

## Διπλωματικές Εργασίες Ακαδημαϊκού Έτους 2011-2012

B. Μεγαλοοικονόμου, Αναπληρωτής Καθηγητής

### **1. Συστήματα Βάσεων Πολυμεσικών Δεδομένων - Τεχνικές αναζήτησης ομοιότητας σχημάτων**

Τα τελευταία χρόνια, η μεγάλη ανάπτυξη στην συλλογή εικόνων και πολυμεσικών δεδομένων, τα οποία είναι διαθέσιμα μέσω διαφόρων τεχνολογιών είχε ως αποτέλεσμα η ερευνητική κοινότητα να στραφεί στην αποδοτική ανάκτηση και ευρετηριοποίηση τους. Οι περισσότερες τεχνικές που έχουν προταθεί στην βιβλιογραφία τις τελευταίες δύο δεκαετίες για την ανάκτηση όμοιων εικόνων με βάση το περιεχόμενό τους χρησιμοποιούν χαρακτηριστικά όπως το σχήμα, το χρώμα και την μορφολογία. Το σχήμα αποτελεί ένα σημαντικό γνώρισμα για την εύρεση παρόμοιων αντικειμένων και είναι αυτό που ξεχωρίζει από τα υπόλοιπα χαρακτηριστικά που έχουν χρησιμοποιηθεί στην βιβλιογραφία και τα οποία συνήθως δεν αποκαλύπτουν την ταυτότητα των αντικειμένων. Η ανάλυση σχημάτων μας δίνει την δυνατότητα να επικεντρωθούμε σε συγκεκριμένες περιοχές (οπτικά μέρη) που έχουν ιδιαίτερο ενδιαφέρον και περιέχουν σημαντική πληροφορία. Τα οπτικά αυτά μέρη μπορεί να είναι διαθέσιμα ακόμη και στην περίπτωση που ένα μεγάλο μέρος του αντικειμένου δεν είναι ορατό. Στην παρούσα εργασία θα ασχοληθούμε με την ομοιότητα σχημάτων που υπάρχουν αποθηκευμένα σε μια βάση δεδομένων. Για την πιο αποδοτική ανάκτησή τους τα σχήματα θα αναπαρασταθούν ως ακολουθίες, κατά συνέπεια η εξαγωγή όμοιων οπτικών μερών θα μετατραπεί σε ταίριασμα υπο-ακολουθιών από βάσεις δεδομένων που περιέχουν υποακολουθίες (που αναπαριστούν οπτικά μέρη αντικειμένων).

Αρχικά, στην παρούσα εργασία θα μελετηθούν διάφορες τεχνικές που έχουν παρουσιασθεί στην βιβλιογραφία για την αναπαράσταση και την ομοιότητα σχημάτων. Στην συνέχεια, κάποιες από αυτές τις τεχνικές θα αξιολογηθούν και θα εφαρμοστούν σε πραγματικά δεδομένα.

Επιθυμητές γνώσεις: Βάσεις Δεδομένων, Εξόρυξη Δεδομένων, Γλώσσες Προγραμματισμού (C, C++, Matlab)

Ενδεικτική Βιβλιογραφία:

1. L. Latecki, V. Megalooikonomou, Q. Wang, D. Yu, «An Elastic Partial Shape Matching Technique», Pattern Recognition, Vol. 40, No. 11, pp. 3069-3080, 2007.
2. D. Kontos and V. Megalooikonomou, «Fast and effective characterization for classification and similarity searches of 2D and 3D spatial region data», Pattern Recognition, Vol. 38, No. 11, pp. 1831-1846, 2005.

### **2. Γεωγραφικά Πληροφοριακά Συστήματα και Ανάλυση Χωρο-χρονικών Δεδομένων**

Περιγραφή:

Ο στόχος της διπλωματικής αυτής είναι η ανάλυση χωρο-χρονικών δεδομένων από περιβαλλοντικές μελέτες και η εξόρυξη γνώσης από αυτά, είτε για πρόβλεψη μελλοντικών τιμών ρύπων-πηγών μολύνσεων ή για συσταδοποίηση ομοίων παρατηρήσεων. Η γνώση αυτή αποτελεί ουσιαστικό εργαλείο τόσο για τους επιστήμονες του περιβάλλοντος όσο και για τις δημόσιες υπηρεσίες για τον

αποτελεσματικό έλεγχο ακτών, παράκτιων εκτάσεων, παραποτάμιων περιοχών, κτλ. Η χρήση λογισμικού Γεωγραφικών Πληροφοριακών Συστημάτων (GIS) θα καταστήσει εφικτή την άμεση προβολή των εξαγόμενων συμπερασμάτων σε πραγματικά δεδομένα και γεγονότα. Αναφορικά με τις τεχνολογίες εξόρυξης δεδομένων, θα μελετηθούν σύγχρονοι αλγόριθμοι ταξινόμησης χωρο-χρονικών δεδομένων καθώς επίσης και τεχνικές συσταδοποίησης και πρόβλεψης ετερογενών δεδομένων.

Επιθυμητές γνώσεις: Εξόρυξη γνώσης, Ανάκτηση πληροφορίας, Κατανεμημένα συστήματα, Βάσεις δεδομένων, Γλώσσες προγραμματισμού (Java, C, C++, Matlab)

Συνεπιβλέποντες: Β. Μεγαλοοικονόμου, Δ. Χριστοδουλάκης

### **3. Αναπαράσταση αντιστοιχίσεων μεταξύ ετερογενών οντολογιών**

Περιγραφή:

Οι οντολογίες ως εννοιολογικές μορφοποιήσεις αποτελούν προϊόντα υποκειμενικής κρίσης, οπότε το ίδιο πεδίο ενδιαφέροντος είναι δυνατόν να περιγραφεί με διαφορετικούς τρόπους, με αποτέλεσμα, οι οντολογίες που αναπτύσσονται να αποτελούν ετερογενείς πηγές γνώσης. Για να επιτευχθεί η ενιαία πρόσβαση στην πληροφορία και η δια-λειτουργικότητα των συστημάτων ή εφαρμογών οι οποίες χρησιμοποιούν τις ετερογενείς οντολογίες, θα πρέπει η γνώση που περιγράφεται στις διάφορες οντολογίες να είναι εναρμονισμένη. Για το λόγο αυτό ένα από τα πιο σημαντικά ερευνητικά θέματα στο χώρο των οντολογιών είναι η ανάπτυξη αλγορίθμων εύρεσης σημασιολογικών ομοιοτήτων μεταξύ δύο ετερογενών οντολογιών. Το πρόβλημα αναφέρεται ως ευθυγράμμιση οντολογιών και έχουν αναπτυχθεί μια πληθώρα από πλατφόρμες και αλγόριθμους που προσπαθούν να επιλύσουν το πρόβλημα με αυτόματο ή ημι-αυτόματο τρόπο. Στα πλαίσια της διπλωματικής εργασίας θα μελετηθούν οι αλγόριθμοι ευθυγράμμισης οντολογιών και θα υλοποιηθεί ένα σύστημα, το οποίο θα δέχεται ως είσοδο δυο διαφορετικές οντολογίες ή δύο οντολογίες και ένα αρχικό σύνολο αντιστοιχίσεων και συνδυάζοντας έτοιμους αλγόριθμους ευθυγράμμισης οντολογιών θα εξάγει αντιστοιχίσεις μεταξύ των οντοτήτων των δύο οντολογιών σε μια σειρά από κατάλληλες μορφές αρχείων οι οποίες μπορούν να αναπαραστήσουν τέτοια πληροφορία, όπως είναι τα αρχεία τύπου C-OWL.

Σκοπός της εργασίας αυτής είναι (α) η εξοικείωση με βασικές έννοιες των οντολογιών και του πεδίου της ευθυγράμμισης οντολογιών, (β) η ανασκόπηση μεθόδων και εργαλείων τα οποία έχουν προταθεί για το πρόβλημα της ευθυγράμμισης οντολογιών, (γ) η υλοποίηση ενός εργαλείου το οποίο θα δέχεται ως είσοδο δύο ετερογενείς οντολογίες και θα εξάγει τις αντιστοιχίσεις μεταξύ τους σε κατάλληλη μορφή, (δ) ο έλεγχος της παραπάνω τεχνολογίας σε ένα απλό σενάριο ευθυγράμμισης οντολογικής γνώσης.

Επιθυμητές γνώσεις: Γλωσσική Τεχνολογία, Εξόρυξη γνώσης, Ανάκτηση πληροφορίας, Τεχνολογίες Διαδικτύου, Βάσεις δεδομένων, Γλώσσες προγραμματισμού (C, C++, Java)

Ενδεικτική Βιβλιογραφία:

[1] <http://www.ontologymatching.org>

[2] *Ontology Alignment: Bringing the Semantic Gap*. Marc Ehrig. Springer Science+Business Media, LLC, 2007.

