Θέματα Πτυχιακών / Διπλωματικών Εργασιών

Υπεύθυνος Καθηγητής: Δημήτριος Γουνόπουλος

Θέμα 1

HTML 2 ER*: Μεταφορά HTML σελίδων στο μοντέλο Ο/Σ*

(From DBs to the Web and back again)

Πάρα πολλά site τα οποία υπάρχουν στο Web αυτή τη στιγμή έχουν δυναμικό περιεχόμενο το οποίο είναι αποθηκευμένο σε βάσεις δεδομένων, σύμφωνα με το γνωστό μοντέλο Ο/Σ (ή παραλλαγές – επεκτάσεις του). Οι βάσεις έχουν ένα συγκεκριμένο σχήμα με το οποίο έχουν μοντελοποιήσει τα δεδομένα τους και οι πληροφορίες που υπάρχουν εκεί ακολουθούν το σχήμα αυτό. Όταν ο χρήστης βλέπει μια δυναμική σελίδα, αυτή τότε γεμίζει με δεδομένα από τη βάση και ακολουθώντας κάποιο συγκεκριμένο template παράγει την τελική σελίδα που δίνεται στο χρήστη.

Ωστόσο, όταν θέλουμε να εξάγουμε δεδομένα από τις σελίδες αυτές ώστε να τις αποθηκεύσουμε σε μία δική μας Β.Δ., πρέπει να δούμε τι πληροφορίες μας ενδιαφέρουν από τη σελίδα και στη συνέχεια να τις αποτυπώσουμε σε ένα δικό μας σχήμα Ο/Σ. Κάθε φορά που εντοπίζουμε καινούρια πληροφορία, πρέπει να αλλάζουμε manually το σχήμα της δικής μας βάσης, ώστε να ανταποκρίνεται στη νέα πληροφορία. Ακόμα και χωρίς να αλλάξουμε το σχήμα της βάσης, η αναγνώριση του τρόπου που είναι δομημένο το site και των πεδίων / πληροφοριών που περιλαμβάνει είναι χρονοβόρα διαδικασία.

Σκοπός της πτυχιακής / διπλωματικής αυτής εργασίας είναι η ανάπτυξη αλγορίθμων και όλου του απαραίτητου μηχανισμού προκειμένου, δοθέντος ενός ιστοτόπου, να παράγει αυτόματα το μοντέλο Ο/Σ που πρέπει να έχει η δική μας βάση. Άρα το πρόβλημα ανάγεται σε εξαγωγή σχήματος από δεδομένα (schema extraction from instances), με τη διαφορά ότι είμαστε μεταξύ ημιδομημένων (HTML) και δομημένων δεδομένων (DBs). Αν και υπάρχει σχετική δουλειά σε XML, RDF και DB δεν φαίνεται να υπάρχει δουλειά που να βρίσκεται στο ενδιάμεσο 2 περιοχών: των ημιδομημένων και των δομημένων δεδομένων. Επίσης, μηχανισμοί όπως το XSLT μετατρέπουν τη δομημένη πληροφορία (π.χ. DB) ή ημιδομημένη πληροφορία (XML) σε ημιδομημένη πληροφορία (HTML). Αντιθέτως, εμείς ενδιαφερόμαστε για την αντίστροφη πορεία.

Επίσης, ερευνητική δουλειά (και αντίστοιχοι αλγόριθμοι) υπάρχουν για τη μεταφορά κλάσεων (π.χ. C++ classes) σε μοντέλο Ο/Σ. Ωστόσο, οι κλάσεις έχουν εγγενή τη μοντελοποίηση του σχήματος και άρα η μεταφορά τους σε Ο/Σ δεν είναι τόσο δύσκολη. Εξακολουθεί να είναι συναφές πεδίο όμως.

Ζητούμενα της διπλωματικής είναι: *i*) Μία σύντομη ανασκόπηση της βιβλιογραφίας αναφορικά με τις δουλειές που υπάρχουν για αυτόματη εξαγωγή σχήματος Ο/Σ ή άλλων ημι-δομημένων σχημάτων από δεδομένα (instances). *ii*) Υλοποίηση αλγορίθμων που βρίσκουν κοινά χαρακτηριστικά μεταξύ σελίδων ή ομάδων σελίδων, ώστε να μπορούν να ενταχθούν σε instances της ίδιας οντότητας. *iii*) Υλοποίηση αλγορίθμων για την μεταφορά των instances σε συγκεκριμένο σχήμα Ο/Σ που να ικανοποιεί τα instances.

Άτομα: 1 - 3 (αναλόγως την έκταση)
Συνιστώμενες Γνώσεις:

- Προγραμματισμός (οποιαδήποτε γλώσσα), για την ανάπτυξη του λογισμικού
- Μοντέλο Ο/Σ, ANSI-SQL, κανονικοποίηση

- HTML
- Αγγλικά (για τη βιβλιογραφία)

Επιπλέον πληροφορίες:

- XStruct: Efficient Schema Extraction from Multiple and Large XML Documents - http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.60.9295
- Extracting Entity-Relationship Schemas from Relational Databases: A Form-Driven Approach - http://www.fing.edu.uy/inco/cursos/tagsi/DBRE_FormDriven_81620184.pdf
- Schema Extraction from XML Collections - http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.4.5668
- Schema Extraction from XML Data: A Grammatical Inference Approach - http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.2.4760
- Link 2 SQL - http://msdn.microsoft.com/en-us/library/bb386976.aspx
- ADO .NET **Entity Framework** - http://en.wikipedia.org/wiki/ADO.NET_Entity_Framework

Θέμα 2

***MailCalendar**: Εξαγωγή στοιχείων συνάντησης (calendar) από κείμενο σε email*

Ένας κλασσικός πλέον τρόπος επικοινωνίας είναι το email (ηλεκτρονικό ταχυδρομείο). Ωστόσο, η χρησιμότητα του email δεν είναι περιορισμένη στην μεταφορά νέων όπως είναι το συμβατικό ταχυδρομείο, κυρίως λόγω του μηδενικού κόστους χρήσης του. Αντιθέτως, πολύ συχνά λειτουργεί ως τρόπος ανταλλαγής σύντομων μηνυμάτων (ανάλογο του SMS των κινητών) , links προς site, συνημμένου υλικού κ.λπ.

Μια επίσης σημαντική χρησιμότητα για άτομα τα οποία χρησιμοποιούν το email τους πολύ είναι το να κανονίζουν συναντήσεις με φίλους / γνωστούς / συνεργάτες κ.λπ. Έτσι, μέσα στο email συχνά περιλαμβάνεται χωρική και χρονική πληροφορία που αφορούν τον τόπο και το χρόνο συνάντησης.

