dblab.ece.ntua.gr/pubs/uploads/DIPL-2006-10.pdf και http://www.dblab.ece.ntua.gr/pubs/uploads/TR-2007-11.pdf και τη διεύθυνση http://casablanca.dblab.ece.ntua.gr/p-miner/.

ΘΕΜΑΤΑ ΔΙΠΛΩΜΑΤΙΚΩΝ ΕΡΓΑΣΙΩΝ

Εργαστήριο Συστημάτων Βάσεων Γνώσεων & Δεδομένων

V: ΔΙΑΔΙΚΤΥΑΚΗ ΥΠΗΡΕΣΙΑ ΙΣΤΟΥ ΣΥΝΑΛΛΑΓΗΣ ΑΝΤΙΚΕΙΜΕΝΩΝ ΚΑΙ ΥΠΗΡΕΣΙΩΝ

ΠΛΗΡΟΦΟΡΙΕΣ: Δ. Σαχαρίδης, 210 772 1402, dsachar@dblab.ntua.gr, Β. Καντερέ, 210 772 1402, verena@dblab.ntua.gr, Θ. Δαλαμάγκας, 210 772 1402, dalamag@dblab.ntua.gr

ΠΕΡΙΛΗΨΗ: Η διπλωματική εργασία θα δημιουργήσει μια υπηρεσία Παγκόσμιου Ιστού που θα για συναλλαγή αγαθών μεταξύ των μελών της.

ΑΤΟΜΑ: 1-2

ΠΛΑΤΦΟΡΜΑ ΕΡΓΑΣΙΑΣ: Ruby on Rails

ΣΥΝΤΟΜΗ ΠΕΡΙΓΡΑΦΗ: Στόχος της διπλωματικής εργασίας είναι η δημιουργία ενός συστήματος συναλλαγής αγαθών στον Ιστό. Η οικονομία της ηλεκτρονικής αυτής *κοινωνίας* (κοινωνία **V**) είναι *μη νομισματική*, δηλαδή, δεν επιτρέπει χρηματικές δοσοληψίες, αλλά βασίζεται ουσιαστικά στην *υπόληψη* των μελών της. Όπως και σε μία αληθινή, τα μέλη της ηλεκτρονικής κοινωνίας σχηματίζουν κοινότητες με βάση κάποια κοινά ενδιαφέροντα ή επιδιώξεις, π.χ., φίλοι κατασκευής γηπέδου καλλιτεχνικού πατινάζ, είτε για διάφορους άλλους λόγους, όπως γεωγραφικούς.

- ✓ Μια συναλλαγή μπορεί να είναι: (α) *ανταλλαγή (exchange/swap)*, (β) *δωρεά (gift)*, (γ) *κοινοχρησία (sharing)* αγαθών.
- ✓ Τα αγαθά μπορεί να είναι: (α) *αντικείμενα* , (β) *υπηρεσίες*.

Ως παράδειγμα συναλλαγής αντικειμένου μπορούμε να θεωρήσουμε έναν χρήστη που διαθέτει ένα φορητό υπολογιστή και θέλει να το (i) ανταλλάξει με τηλεόραση πλάσματος, ή (ii) να το δωρίσει σε κάποιο μέλος ή κοινότητα, είτε (iii) να το μοιράζεται με μέλη σε κάποια κοινότητα που ανήκει, όπως, π.χ., με τους συναδέλφους του. Επίσης, μπορούν να πραγματοποιηθούν συναλλαγές υπηρεσιών: για παράδειγμα, ένας φοιτητής ικανός με τους υπολογιστές επιθυμεί να προσφέρει την τεχνογνωσία του με αντάλλαγμα φροντιστήριο σε κάποιο μάθημα από συμφοιτητή του. Τέλος συναλλαγές μπορούν να πραγματοποιηθούν μεταξύ υπηρεσιών και αντικειμένων: για παράδειγμα, κάποιος διαθέτει δύο εισιτήρια για μία παράσταση αλλά δεν έχει αυτοκίνητο, οπότε ανταλλάσει το ένα με την μεταφορά του από και προς το συναυλιακό χώρο.