[3] *Ontology matching*. Jerome Euzenat and Pavel Schvaiko. Springer-Verlag, Berlin Heidelberg (DE), 2007.

Συνεπιβλέποντες: Β. Μεγαλοοικονόμου, Α. Καμέας (EAITY και ΕΑΠ)

#### 4. Ανάπτυξη Ιατρικής Βάσης Δεδομένων Κυτταρομετρίας Ροής

Περιγραφή:

Η κυτταρομετρία ροής αποτελεί μια ραγδαία εξελισσόμενη τεχνική, που χρησιμοποιείται για να μετράει ταυτόχρονα και κατόπιν να αναλύει πολλαπλά φυσικά ή/και χημικά χαρακτηριστικά μικροσκοπικών σωματιδίων, συνήθως κυττάρων. Τις τελευταίες δεκαετίες, η χρηστική αξία της κυτταρομετρίας ροής έχει εκτιμηθεί σημαντικά σε ποικίλες βιολογικές και ιατρικές εφαρμογές. Μια βασική εφαρμογή της αφορά την αιματολογία, στην οποία χρησιμοποιείται για το χαρακτηρισμό του ανοσοφαινότυπου δειγμάτων αίματος και συμβάλλει έτσι στη διάγνωση αιματολογικών κακοηθειών. Οι ιδιότητες που μετρούνται, σε μία εξέταση κυτταρομετρίας ροής, περιλαμβάνουν το σχετικό μέγεθος του κάθε σωματιδίου, τη σχετική εσωτερική πολυπλοκότητά του, καθώς και τη σχετική ένταση φθορισμού (fluorescence) του.

Δεδομένου ότι οι μετρήσεις κυτταρομετρίας ροής αφορούν πολλαπλές τιμές για κάθε κύτταρο ξεχωριστά, το πλήθος των παραγόμενων δεδομένων ανέρχεται σε δεκάδες χιλιάδες τιμές ανά εξέταση. Επιπλέον οι τιμές κάθε εξέτασης μεταβάλλονται ανάλογα με τις συνθήκες που επικρατούν κατά τη διεξαγωγή της μέτρησης καθώς και τις ρυθμίσεις του εξοπλισμού (ένταση του laser, ποσοστό ενίσχυσης, αντιγόνα που χρησιμοποιήθηκαν, κ.α.). Στόχος της εργασίας είναι η ανάπτυξη μίας εφαρμογής που θα ομογενοποιεί και θα κανονικοποιεί τα δεδομένα κυτταρομετρίας ροής και θα τα εισάγει σε μία σχεσιακή βάση δεδομένων ώστε να καταστεί δυνατή η συστηματική, δομημένη και συνδυαστική ανάλυση πολλαπλών εξετάσεων κυτταρομετρίας ροής (τιμές εκατοντάδων εκατομμυρίων κυττάρων). Η εφαρμογή αυτή θα πρέπει να είναι ικανή για:

- Την εκτέλεση συντακτικά απλών και διαισθητικών ερωτημάτων σχετικών με επιστημονικές ερωτήσεις που ανακύπτουν κατά την κλινική καθημερινότητα καθώς και κατά την έρευνα.
- Άμεση συσχέτιση των ερωτημάτων που απευθύνονται στη βάση δεδομένων με τα κλινικά, εργαστηριακά και ερευνητικά ερωτήματα της πραγματικής ζωής.
- Απαντήσεις σε ερωτήματα σε πραγματικό χρόνο.
- Αποθήκευση των δεδομένων χωρίς απώλεια πληροφοριών, όπως συμβαίνει όταν αποθηκεύεται η μέση τιμή ενός αντιγόνου αντί των μεμονωμένων τιμών κάθε κυττάρου.
- Διατήρηση των δεδομένων σε τέτοια μορφή ώστε η διασύνδεσή τους με δεδομένα προερχόμενα από διαφορετικούς εργαστηριακούς ή κλινικούς χώρους να απαιτεί την ελάχιστη δυνατή προσπάθεια.
- Δυνατότητα, εκτός από την εξέταση μεμονωμένων περιστατικών – εξετάσεων, μαζικής επεξεργασίας μέρους ή του συνόλου των δεδομένων.

Επιθυμητές γνώσεις: Βάσεις Δεδομένων, Εξόρυξη Δεδομένων, Γλώσσες Προγραμματισμού (C, C++, C#, Matlab, Python)

Ενδεικτική Βιβλιογραφία:

[1] J. Drakos, M. Karakantza, N.C. Zoumbos, J. Lakoumentas, G.C. Nikiforidis, G.C Sakellaropoulos, “A perspective for biomedical data integration: Design of databases for flow cytometry”, BMC Bioinformatics, 2008, 9:99 doi:10.1186/1471-2105-9-99.

[2] F. Hahne, A.H. Khodabakhshi, A. Bashashati, C.-J. Wong, R.D. Gascoyne, A.P. Weng, V. Seyfert-Margolis, K. Bourcier, A. Asare, T. Lumley, R. Gentleman, R.R. Brinkman, “Per-Channel Basis Normalization Methods for Flow Cytometry Data”, Cytometry Part A, 77A: 121-131, 2010.

Συνεπιβλέποντες: Β. Μεγαλοοικονόμου, Μ. Καρακάντζα (Ιατρική Σχολή)

## **5. Εφαρμογή Τεχνικών Εξόρυξης σε Πολυδιάστατα Δεδομένα Κυτταρομετρίας Ροής**

Περιγραφή:

Η κυτταρομετρία ροής χρησιμοποιείται για την ταυτόχρονη μέτρηση και ανάλυση πολλαπλών φυσικών ή/και χημικών χαρακτηριστικών μικροσκοπικών σωματιδίων, συνήθως κυττάρων. Σημαντική τεχνολογική πρόοδος στο υλικό/πειραματικά όργανα και την ανάπτυξη φθορίζοντων ιχνηθετών και υποστρωμάτων, έχουν καταστήσει δυνατή την παραγωγή πολύ σύνθετων συνόλων δεδομένων (και μεγάλου αριθμού παραμέτρων) που απαιτούν την ανάπτυξη προηγμένων εργαλείων ανάλυσης. Αν και ο αριθμός των μεταβλητών που μετριοούνται ταυτόχρονα μπορεί να αυξηθεί από τους διαφορετικούς δείκτες που χρησιμοποιούνται στην ανάλυση, από τις συνθήκες που επικρατούν κατά τη διεξαγωγή της μέτρησης (π.χ., χρόνος υποκίνησης, συγκέντρωση του ερεθίσματος) ή από τα χρονικά σημεία σε ένα in-vitro πείραμα ή κλινική δοκιμή τα δεδομένα αυτά δεν μπορούν να αξιοποιηθούν κατάλληλα από τους χρήστες με αποτέλεσμα την απώλεια σημαντικής πληροφορίας. Μέχρι σήμερα η ανάλυση βασίζεται σε επιλογή από τον χρήστη δυάδων παραμέτρων που απεικονίζονται δυσδιάστατα. Την ανάλυση της πρώτης δυάδας, ακολουθεί δεύτερη και ούτω καθεξής. Αυτή η διαδοχική διπαραμετρική ανάλυση είναι χρονοβόρα, απαιτεί μεγάλη εμπειρία και δεν αναδεικνύει όλες τις σχέσεις των δεδομένων.

Αρκετές προσπάθειες έχουν γίνει για να απλοποιηθεί η ανάλυση. Αυτές μπορούν να διαιρεθούν κατά προσέγγιση σε δύο κύριες κατηγορίες: εποπτευόμενες (supervised) και μη εποπτευόμενες (unsupervised). Οι περισσότερες από αυτές τις νέες προσεγγίσεις είναι κυρίως explorative και όχι ποσοτικές. Τα ιστόγραμμα και οι γραφικές παραστάσεις σημείων είναι πολύ απλοί και διαισθητικοί τρόποι για την ανάλυση δεδομένων κυτταρομετρίας ροής. Όσο περιλαμβάνουμε στην ανάλυση όλο και περισσότερες παραμέτρους, ο αριθμός των πιθανών συνδυασμών ( $2^n$ , όπου το  $n$  είναι ο αριθμός παραμέτρων) αυξάνεται εκθετικά. Κατά συνέπεια, απαιτείται απλοποίηση των συνόλων δεδομένων. Αλγόριθμοι συσταδοποίησης έχουν χρησιμοποιηθεί για την εύρεση ομοιοτήτων και διαφορών μεταξύ των δειγμάτων. Επίσης δεδομένου ότι τα δεδομένα κυτταρομετρίας ροής είναι υψηλής διαστατικότητας τεχνικές όπως η PCA έχουν εφαρμοστεί για μειώσουν τον αριθμό των διαστάσεων. Στη παρούσα εργασία θα γίνει μελέτη των τεχνικών που έχουν προταθεί στην βιβλιογραφία για την ανάλυση δεδομένων κυτταρομετρίας ροής και θα υλοποιηθούν κάποιες από αυτές. Επίσης θα μελετηθεί η χρήση τους σε πραγματικά δεδομένα.