Η πτυχιακή αυτή έχει ως σκοπό την ανάπτυξη μηχανισμού (ή μηχανισμών) / αλγορίθμου για την αναγνώριση χωρικής και χρονικής πληροφορίας που υπάρχει μέσα στα email και τη σύσταση για προσθήκη σε ένα ημερολόγιο / ατζέντα (calendar). Η πληροφορία μπορεί να είναι συγκεκριμένη, σε format που να δηλώνει ρητά την ημερομηνία π.χ. “7 / 03 / 2012, 16:00” ή σε αφηρημένο format που να χρειάζεται περισσότερη ανάλυση του context του μηνύματος, π.χ. “συνάντηση αύριο στο γραφείο μου τη γνωστή ώρα”.

Αν και το συγκεκριμένο format είναι το πιο εύκολο να βρεθεί, είναι και το λιγότερο χρησιμοποιούμενο αφού πολλές φορές τα email έχουν απλό, καθημερινό λόγο. Οπότε εμείς στοχεύουμε σε περιπτώσεις όπως η 2η ή τουλάχιστον σε περιπώσεις με πιο ελεύθερη έκφραση. Ο συσχετισμός και με άλλες πληροφορίες, όπως π.χ. Παλαιότερα events μπορούν να βοηθήσουν προς αυτή την κατεύθυνση.

Άτομα: 1
Συνιστώμενες Γνώσεις:

- Προγραμματισμός (οποιαδήποτε γλώσσα), για την ανάπτυξη του λογισμικού
- Αγγλικά (για τη βιβλιογραφία)

Επιπλέον πληροφορίες:

- http://email.about.com/od/gmailtips/qt/et_mailtoevent.htm
- http://googlesystem.blogspot.com/2007/10/create-google-calendar-events-from.html
- http://help.yahoo.com/tutorials/cal/cal/cal_addevent4.html

Θέμα 3

***GraphicCalendar**: Συγχώνευση calendars από φωτογραφικό υλικό*

Συχνά ο κόσμος οργανώνει το χρόνο του με χρήση κάποιας ατζέντας, είτε αυτή είναι online είτε όχι. Εκεί μπορεί να σημειώνει συναντήσεις, tasks για να κάνει, διακοπές, εξόδους κ.λπ. Συχνά, προκειμένου να συντονίζεται με άλλα άτομα ανταλάσσει πληροφορία που περιλαμβάνει τη διαθεσιμότητα του σε ημέρες και ώρες. Είναι, όμως, πρακτικά αδύνατο να γράψει πότε μπορεί και πότε όχι για κάθε πιθανή μέρα της εβδομάδας όταν θέλει να κανονίσει συναντήσεις με άλλα άτομα.

Επίσης, με τη χρήση των σύγχρονων τηλεφωνικών συσκευών, πολύ συχνά βγάζουμε φωτογραφίες, ακόμα και για σημειώσεις κ.λπ. Δεν είναι απίθανο λοιπόν να χρησιμοποιούμε τα κινητά για να φωτογραφίζουμε εκτυπωμένα calendars ή calendars από οθόνες υπολογιστών, προκειμένου να τα συγκρίνουμε με τη δική μας διαθεσιμότητα.

Η πτυχιακή αυτή έχει ως σκοπό την ανάπτυξη αλγορίθμων προκειμένου να βρίσκει και να προτείνει διαθεσιμότητες (δλδ, διαθέσιμες ημέρες και ώρες) βάσει του δικού μας προγράμματος το οποίο είναι σε μία μορφή (π.χ. Ηλεκτρονική) και ενός άλλου το οποίο έχουμε σε format φωτογραφίας (jpg, bmp κ.λπ.). Ακόμα μεγαλύτερο ενδιαφέρον παρουσιάζει το να προσπαθούμε να βρίσκουμε διαθεσιμότητες έχοντας ένα σύνολο από τέτοιες φωτογραφίες, τις οποίες προσπαθούμε να κάνουμε match. Σε αυτή την περίπτωση ακόμα και το δικό μας πρόγραμμα (calendar) μπορεί να είναι κάποια από τις φωτογραφίες. Ο τρόπος αυτός για την εύρεση διαθεσιμότητας είναι ανάλογος της εποπτικής διερεύνησης την οποία κάνουμε εμείς ως άνθρωποι όταν προσπαθούμε να λύσουμε το πρόβλημα αυτό.

Άτομα: 1
Συνιστώμενες Γνώσεις:

- Προγραμματισμός (οποιαδήποτε γλώσσα), για την ανάπτυξη του λογισμικού
- Χειρισμός εικόνων (βλ. Pixel), αναγνώριση προτύπων σε εικόνες

Θέμα 4

Έυρεση καταλληλότερου site για την επίλυση προβλημάτων

Έστω ότι ένας χρήστης αντιμετωπίζει κάποιο πρόβλημα το οποίο θέλει να επιλύσει. Κατηγορίες τέτοιων προβλημάτων μπορεί να είναι τεχνικού (επιδιόρθωση τηλεόρασης, υπολογιστή, βρύσης, κ.λπ.) ή άλλου περιεχομένου. Η εξέυρεση συγκεκριμένης λύσης στο net δεν είναι εφικτή, κυρίως επειδή πρέπει να γίνει συνδυασμός πληροφορίας από διαφορετικές πηγές, ή γιατί το ερώτημα του χρήστη δεν περιλαμβάνει τα κατάλληλα keywords.

Με βάση το προηγούμενο setting, δεν έχουμε κάποιο "καλό" site να προτείνουμε ώστε να χρησιμοποιήσει για την επίλυση του προβλήματός του. Ωστόσο, υπάρχουν αρκετά site όπου ένας χρήστης μπορεί να υποβάλλει το ερώτημα του (newsgroup, forums, etc), και να πάρει απάντηση. Το ζητούμενο, λοιπόν, είναι, με βάση κάποιο συγκεκριμένο ερώτημα χρήστη, διαμορφωμένο σε keywords ή φυσική γλώσσα, να βρίσκουμε το καταλληλότερο site όπου μπορεί να διατυπώσει το ερώτημά του. Τα προτεινόμενα site μπορεί να αφορούν τελειώς διαφορετικά site, ή και υπο-κατηγορίες ενός ευρύτερου site, όπως π.χ. Το Yahoo! Answers. Στη 2η περίπτωση, θέλουμε να προτείνουμε το συγκεκριμένο site που σχετίζεται με τα ερωτήματα του χρήστη, και όχι το γενικό Yahoo! Answers.

Επίσης, σημαντικό ρόλο στην πρόταση / εύρεση του καταλληλότερου site παίζει και το κατά πόσο ευχαριστημένοι φαίνεται να είναι οι χρήστες που έχουν ήδη λάβει απάντηση. Οπότε, χρειάζομαστε και μετρικές που να αξιολογούν την ποιότητα των απαντήσεων σε συγκεκριμένη και ευρεία κλίμακα.