Κύριο συστατικό της κοινωνίας **V** είναι η *υπόληψη* των μελών της, εφόσον δεν επιτρέπονται οι χρηματικές δοσοληψίες. Η υπόληψη ενός μέλους είναι μια συνολική βαθμολογία που αντικατοπτρίζει πόσο καλός πολίτης της κοινωνίας **V** είναι. Μπορεί να προκύψει από τη συνάθροιση βαθμολογιών των συναλλαγών του μέλους.

Έχουν ήδη πραγματοποιηθεί τα εξής: (α) Ανάλυση απαιτήσεων συστήματος. (β) Σχεδίαση Βάσης Δεδομένων/Συστήματος, (γ) Επιμέρους υλοποιήσεις.
Η νέα διπλωματική θα επεκτείνει την υπάρχουσα υλοποίηση και θα παρουσιάσει ένα πλήρως λειτουργικό σύστημα συναλλαγής αντικειμένων και υπηρεσιών.

ΘΕΜΑΤΑ ΔΙΠΛΩΜΑΤΙΚΩΝ ΕΡΓΑΣΙΩΝ

Εργαστήριο Συστημάτων Βάσεων Γνώσεων & Δεδομένων

ΥΛΟΠΟΙΗΣΗ ΜΗΧΑΝΙΣΜΟΥ ΕΡΩΤΑΠΟΚΡΙΣΕΩΝ ΓΙΑ ΔΙΚΤΥΟ ΟΜΟΤΙΜΩΝ ΒΑΣΕΩΝ (PEER-2-PEER DATABASES)

ΠΛΗΡΟΦΟΡΙΕΣ: Βηρένα Καντερέ, 210 772 1402, vkante@dblab.ntua.gr

ΠΕΡΙΛΗΨΗ: Η διπλωματική εργασία στοχεύει στην υλοποίηση κόμβων ομότιμων βάσεων που μπορούν να συνδέονται μεταξύ τους με αντιστοιχίσεις (mappings) και να θέτουν ερωτήσεις (queries) που να προωθούνται από τη μία στην άλλη. Μέσω των ερωτήσεων που θέτουν οι βάσεις αποσκοπούν στο να βρουν βάσεις με κοινά ενδιαφέροντα και να συνδεθούν μαζί τους.

ΑΤΟΜΑ: 1

ΠΛΑΤΦΟΡΜΑ ΕΡΓΑΣΙΑΣ: JAVA

ΣΥΝΤΟΜΗ ΠΕΡΙΓΡΑΦΗ: Θεωρούμε ένα σύστημα ομοτίμων όπου κάθε ομότιμος κατέχει μια σχεσιακή βάση δεδομένων. Οι ομότιμοι που είναι απευθείας συνδεδεμένοι μεταξύ τους, έχουν ανά δύο αντιστοιχίσεις (mappings) πάνω σε μέρος των σχημάτων τους. Όταν ένας ομότιμος θέτει μια ερώτηση στο δίκτυο ομοτίμων, αυτή προωθείται σε κάποιο μονοπάτι κόμβων. Η ερώτηση πρέπει να μεταφράζεται σε κάθε κόμβο του μονοπατιού στο τοπικό σχήμα, έτσι ώστε να απαντηθεί από τον τοπικό κόμβο, αλλά και για να προωθηθεί παρακάτω. Έτσι όμως η ερώτηση χάνει σε κάθε κόμβο πληροφορία που δεν μπορεί να μεταφραστεί τοπικά, με αποτέλεσμα να μην μπορεί πολλές φορές να φτάσει μακριά. Για το λόγο αυτό, συχνά η ερώτηση δεν φτάνει σε κόμβους που μπορούν να την απαντήσουν ικανοποιητικά, ή όταν φτάνει σε αυτούς έχει χάσει σχετική πληροφορία.

Ο σκοπός της διπλωματικής αυτής είναι να υλοποιηθεί μια υπάρχουσα τεχνική για την αντιμετώπιση του παραπάνω προβλήματος. Συγκεκριμένα, η τεχνική προτείνει μια επαναληπτική μέθοδο μέσω της οποίας ένας κόμβος μαθαίνει σταδιακά για τα περιεχόμενα ενός απομακρυσμένου κόμβου. Κατά τη διαδικασία της μάθησης χτίζονται σταδιακά αντιστοιχίσεις μεταξύ των σχημάτων των απομακρυσμένων κόμβων. Δύο τέτοιοι κόμβοι μπορούν σε κάποιο σημείο να αποφασίσουν ότι έχουν κοινά ενδιαφέροντα – δηλαδή έχουν παρόμοια σχήματα, και να συνδεθούν απευθείας χρησιμοποιώντας τις αντιστοιχίσεις που χτίσανε κατά τη γνωριμία τους.