Επιθυμητές γνώσεις: Βάσεις Δεδομένων, Εξόρυξη Δεδομένων, Γλώσσες Προγραμματισμού (C, C++, C#, Matlab, Python)

Ενδεικτική Βιβλιογραφία:

[1] E. Lugli, M. Roederer, A. Cossarizza, “Data Analysis in Flow Cytometry: The Future Just Started”, Cytometry, Part A, 77A: 705-713, 2010.

[2] Ali Bashashati and Ryan R. Brinkman, «A Survey of Flow Cytometry Data Analysis Methods» Advances in Bioinformatics, Volume 2009, Article ID 584603, 19 pages, doi:10.1155/2009/584603.

Συνεπιβλέποντες: Β. Μεγαλοοικονόμου, Μ. Καρακάντζα (Ιατρική Σχολή)

## 6. Μελέτη στατιστικών ιδιοτήτων πραγματικών γραφημάτων

Περιγραφή:

Τα τελευταία χρόνια έχει παρατηρηθεί ιδιαίτερο ενδιαφέρον στη μελέτη γραφημάτων που προκύπτουν από τεχνολογικές, κοινωνικές και επιστημονικές δραστηριότητες. Χαρακτηριστικά παραδείγματα αποτελούν το γράφημα του Διαδικτύου (οι κόμβοι αναπαριστούν δρομολογητές και οι ακμές συνδέσεις μεταξύ αυτών), το γράφημα του Παγκοσμίου Ιστού (οι κόμβοι αντιστοιχούν σε σελίδες και οι ακμές σε υπερσυνδέσμους μεταξύ των σελίδων), κοινωνικά δίκτυα (π.χ. Facebook, Flickr), δίκτυα ετεροαναφορών (citation networks) σε επιστημονικές εργασίες (οι κόμβοι αντιστοιχούν σε επιστημονικές εργασίες και οι ακμές υποδηλώνουν αναφορά της μιας εργασίας στην άλλη), κ.α.. Βασικό συστατικό στην κατανόηση της δομής τέτοιου είδους γραφημάτων, αποτελεί η εύρεση και μελέτη στατιστικών και δομικών ιδιοτήτων που εμφανίζονται σε αυτά. Συνήθως οι ιδιότητες αυτές είναι στατικές, δηλαδή προκύπτουν από τη μελέτη ενός στιγμιότυπου του γραφήματος για κάποια χρονική στιγμή. Χαρακτηριστικά παραδείγματα τέτοιου είδους ιδιοτήτων αποτελεί η power-law κατανομή των βαθμών των κόμβων (degree distribution) και η μικρή διάμετρος (φαινόμενο του μικρού κόσμου (small-world phenomenon) ή six degrees of separation). Ωστόσο, πολλά από τα γραφήματα αυτά είναι δυναμικά, δηλαδή εξελίσσονται στο χρόνο, κάτι που δημιουργεί την ανάγκη για την εύρεση και μελέτη δυναμικών ιδιοτήτων. Η μελέτη των ιδιοτήτων αυτών μπορεί να χρησιμοποιηθεί σε διάφορες πρακτικές εφαρμογές, όπως καθορισμός ομοιότητας μεταξύ δύο γραφημάτων, ανίχνευση ανωμαλιών (anomaly detection) και εύρεση κοινοτήτων (community discovery).

Στα πλαίσια της διπλωματικής αυτής, αρχικά θα μελετηθούν διάφορες στατιστικές ιδιότητες πραγματικών γραφημάτων (τόσο στατικές όσο και δυναμικές), που έχουν παρουσιασθεί στη βιβλιογραφία. Στη συνέχεια, ορισμένες από τις ιδιότητες αυτές θα εξετασθούν σε πραγματικά γραφήματα διαφόρων τύπων (π.χ. γραφήματα με βάρη στις ακμές). Τέλος, θα γίνει μελέτη των εφαρμογών στις οποίες μπορούν να χρησιμοποιηθούν οι ιδιότητες αυτές.

Επιθυμητές γνώσεις: Εξόρυξη γνώσης, Θεωρία γραφημάτων, Πιθανότητες, Γραμμική Άλγεβρα, Γλώσσες προγραμματισμού (Matlab, Python)

Ενδεικτική Βιβλιογραφία:

[1] M. Faloutsos, P. Faloutsos, and C. Faloutsos. *On Power-Law Relationships of the Internet Topology*. In *ACM SIGCOMM*, 1999.

[2] C. E. Tsourakakis. Fast Counting of Triangles in Large Real Networks, without counting: Algorithms and Laws. In *IEEE ICDM*, Pisa, Italy, 2008.

## 7. Εύρεση μοντέλων για χρονικά εξελισσόμενα γραφήματα

Περιγραφή:

Πρόσφατα, η μελέτη πολύπλοκων γραφημάτων όπως η τοπολογία του Διαδικτύου, το γράφημα του Παγκοσμίου Ιστού (WWW), κοινωνικά δίκτυα (π.χ. Facebook), κ.α., έχει αποκτήσει ιδιαίτερο ερευνητικό ενδιαφέρον. Ένα βασικό χαρακτηριστικό των γραφημάτων αυτών αποτελεί το γεγονός ότι εξελίσσονται στο χρόνο, δηλαδή με την πάροδο του χρόνου νέοι κόμβοι και ακμές μπορεί να προστίθενται (ή και να διαγράφονται). Η έρευνα στο πεδίο αυτό έχει εστιάσει κυρίως στην εύρεση προτύπων που εμφανίζονται

σε τέτοιου είδους χρονικά εξελισσόμενα γραφήματα και πολλά ενδιαφέροντα αποτελέσματα έχουν προκύψει (π.χ. διάμετρος που τείνει να μειώνεται με την πάροδο του χρόνου). Ένα βασικό ερώτημα που προκύπτει είναι πώς μπορούμε να παράγουμε συνθετικά, αλλά ταυτόχρονα ρεαλιστικά, χρονικά εξελισσόμενα γραφήματα. Με άλλα λόγια, ενδιαφερόμαστε για το σχεδιασμό μηχανισμών και μοντέλων (graph generators) που θα παράγουν γραφήματα με πρότυπα παρόμοια αυτών που έχουν παρατηρηθεί σε πραγματικά γραφήματα. Κάτι τέτοιο είναι ιδιαίτερα σημαντικό για διάφορους λόγους, ένας από τους οποίους είναι η δυνατότητα να παράγουμε συνθετικά γραφήματα για προσομοιώσεις, αξιολόγηση αλγορίθμων, κ.α., όταν πραγματικά δεδομένα είναι δύσκολο ή ακόμα και αδύνατο να συλλεχθούν.

Στα πλαίσια της διπλωματικής αυτής, αρχικά θα μελετηθούν τα διάφορα μοντέλα παραγωγής χρονικά εξελισσόμενων γραφημάτων που έχουν παρουσιασθεί στη βιβλιογραφία. Στη συνέχεια, ορισμένα από τα μοντέλα αυτά, θα αξιολογηθούν και θα συγκριθούν πειραματικά.

Επιθυμητές γνώσεις: Εξόρυξη γνώσης, Θεωρία γραφημάτων, Πιθανότητες, Γραμμική Άλγεβρα, Γλώσσες προγραμματισμού (Matlab, Python).

Ενδεικτική Βιβλιογραφία:

[1] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In ACM SIGKDD, 2005.

[2] J. Leskovec, D. Chakrabarti, J. M. Kleinberg, and C. Faloutsos. Realistic, mathematically tractable graph generation and evolution, using Kronecker multiplication. In PKDD, Porto, Portugal, 2005.

