

Εργαστήριο Βάσεων Δεδομένων

Θέματα Διπλωματικών Εργασιών 2011-2012

Τελευταία Ανανέωση: 01/11/2011

Contents

Θέμα 1	2
Σχεδιασμός της απαραίτητης υποδομής για την κατασκευή ενός κατακεμημένου προσκομιστή ιστοσελίδων υψηλής απόδοσης.....	2
Θέμα 2	3
Εξόρυξη γλωσσικών πόρων από το διαδίκτυο και αυτόματη κατηγοριοποίησή τους, με τη βοήθεια εστιασμένης αναζήτησης.	3
Θέμα 3	4
Αυτόματη δημιουργία λεξιλογίου με τη βοήθεια εργαλείων εστιασμένης αναζήτησης και επεξεργασίας κειμένου.....	4
Θέμα 4	5
Μελέτη και υλοποίηση μεθόδων αξιολόγησης αυτόματων περιλήψεων.....	5

Θέμα 1

Σχεδιασμός της απαραίτητης υποδομής για την κατασκευή ενός καταναμημένου προσκομιστή ιστοσελίδων υψηλής απόδοσης.

Οι προσκομιστές ιστοσελίδων αποτελούν ένα από τα βασικότερα υποσυστήματα μίας μηχανής αναζήτησης. Η εκρηκτική ανάπτυξη του Διαδικτύου καθιστά εξαιρετικά δύσκολη την δεικτοδότηση της συνολικής διαθέσιμης πληροφορίας. Για αυτό τον λόγο η αποτελεσματικότητα μίας μηχανής αναζήτησης εξαρτάται άμεσα από την απόδοση του συστήματος προσκόμισης ιστοσελίδων η οποία θα πρέπει να περιορίζεται κατά το ελάχιστο από την ικανότητα του προσκομιστή να εναρμονίζεται με συγκεκριμένους κανόνες δεοντολογίας (politeness policy). Αντικείμενο της παρούσας εργασίας είναι ο σχεδιασμός και η υλοποίηση των επιμέρους υποσυστημάτων ενός καταναμημένου προσκομιστή ιστοσελίδων με τα ζητούμενα χαρακτηριστικά.

Προαπαιτούμενα: Καταναμημένα συστήματα I, Βάσεις δεδομένων I, Βάσεις δεδομένων II, Δομές δεδομένων, Πολύ καλή γνώση C# (ή Java)

Επιθυμητές γνώσεις: Σχεδιασμός εφαρμογών που ακολουθούν το μοντέλο client-server, Εξοικείωση με δομές όπως Btree/Hash tables/Trie,

Σχετική βιβλιογραφία:

1. wiki:
 - a. http://en.wikipedia.org/wiki/Web_crawler
 - b. http://en.wikipedia.org/wiki/Distributed_web_crawling
 - c. <http://en.wikipedia.org/wiki/Robots.txt>
 - d. http://en.wikipedia.org/wiki/Web_crawler#Politeness_policy
2. Σχετική έρευνα:
 - a. [Design and implementation of a high performance distributed web spider](#)
 - b. [Mercator: A scalable, Extensible Web Crawler \(παρουσίαση\)](#)
 - c. [The anatomy of a large-scale Hypertextual Web Search engine](#)
 - d. [Efficient crawling through URL ordering](#)
 - e. [High performance large scale web spider architecture](#)
 - f. [Distributed High-performance Web-crawlers: A Survey of the State of the Art](#)
3. Tutorials:
 - a. <http://support.microsoft.com/kb/307445>
 - b. <http://www.codeproject.com/KB/IP/csremoteevents1.aspx>
 - c. <http://generally.wordpress.com/2007/05/31/a-simple-remoting-example-in-c/>
 - d. <http://www.csharp-station.com/Articles/IntroducingDotNetRemoting.aspx>
 - e. http://www.jot.fm/issues/issue_2004_01/column8.pdf
 - f. <http://www.csharpfriends.com/articles/getarticle.aspx?articleid=62>
 - g. <http://www.codeproject.com/KB/IP/Crawler.aspx>
 - h. http://www.example-code.com/csharp/spider_simpleCrawler.asp

Επιβλέπων Καθηγητής: Καθ. Χριστοδουλάκης Δημήτριος

Υπεύθυνος Διδακτορικός: Λευτέρης Κοζανίδης

email επικοινωνίας: kozanid@ceid.upatras.gr

Θέμα 2

Εξόρυξη γλωσσικών πόρων από το διαδίκτυο και αυτόματη κατηγοριοποίησή τους, με τη βοήθεια εστιασμένης αναζήτησης.