Άτομα: 1 - 2
Συνιστώμενες Γνώσεις:
- Προγραμματισμός (οποιαδήποτε γλώσσα), για την ανάπτυξη του λογισμικού
- Αγγλικά (για τη βιβλιογραφία)

Θέμα 5

Βέλτιστη τοποθεσία

Ένα κλασσικό και, ταυτόχρονα, πρακτικό πρόβλημα που συχνά καλείται μια επιχείρηση να επιλύσει είναι η επιλογή του σημείου όπου θα ανοίξει το νέο της κατάστημα / βιοτεχνία / εργοστάσιο κ.λπ. Γι' αυτό συνήθως γίνεται μια έρευνα αγοράς της περιοχής και αξιολόγηση του ανταγωνισμού, με σκοπό την κατανόηση του πόσο καλά μπορεί να καλύψει τις ανάγκες της περιοχής αυτής η νέα επιχείρηση.

Στοιχεία τα οποία χρησιμοποιούνται είναι η ζήτηση που υπάρχει για το προϊόν που προσφέρεται, αλλά και ο αντίστοιχος ανταγωνισμός που προσφέρει το ίδιο ή παραπλήσια προϊόντα, το κόστος και τα αντίστοιχα οφέλη της τοποθεσίας κ.λπ. Σημαντικό ρόλο εδώ παίζει η γεωγραφική τοποθεσία τόσο των ανταγωνιστών όσο και των ενδιαφερόμενων για το προϊόν, οπότε πρέπει να λαμβάνονται υπόψη και χωρικά δεδομένα.

Σκοπός της διπλωματκής αυτής εργασίας είναι η ανάπτυξη κατάλληλων τεχνικών / αλγορίθμων που θα προτείνουν πιθανές τοποθεσίες, είτε ως σημεία είτε ως ευρύτερες περιοχές, δεδομένων ορισμένων χαρακτηριστικών που επηρεάζουν την ποιότητά τους. Για την καλύτερη μοντελοποίηση του προβλήματος, μπορούν να χρησιμοποιηθούν μαθηματικά ή οικονομικά μοντέλα αναφορικά με προσφορά και ζήτηση.

Άτομα: 1 - 2
Συνιστώμενες Γνώσεις:
- Προγραμματισμός (οποιαδήποτε γλώσσα), για την ανάπτυξη του λογισμικού
- Αγγλικά (για τη βιβλιογραφία)

Θέμα 6

Χρήση skyline ερωτημάτων για την καλύτερη αντιπροσώπευση συστάδων (cluster)

Ένα σημαντικό ζητούμενο κατά την δημιουργία clusters (ή και μετά την ολοκλήρωση της διαδκασίας) είναι η εύρεση αντιπροσωπευτικών στοιχείων, τα οποία θα μπορούν να περιγράψουν όσο το δυνατόν καλύτερα τα clusters που δημιουργήθηκαν. Τα στοιχεία αυτά θα πρέπει, πρακτικά, να περιγράφουν καλύτερα από τα υπόλοιπα, βάσει ορισμένων χαρακτηριστικών το cluster στο οποίο ανήκουν.

Σε πολυδιάστατα δεδομένα, ένας τρόπος για την εύρεση του ποιο σημείο / αντικείμενο είναι καλύτερο από τα υπόλοιπα είναι τα ερωτήματα τύπου skyline. Το αποτέλεσμα που δίνουν αυτά τα ερωτήματα είναι τα στοιχεία του συνόλου για τα οποία δεν μπορεί να βρεθεί κανένα άλλο σημείο που να υπερτερεί σε όλες τις διαστάσεις. Αν κάποιο σημείο “χάνει” σε όλες τις διαστάσεις τότε δεν ανήκει στο skyline.

Η ιδέα της πτυχιακής / διπλωματικής αυτής, λοιπόν, είναι να αξιολογηθεί κατά πόσο τα skyline ερωτήματα μπορούν να δώσουν σημεία / στοιχεία τα οποία να αντιπροσωπεύουν καλύτερα τα clusters. Τα ερωτήματα τύπου skyline [1] μπορούν κάλλιστα να χρησιμοποιηθούν και κατά την εκτέλεση του αλγορίθμου, για την επιλογή των αντιπροσώπων, προτού προχωρήσει στην επόμενη φάση ο αλγόριθμος π.χ. Στον k-Means.

Υπάρχει ήδη δουλειά [2] που έχει δείξει ότι η αξιοποίηση των skyline ερωτημάτων σε γράφους, έχει ως αποτέλεσμα ενδιαφέροντες υπογράφους με σημαντική δομή. Στηριζόμενοι σε αυτή την παρατήρηση, θέλουμε να δούμε κατά πόσο μπορούμε να χρησιμοποιήσουμε τα skyline ερωτήματα στην εξόρυξη γνώσης πιο “κλασσικού” τύπου, δηλαδή στο clustering. Αυτός ακριβώς είναι και ο σκοπός της πτυχιακής / διπλωματικής εργασίας. Η αξιολόγηση μπορεί να γίνει σε διάφορα dataset, με μεγαλύτερο ενδιαφέρον να πραγματοποιηθεί σε documents, ώστε να εξετάσουμε αν με χρήση αυτής της τεχνικής θα πάρουμε terms που θα περιγράφουν πιο καλά τα clusters από τις κατηγορίες που θα έχουν δημιουργηθεί.

Άτομα: 1 - 2

Συνιστώμενες Γνώσεις:

- Προγραμματισμός (οποιαδήποτε γλώσσα), για την ανάπτυξη του λογισμικού
- Αγγλικά (για τη βιβλιογραφία)

Επιπλέον Πληροφορίες:

[1] Stephan Borzsonyi, Donald Kossmann, and Konrad Stocker. 2001. The Skyline Operator. In Proceedings of the 17th International Conference on Data Engineering. IEEE Computer Society, Washington, DC, USA, 421-430., http://dl.acm.org/citation.cfm?id=656550

[2] SkyGraph: An Algorithm for Important Subgraph Discovery in Relational Graphs, by Apostolos N. Papadopoulos, Apostolos Lyritsis, and Yannis Manolopoulos., http://delab.csd.auth.gr/papers/ECMLPKDD08plm.pdf

Θέμα 7

Βαθμονόμηση δρομολογίων (Ranking Itineraries)

Μια πρόσφατη δουλειά [1] προσπαθεί να προτείνει δρομολόγια στους χρήστες, με βάση συγκεκριμένα ενδιαφέροντα τα οποία μπορεί να έχουν. Η ιδέα εκεί είναι να επιτρέπουν στο χρήστη να αλλάζει δυναμικά τα points of interest (POIs) τα οποία θέλει να επισκεφτεί, και τα οποία αρχικά του έχουν προταθεί από το σύστημα. Αν στον χρήστη δεν αρέσουν κάποια POIs, αυτός τα αποεπιλέγει και στη θέση τους μπορούν να μπουν καινούρια.