## **8. Μελέτη και εφαρμογή τεχνικών εξόρυξης γνώσης στα πλαίσια του διαδικτύου των αντικειμένων (*internet of things*)**

Περιγραφή:

Η ραγδαία ανάπτυξη του κλάδου των δικτύων αισθητήρων σε συνδυασμό με την δυνατότητα διαδικτύωσης όλο και περισσότερων συσκευών έχουν συμβάλει στην ανάπτυξη ενός ανερχόμενου πεδίου, του Διαδικτύου των Αντικειμένων (*Internet of Things*). Το *Internet of Things* αναφέρεται στη δημιουργία ενός ενιαίου διαδικτύου τρισεκατομμυρίων κόμβων, στο οποίο θα συνδέονται, αντίθετα με τα σημερινά δεδομένα, κάθε είδους αντικείμενα, από απλές καθημερινές συσκευές και αισθητήρες μέχρι super computers και computer clusters. Από τη σκοπιά της Εξόρυξης Γνώσης, η διαχείριση και ανάλυση του όγκου των δεδομένων που θα δημιουργήσει το *Internet of Things* είναι προφανές ότι δε μπορεί να πραγματοποιηθεί χρησιμοποιώντας τις υπάρχουσες τεχνικές και μεθόδους. Δημιουργείται λοιπόν η ανάγκη εύρεσης νέων αλγορίθμων που θα δώσουν λύση σε αναδυόμενα προβλήματα όπως ο εντοπισμός γεγονότων από την αλληλεπίδραση μεγάλου πλήθους συσκευών, η πραγματικού χρόνου γεωγραφική παρακολούθηση δισεκατομμυρίων αντικειμένων και η αποδοτική οργάνωση της ακατάπαυστης ροής δεδομένων που δημιουργούν τα συνδεδεμένα αντικείμενα στο διαδίκτυο. Τα δεδομένα που προκύπτουν από ένα τέτοιο δίκτυο είναι υψηλής διαστατικότητας λόγω της συμμετοχής πολλών μεταβλητών για την εξαγωγή χρήσιμων αποτελεσμάτων. Επίσης, ο συνδυασμός της συνεχούς ροής των δεδομένων και της εισαγωγής χωρικής πληροφορίας που σχετίζεται με τη θέση των αντικειμένων του δικτύου, προσδίδουν στα τελικά δεδομένα χωροχρονικό χαρακτήρα. Στόχος της διπλωματικής αυτής εργασίας είναι η μελέτη των προβλημάτων που προκύπτουν στην διαχείριση των δεδομένων από τους κόμβους του Internet of Things καθώς και η εξαγωγή χρήσιμης πληροφορίας από τέτοιου είδους δεδομένα.

Επιθυμητές γνώσεις: Εξόρυξη γνώσης, Βάσεις Δεδομένων, Πιθανότητες, Γραμμική Άλγεβρα, Επεξεργασία Σημάτων, Γλώσσες προγραμματισμού (Matlab,C++).

Ενδεικτική Βιβλιογραφία:

[1] Minnen, D., Isbell, C., Essa, I., and Starner, T. 2007. Detecting Subdimensional Motifs: An Efficient Algorithm for Generalized Multivariate Pattern Discovery. In Proceedings of the 2007 Seventh IEEE international Conference on Data Mining (October 28 - 31, 2007). ICDM. IEEE Computer Society, Washington, DC, 2007.

[2] Shen Bin, Liu Yuan, Wang Xiaoyi, Research on Data Mining Models for the Internet of Things, in IASP '10: International Conference on Image Analysis and Signal Processing, 2010.

### **9. Ανάλυση χαρτών έκφρασης γονιδίων και λειτουργίας τους**

Το συγκεκριμένο θέμα ασχολείται με τον έλεγχο της υπόθεσης ότι τα γονίδια με παρόμοιους χάρτες έκφρασης παρουσιάζουν παρόμοια λειτουργία. Προκειμένου να προσδιοριστεί η σχέση μεταξύ χαρτών γονιδιακής έκφρασης και γονιδιακής λειτουργίας μπορούν καταρχήν να εντοπιστούν γονίδια με παρόμοιους χάρτες έκφρασης και κατόπιν να ελεγχθεί η ομοιότητα των αντίστοιχων γονιδιακών λειτουργιών. Ο υπολογισμός της ομοιότητας των γονιδιακών χαρτών έκφρασης μπορεί να βασιστεί σε διάφορα χαρακτηριστικά τα οποία μπορούν να εξαχθούν από τους χάρτες ενώ η ομοιότητα των γονιδιακών λειτουργιών μπορεί να υπολογιστεί με βάση την μέση λειτουργική απόσταση της γονιδιακής οντολογίας. Για το συγκεκριμένο θέμα υπάρχει διαθέσιμο ένα σύνολο σύνολο δεδομένων, το οποίο περιέχει πληροφορίες για περισσότερα από 20.000 γονίδια. Μεταξύ άλλων η διπλωματική αυτή θα εστιάσει στην μελέτη της σχετικής βιβλιογραφίας, στην μελέτη και χρήση διαφόρων τεχνικών για εξαγωγή χαρακτηριστικών από τους χάρτες έκφρασης γονιδίων, στην μελέτη και χρήση διαφορετικών μετρικών ομοιότητας χαρτών έκφρασης και γονιδιακών λειτουργιών και στην μελέτη και χρήση της γονιδιακής οντολογίας (Gene Ontology).

Επιθυμητές γνώσεις: Εξόρυξη γνώσης, Βιοπληροφορική, Επεξεργασία Σημάτων, Επεξεργασία Εικόνας, Γλώσσες προγραμματισμού (Matlab, C, C++)

Ενδεικτική Βιβλιογραφία:

[1] Brown VM, Ossadtchi A, Khan AH, Cherry SR, Leahy RM, Smith DJ.: High-throughput imaging of brain gene expression. Genome Res, 2002. 12(2): p. 244-54.

[2] Velculescu, V.E., Zhang, L., Vogelstein, B., and Kinzler, K.W. 1995: Serial analysis of gene expression. Science 270, p.484-487.

### **10. Ανάπτυξη Εργαλείων Πρωτεομικής Ανάλυσης και Οπτικοποίησης Αποτελεσμάτων**

Περιγραφή:

Η πρωτεομική ανάλυση διακρίνεται σε δύο στάδια: (1) τον διαχωρισμό των πρωτεϊνών και (2) την αναγνώριση των πρωτεϊνών μέσω τεχνικών όπως η φασματομετρία μάζας. Οι κλασσικές προσεγγίσεις

πρωτεομικής ανάλυσης που συνήθως χρησιμοποιούνται στην πράξη είναι ο διαχωρισμός των πρωτεϊνών με διδιάστατη ηλεκτροφόρηση (2D – gel electrophoresis, 2DGE) ή υγρή χρωματογραφία (Liquid Chromatography - LC) και η ταυτοποίησή τους με τεχνικές φασματομετρίας μάζας (mass spectrometry). Στην διπλωματική αυτή θα μελετηθούν διάφορα λογισμικά πακέτα και εργαλεία που χρησιμοποιούνται στην πρωτεομική ανάλυση. Η μελέτη θα εστιάσει στις δυνατότητες των λογισμικών πακέτων ως προς τα στάδια της συγκέντρωσης και μετα-ανάλυσης των πρωτεομικών δεδομένων. Θα αναπτυχθούν εργαλεία λογισμικού για την προεπεξεργασία εικόνων πρωτεομικής ανάλυσης και την ανακάλυψη συσχετίσεων σε τέτοιες εικόνες με τελικό στόχο την σύγκριση των πρωτεομάτων διαφορετικών βιολογικών καταστάσεων (παθολογικό, φυσιολογικό) στοχεύοντας έτσι στον εντοπισμό πρωτεϊνών οι οποίες συμμετέχουν σε διαφορετικές φυσιολογικές καταστάσεις. Η διπλωματική αυτή θα ασχοληθεί επίσης με την οπτικοποίηση των αποτελεσμάτων της πρωτεομικής ανάλυσης.

Επιθυμητές γνώσεις: Εξόρυξη γνώσης, Βιοπληροφορική, Επεξεργασία Σημάτων, Επεξεργασία Εικόνας, Γλώσσες προγραμματισμού (Matlab, C, C++).

Ενδεικτική Βιβλιογραφία:

[1] D. Tsagktrasoulis, P. Zerefos, G. Loudos, A. Vlahou, M. Baumann, S. Kossida, “ 'Brukin2D': a 2D visualization and comparison tool for LC-MS data”, BMC Bioinformatics 2009, 10(Suppl 6):S12.

