ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ –
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΕΡΕΥΝΗΤΙΚΗ ΟΜΑΔΑ ΕΞΟΡΥΞΗΣ ΓΝΩΣΗΣ & ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ
(DB-NET)

http://www.db-net.aueb.gr

ΥΠΕΥΘΥΝΟΣ: ΑΝ. ΚΑΘΗΓΗΤΗΣ Μ. ΒΑΖΙΡΓΙΑΝΝΗΣ

(mvazirg@aueb.gr)

ΔΙΠΛΩΜΑΤΙΚΕΣ ΕΡΓΑΣΙΕΣ 2010

Α/Α	ΤΙΤΛΟΣ	ΠΕΡΙΓΡΑΦΗ
1.	**Model learning & mining the spectral structure of the web graph**	Στόχος της εργασίας είναι η μάθηση μοντέλων πρόβλεψης κατάταξης ιστοσελιδων στον παγκόσμιο ιστό. Θεωρούμε μια σειρά από παρατηρήσεις (σελίδες, κατάταξη και ένα πλήθος χαρακτηριστικών) στο χρόνο. Με βάση τα δεδομένα εκπαίδευσης θα γίνει εκπαίδευση μοντέλων κατάταξης (στην βάση πολλών χαρακτηριστικών) σελίδων σε κατηγορίες οι οποίες προβλέπουν την κατάταξη της σελίδας. Προβλέπεται η προ επεξεργασία των δεδομένων με spectral μετασχηματισμούς (PCA, SVD). Μέρος της εργασίας θα αφιερωθεί στην συλλογή και διαχείριση δεδομένων από τον παγκόσμιο ιστό Προαπαιτούμενα: - Καλή γνώση μαθηματικών και ειδικότερα γραμμικής άλγεβρας και αναλυτικής γεωμετρίας. Πολύ καλή γνώση προγραμματισμού και ειδικότερα του περιβάλλοντος MATLAB. *Σχετική δημοσίευση:* - Polyxeni Zacharouli, Michalis Titsias, Michalis Vazirgiannis: «Web Page Rank Prediction with PCA and EM Clustering», WAW 2009: 104-115.
2.	**Real Web time personalization**	Η μεγάλη κινητικότητα στο χώρο της εξατομίκευσης αποδεικνύει το μεγάλο ερευνητικό και κυρίως εμπορικό ενδιαφέρον για την αναζήτηση νέων και τη βελτίωση των υπαρχόντων μεθόδων παροχής προσωποποιημένων υπηρεσιών στο διαδίκτυο. Η πρωτοτυπία της προτεινομενης εργασίας εγκειται στη υλοποίησης αυτού του είδους των υπηρεσιών από τον πάροχο διαδικτύου και όχι από έναν μεμονωμένο δικτυακό τόπο. Μια τέτοια υπηρεσία θα έχει ιδιαίτερα βελτιωμένα αποτελέσματα σε σχέση με τις υπάρχουσες. Αυτό συμβαίνει διότι οι συστάσεις (recommendations) που γίνονται στο χρήστη γίνονται