ΥΛΟΠΟΙΗΣΗ ΣΥΣΤΗΜΑΤΟΣ ΓΙΑ ΤΗΝ ΑΝΤΑΛΛΑΓΗ ΔΕΔΟΜΕΝΩΝ ΣΕ ΔΙΚΤΥΑ ΟΜΟΤΙΜΩΝ ΜΕ ΧΡΗΣΗ ΟΝΤΟΛΟΓΙΩΝ

ΠΛΗΡΟΦΟΡΙΕΣ: Δ. Σκούτας, 210 772 1436, dskoutas@dblab.ntua.gr , Β. Καντερέ, 210 772 1402, verena@dblab.ntua.gr

ΠΕΡΙΛΗΨΗ: Η διπλωματική εργασία στοχεύει στην υλοποίηση και πειραματική αξιολόγηση ενός μηχανισμού που θα βασίζεται στη χρήση οντολογιών για την αποδοτικότερη ανταλλαγή δεδομένων σε δίκτυα ομοτίμων.

ΑΤΟΜΑ: 1

ΠΛΑΤΦΟΡΜΑ ΕΡΓΑΣΙΑΣ: Jena [6], Pellet [7]

ΣΥΝΤΟΜΗ ΠΕΡΙΓΡΑΦΗ: Ο Σημασιολογικός Ιστός αποτελεί μια εξελισσόμενη προσπάθεια για τον εμπλουτισμό του περιεχομένου του Ιστού με σημασιολογική πληροφορία, προκειμένου να καταστεί δυνατή η χρήση και επεξεργασία του από πράκτορες λογισμικού, διευκολύνοντας εργασίες όπως η αναζήτηση και η ολοκλήρωση πληροφορίας. Κεντρικό ρόλο στην προσπάθεια αυτή έχουν οι οντολογίες. Μία οντολογία αποτελεί ένα τυπικό μοντέλο αναπαράστασης των εννοιών ενός πεδίου ενδιαφέροντος, των ιδιοτήτων αυτών των εννοιών και των μεταξύ τους σχέσεων. Η OWL [8] έχει προταθεί από το W3C ως γλώσσα για την αναπαράσταση οντολογιών στον Σημασιολογικό Ιστό. Επίσης υπάρχουν εργαλεία που παρέχουν δυνατότητες διαχείρισης (π.χ. Jena) και συλλογιστικής (π.χ. Pellet) σε OWL οντολογίες.
Από την άλλη πλευρά, τα δίκτυα ομοτίμων έχουν γίνει ιδιαίτερα δημοφιλή τα τελευταία χρόνια ως πλατφόρμες για την ανταλλαγή αρχείων και δεδομένων σε κατανεμημένα συστήματα μεγάλης κλίμακας. Η διπλωματική αυτή θα ασχοληθεί με αδόμητα δίκτυα ομοτίμων, όπου κάθε ομότιμος διαθέτει μία τοπική σχεσιακή βάση δεδομένων και αντιστοιχήσεις μεταξύ του σχήματος της τοπικής του βάσης και των σχημάτων των ομοτίμων με τους οποίους είναι απευθείας συνδεδεμένος. Η ανταλλαγή δεδομένων γίνεται μέσω SQL ερωτημάτων, τα οποία, καθώς διαδίδονται από έναν ομότιμο σε άλλο, μεταγράφονται σύμφωνα με αυτές τις αντιστοιχήσεις.
Για την αύξηση της αποδοτικότητας της ανταλλαγής δεδομένων μέσω αυτής της διαδικασίας μεταγραφής ερωτημάτων, θεωρούμε την ύπαρξη μιας οντολογίας που περιγράφει το πεδίο ενδιαφέροντος για το συγκεκριμένο δίκτυο ομοτίμων. Οι ομότιμοι χρησιμοποιούν την οντολογία για να περιγράψουν σημασιολογικά τα σχήματα των τοπικών τους βάσεων, μέσω του καθορισμού αντιστοιχήσεων μεταξύ των όρων των σχημάτων και των όρων της οντολογίας. Αυτή η επιπρόσθετη σημασιολογική πληροφορία μπορεί να διευκολύνει τον προσδιορισμό του βαθμού ομοιότητας μεταξύ ομοτίμων, καθώς και της ομοιότητας των ερωτημάτων που προκύπτουν από τη διαδικασία μεταγραφής σε σχέση με τα αρχικά.
Σκοπός της διπλωματικής θα είναι η υλοποίηση ενός μηχανισμού που θα αξιοποιεί τα παραπάνω προκειμένου να καταστήσει αποδοτικότερη την ανταλλαγή δεδομένων μεταξύ των ομοτίμων.