Συνεπιβλέποντες: Β. Μεγαλοοικονόμου, Σ. Κοσσίδα (ΙΙΒΕΑΑ, Ακαδημία Αθηνών)

## ***11. Ενοποίηση κατηγοριοποιητών***

Περιγραφή:

Η ακρίβεια ενός μοντέλου κατηγοριοποίησης είναι σημαντική διότι αντικατοπτρίζει την αξιοπιστία του κατηγοριοποιητή όταν εφαρμοστεί σε νέα δεδομένα. Για την αύξηση της ακρίβειας των κατηγοριοποιητών έχουν προταθεί διάφορες τεχνικές όπως οι στρατηγικές εμφωλίας (bagging) και ενδυνάμωσης (boosting). Η τεχνική της εμφωλίας στηρίζεται στην πλειοψηφούσα απόφαση των επιμέρους κατηγοριοποιητών ενώ η τεχνική της ενδυνάμωσης αποδίδει συντελεστές βαρύτητας στους επιμέρους κατηγοριοποιητές ώστε ο κάθε ένας τους να συμμετέχει στην τελική απόφαση ανάλογα με την ακρίβειά του. Ο σκοπός της παρούσας διπλωματικής εργασίας είναι η βιβλιογραφική καταγραφή των υπαρχόντων τεχνικών για την ενοποίηση κατηγοριοποιητών, η υλοποίηση των μεθοδολογιών εμφωλίας και ενδυνάμωσης καθώς και η διερεύνηση νέων παρόμοιων τεχνικών. Τέλος, προτείνεται η εφαρμογή των ενοποιημένων μοντέλων σε πραγματικά δεδομένα με σκοπό την μεγιστοποίηση της ακρίβειας των επιμέρους κατηγοριοποιητών.

Επιθυμητές γνώσεις: Εξόρυξη γνώσης, Ανάκτηση πληροφορίας, Γλώσσες προγραμματισμού (C, C++, Matlab).

## ***12. Μελέτη και κατηγοριοποίηση ιατρικών εικόνων με χρήση τεχνικών ανάλυσης υφής***

Περιγραφή:

Η ανάλυση υφής αποτελεί μια από τις σημαντικότερες τεχνικές ανάλυσης ιατρικών εικόνων για την εξαγωγή χρήσιμης ιατρικής πληροφορίας. Αρκετές μέθοδοι έχουν παρουσιαστεί στη διεθνή βιβλιογραφία οι οποίες αποσκοπούν στη βελτίωση της ικανότητας ανίχνευσης παθολογικών ευρημάτων σε ιατρικές

εικόνες αλλά και στην υποβοήθηση της αξιολόγησης παθολογικών ευρημάτων κατά τη διαγνωστική διαδικασία μέσα από την εξόρυξη χαρακτηριστικών υφής. Η παρούσα διπλωματική περιλαμβάνει εκτενή βιβλιογραφική ανασκόπηση και παρουσίαση των βασικών τεχνικών ανάλυσης υφής με έμφαση στην ανάλυση ιατρικών εικόνων. Οι πηγές πληροφορίας θα προέρχονται κυρίως από το διαδίκτυο (σχετικές ιστοσελίδες, δημοσιευμένες εργασίες σε ηλεκτρονική μορφή κ.λπ.). Η εργασία περιλαμβάνει επίσης την ανάπτυξη αλγορίθμων για την ανάλυση υφής σε ιατρικές εικόνες σε περιβάλλον προγραμματισμού Matlab, C++, Java.

Επιθυμητές γνώσεις: Εξόρυξη γνώσης, Βιοπληροφορική, Επεξεργασία Σημάτων, Επεξεργασία Εικόνας, Γλώσσες προγραμματισμού (Matlab, C, C++).

Ενδεικτική Βιβλιογραφία:

[1] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural Features of Image Classification," IEEE Transactions on Systems, Man and Cybernetics, Vol. 3- 6, pp. 610-621, 1973.

[2] K. Sikka, T.M. Deserno. "Segmentation of Ultrasound Image Based on Texture Feature and Graph Cut", CSSE, Vol. 1, pp.795-798, 2008.

[3] H. Li, M.L. Giger, O.I. Olopade, etc, "Computerized texture analysis of mammographic parenchymal patterns of digitized mammograms," Acad Radiol, Vol. 12, pp. 863-873, 2005.

[4] A. Bhattacharya, V. Ljosa, J.-Y. Pan, M. R. Verardo, H. Yang, C. Faloutsos and A.K. Singh, "ViVo: Visual vocabulary construction for mining biomedical images" Proc. Fifth IEEE International Conference on Data Mining (ICDM), pp. 50-57, Nov. 2005.

### ***13. Μελέτη και ανάπτυξη τεχνικών κατάτμησης ιατρικών εικόνων***

Περιγραφή:

Τα τελευταία χρόνια, ο ρόλος και η συμβολή της ιατρικής απεικόνισης στη διαγνωστική και θεραπευτική διαδικασία έχει ενισχυθεί σημαντικά λόγω της προόδου της επιστήμης των υπολογιστών. Ένα κρίσιμο ζήτημα όταν αναλύουμε ιατρικές εικόνες είναι η ακριβής ανίχνευση του περιγράμματος των περιοχών ενδιαφέροντος. Αυτή η διαδικασία, ονομαζόμενη ως τμηματοποίηση ή κατάτμηση (image segmentation), βασίζεται στην αναγνώριση των ευδιάκριτων ορίων μεταξύ διαφορετικών ιστών. Στα πλαίσια της παρούσας διπλωματικής εργασίας θα μελετηθούν μεθοδολογίες της σύγχρονης βιβλιογραφίας που έχουν προταθεί για το ζήτημα της κατάτμησης των ιατρικών εικόνων. Επίσης, η διπλωματική εργασία περιλαμβάνει την ανάπτυξη μιας πρωτότυπης τεχνικής κατάτμησης ή την επέκταση υπάρχουσών τεχνικών σε περιβάλλον προγραμματισμού Matlab C++, ή Java.

Επιθυμητές γνώσεις: Εξόρυξη γνώσης, Βιοπληροφορική, Επεξεργασία Σημάτων, Επεξεργασία Εικόνας, Γλώσσες προγραμματισμού (Matlab, C, C++, Java).

Ενδεικτική Βιβλιογραφία:

[1] Y. Boykov, M. P. Jolly. "Interactive Graph cuts for optimal Boundary and Region Segmentation of Object in N-D Images", ICCV, Vol. I, pp.105-112, 2001.

[2] E.N. Mortensen and W.A. Barrett. "Interactive segmentation with intelligent scissors," Graphical Models and Image Processing, Vol.60, pp.349-384, 1998.

[3] Y. Zheng, S. Baloch, S. Englander, etc, "Segmentation and Classification of Breast Tumor Using Dynamic Contrast-Enhanced MR Images," MICCAI, Vol. 4792, pp. 393–401, 2007.

#### **14. Πιθανοτική ανίχνευση κόμβων διακλάδωσης σε εικόνες δενδρικών δομών**

Περιγραφή:

Οι δομές διακλάδωσης είναι παρούσες σε ποικίλα βιοϊατρικά πλαίσια, συμπεριλαμβανομένων των αγγειακών, νευρικών, βρογχικών, και γαλακτοφόρων δικτύων του ανθρώπινου σώματος. Πολλές ιδιότητες αυτών των δομών έχουν μελετηθεί από ερευνητές και οι αλλαγές αυτών των δομών έχουν συνδεθεί με αλλαγμένη λειτουργία ή/και παθολογία. Εντούτοις, αν και οι δομές αυτές εμφανίζονται συχνά στη φύση και οι κανόνες ανάπτυξής τους έχουν μελετηθεί αρκετά, υπάρχουν ακόμα πολλά ανοιχτά προβλήματα στην κατάτμηση και ανάλυση τέτοιων δομών: οι εικόνες των φυσικών και βιοϊατρικών δομών διακλάδωσης περιλαμβάνουν συχνά σύνθετα περίχωρα (surroundings) που μπορεί μερικώς να κρύβουν τις δομές διακλάδωσης. Οι προβολές τρισδιάστατων δομών διακλάδωσης μπορούν επίσης να προκαλέσουν τις επικαλύψεις μεταξύ των κλάδων λόγω της απώλειας βάθους. Επιπλέον, οι μονάδες που χρησιμοποιούνται για την απόκτηση των εικόνων διαφέρουν στο βαθμό ευαισθησίας τους στην απεικόνιση του δέντρου. Σε ορισμένες μορφές απεικόνισης η τοπολογία διακλάδωσης μιας δενδρικής δομής μπορεί να είναι μόλις ορατή από μια εικόνα. Το μέγιστο βάθος της δενδρικής δομής που συλλαμβάνεται στην εικόνα μπορεί επίσης να ποικίλει, ανάλογα με τη δυνατότητα των οργάνων να εξάγουν μια δομή διακλάδωσης από τα σύνθετα περίχωρά της. Σε αυτή τη διπλωματική θα εξετασθούν θεωρητικά και πειραματικά διάφορες τεχνικές από την περιοχή της μηχανικής μάθησης που μπορούν να εφαρμοστούν στην πιθανοτική ανίχνευση κόμβων διακλάδωσης σε εικόνες δενδρικών δομών καθώς και δικτύων.