		βάση της συνολικής διαδικτυακής συμπεριφοράς του. Το αντικείμενο της εργασίας είναι η σχεδίαση και ανάπτυξη ενός συστήματος που θα δίνει τη δυνατότητα εξατομίκευσης στον παγκόσμιο ιστό σε πραγματικό χρόνο από τον πάροχο διαδικτύου, δηλαδή την προσαρμογή του περιεχομένου των ιστοσελίδων ενός δικτυακού τόπου, λαμβάνοντας υπόψη τη συμπεριφορά του χρήστη *καθώς* αυτός περιηγείται στο διαδίκτυο. Με τον τρόπο αυτό ο χρήστης υποβοηθείται κατά την περιήγηση του, ακολουθώντας τις προτάσεις με επόμενες ιστοσελίδες που πιθανότατα θα ενδιαφέρεται να επισκεφθεί, ενώ παράλληλα εξυπηρετείται και η εμπορική διαδικασία με την προβολή στο χρήστη πληροφοριών όπως προσφερόμενες υπηρεσίες, διαφημίσεις, προϊόντα που είναι πιθανόν να τον ενδιαφέρουν. Προαπαιτούμενα: • Προγραμματισμός Java, γνώση τεχνολογιών Διαδικτύου και χειρισμού κειμένου. Σχετικές Δημοσιεύσεις: • Magdalini Eirinaki, Michalis Vazirgiannis: Web site personalization based on link analysis and navigational patterns. ACM Trans. Internet Techn. 7(4): (2007)
3.	**Συλλογή δεδομένων από τον Κρυφό Ιστό (Deep Web Crawling)** **(Τεχνική συν επίβλεψη: Δρ. Β. Πλαχούρας)**	Το μεγαλύτερο μέρος της διαθέσιμης πληροφορίας στον Παγκόσμιο Ιστό είναι αποθηκευμένο σε βάσεις δεδομένων, οι οποίες αποτελούν τον Κρυφό Ιστό (Hidden ή Deep Web), και είναι προσβάσιμες μέσω διεπαφών βασισμένων σε φόρμες. Οι διεπαφές αυτές είναι άμεσα κατανοητές από τους χρήστες, αλλά όχι από το λογισμικό που συλλέγει και επεξεργάζεται αυτόματα ιστοσελίδες για τις μηχανές αναζήτησης στον Παγκόσμιο Ιστό, με αποτέλεσμα μεγάλο μέρος της διαθέσιμης πληροφορίας να μην ευρετηριάζεται από τις μηχανές αναζήτησης. Στόχος της διπλωματικής εργασίας είναι α) η συγκέντρωση και η σύγκριση των προτεινόμενων μεθόδων στη βιβλιογραφία για την αυτόματη εξαγωγή πληροφορίας από βάσεις δεδομένων στον Κρυφό Ιστό, και β) η υλοποίηση και επέκταση των παραπάνω μεθόδων για την αυτόματη εξαγωγή πληροφορίας από τις βάσεις δεδομένων. Προαπαιτούμενα:

		• προγραμματισμός Java, γνώση τεχνολογιών Διαδικτύου και χειρισμού κειμένου. Σχετικές Δημοσιεύσεις: • Versioned Corpora. In Proceedings of the ECIR 2008 Workshop on Efficiency Issues on Information Retrieval (EIIR), 2008.
4.	**Ευρετηρίαση και συμπίεση αρχειοθετημένου ιστοπεριεχομένου (Indexing versioned document collections)** **(Τεχνική συν επίβλεψη: Δρ. Β. Πλαχούρας)**	Οι μηχανές αναζήτησης στον Παγκόσμιο Ιστό ευρετηριάζουν μόνο την πιο πρόσφατη έκδοση των ιστοσελίδων, αγνοώντας αλλαγές μεταξύ των διαδοχικών εκδόσεών τους. Υπάρχουν όμως πολλά παραδείγματα εφαρμογών όπου είναι απαραίτητη η ευρετηρίαση των διαφορετικών εκδόσεων της ίδιας ιστοσελίδας, όπως το Internet Archive (http://www.archive.org), που διατηρεί ιστορικές συλλογές με το περιεχόμενο ιστοσελίδων, και η Wikipedia (http://www.wikipedia.org) που διατηρεί τις διαδοχικές εκδόσεις του περιεχομένου της. Στην απλούστερη περίπτωση ευρετηρίασης, οι διαφορετικές εκδόσεις αποθηκεύονται ως διαφορετικά κείμενα. Το μειονέκτημά σε αυτή την περίπτωση είναι ότι δεν επιτυγχάνεται βέλτιστη συμπίεση από την εκμετάλλευση των διαφορών μεταξύ των διαδοχικών εκδόσεων. Στόχος της διπλωματικής είναι η υλοποίηση, σύγκριση, και πιθανή εξέλιξη προτεινόμενων μεθόδων από τη βιβλιογραφία για την ευρετηρίαση διαφορετικών εκδόσεων κειμένων. Η υλοποίηση θα βασιστεί σε κάποια από τις υπάρχουσες πλατφόρμες, για παράδειγμα Lucene ή Terrier, καθώς και σε δεδομένα από τη Wikipedia ή διαδοχικές εκδόσεις ιστότοπων. Προαπαιτούμενα: • προγραμματισμός Java, γνώσεις τεχνολογιών χειρισμού κειμένου. Σχετικές αναφορές: 1) J. He, H. Yan, T. Suel. Compact full-text indexing of versioned document collections. In Proceedings of the 18th ACM conference on Information and knowledge management, pp 415-424, 2009. 2) K. Berberich, S. Bedathur, G. Weikum. Tunable Word-Level Index Compression for