Η ιδέα σε εμάς είναι να δούμε το πρόβλημα από μια διαφορετική σκοπιά: αυτή της αξιολόγησης της ποιότητας των δρομολογίων. Να βασιστούμε σε συγκεκριμένα κριτήρια και να ορίσουμε ποια δρομολόγια είναι καλύτερα από τα άλλα. Στη δουλειά του [1] κάνουν ένα πρώτο μικρό βήμα προς αυτή την κατεύθυνση, θεωρώντας ότι όσο περισσότερα POIs καλύψει σε μία διαδρομή τόσο πιο ευχαρηστιμένος θα είναι ο χρήστης. Κάτι τέτοιο φυσικά δεν είναι υποχρεωτικό, αφού λιγότερα σημεία μπορεί να είναι πιο ευχάριστα για το χρήστη, κάτι το οποίο έχει να κάνει φυσικά με το κάθε άτομο.

Το ζητούμενο της πτυχιακής / διπλωματικής αυτής εργασίας είναι η ανάπτυξη αλγορίθμων αξιολόγησης και βαθμονόμησης των πιθανών δρομολογίων, και η σύσταση αυτών στους χρήστες. Εκτός από την αξιολόγηση των δρομολογίων, ζητούμενο είναι οι αλγόριθμοι να είναι γρήγοροι, και κατά προτίμηση δυναμικοί στη φύση τους, ώστε να μπορούν εύκολα να ενσωματώνουν νέα σημεία, όπως έκαναν και στο [1].

Άτομα: 1 - 2
Συνιστώμενες Γνώσεις:
- Προγραμματισμός (οποιαδήποτε γλώσσα), για την ανάπτυξη του λογισμικού
- Αγγλικά (για τη βιβλιογραφία)

Επιπλέον Πληροφορίες:
[1] Senjuti Basu Royz, Gautam Das, Sihem Amer-Yahiay, Cong Yu, *Interactive itinerary planning*, http://www.eecs.umich.edu/~congy/work/icde11b.pdf

Θέμα 8

Αυτόματη εξαγωγή χαρακτηριστικών από reviews

Στο web σήμερα, τα περισσότερα site δίνουν τη δυνατότητα στους χρήστες να καταγράφουν την άποψή τους για τα προϊόντα. Η πληροφορία αυτή είναι ιδιαίτερα χρήσιμη για μελλοντικούς χρήστες, ώστε να αξιολογήσουν αν ένα προϊόν είναι καλύτερο ή χειρότερο ποιοτικά από ένα άλλο. Ο λόγος είναι ότι η πληροφορία αυτή βασίζεται στη χρήση του προϊόντος, και όχι στη διαφήμισή του.

Βασική ιδέα και ζητούμενο της πτυχιακής αυτής είναι η ανάπτυξη αλγορίθμων που να προσπαθούν να εξάγουν συγκεκριμένα χαρακτηριστικά στα οποία στηρίζονται οι χρήστες, προκειμένου να κάνουν την αξιολόγησή τους. Για παράδειγμα, για μία φωτογραφική μηχανή, ενδιαφέρει η ποιότητα της φωτογραφίας, το βάρος της κ.λπ. Αντίστοιχα, για ένα laptop ενδιαφέρει το μέγεθος της οθόνης, ο δίσκος, η αυτονομία της μπαταρίας κ.λπ., ανάλογα βέβαια τη χρήση που του γίνεται.

Η κριτική που κάνουν οι χρήστες μπορεί να είναι θετική η αρνητική. Ωστόσο, αυτό δεν ενδιαφέρει αποκλειστικά, γιατί από τη στιγμή που ένα χαρακτηριστικό αναφέρεται ως σημείο κριτικής (θετικής ή αρνητικής) σημαίνει ότι είναι κάτι που ενδιαφέρει το χρήστη και άρα πρέπει να ληφθεί υπόψη. Η ιδέα είναι να χρησιμοποιηθούν text mining τεχνικές (τεχνικές εξόρυξης πληροφορίας αό κείμενο), ώστε να βρούμε ποια είναι εκείνα τα συγκεκριμένα χαρακτηριστικά που ενδιαφέρουν το χρήστη για ένα αντικείμενο. Μεγαλύτερο ενδιαφέρον παρουσιάζει φυσικά η εξαγωγή τέτοιων στοιχείων σε μακροσκοπικό επίπεδο, ώστε να μπορούμε να πούμε (με αυτόματο τρόπο) ότι όταν πρόκειται για τηλεόραση ενδιαφέρει η διαγώνιός της και η γωνία όρασης.

Επιπλέον, με την εξαγωγή συγκεκριμένων χαρακτηριστικών μπορούμε κατόπιν να ομαδοποιήσουμε τους χρήστες αναφορικά με τα κριτήρια που θέτουν όταν αξιολογούν, ώστε να δούμε πόσο σημαντικό ρόλο παίζουν τα χαρακτηριστικά για διάφορες ομάδες χρηστών. Τέτοια πληροφορία μας ενδιαφέρει ώστε να μπορούμε για παράδειγμα να κάνουμε καλύτερη προώθηση προϊόντων.

Άτομα: 1 - 2

Συνιστώμενες Γνώσεις:

- Προγραμματισμός (οποιαδήποτε γλώσσα), για την ανάπτυξη του λογισμικού
- Αγγλικά (για τη βιβλιογραφία)

Θέμα 9

Επιτυχημένες διακοπές!

Στο web σήμερα υπάρχει πληθώρα πληροφορίας, σχεδόν εφ' όλης της ύλης. Ακόμα περισσότερο υλικό όμως λόγω της κινητικότητας των ίδιων των χρηστών τα τελευταία χρόνια, αφού από απλοί καταναλωτές πληροφορίας έχουν γίνει δημιουργοί: μπορούν και γράφουν σε blogs, ανεβάζουν φωτογραφίες, videos κ.λπ με πρακτικά μηδενικό κόστος.

Συχνά, όταν κάποιος προγραμματίζει τις διακοπές του, ή γενικότερα ένα ταξίδι, θέλει να επισκεφθεί ένα μέρος όπου να μπορεί να κάνει ποικίλα πράματα του ενδιαφέροντός του. Έτσι, στην αναζήτηση προορισμού, το web είναι πλούσια πηγή πληροφορίας, και όπως ήδη ειπώθηκε, αυτό ισχύει κατά βάση επειδή πολλοί άλλοι χρήστες δίνουν πληροφορίες για τα μέρη που έχουν ήδη πάει. Ακόμα και αν έχει ήδη αποφασίσει στον προορισμό, μπορεί να αναζητήσει πληροφορίες για συγκεκριμένα μέρη, π.χ. διασκέδαση, παραλίες, ξενοδοχεία, τοπία, αρχαιολογικοί τόποι και μνημεία κ.λπ.