Επιθυμητές γνώσεις: Εξόρυξη γνώσης, Μηχανική μάθηση, Ανάκτηση πληροφορίας, Επεξεργασία Σημάτων, Επεξεργασία Εικόνας, Γλώσσες προγραμματισμού (C, C++, Matlab),

Ενδεικτική Βιβλιογραφία:

[1] H. Ling, M. Barnathan, V. Megalooikonomou, P. Bakic, A. Maidment, «Probabilistic Branching Node Detection Using Hybrid Local Features», Proceedings of the 6th IEEE International Symposium on Biomedical Imaging (ISBI), Boston, MA, 2009, pp. 233-236.

#### **15. Μελέτη και υλοποίηση δομών ευρετηρίου για την διαχείριση και αποδοτική ανάκτηση πολυδιάστατων ακολουθιών**

Περιγραφή:

Τα τελευταία χρόνια, ο μεγάλος όγκος των πολυδιάστατων ακολουθιών (χρονοσειρών), που προέρχονται από πολλούς διαφορετικούς κλάδους της επιστήμης και της τεχνολογίας, έχει στρέψει το ενδιαφέρον των ερευνητών στην εύρεση τρόπων για την αποδοτική οργάνωση και διαχείρισή τους. Χαρακτηριστικά παραδείγματα τέτοιων πολυδιάστατων δεδομένων αποτελούν το βίντεο (ακολουθία από frames όπου το καθένα μπορεί να περιλαμβάνει διάφορα χαρακτηριστικά όπως χρώμα, σχήμα κτλ), οι ιατρικές εικόνες/ιατρικά σήματα (π.χ., ακολουθίες λειτουργικής μαγνητικής τομογραφίας (fMRI)), τα

χωροχρονικά δεδομένα που λαμβάνονται από αισθητήρες (π.χ., για περιβαλλοντικές μελέτες ή μετεωρολογικές προβλέψεις) και πολλά άλλα. Βασική προϋπόθεση για την εξόρυξη χρήσιμης πληροφορίας από βάσεις πολυδιάστατων ακολουθιών είναι η οργάνωσή τους με τέτοιο τρόπο ώστε να επιτρέπονται γρήγορες αναζητήσεις. Η δημιουργία ενός ευρετηρίου που να μπορεί να απορρίπτει όλα τα άσχετα ως προς το ερώτημα δεδομένα, ενώ ταυτόχρονα να υποδεικνύει μόνο τις πιθανές απαντήσεις αποτελεί μια κλασσική τεχνική για την ανάκτηση τέτοιου είδους δεδομένων. Για να επιτευχθεί αυτό θα πρέπει να καθοριστεί μία μετρική απόστασης/ομοιότητας ικανής να αποτυπώσει την απόσταση/ομοιότητα των χρονοσειρών σε όλες τις διαστάσεις.

Στα πλαίσια αυτής της διπλωματικής, θα μελετηθούν διάφορες μετρικές ομοιότητας πολυδιάστατων ακολουθιών που έχουν προταθεί στη βιβλιογραφία. Επίσης, θα μελετηθούν δομές δεδομένων και τεχνικές δεικτοδότησης που μπορούν να χρησιμοποιηθούν σε τέτοιου είδους δεδομένα. Στη συνέχεια, θα επιλεχθούν ορισμένες από αυτές για να αξιολογηθούν πειραματικά σε πραγματικά δεδομένα.

Επιθυμητές γνώσεις: Βάσεις Δεδομένων, Εξόρυξη Δεδομένων, Δομές Δεδομένων, Ανάκτηση Πληροφορίας, Γλώσσες Προγραμματισμού (Matlab, C)

Ενδεικτική Βιβλιογραφία:

[1] A. Guttman. R-trees: a dynamic index structure for spatial searching. Proceedings of ACM SIGMOD Int'l Conference on Management of Data, pages 47-57, Boston, Massachusetts, June, 1984.

[2] M. Vlachos, M. Hadjieleftheriou, D. Gunopulos, and E. Keogh. Indexing multidimensional time-series. The VLDB Journal, 15:1–20, 2006. 10.1007/s00778-004-0144-2.

[3] L.J. Latecki, Qiang Wang, S. Koknar-Tezel, V. Megalooikonomou. Optimal Subsequence Bijection. ICDM 2007. Seventh IEEE International Conference on Data Mining, 2007.

## ***16. Ανάλυση δεδομένων απο τον ανθρώπινο εγκέφαλο***

Περιγραφή:

Αντικείμενο αυτής της εργασίας είναι η μελέτη τεχνικών για την ανάλυση δεδομένων που προέρχονται από συστήματα απεικόνισης της λειτουργίας του ανθρώπινου εγκεφάλου όπως το ηλεκτροεγκεφαλογράφημα (EEG). Τα δεδομένα που μελετώνται προέρχονται από διαφορετικές περιοχές του εγκεφάλου και επίσης εξελίσσονται χρονικά. Σκοπός των τεχνικών ανάλυσης είναι η ανίχνευση συγκεκριμένων μορφών αυτών των σημάτων (όπως για παράδειγμα τα συμπλέγματα -K, ή οι άτρακτοι στο EEG), η ανακάλυψη συσχετίσεων μεταξύ αυτών, η ανακάλυψη ομοιοτήτων, προτύπων ή κανόνων συσχετίσεων ακολουθιών (sequence association rules), η ομαδοποίηση, η ταξινόμηση τους, κ.λ.π. Η αναπαράσταση επίσης αυτών των πολυδιάστατων χρονοσειρών αποτελεί ένα άλλο σημαντικό πρόβλημα που θα μελετηθεί σε αυτή την διπλωματική εργασία μαζί με το θέμα της ανάλυσής τους. Στα πλαίσια αυτής της διπλωματικής θα μελετηθούν τεχνικές που έχουν προταθεί στην βιβλιογραφία και θα υλοποιηθούν κάποιες απο αυτές. Προαιρετικά μπορεί να σχεδιαστεί και να υλοποιηθεί μια νέα τεχνική που να βελτιώνει σε κάποιο τομέα τις υπάρχουσες τεχνικές.

Επιθυμητές γνώσεις: Εξόρυξη γνώσης, Ανάκτηση πληροφορίας, Επεξεργασία Σημάτων, Βάσεις δεδομένων, Γλώσσες προγραμματισμού (C, C++, Matlab)

Συνεπιβλέποντες: Β. Μεγαλοοικονόμου, Γ. Κωστόπουλος (Εργ. Νευροφυσιολογίας)

Ενδεικτική Βιβλιογραφία:

[1] I. Bankman, V. Sigillito, R. Wise, and P. Smith. Feature based detection of the k-complex wave in the human electroencephalogram using neural networks. Biomedical Engineering, IEEE Transactions on, 39(12):1305 –1310, dec. 1992.

[2] S. Devuyt, T. Dutoit, P. Stenuit, and M. Kerkhofs. Automatic k-complexes detection in sleep eeg recordings using likelihood thresholds. In Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE, pages 4658 –4661, 31 2010-sept. 4 2010.

### ***17. Σχεδιασμός και Υλοποίηση Γεωγραφικά Προσανατολισμένων Ευρετηρίων Δεδομένων Διαδικτύου***

Περιγραφή:

Μια σημαντική παράμετρος της αναζήτησης πληροφορίας στον Παγκόσμιο Ιστό είναι η τοπικότητα των αποτελεσμάτων ανάκτησης. Μέχρι σήμερα, το πρόβλημα της ανάκτησης πληροφορίας για γεωγραφικά δεδομένα αντιμετωπίζεται με τεχνικές αυτόματου εντοπισμού της γεωγραφικής διάστασης των ερωτημάτων. Ωστόσο, η αναγνώριση γεωγραφικών ερωτημάτων απαιτεί χρόνο και είναι μια κοπιαστική διεργασία εφόσον προϋποθέτει την επεξεργασία μεγάλου όγκου δεδομένων για την επίλυση του γεωγραφικού προσδιορισμού των λέξεων στα ερωτήματα των χρηστών. Επιπλέον, τα γεωγραφικά ερωτήματα ακόμα κι αν αναγνωριστούν επιτυχώς δεν απαντώνται εξ' ορισμού στα αποτελέσματα ανάκτησης εκτός κι αν αναζητηθούν σε ένα γεωγραφικά ενημερωμένο ευρετήριο δεδομένων.