Στην παρούσα διπλωματική εργασία στόχος είναι η δημιουργία ενός σώματος κειμένων, ενός corpus, το οποίο θα προκύψει από τη χρήση ενός focused crawler, τον οποίο και θα εκπαιδεύσουμε. Ο crawler είναι το εργαλείο που θα χρησιμοποιήσουμε για να αναζητήσει στο διαδίκτυο με μεθοδικό τρόπο, ιστοσελίδες που ανταποκρίνονται στα κριτήρια που a priori έχουμε θέσει για την αναζήτησή μας. Ένα τέτοιο κριτήριο θα μπορούσε να είναι κάποια λέξη-κλειδί η οποία θα πρέπει να υπάρχει στο περιεχόμενο των σελίδων που επιλέγει να κατεβάσει το εργαλείο μας.

Το αποτέλεσμα της αναζήτησης του crawler θα συνιστά ένα corpus, το οποίο εν συνεχεία θα ταξινομήσουμε αυτόματα σε υποκατηγορίες. Για παράδειγμα, εάν από την εστιασμένη μας αναζήτηση προκύψει ένα corpus με άρθρα από εφημερίδες, η επιμέρους κατηγοριοποίηση αυτών των δεδομένων μπορεί να είναι ανά τομέα(οικονομικά άρθρα-πολιτικά-διεθνή) είτε χρονική(ημερομηνία δημοσίευσης είδησης). Από τα αποτελέσματα της κατηγοριοποίησης θα προκύψουν υποσύνολα κειμένων, τα οποία θα συνιστούν δεδομένα αντιπροσωπευτικά ενός κειμενικού είδους για έναν συγκεκριμένο τομέα η χρονική περίοδο.

Ενδεικτική Βιβλιογραφία:

http://en.wikipedia.org/wiki/Corpus_linguistics

http://en.wikipedia.org/wiki/Web_crawler

http://en.wikipedia.org/wiki/Document_classification

Προ-απαιτούμενα μαθήματα: Γλωσσική Τεχνολογία

Επιθυμητά προσόντα: Καλή γνώση σε γλώσσα προγραμματισμού της επιλογής σας

Επιβλέπων Καθηγητής: Καθ. Χριστοδουλάκης Δημήτριος

Υπεύθυνη Διδακτορικός: Βασιλική Σιμάκη

email επικοινωνίας: simaki@ceid.upatras.gr

Θέμα 3

Αυτόματη δημιουργία λεξιλογίου με τη βοήθεια εργαλείων εστιασμένης αναζήτησης και επεξεργασίας κειμένου.

Στόχος αυτής της εργασίας είναι η εξαγωγή ενός λεξιλογίου που θα περιέχει αντιπροσωπευτικούς όρους για τον τομέα που θα επιλεγεί. Αρχικά, με τη βοήθεια ενός focused crawler, ο οποίος θα επιλέξει με βάση συγκεκριμένα κριτήρια, θα συλλεχθούν τα δεδομένα που θα αποτελέσουν το corpus που θα χρησιμοποιηθεί. Αυτό το σύνολο κειμένων θα επεξεργαστεί και μετά από μια σειρά εργασιών και εφαρμογών, και με βάση τα κριτήρια επιλογής που θα θέσουμε, θα εξάγουμε τους αντιπροσωπευτικότερους όρους του επιλεγμένου τομέα. Για παράδειγμα, εάν επιλέξουμε τον τομέα τεχνολογία, ένας αντιπροσωπευτικός όρος που θα μπορούσε να προκύψει θα είναι ο όρος “υπολογιστής”, ο οποίος θα μπει στο λεξιλόγιο που θα δημιουργηθεί.

Στη συνέχεια οι 20 πιο αντιπροσωπευτικοί όροι του εστιασμένου (από άποψη τομέα) corpus, θα συνθέσουν το domain terminology (ορολογία του τομέα) και θα δίνουν όσες το δυνατόν περισσότερες πληροφορίες. Με τη βοήθεια του WordNet και άλλων εφαρμογών θα μπορούμε να εμφανίζουμε στοιχεία όπως τον ορισμό, τη φρασεολογία, τα συνώνυμα-αντώνυμα, κλπ. των όρων του λεξιλογίου μας.