Η ιδέα της πτυχιακής / διπλωματικής αυτής είναι η δημιουργία τεχνικών / αλγορίθμων, που αξιοποιούν την υπάρχουσα πληροφορία του web, προκειμένου να προτείνουν στο χρήστη πιθανούς προορισμούς, αλλά και σημεία ενδιαφέροντος για ένα συγκεκριμένο προορισμό. Οι πηγές πληροφορίας μπορούν να είναι δεδομένες στα πλαίσια της πτυχιακής / διπλωματικής, ώστε να δίνουν "προβλέψιμη" είσοδο, αναφορικά με τη δομή του περιεχομένου τους. Σημαντικότερο, όμως, είναι να υπάρχει αξιοποίηση ποικίλων πηγών, διαφορετικού τύπου, όπως π.χ. Τουριστικοί οδηγοί, blogs, εικόνες (π.χ. Flickr), videos (youtube) και αντίστοιχα σχόλια.

Φυσικά, δεδομένου αυτού του πλουραρισμού, θα πρέπει οι τεχνικές που θα αναπτυχθούν να φιλτράρουν τι αρέσει γενικά από το τι αρέσει σε μεμονωμένα άτομα (ελλείψει εξατομίκευσης), ώστε το περιεχόμενο να είναι υψηλής ποιότητας. Το μεγάλο ζητούμενο, πρακτικά, είναι να μπορεί να συνδυαστεί πληροφορία από διαφορετικές πηγές δεδομένων σε σχεδόν πραγματικό χρόνο.

Άτομα: 1 - 3

Συνιστώμενες Γνώσεις:

- Προγραμματισμός (οποιαδήποτε γλώσσα), για την ανάπτυξη του λογισμικού
- Αγγλικά (για τη βιβλιογραφία)

Θέμα 10

Εύρεση συναδέλφων

Μια σημαντική πλατφόρμα κοινωνικής δικτύωσης σήμερα είναι το LinkedIn (www.linkedin.com) το οποίο χρησιμοποιείται για την δημιουργία επαφών επαγγελματικού προσανατολισμού. Φυσικά, παρόμοια site υπάρχουν και αλλού, απλά αυτό είναι το πιο διαδεδομένο.

Ένα βασικό ερώτημα που τίθεται σε τέτοιου τύπου site είναι η επιλογή των ατόμων για μια συγκεκριμένη δουλειά. Ωστόσο, η επιλογή των ατόμων είναι πολύ δύσκολο πρόβλημα, τόσο από πλευράς ορισμού συγκεκριμένων κριτηρίων με βάση τα οποία θα γίνει η επιλογή, όσο και από πλευράς πολυπλοκότητας [1], [2].

Στόχος της πτυχιακής / διπλωματικής αυτής είναι να αξιολογηθούν οι υπάρχουσες τεχνικές [3] και να εντοπιστούν προβλήματα τα οποία παρουσιάζουν. Επίσης, η ιδέα είναι να αξιολογηθούν οι τεχνικές σε πιο ευρεία κλίμακα, και σε περιπτώσεις όπου δεν είναι γνωστές οι συσχετίσεις των μελών (δεν είναι δηλαδή γνωστός όλος ο γράφος κοινωνικής δικτύωσης).

Άτομα: 1 - 3
Συνιστώμενες Γνώσεις:

- Προγραμματισμός (οποιαδήποτε γλώσσα), για την ανάπτυξη του λογισμικού
- Αγγλικά (για τη βιβλιογραφία)

[1] Theodoros Lappas, Kun Liu, Evimaria Terzi. Finding a team of experts in social networks. In Proceedings of KDD'2009. pp.467~476, http://www.cs.ucr.edu/~tlappas/team-formation.pdf

[2] Mauro Sozio, Aristides Gionis. The community-search problem and how to plan a successful cocktail party. In Proceedings of KDD'2010. pp.939~948, http://dl.acm.org/citation.cfm?id=1835923

[3] Theodoros Lappas, Kun Liu and Evimaria Terzi, "A Survey of Algorithms and Systems for Expert Location in Social Networks", In Social Network Data Analytics, http://www.springerlink.com/content/vk7410u57655w7v8/fulltext.pdf

Θέμα 11

Βλέποντας τα πράματα σε ευρύτερο πλαίσιο

Στα περισσότερα site σήμερα οι χρήστες μπορούν να αφήνουν τη γνώμη τους για τα προϊόντα τα οποία εμπορεύεται το site. Τα σχόλια μπορεί να είναι θετικά ή αρνητικά, ανάλογα με την εμπειρία του κάθε χρήστη. Ωστόσο, κάτι το οποίο σπάνια λαμβάνεται υπόψη είναι το γεγονός ότι οι χρήστες επιλέγουν τα προϊόντα έχοντας κατά νου μια συγκεκριμένη χρήση που θέλουν να τους κάνουν. Για παράδειγμα, ένα laptop μπορεί να είναι πολύ υψηλών προδιαγραφών (σκληρός δίσκος, οθόνη, RAM) αλλά δεν ενδείκνυται για συχνή μεταφορά / ταξίδια. Αντίστροφα, μια φωτογραφική μηχανή μπορεί να είναι πολύ καλή για καθημερινές φωτογραφίες αλλά όχι για επαγγελματική χρήση, ακόμα και αν είναι αισθητά πιο ακριβή από τις υπόλοιπες.

Πρακτικά, αυτό σημαίνει ότι τα προϊόντα είναι αξιόλογα ή όχι, θεωρώντας μια πολύ συγκεκριμένη χρήση τους. Υπό αυτό το πρίσμα, είναι άδικη η άμεση σύγκρισή τους με άλλα προϊόντα τα οποία δεν είναι στοχευμένα για την ίδια χρήση. Αντίστροφα, αν θεωρείται πως ήδη ανήκουν σε μια κατηγορία, χωρίς να έχουν εμφανείς διαφορές μεταξύ τους, τότε ενδεχομένως ο διαχωρισμός τους σε κατηγορίες να ήταν ατυχής.

Με βάση τα σχόλια των χρηστών μπορούμε να βρούμε ποιος είναι ο σκοπός για τον οποίο είναι καλό το προϊόν? Για παράδειγμα, μπορούμε να βρούμε από τα σχόλια αν ένα laptop είναι καταλληλότερο για ταξίδια ή για “βαρειά” χρήση? Ή αν ένα εστιατόριο είναι καλύτερο για να πας με φίλους ή αν είναι καλύτερο για οικογενειακές καταστάσεις?