Σκοπός αυτής της διπλωματικής εργασίας είναι να οριστεί το πρόβλημα της αναζήτησης γεωγραφικά προσανατολισμένης πληροφορίας από την οπτική των μηχανών αναζήτησης και να υλοποιηθεί μια τεχνική για τον χειρισμό γεωγραφικών ερωτημάτων, η οποία θα ενσωματώνει ένα χωρικό ευρετήριο. Επιπλέον, θα πρέπει να σχεδιαστούν και να υλοποιηθούν τεχνικές βαθμολόγησης της συσχέτισης (relevance) μεταξύ των ερωτημάτων αναζήτησης και των δεδομένων ανάκτησης, η οποία θα λαμβάνει υπόψη χωρικές παραμέτρους στα στοιχεία των URLs και του κειμένου αγκύρωσης των σελίδων για την ταξινόμηση των αποτελεσμάτων ανάκτησης.

Επιθυμητές γνώσεις: Δομές Δεδομένων, Ανάκτηση Πληροφορίας, Αλγόριθμοι, Βάσεις Δεδομένων I & II, Γλωσσική Τεχνολογία, Εξόρυξη Δεδομένων και Αλγόριθμοι Μάθησης, Τεχνολογίες Διαδικτύου

Συνεπιβλέποντες: Β. Μεγαλοοικονόμου, Σ. Στάμου

### ***18. Εξόρυξη γνώσης από αποθήκες πληροφορίας σε κινητές συσκευές για την αυτόματη αναγνώριση μελλοντικού πλαισίου δραστηριοτήτων του χρήστη***

Περιγραφή:

Σε κάθε κινητή συσκευή βρίσκεται πληθώρα αποθηκευμένων προσωπικών πληροφοριών, οι οποίες μπορούν να δώσουν στην συσκευή επίγνωση του πλαισίου (context) υπό το οποίο τελεί ο χρήστης, αλλά και πληροφορίες των πιθανών δραστηριοτήτων του χρήστη στο άμεσο μέλλον, οι οποίες μπορεί να περιλαμβάνουν χρήση της συσκευής. Για παράδειγμα, ο χρήστης μπορεί να έχει προσκληθεί σε κάποια εκδήλωση μέσω κάποιου κοινωνικού δικτύου και να έχει αποδεχθεί την πρόσκληση. Όμως δεν έχει

περάσει το γεγονός στο ηλεκτρονικό του ημερολόγιο. Κάποιος φίλος του (ο οποίος και ο ίδιος έχει αποδεχθεί την πρόσκληση μέσω κοινωνικού δικτύου) έχει στείλει ένα sms ρωτώντας τι ώρα θα πάνε στο πάρτυ, και ο χρήστης έχει απαντήσει σχετικά. Μια κινητή συσκευή η οποία έχει την ικανότητα να κατηγοριοποιήσει και να συσχετίσει αυτόματα τα διάφορα κομμάτια πληροφορίας από τις ετερογενείς πηγές, μπορεί να τα συνδυάσει με άλλη γνώση (π.χ. την τοποθεσία του χρήστη) ώστε να προσφέρει χρήσιμες υπηρεσίες, όπως το κοντινότερο ανοιχτό ζαχαροπλαστείο, την ώρα αναχώρησης από την στάση του λεωφορείου που πιθανώς να θέλει να πάρει ο χρήστης για να πάει στην εκδήλωση, την συχνότητα των λεωφορείων σε περίπτωση που το χάσει κτλ.

Η εργασία αυτή αφορά την ανάπτυξη λογισμικού για την συγκέντρωση ετερογενούς πληροφορίας (χωρικής, χρονικής, δραστηριότητας, κοινωνικής) μέσω κινητών συσκευών (Android), την διαχείριση, ανάλυση (πιθανοτική κατηγοριοποίηση) και συσχέτισή της με άλλη πληροφορία και 1) την χρήση της συνδυασμένης γνώσης σαν βάση για την απόκτηση περαιτέρω πληροφορίας από διάχυτες πηγές πληροφόρησης (π.χ. web) ή 2) την παρουσίαση των συσχετισμένων πληροφοριών στον χρήστη με οπτικοποιημένη μορφή.

Επιθυμητές γνώσεις: Εξόρυξη γνώσης, Μηχανική μάθηση, Ανάκτηση πληροφορίας, Γλώσσες προγραμματισμού (Java),

Ενδεικτική Βιβλιογραφία:

- [1] Dexter H. Hu, Fan Dong, Cho-Li Wang, "A Semantic Context Management Framework on Mobile Device," Embedded Software and Systems, Second International Conference on, pp. 331-338, 2009 International Conference on Embedded Software and Systems, 2009
- [2] Claudio Bettini, Oliver Brdiczka, Karen Henriksen, Jadwiga Indulska, Daniela Nicklas, Anand Ranganathan, Daniele Riboni, A survey of context modelling and reasoning techniques, Pervasive and Mobile Computing, Volume 6, Issue 2, April 2010, Pages 161-180, ISSN 1574-1192, 10.1016/j.pmcj.2009.06.002.
- [3] Karen Church and Barry Smyth. 2009. Understanding the intent behind mobile information needs. In Proceedings of the 14th international conference on Intelligent user interfaces (IUI '09). ACM, New York, NY, USA, 247-256. <http://doi.acm.org/10.1145/1502650.1502686>.

Συνεπιβλέποντες: Β. Μεγαλοοικονόμου, Α. Κομνηνός

## ***19. Ανάλυση δεδομένων αισθητήρων κινητών συσκευών για την αυτόματη αναγνώριση φυσικής δραστηριότητας.***

Περιγραφή:

Ένα από τα μεγαλύτερα προβλήματα στην παρακολούθηση ιατρικών περιπτώσεων είναι η ικανότητα των θεραπειών να γνωρίζουν τα πραγματικά επίπεδα φυσικής δραστηριότητας των ασθενών τους. Υπάρχουν ειδικές συσκευές (πολλαπλά επιταχυνσιόμετρα) οι οποίες φέρονται από τους χρήστες για ορισμένα χρονικά διαστήματα με σκοπό να χτιστεί μια εικόνα της φυσικής τους δραστηριότητας, όμως αυτές είναι περιορισμένων δυνατοτήτων και επίσης δεν μπορούν να δανείζονται στους ασθενείς για μεγάλα χρονικά διαστήματα ή να φοριούνται συνεχώς. Οι περισσότερες κινητές συσκευές σήμερα διαθέτουν αισθητήρες όπως επιταχυνσιόμετρα τριών διαστάσεων, μαγνητικές πυξίδες και GPS. Σκοπός της πτυχιακής αυτής είναι ο συνδυασμός πληροφοριών από τις διάφορες αυτές πηγές αλλά και έμμεσες πηγές (όπως π.χ ο ρυθμός ανίχνευσης νέων δικτύων wi-fi που μπορεί να υποδηλώνει κάποιον ταχέως κινούμενο χρήστη)

για την εξαγωγή συμπερασμάτων για την φυσική δραστηριότητα (περπάτημα, τρέξιμο, γυμναστική, άθληση κτλ) στην οποία εμπλέκεται ο χρήστης κατά την διάρκεια της ημέρας. Επίσης, σκοπός της διπλωματικής δεν είναι μόνο η αναγνώριση και κατηγοριοποίηση έντονης φυσικής δραστηριότητας, αλλά και ήπιων δραστηριοτήτων όπως μετακίνηση από δωμάτιο σε δωμάτιο, κάθισμα ή ανασήκωμα από καθιστή ή οριζόντια στάση του σώματος κτλ. Η υλοποίηση προβλέπεται να γίνεται σε κινητή συσκευή με το λειτουργικό Android.