[6] http://jena.sourceforge.net/
[7] http://pellet.owldl.com/
[8] Web Ontology Language - http://www.w3.org/TR/owl-features/

ΥΛΟΠΟΙΗΣΗ ΣΥΣΤΗΜΑΤΟΣ ΑΝΑΚΑΛΥΨΗΣ ΚΑΙ ΚΑΤΑΤΑΞΗΣ ΥΠΗΡΕΣΙΩΝ ΤΟΥ ΣΗΜΑΣΙΟΛΟΓΙΚΟΥ ΙΣΤΟΥ

ΠΛΗΡΟΦΟΡΙΕΣ: Δ. Σκούτας, 210 772 1436, dskoutas@dblab.ntua.gr

ΠΕΡΙΛΗΨΗ: Η διπλωματική εργασία στοχεύει στην υλοποίηση και πειραματική αξιολόγηση ενός συστήματος για την ανακάλυψη και κατάταξη υπηρεσιών του Σημασιολογικού Ιστού, σύμφωνα με αιτήματα χρηστών, τα οποία επίσης περιγράφονται σημασιολογικά μέσω κοινής οντολογίας.

ΑΤΟΜΑ: 1

ΠΛΑΤΦΟΡΜΑ ΕΡΓΑΣΙΑΣ: Jena [9], Pellet [10]

ΣΥΝΤΟΜΗ ΠΕΡΙΓΡΑΦΗ: Ο Σημασιολογικός Ιστός αποτελεί μια εξελισσόμενη προσπάθεια για τον εμπλουτισμό του περιεχομένου του Ιστού με σημασιολογική πληροφορία, προκειμένου να καταστεί δυνατή η χρήση και επεξεργασία του από πράκτορες λογισμικού, διευκολύνοντας εργασίες όπως η αναζήτηση και η ολοκλήρωση πληροφορίας. Κεντρικό ρόλο στην προσπάθεια αυτή έχουν οι οντολογίες. Μία οντολογία αποτελεί ένα τυπικό μοντέλο αναπαράστασης των εννοιών ενός πεδίου ενδιαφέροντος, καθώς επίσης των ιδιοτήτων αυτών των εννοιών και των μεταξύ τους σχέσεων. Η OWL [11] (Web Ontology Language) έχει προταθεί από το W3C ως γλώσσα για την αναπαράσταση οντολογιών στον Σημασιολογικό Ιστό. Επίσης υπάρχουν εργαλεία που παρέχουν δυνατότητες διαχείρισης και συλλογιστικής σε OWL οντολογίες (π.χ. Jena και Pellet, αντίστοιχα).

Η χρήση οντολογιών μπορεί να επεκταθεί για τη σημασιολογική περιγραφή υπηρεσιών του Ιστού. Στην περίπτωση αυτή, οι διάφορες παράμετροι της υπηρεσίας, όπως παράμετροι εισόδου και εξόδου, περιγράφονται σημασιολογικά μέσω της αντιστοίχησής τους σε έννοιες που ορίζονται στην οντολογία. Η πληροφορία αυτή μπορεί να χρησιμοποιηθεί για να καταστήσει τη διαδικασία αναζήτησης και επιλογής υπηρεσιών πιο αποδοτική και αποτελεσματική.

Σκοπός αυτής της διπλωματικής θα είναι η υλοποίηση και πειραματική αξιολόγηση ενός μηχανισμού για την κατάταξη υπηρεσιών σύμφωνα με αιτήματα χρηστών. Θεωρούμε ότι τόσο οι διαθέσιμες υπηρεσίες όσο και τα αιτήματα των χρηστών περιγράφονται σημασιολογικά μέσω μιας κοινής οντολογίας. Η κατάταξη των υπηρεσιών θα γίνεται με κριτήριο τη σχέση των αντίστοιχων εννοιών σε αυτή την οντολογία.