Ενδεικτική Βιβλιογραφία:

http://en.wikipedia.org/wiki/Web_crawler

<http://wordnet.princeton.edu/>

http://en.wikipedia.org/wiki/Terminology_extraction

Προ-απαιτούμενα μαθήματα: Γλωσσική Τεχνολογία

Επιθυμητά προσόντα: Καλή γνώση σε γλώσσα προγραμματισμού της επιλογής σας

Επιβλέπων Καθηγητής: Καθ. Χριστοδουλάκης Δημήτριος

Υπεύθυνη Διδακτορικός: Βασιλική Σιμάκη

email επικοινωνίας: simaki@ceid.upatras.gr

Θέμα 4

Μελέτη και υλοποίηση μεθόδων αξιολόγησης αυτόματων περιλήψεων.

Το ερευνητικό ενδιαφέρον για την αυτόματη εξαγωγή περιλήψεων από κείμενα ξεκίνησε νωρίς (από το 1958) και μέχρι σήμερα έχουν προταθεί πολλές τεχνικές και διάφορες προσεγγίσεις. Η έλευση του Παγκόσμιου Ιστού και η διαρκώς αυξανόμενη δημοτικότητα του έχει στρέψει το ενδιαφέρον της ερευνητικής κοινότητας στο σχεδιασμό εύρωστων τεχνικών αυτόματης εξαγωγής περιλήψεων από τις ιστοσελίδες του Διαδικτύου.

Το ζήτημα της αξιολόγησης των μεθόδων εξαγωγής αυτόματων περιλήψεων αποτελεί επίσης αντικείμενο έρευνας πολλών δεκαετιών, χωρίς να έχει προκύψει μια ομόφωνα αποδεκτή τεχνική. Υπάρχουν αρκετές προκλήσεις που πρέπει να αντιμετωπιστούν κατά τον σχεδιασμό μεθόδων αξιολόγησης αυτόματων περιλήψεων, με κυριότερη πρόκληση το γεγονός ότι η τελική πληροφορία πρέπει να είναι αναγνώσιμη και κατανοητή από τον άνθρωπο και όσο το δυνατόν κοντύτερα στις ανάγκες του. Η εκτίμηση της ποιότητας των περιλήψεων με βάση τις ανάγκες του ανθρώπου χωρίς την παρέμβασή του, ή με τη μικρότερη δυνατή παρέμβαση, είναι ένα πολύπλοκο πρόβλημα και οι μέθοδοι που έχουν προταθεί στη βιβλιογραφία είναι ποικίλες.

Στα πλαίσια της διπλωματικής εργασίας θα μελετηθεί η σχετική βιβλιογραφία και θα περιγραφούν οι κατηγορίες μεθόδων αξιολόγησης που έχουν προταθεί. Θα πραγματοποιηθεί επίσης συγκριτική μελέτη των κυριότερων μεθόδων αξιολόγησης αυτόματων περιλήψεων, λαμβάνοντας υπόψη ποικίλους παράγοντες, όπως τις ιδιαιτερότητες των κειμένων εισόδου, τις απαιτήσεις του πεδίου εφαρμογής κλπ. Τέλος θα σχεδιαστεί και θα υλοποιηθεί πρότυπο σύστημα αξιολόγησης αυτόματων περιλήψεων με χρήση μιας ή περισσότερων από τις μεθόδους που μελετήθηκαν.

Ενδεικτική Βιβλιογραφία:

http://en.wikipedia.org/wiki/Automatic_summarization

<http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings2/sum-mani.pdf>

<http://berouge.com/default.aspx>

Keywords: automatic summarization evaluation

Προ-απαιτούμενα μαθήματα: Γλωσσική Τεχνολογία

Επιθυμητά προσόντα: Καλή γνώση σε γλώσσα προγραμματισμού της επιλογής σας

Επιβλέπων Καθηγητής: Καθ. Χριστοδουλάκης Δημήτριος

Υπεύθυνη Διδακτορικός: Βιβή Τζέκου

email επικοινωνίας: tzekou@ceid.upatras.gr