Επιθυμητές γνώσεις: Επεξεργασία Σημάτων, Γλώσσες προγραμματισμού (Java),

Ενδεικτική Βιβλιογραφία:

- [1] Yu-Jin Hong, Ig-Jae Kim, Sang Chul Ahn, Hyoung-Gon Kim, Mobile health monitoring system based on activity recognition using accelerometer, Simulation Modelling Practice and Theory, Volume 18, Issue 4, April 2010, Pages 446-455, ISSN 1569-190X, 10.1016/j.simpat.2009.09.002.
- [2] Tomas Brezmes, Juan-Luis Gorricho and Josep Cotrina (2009): Activity Recognition from Accelerometer Data on a Mobile Phone, in Distributed Computing, Artificial Intelligence, Bioinformatics, Soft Computing, and Ambient Assisted Living, Springer Lecture Notes in Computer Science, 2009, Volume 5518/2009, 796-799, DOI: 10.1007/978-3-642-02481-8\_120
- [3] Jennifer R. Kwapisz, Gary M. Weiss, and Samuel A. Moore. 2011. Activity recognition using cell phone accelerometers. SIGKDD Explor. Newsl. 12, 2 (March 2011), 74-82. DOI=10.1145/1964897.1964918 <http://doi.acm.org/10.1145/1964897.1964918>
- [4] Alberto G. Bonomi (2011) Physical Activity Recognition Using a Wearable Accelerometer, in Sensing Emotions, Philips Research Book Series, 2011, Volume 12, 41-51, DOI: 10.1007/978-90-481-3258-4\_3

Συνεπιβλέποντες: Β. Μεγαλοοικονόμου, Α. Κομνηνός

## ***20. Κατανεμημένο μοντέλο περιρρέουσας νοημοσύνης σε προσωπικά περιβάλλοντα, για την αναγνώριση αλληλεπίδρασης ετερογενών συσκευών με το χρήστη και την εφαρμογή βέλτιστων στρατηγικών απόσπασης της προσοχής του***

Περιγραφή:

Όσοι εργάζονται σε κάποιο γραφείο δημιουργούν ένα προσωπικό χώρο πληροφορίας, ο οποίος περιλαμβάνει (εκτός από αρκετό χαρτί) διάφορες ηλεκτρονικές συσκευές. Μέσω αυτών οι εργαζόμενοι δέχονται ροές πληροφοριών ή δημιουργούν οι ίδιοι τέτοιες ροές. Έτσι το τηλέφωνο μπορεί να ειπωθεί σαν μια συσκευή εισόδου-εξόδου ηχητικής πληροφορίας, το ραδιόφωνο σαν μια πηγή εισόδου, ο υπολογιστής και το κινητό σαν πηγές εισόδου-εξόδου πολυμεσικής πληροφορίας με πολυτροπικά μέσα αλληλεπίδρασης, ακόμα και η πόρτα ενός γραφείου μπορεί να θεωρηθεί σαν κάποια πηγή εισόδου πληροφορίας (θεωρώντας τους ανθρώπους που μπαίνουν σαν πηγές). Ορισμένες φορές η ροή των πληροφοριών προς τον χρήστη είναι τόσο μεγάλη ώστε ο ίδιος να υπερφορτώνεται και να αδυνατεί να διαχειριστεί τον όγκο, οδηγούμενος σε συμπεριφορές αποκοπής των ροών οι οποίες μπορεί να οδηγήσουν σε απώλεια πληροφορίας ή σημαντικών γεγονότων. Επί παραδείγματι, όταν μιλάμε στο τηλέφωνο συνήθως χαμηλώνουμε τη μουσική, στρέφουμε το βλέμμα μακριά από την οθόνη ή αν κάποιος μπει στο γραφείο μας την ώρα εκείνη του κάνουμε κάποια χειρονομία να καθήσει ή να φύγει. Στο μεταξύ, κάποια

γεγονότα όπως ειδοποίηση για κάποιο email ή κάποιο μήνυμα στο κινητό μπορεί να περάσουν απαρατήρητα και να περάσει πολλή ώρα μέχρι να θυμηθούμε ότι πρέπει να δράσουμε γι'αυτά. Σκοπός της διπλωματικής είναι να ερευνήσει αν είναι δυνατόν να δημιουργηθεί ένα ad-hoc μοντέλο περιρρέουσας νοημοσύνης μεταξύ των ετερογενών συσκευών που βρίσκονται σε ένα γραφείο και οι οποίες αποτελούν ένα τοπικό internet of things, ώστε να μειώνεται στο ελάχιστο ο ανταγωνισμός μεταξύ τους για την προσοχή του χρήστη με τέτοιο τρόπο που ο χρήστης να μπορεί να αφοσιωθεί μεν στην δραστηριότητα που είναι απασχολημένος αλλά αφ' ετέρου να μπορεί πολύ εύκολα με μία ματιά, αναλόγως τη συσκευή που χρησιμοποιεί, να έχει πολυτροπική ενημέρωση για τις ροές πληροφοριών που πιθανώς έχει χάσει από τις άλλες συσκευές. Η διπλωματική αυτή μπορεί να υλοποιηθεί με την επαύξηση πραγματικών συσκευών με κατάλληλο hardware ή την εξομοίωση της συμπεριφοράς ενός τέτοιου περιβάλλοντος.

Επιθυμητές γνώσεις: Εξόρυξη γνώσης, Μηχανική μάθηση, Ανάκτηση πληροφορίας, Γλώσσες προγραμματισμού (C, C++, Java),

Ενδεικτική Βιβλιογραφία:

- [1] Jan-Patrick Elsholz, Guido de Melo, Marc Hermann, Michael Weber, Designing an extensible architecture for Personalized Ambient Information, Pervasive and Mobile Computing, Volume 5, Issue 5, October 2009, Pages 592-605, ISSN 1574-1192, 10.1016/j.pmcj.2009.06.005.
- [2] Martijn H. Vastenburg, David V. Keyson, Huib de Ridder, Considerate home notification systems: A user study of acceptability of notifications in a living-room laboratory, International Journal of Human-Computer Studies, Volume 67, Issue 9, September 2009, Pages 814-826, ISSN 1071-5819, 10.1016/j.ijhcs.2009.06.002.
- [3] Edward R. Sykes, Interruptions in the workplace: A case study to reduce their effects, International Journal of Information Management, Volume 31, Issue 4, August 2011, Pages 385-394, ISSN 0268-4012, 10.1016/j.ijinfomgt.2010.10.010.

Συνεπιβλέποντες: Β. Μεγαλοοικονόμου, Α. Κομνηνός

## **21. Μοντέλα πρόβλεψης σεισμικότητας, ανάλυση και εφαρμογή στον Ελληνικό χώρο**

Περιγραφή:

Βασικός στόχος της Σεισμολογίας, πέρα από την παρατήρηση της κατανομής των σεισμών στο χώρο και στο χρόνο είναι και η πρόγνωση των σεισμών. Αν και ο στόχος της πρόγνωσης είναι ακόμα πολύ δύσκολο να επιτευχθεί εντούτοις έχουν προταθεί μοντέλα πρόβλεψης της σεισμικότητας τα οποία βασίζονται σε σεισμικούς καταλόγους (κατανομή στο χώρο και στο χρόνο των σεισμικών μεγεθών). Τα μοντέλα αυτά βασίζονται σε κάποιες παραδοχές για τη γένεση των σεισμικών γεγονότων (π.χ. μοντέλο ETAS, Epidemic-Type Aftershock Sequences στο μοντέλο των Gutenberg - Richter, Omori, αλληλεπίδραση σεισμών κλπ). Στα πλαίσια της διπλωματικής, θα μελετηθεί η δυνατότητα εφαρμογής μεθόδων/μοντέλων πρόβλεψης της σεισμικότητας χρησιμοποιώντας δεδομένα του Ελληνικού καταλόγου (<http://www.gein.noa.gr>, <http://geophysics.geo.auth.gr/ss/>, <http://seismo.geology.upatras.gr/>).

Επιθυμητές γνώσεις: Βάσεις Δεδομένων, Εξόρυξη Δεδομένων και Αλγόριθμοι Μάθησης, Ανάκτηση Πληροφορίας, Γλώσσες Προγραμματισμού (Matlab, C)

Ενδεικτική Βιβλιογραφία:

[1] Jordan, T. H. (2006), Earthquake predictability, brick by brick, Seismol. Res. Lett., 77, 3-6.

[2] Lombardi, A., & Marzocchi, W. (2010). The ETAS model for daily forecasting of Italian seismicity in the CSEP experiment. Annals Of Geophysics, 53(3), 155-164. doi:10.4401/ag-4848

[3] Web Pages: <http://www.cseptesting.org/>

Συνεπιβλέποντες: Β. Μεγαλοοικονόμου, Ε. Σώκος (Τμήμα Γεωλογίας)

---

Επιπλέον πιθανά θέματα για διπλωματική εργασία μπορούν να διερευνηθούν σε συνενόηση με τον διδάσκοντα.

Διευκρινήσεις για τα θέματα δίνονται από τον διδάσκοντα ([vasilis@ceid.upatras.gr](mailto:vasilis@ceid.upatras.gr)).

Αιτήσεις με email στην ηλεκτρονική διεύθυνση [vasilis@ceid.upatras.gr](mailto:vasilis@ceid.upatras.gr)

- απλή αίτηση όπου θα αναγράφονται το πολύ μέχρι 2 θέματα με σειρά προτίμησης
- αντίγραφο αναλυτικής βαθμολογίας (scanned αφού η αίτηση θα σταλεί ηλεκτρονικά).