[9] http://jena.sourceforge.net/

[10] http://pellet.owldl.com/

[11] http://www.w3.org/TR/owl-features/

O2Omap: ΕΡΓΑΛΕΙΟ ΟΡΙΣΜΟΥ ΑΝΤΙΣΤΟΙΧΙΣΕΩΝ(MAPPINGS) ΜΕΤΑΞΥ ΟΝΤΟΛΟΓΙΩΝ

ΠΛΗΡΟΦΟΡΙΕΣ: Α. Δημητρίου, 210 7723415, angela@dblab.ntua.gr

ΠΕΡΙΛΗΨΗ: Η διπλωματική εργασία στοχεύει στην υλοποίηση ενός εργαλείου, μέσω του οποίου ο χρήστης θα μπορεί να ορίζει αντιστοιχίσεις (mappings) μεταξύ δύο οντολογιών και θα τις χρησιμοποιεί για μετάφραση ερωτημάτων (query reformulation). Το εργαλείο αυτό θα έχει τη δυνατότητα να ενσωματωθεί στο σύστημα διαχείρισης οντολογιών ***Protégé***, ως plugin.

ΑΤΟΜΑ: 1-2

ΠΛΑΤΦΟΡΜΑ ΕΡΓΑΣΙΑΣ: Java

ΣΥΝΤΟΜΗ ΠΕΡΙΓΡΑΦΗ:

Ο όγκος της πληροφορίας, που διατίθεται στον παγκόσμιο ιστό είναι τερράστιος. Η χρήση οντολογιών στοχεύει στην οργάνωση του μεγάλου αυτού συνόλου δεδομένων, έτσι ώστε να δίνεται η δυνατότητα στον καθένα να αναζητά οτιδήποτε τον ενδιαφέρει, με τρόπους πιο έξυπνους από την απλή αναζήτηση με χρήση λέξεων κλειδιών. Οι οντολογίες μπορούν, εξάλλου, να διευκολύνουν την επικοινωνία εσωτερικά σε ένα δίκτυο ομότιμων κόμβων (δηλ. p2p συστημάτων), ώστε να επιτυγχάνεται αποδοτικότερα ο διαμοιρασμός (ημι)δομημένης πληροφορίας μεταξύ τους. Στην περίπτωση που η αποθηκευμένη πληροφορία στο δίκτυο έχει διαφορετική δομή από κόμβο σε κόμβο, είναι αναγκαία η μετάφραση των ερωτημάτων που απευθύνονται στον αρχικό κόμβο, πριν αυτά ταξιδέψουν προς κάποιο γειτονικό του. Ρωτώντας, π.χ. για «αυτοκίνητα» σε έναν κόμβο, να επιστρέφουν απαντήσεις και από κάποιον απομακρυσμένο, ο οποίος έχει καταχωρημένα «επιβατικά οχήματα».

Μια οντολογία αποτελείται από δύο μέρη: το σχήμα και τα δεδομένα της. Το σχήμα της συνίσταται σε έννοιες, οι οποίες υπακούουν σε κάποια ιεραρχία μεταξύ τους, και σε συσχετίσεις(ρόλους) ανάμεσα σε αυτές. Τα δεδομένα αποτελούν τα στιγμιότυπα αυτών των δομικών στοιχείων. Οι πιο συνήθεις γλώσσες ορισμού οντολογιών είναι οι RDF(S) (http://www.w3.org/RDF/) και OWL (http://www.w3.org/2004/OWL/) .

Το Protégé (http://protege.stanford.edu/) είναι μία βιβλιοθήκη ανοιχτού κώδικα, σε γλώσσα προγραμματισμού Java. Είναι κατάλληλο για κατασκευή και διαχείριση οντολογιών, ενώ οι δυνατότητες αυτές δίνονται στο χρήστη μέσω γραφικών εργαλείων. Η κοινωνία ανάπτυξης του Protégé είναι ενεργή και συμπληρώνει συνεχώς την κύρια βιβλιοθήκη με επιπλέον εργαλεία. Η δημιουργία ενός ανάλογου θα είναι και το αποτέλεσμα της παρούσας διπλωματικής.