Σε δεύτερο επίπεδο, μπορούμε να εξάγουμε τα χαρακτηριστικά εκείνα τα οποία οριοθετούν πότε ένα προϊόν / αντικείμενο προς αξιολόγηση είναι κατάλληλο για μια περίπτωση Α (π.χ. Ταξίδια) και πότε για μια περίπτωση Β (π.χ. Παιχνίδια)? Για παράδειγμα, μπορούμε με αυτόματο τρόπο να βρίσκουμε συσχετίσεις της μορφής:

Ταξίδια => Μπαταρία, βάρος
Παιχνίδια => Γραφικά, CPU, RAM

Σκοπός της πτυχιακής / διπλωματικής είναι η ανάπτυξη κατάλληλων τεχνικών / αλγορίθμων για την επίλυση των πιο πάνω προβλημάτων.

Άτομα: 1 - 2
Συνιστώμενες Γνώσεις:

- Προγραμματισμός (οποιαδήποτε γλώσσα), για την ανάπτυξη του λογισμικού
- Αγγλικά (για τη βιβλιογραφία)

Θέμα 12

*Μάθε **πως** να ρωτάς!*

Συχνά για να απαντήσουμε τα ερωτήματά μας, απευθυνόμαστε είτε σε ειδικούς, είτε σε άτομα τα οποία γνωρίζουμε, είτε άλλες φορές σε newsgroups / forums που υπάρχουν στο web. Στην τελευταία περίπτωση όμως δεν είναι πάντα σίγουρο ότι θα πάρουμε απάντηση, ή ότι η απάντηση που θα λάβουμε θα είναι και ικανοποιητική. Ωστόσο, υπάρχουν πολλές άλλες ερωτήσεις στις οποίες οι απαντήσεις είναι πολύ ικανοποιητικές.

Το ζητούμενο αυτής της πτυχιακής / διπλωματικής είναι να βρούμε συγκεκριμένα, ποιοτικά ή ποσοτικά χαρακτηριστικά, τα οποία επηρεάζουν το αν θα δοθεί απάντηση και την ποιότητα αυτής, μετρώντας κατά πόσο ευχαριστημένος ήταν ο χρήστης ή όχι. Ακόμα μεγαλύτερο ενδιαφέρον παρουσιάζει η αντιμετώπιση του προηγούμενου προβλήματος στο twitter και στο πως οι πληροφορία διαχέεται στο (κοινωνικό) δίκτυο και το τελικό ζητούμενο είναι η ανάπτυξη τεχνικών που να δουλεύουν σε αυτό το πεδίο (domain).

Άτομα: 1 - 2
Συνιστώμενες Γνώσεις:

- Προγραμματισμός (οποιαδήποτε γλώσσα), για την ανάπτυξη του λογισμικού
- Αγγλικά (για τη βιβλιογραφία)

Θέμα 13

Τι ήταν εκείνο που σου άρεσε τελικά?

Το YouTube (http://www.youtube.com) είναι μια από τις πλέον διαδεδομένες υπηρεσίες που υπάρχουν σήμερα στο Internet. Οι χρήστες όχι μόνο τη χρησιμοποιούν για να ανεβάσουν videos, να επικοινωνήσουν και να διασκεδάσουν . Ωστόσο, ένα βασικό πρόβλημα των σχολίων που υπάρχουν στο YouTube είναι ότι τις περισσότερες φορές είναι πολύ δύσκολο να καταλάβεις αν αφορά το ίδιο το video, κάποιον άλλο χρήστη, είτε πρόκειται γι' αυτόν που ανέβασε το video, είτε για κάποιον άλλο που σχολίασε, ή αν είναι κάτι τελείως ανεξάρτητο.

Επίσης, δεν είναι λίγες οι φορές που τα σχόλια είναι αντιφατικά: “great song, lousy video”, εννοώντας ότι το τραγούδι είναι πολύ ωραίο, αλλά ότι η εικόνα ήταν πολύ κακής ποιότητας, επειδή ήταν π.χ. από κινητό τηλέφωνο. Εναλλακτικά, μπορεί να σημαίνει ότι το τραγούδι αποτελούσε απλή “επένδυση” του video, και τα δύο δεν συνδέονται με κάποιο τρόπο. Στην περίπτωση αυτή, στον χρήστη άρεσε το τραγούδι αλλά το περιεχόμενο του video (ανεξάρτητο του τραγουδιού) δεν του άρεσε καθόλου.

Για τους παραπάνω λόγους, είναι πολύ σημαντικό να υπάρχουν μηχανισμοί που θα αναγνωρίζουν:

α) Ποιον / Τι αφορά το σχόλιο του χρήστη? Αναφέρεται στο περιεχόμενο του video ή σε μία πολύ συγκεκριμένη πτυχή του (π.χ. Αν είναι slideshow μπορεί να του αρέσει **μια** συγκεκριμένη εικόνα) ? Απαντάει / Σχολιάζει κάποιον / κάτι άλλο?

β) Δεδομένου του σε ποιον / τι αναφέρεται το σχόλιο του χρήστη, μπορούμε να βρούμε αν μιλάει θετικά / αρνητικά, αν συμφωνεί ή διαφωνεί?

γ) Μπορούμε να κάνουμε όλη αυτή τη διαδικασία, χωρίς να χρειάζεται να κάνουμε σημαντικό NLP, κυρίως τη στιγμή που τα σχόλια είναι σχετικά μικρά σε έκταση?

Μόνο αν έχουμε τέτοιους μηχανισμούς στη διάθεσή μας μπορούμε να αξιοποιήσουμε καλύτερα το περιεχόμενο των σχολίων που υπάρχουν στο YouTube, τα οποία μέχρι στιγμής δεν έχουν αξιοποιηθεί στο έπακρο. Στόχος της πτυχιακής / διπλωματικής είναι η επίλυση των πιο πάνω προβλημάτων.

Άτομα: 1 - 2
Συνιστώμενες Γνώσεις:
- Προγραμματισμός (οποιαδήποτε γλώσσα), για την ανάπτυξη του λογισμικού
- Αγγλικά (για τη βιβλιογραφία)

Θέμα 14

Αυτόματη δημιουργία ερωτημάτων χρηστών
(Automatic Query Log Extraction)

Ένα τυπικό πρόβλημα στην αξιολόγηση τεχνικών μεταξύ διαφορετικών μηχανών αναζήτησης είναι ότι, λόγω του κλειστού χαρακτήρα των εταιριών τους, τα δεδομένα που έχουν στη διάθεσή τους είναι διαφορετικά. Αυτό ισχύει ακόμα και για την πιο απλή πληροφορία, που προκύπτει από το crawling των σελίδων.