5.	**Αυτοματοποιημένη δημιουργία διαφημιστικής καμπάνιας**	Είναι πολύ σημαντική πλέον η διαφήμιση στον Παγκόσμιο Ιστό σε οικονομικά μεγέθη. Η ανάπτυξη μιας καμπανιάς είναι μια σύνθετη διαδικασία η οποία εμπλέκει την επιλογή λέξεων κλειδιών, γεωγραφικών, γλωσσικών, χρονικών και άλλων περιορισμών (όπως τιμές σε δημοπρασίες κλπ). Η καλή σχεδίαση μιας καμπανιάς μπορεί να είναι ιδιαίτερα χρονοβόρα καθώς πρέπει να ικανοποιηθούν διάφοροι περιορισμοί (προϋπολογισμός, CPC, CPA, CPM κλπ). Επίσης είναι αναγκαία η ημιαυτόματη δημιουργία μαζικών αναφορών για την παρακολούθηση μιας καμπανιάς στην βάση των services που παρέχονται από το Google Analytics. Στόχος της εργασίας θα είναι να αναπτυχτεί μια μεθοδολογία και ένα εργαλείο για την ημιαυτόματη ανάπτυξη διαφημιστικής καμπάνιας adwords στην βάση των παραπάνω χαρακτηριστικών. Για την ανάπτυξη θα γίνει χρήση των services παρέχει το Goggle στα σχετικά APIs Σχετικά Links: http://code.google.com/apis/adwords/ http://blog.programmableweb.com/2009/04/23/google-analytics-api-released-now-get-your-web-site-metrics-via-code/
6.	**Word Sense Disambiguation For information Retrieval (Τεχνική συν επίβλεψη: Β. Πλαχούρας)**	Ο στόχος της εργασίας θα είναι η εφαρμογή και επέκταση μηχανισμού αποσαφήνισης λέξεων που έχει αναπτυχθεί από την ομάδα μας [1] και η επέκταση και εφαρμογή της σε περιβάλλον ανάκτησης πληροφορίας με στόχο την βελτίωση της ποιότητας ανάκτησης. Προαπαιτούμενη γνώση: - αλγόριθμοι ανάλυσης γράφων, εξοικείωση με χειρισμό κειμένου - C/C++ ή Java, Βάσεις Δεδομένων *Αναφορές* [1] D. Mavroeidis, G. Tsatsaronis, M. Vazirgiannis, M. Theobald, G. Weikum, "Word Sense Disambiguation for Exploiting Hierarchical Thesauri in Text Classification",in the proceedings of the ECML/PKDD 2005 Conference, Portugal