Η εργασία στοχεύει στη σχεδίαση και στην υλοποίηση ενός plugin του Protégé, που θα δίνει τη δυνατότητα ορισμού αντιστοιχίσεων μεταξύ δύο οντολογιών. Σκοπός είναι να χρησιμοποιηθούν οι αντιστοιχίσεις αυτές στη μετάφραση ερωτημάτων, που προορίζονται για μία οντολογία, σε κατάλληλη μορφή που μπορεί να δεχτεί απαντήσεις από μια δεύτερη. Στο παραπάνω παράδειγμα δηλ., θα πρέπει να ορίσει κανείς την αντιστοίχιση O_1:«αυτοκίνητο» ↔ O_2:«επιβατικό όχημα», ώστε να επιστρέψουν πληροφορίες και από τις δύο πηγές (δηλ. τις οντολογίες O_1 και O_2).Οι αντιστοιχίσεις αυτές θα είναι τριών κατηγοριών:

(α) Απλές 1-1 αντιστοιχίσεις: Ο χρήστης θα μπορεί να αντιστοιχίσει δύο έννοιες ή δύο ρόλους, που ανήκουν σε δύο διαφορετικές οντολογίες, μεταξύ τους.
(β) Σύνθετες αντιστοιχίσεις: Ο χρήστης θα μπορεί να ορίσει αντιστοιχίσεις ανάμεσα σε πολύπλοκες φόρμουλες, αποτελούμενες από πολλαπλές έννοιες ή ρόλους. Για παράδειγμα, θα είναι δυνατός ο ορισμός O_1: «διθέσιο» ∩ «ανοιχτό» ↔ O_2 «σπορ» ∩ «κάμπριο».
(γ) Αντιστοιχίσεις με χρήση συναρτήσεων: Πολλές φορές για έννοιες που αποτελούν χαρακτηριστικά (attributes), και όχι κατηγορία από στιγμιότυπα, χρειάζεται η μετατροπή από μία μορφή σε άλλη. Π.χ. O_1: calculateAge(«έτος κατασκευής») → O_2 : «ηλικία». Σε αυτήν την περίπτωση απαραίτητη είναι και η αντίστροφη αντιστοίχιση, δηλ. O_1: «έτος κατασκευής» ← O_2: findYear(«ηλικία»).

Για την αξιοποίηση της λειτουργικότητας που θα παρέχει το O2Omap θα δημιουργηθεί το κατάλληλο περιβάλλον θέσης και απάντησης ερωτημάτων. Ο χρήστης, μέσω γραφικής διαπροσωπείας (GUI), θα ορίζει ερωτήματα με βάση μια οντολογία της επιλογής του. Στη συνέχεια, ένας αλγόριθμος μετάφρασης, ο οποίος θα συμβουλεύεται τις προϋπάρχουσες αντιστοιχίσεις (mappings) θα μεταφράζει το ερώτημα σε μορφή που θα μπορεί να απαντηθεί από μια δεύτερη οντολογία.

Η διπλωματική αυτή επομένως, θα ασχοληθεί με τα παρακάτω θέματα:

6. Μελέτη οντολογιών και των γλωσσών OWL και RDF(S)
7. Γνωριμία με Protégé API και μελέτη του τρόπου υλοποίησης plugin σε αυτό (http://protege.stanford.edu/doc/dev.html)
8. Σχεδιασμός εργαλείου ορισμού αντιστοιχίσεων των τριών παραπάνω κατηγοριών.
9. Ειδικά για την κατηγορία (γ) θα δημιουργηθεί μια προκαθορισμένη λίστα συναρτήσεων πρόχειρων στο χρήστη (συναρτήσεις χειρισμού ημερομηνιών, αριθμών, διαχείρισης ακολουθιών χαρακτήρων, μετατροπής μονάδων μέτρησης και νομισμάτων κ.α.). Επιπλέον, θα δίνεται η δυνατότητα στο χρήστη να καταχωρήσει τη δική του συνάρτηση μέσω μιας εκτελέσιμης βιβλιοθήκης, η οποία θα συνοδεύεται από κατάλληλη περιγραφή (σύντομη περιγραφή, ορίσματα, αποτελέσματα).
10. Σχεδιασμός εφαρμογής ορισμού και μετάφρασης ερωτημάτων αλλά και λήψης απαντήσεων.
11. Υλοποίηση O2Omap