Την κατάσταση δυσχεραίνει ακόμα περισσότερο το γεγονός ότι τα ερωτήματα χρηστών είναι τελείως διαφορετικά. Αυτό δεν δυσκολεύει μόνο την άμεση σύγκριση μεταξύ δύο μηχανών αναζήτσης ως προς το περιεχόμενό τους, αλλά και την σύγκριση της αποτελεσματικότητάς τους, αφού δεν διαθέτουν τα ερωτήματα των χρηστών. Επίσης, η σύγκριση νέων τεχνικών με ήδη υπάρχουσες απαιτεί την άμεση εμπλοκή χρηστών, μια διαδικασία που είναι αρκετά χρονοβόρα.

Για τους προηγούμενους λόγους, είναι ιδιαίτερα ενδιαφέρον να μπορούμε να δημιουργήσουμε πραγματικά ερωτήματα χρηστών, ώστε να τα χρησιμοποιούμε ως benchmark για την διεξαγωγή ακόλουθης έρευνας. Ελλείψει όλης της υπόλοιπης υποδομής, μια ιδιαίτερα ελκυστική εναλλακτική είναι να μπορούμε να παράγουμε αυτόματα logs (αρχεία καταγραφής πληροφοριών) τα οποία να πλησιάζουν αρκετά τα ερωτήματα που πραγματικά έχουν γίνει.

Για τη δημιουργία των logs μπορούμε να χρησιμοποιούμε μια σειρά στατιστικών στοιχείων, τα οποία πρέπει να ικανοποιούνται από το τελικό αποτέλεσμα. Για παράδειγμα, ένα τέτοιο στατιστικό μπορεί να λέει ότι “ο μέσος όρος λέξεων ανά ερώτημα είναι 2.4 λέξεις”, ή πιο περίπλοκα στοιχεία όπως “το 10% των ερωτημάτων ειναι URLs” κ.λπ. Η ύπαρξη των τεχνικών αυτών με το συγκεκριμένο τρόπο, θα δίνει τη δυνατότητα σε μηχανές αναζήτησης να κάνουν διαθέσιμα στοιχεία από τα ερωτήματα που τους έχουν γίνει, χωρίς να δίνονται πραγματικά δεδομένα.

Το ζητούμενο αυτής της πτυχιακής / διπλωματικής είναι η ανάπτυξη κατάλληλων τεχνικών / αλγορίθμων, γρήγορων και αποτελεσματικών, με σκοπό την αυτόματη δημιουργία τέτοιων ιστορικών από ερωτήματα χρηστών, με χρήση των δοθέντων στατιστικών στοιχείων που δόθηκαν προηγουμένως.

Άτομα: 2 - 3
Συνιστώμενες Γνώσεις:
- Προγραμματισμός (οποιαδήποτε γλώσσα), για την ανάπτυξη του λογισμικού
- Αγγλικά (για τη βιβλιογραφία)

Θέμα 15

PageRank με συναίσθημα!

Μια βασική τεχνική που χρησιμοποιείται για τη βαθμονόμηση (ranking) σελίδων στο web είναι η χρήση του PageRank αλγορίθμου. Η ιδέα είναι ότι κάποιο link πάντα κάνει “reference” μια άλλη σελίδα. Το ίδιο, βέβαια, μπορεί να ισχύσει και για άλλες κατηγορίες, όπως π.χ. blogs, citations , κ.λπ.

Μια παράμετρος που δεν λαμβάνεται υπόψη, όμως, στην αρχική παραλλαγή του αλγορίθμου είναι αν η σελίδα από όπου υπάρχει το link μιλάει θετικά η αρνητικά για το περιεχόμενο της σελίδας προς την οποία δείχνει. Για παράδειγμα, το ότι μιλάω θετικά σημαίνει ότι έχω εμπιστοσύνη προς τη σελίδα αυτή. Αντιθέτως, αν μιλάω αρνητικά, σημαίνει ότι δεν μου αρέσει για αρκετούς λόγους, ωστόσο το link μπαίνει για να γνωρίζει κάποιος για ποιον μιλάω.

Μπορούμε, λοιπόν, να μιλήσουμε για μια παραλλαγή του PageRank, όπου θα λαμβάνουν υπόψη το αν η αρχική σελίδα μιλάει θετικά ή αρνητικά για τη σελίδα στην οποία δείχνει. Έτσι, θέλουμε να δούμε πως διαφοροποιείται το ranking (διάταξη) των σελίδων, κάτω από το συγκεκριμένο πρίσμα.

Σκοπός της πτυχιακής / διπλωματικής είναι η ανάπτυξη κατάλληλων τεχνικών / αλγορίθμων για την αξιολόγηση της παραλλαγής του PageRank αλγορίθμου που διατυπώθηκε προηγουμένως. Θα χρησιμοποιηθούν επίσης sentiment analysis τεχνικές για την κατανόηση αν η αρχική σελίδα μιλάει θετικά ή αρνητικά για εκείνη προς την οποία δείχνει.

Άτομα: 1 - 2
Συνιστώμενες Γνώσεις:
- Προγραμματισμός (οποιαδήποτε γλώσσα), για την ανάπτυξη του λογισμικού
- Αγγλικά (για τη βιβλιογραφία)

Θέμα 16

Αξιοποίηση Χρονικής πληροφορίας στο web (Now Web)

Ένα από τα αρνητικά του PageRank αλγορίθμου, για τα οποία έχει δεχθεί κριτική αρκετές φορές, είναι ότι ότι οι καινούριες σελίδες κατατάσσονται χαμηλότερα από άλλες τις οποίες το search engine τις έχει ανακτήσει καιρό πριν. Έτσι, οι νέες σελίδες αργούν να γίνουν σύντομα γνωστές, ενώ για τα ίδια θέματα εμφανίζονται πιο ψηλά παλαιότερες.

Το ζητούμενο είναι όμως να μπορούμε να κατανοήσουμε αν μία σελίδα σχετίζεται με χρονική πληροφορία, π.χ. Ημερομηνίες, ή χρονολογίες συμβάντων, και να την αξιοποιήσουμε καλύτερα για να βελτιώσουμε τη βαθμονόμηση (ranking) των αντίστοιχων εγγράφων. Για παράδειγμα, αν πρόκειται για ένα άρθρο από κάποια εφημερίδα, που συζητάει για ένα συμβάν (π.χ. Οικονομία, πολιτική, τέχνες), θα ήταν παράδοξο να είναι πιο ψηλά στην κατάταξη από ένα άλλο που είναι πιο πρόσφατο και συζητάει για το ίδιο ακριβώς θέμα.

Μια ιδέα για την αντιμετώπιση αυτού του προβλήματος είναι ότι οι σελίδες που σχετίζονται με χρονική πληροφορία, π.χ. πρόκειται για κάποιο άρθρο με νέα κ.λπ, να χάνουν σε δημοτικότητα με την πάροδο του χρόνου, ανεξάρτητα αν συνεχίζουν να αυξάνουν ή όχι. Έτσι, νεότερες σελίδες που έχουν πρόσφατα ανακτηθεί από τη μηχανή αναζήτησης θα έχουν τη δυνατότητα να συναγωνιστούν τις υπόλοιπες, κι ας μην έχουν καταφέρει να λάβουν μεγάλο PageRank value.