7.	**Αξιολόγηση πρωτεϊνικών αλληλεπιδράσεων (Τεχνική συν επίβλεψη: ΥΔ. Μ. Καρκαλή)**	Το πρόβλημα της αξιοπιστίας των πρωτεϊνικών αλληλεπιδράσεων που έχουν παρατηρηθεί παρουσιάστηκε με τη χρήση των high-throughput πειραμάτων που έδιναν μεγάλο αριθμό αλληλεπιδράσεων αλλά και μεγάλο αριθμό false positives. Σχεδόν αμέσως παρουσιάστηκε και το θέμα της αξιολόγησης αυτών των αλληλεπιδράσεων ώστε να μπορούν απομακρυνθούν τα false positives και να δημιουργηθεί ένας γράφος αξιόπιστων αλληλεπιδράσεων που θα μπορούσαν στην συνέχεια να αξιοποιηθούν για περαιτέρω ανάλυση. Στην εργασία θα γίνει μελέτη πάνω στις ήδη υπάρχουσες μεθόδους αξιολόγησης για τους αλγορίθμους τεχνητής μάθησης, τα χαρακτηριστικά γνωρίσματα και τα σύνολα εκπαίδευσης που χρησιμοποιούν και θα προταθεί μια νέα μέθοδος αξιολόγησης με δυνατότητα εφαρμογής σε ενοποιημένες βάσεις πρωτεϊνικών αλληλεπιδράσεων για τον οργανισμό Yeast. Η εργασία θα στηριχθεί σε προηγούμενη δουλεια πάνω στο θέμα με σκοπό την βέλτιστη επιλογή χαρακτηριστικών γνωρισμάτων και την βέλτιστη επιλογή και παραμετροποίηση αλγορίθμων τεχνητής μάθησης. Προαπαιτούμενη γνώση: C/C++ ή Java , Βάσεις Δεδομένων,
8.	**Group Formation and Evolution in Social Networks (Τεχνική συν επίβλεψη: ΥΔ. Ν. Σαλαμάνος)**	Το αντικείμενο της εργασίας είναι η ανάπτυξη μοντέλων που ερμηνεύουν την εμφάνιση ομάδων (*group, communities*) σε social networks καθώς η μελέτη της εξέλιξη τους (*evolution*). Βασικά ερωτήματα είναι: α) υπάρχει κάποιο μοντέλο που προσομοιάζει την τάση των user να συμμετέχουν σε κάποιο *group*? β) πως συνδέεται η τοπολογία και η σημασιολογία του δικτύου με την εξέλιξη των *group*? Θα μελετηθούν συγκεκριμένα data-sets από social networks και θα αναπτυχθούν μοντέλα για την ερμηνείας τους. Προαπαιτούμενα: • Matlab (ή η διάθεση να μάθετε), Java. Σχετικές Αναφορές: 1. L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. *Group formation in large social networks*: Membership, growth, and evolution. In Proc. 12th KDD, pages 44{54, 2006. (http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.62.737). 2. *Algorithmic Game Theory*, (Chapter 24) (http://www.cambridge.org/journals/nisan/downloads/Nisan_Non-printable.pdf) 3. Matthew O. Jackson, *Social and Economic Networks*,. Princeton University Press (2008).
9.	**Evolution of the Two sided Markets**	*Two sided Markets* προκύπτουν όταν ένας αριθμός από online πλατφόρμες (π.χ. websites για: φωτογραφίες, video, question-answering, auctions κ.λ.π.), ανταγωνίζονται μεταξύ τους στο να προσελκύσουν users. Οι users ανήκουν συνήθως σε δύο κατηγορίες και κερδίζουν περισσότερο όταν αλληλεπιδρούν με user από την άλλη κατηγορία. Για

	(Τεχνική συν επίβλεψη: ΥΔ. Ν. Σαλαμάνος)	παράδειγμα web sites για πλειστηριασμούς προϊόντων συνδέουν αγοραστές με πωλητές. Οι αγοραστές προτιμούν sites με μεγάλο αριθμό πωλητών (ποικιλία προϊόντων) ενώ οι πωλητές sites με μεγάλο αριθμό αγοραστών. Εμφανίζεται συνεπώς το *network effect* (http://en.wikipedia.org/wiki/Network_effect) όπου η μία κατηγορία χρηστών επηρεάζει τις επιλογές της άλλης. Το αντικείμενο του θέματος είναι η μελέτη της εξέλιξης (*evolution*) των *two sided markets* με την ανάπτυξη μοντέλων που θα ερμηνεύουν την επιλογή πλατφόρμας από τους users. Προαπαιτούμενα: • Matlab (ή η διάθεση να μάθετε), Java. Σχετικές Αναφορές: 1. Evolution of Two-Sided Markets Kumar, R.; Lifshits, Y.; Tomkins, A. , WSDM, (2010). (http://yury.name/papers/kumar2010evolution.pdf) 2. J. Rochet and J. Tirole. Two-sided markets: A progress report. The RAND Journal of Economics, 35(3):645{667, 2006.