Σκοπός της πτυχιακής / διπλωματικής είναι η ανάπτυξη κατάλληλων τεχνικών / αλγορίθμων που θα μοντελοποιούν το παραπάνω πρόβλημα και θα κατατάσσουν τα έγγραφα με βάσει την παραλλαγή αυτήτου PageRank, όπου η χρονική πληροφορία θα παίζει σημαντικό ρόλο.

Άτομα: 1 - 2
Συνιστώμενες Γνώσεις:
- Προγραμματισμός (οποιαδήποτε γλώσσα), για την ανάπτυξη του λογισμικού
- Αγγλικά (για τη βιβλιογραφία)

Θέμα 17

Απομόνωση διπλότυπης πληροφορίας

Λόγω του τρόπου με τον οποίο δουλεύουν τα περισσότερα search engines, είναι σχεδόν αναπόφευκτό ότι θα υπάρχουν διπλότυπες εγγραφές στα αποτελέσματα μιας αναζήτησης. Τους χρήστες όμως σπάνια τους ενδιαφέρει ποιος θα τους δώσει την πληροφορία, αρκεί να είναι η πληροφορία που ζητάνε. Πρακτικά, αυτό σημαίνει πως όταν 2 site δίνουν ακριβώς την ίδια πληροφορία (ή με ελάχιστες διαφορές), θα έπρεπε να υπάρχει μόνο το ένα από τα 2. Ο λόγος είναι ότι οι χρήστες βλέπουν έτσι κι αλλιώς μόνο τα πρώτα top-10 αποτελέσματα και το να καταλαμβάνονται 2 θέσεις με το ίδιο ακριβώς περιεχόμενο δεν προσφέρει απολύτως τίποτα. Αντιθέτως, αυτό έχει αρνητικές συνέπειες για την εμπειρία του χρήστη.

Έτσι, το ζητούμενο είναι να μπορούμε να αφαιρούμε τα αποτελέσματα εκείνα τα οποία έχουν ακριβώς την ίδια πληροφορία, αλλά αυτή προέρχεται από διαφορετικές πηγές. Η διαδικασία αυτή μπορεί να γίνεται κατά τη διάρκεια του ερωτήματος ή να οφείλεται σε πρότερη επεξεργασία της πληροφορίας που έχουμε στη διάθεσή μας (π.χ. Document-document similarity matrix). Παραπλήσια σκοπιά του προβλήματος αυτού είναι να αυξάνεται το diversity στα αποτελέσματα, το οποίο τον τελευταίο καιρό έχει αποκτήσει σημαντικό ενδιαφέρον.

Σκοπός της πτυχιακής / διπλωματικής αυτής είναι η διερεύνηση πιθανών τρόπων και μεθοδολογιών για την απομάκρυνση διπλότυπων κατά την ανάκτηση αποτελεσμάτων από το search engine. Ταυτόχρονα, θα γίνει και επιλογή συγκεκριμένων τεχνικών για την αξιολόγηση αναφορικά με την αποτελεσματικότητά τους, για την αντιμετώπιση του προβλήματος που παρουσιάστηκε προηγουμένως.

Άτομα: 1 - 3
Συνιστώμενες Γνώσεις:
- Προγραμματισμός (οποιαδήποτε γλώσσα), για την ανάπτυξη του λογισμικού
- Αγγλικά (για τη βιβλιογραφία)

Θέμα 18

Ροές στο wikipedia!

Το wikipedia συχνά χρησιμοποιείται σήμερα ως βασική πηγή πληροφορίας για μεγάλη θεματολογία. Πρόκειται για μία σύγχρονη ηλεκτονική εγκυκλοπαίδεια, η οποία όμως προσπαθεί να εξηγήσει όσο πιο περιεκτικά το κάθε αντικείμενο, μέσα σε μία σελίδα. Τις περισσότερες φορές όμως, χρησιμοποιεί references / links προς άλλες σελίδες μέσα στο wikipedia, προκειμένου να συσχετίσει το περιεχόμενο των 2 σελίδων.

Ένας από τους λόγους που γίνεται αυτή η συσχέτιση των σελίδων του wikipedia, είναι επειδή η μία από τις 2 παρέχει υπόβαθρο για την κατανόηση της άλλης. Αυτό μπορεί να αφορά το κομμάτι των μαθηματικών ή ακόμα και βιογραφικά στοιχεία για καλλιτέχνες. Για μία πληρέστερη άποψη στο αντικείμενο, ο χρήστης θα χρειαστεί να διαβάσει και τα άλλα site. Φυσικά, είναι πρακτικά αδύνατο να διαβάσει όλη την πληροφορία σε όλα τα site που συνδέονται μεταξύ τους, αφού αυτά μπορεί να είναι εκθετικά πολλά.

Η ιδέα είναι να χρησιμοποιήσουμε τεχνικές που να αιξολογούν την εγγύτητα των ιστοσελίδων στο wikipedia, προκειμένου να ορίσουμε τι πρέπει να θεωρείται “υπόβαθρο” και τι όχι. Στην περίπτωση 2 σελίδων που είναι υπόβαθρο, οι χρήστες θα μπορούν να διαβάζουν με πιο συνεκτικό τρόπο την ίδια πληροφορία, και μάλιστα με μία αλληλουχία που να διευκολύνει την κατανόηση του τελικού κειμένου.

Στόχος της πτυχιακής / διπλωματικής είναι η ανάπτυξη τεχνικών / αλγορίθμων που θα αποφασίζουν αν μία σελίδα πρέπει να θεωρηθεί “υπόβαθρο” για μία άλλη και να την προτείνουν καταλλήλως. Για παράδειγμα, δοθέντος ενός topic / ερωτήματος του χρήστη, βρίσκουμε όλα τα σχετικά documents που απαντάνε σε αυτό και κατόπιν φτιάχνουμε μια “λογική” ακολουθία με την οποία ο χρήστης να μπορεί να διαβάσει τα documents, ώστε να έχει και καλύτερη αντίληψη του θέματος. Προβλήματα τα οποία μπορεί να προκύψουν εδώ είναι precision, recall, αλλά και topic-drifting, στην προσπάθεια να επεκτείνουμε το υπόβαθρο στο οποίο πρέπει να στηριχτεί ο χρήστης.

Άτομα: 1 - 3
Συνιστώμενες Γνώσεις:
- Προγραμματισμός (οποιαδήποτε γλώσσα), για την ανάπτυξη του λογισμικού
- Αγγλικά (για τη βιβλιογραφία)