Διπλωματικές Εργασίες Ακαδημαϊκού Έτους 2013-2014

Β. Μεγαλοοικονόμου, Καθηγητής

1. Εύρεση ομοιότητας δενδρικών δομών μέσω μοντελοποίησης ως ακολουθίες

Η δενδρική δομή (tree structure) είναι ένας τρόπος να παραστήσουμε γραφικά την ιεραρχία μιας δομής. Στην παρούσα διπλωματική εργασία θα μελετηθούν μεθοδολογίες αναπαράστασης των δενδρικών δομών ως ακολουθίες συμβόλων οι οποίες κωδικοποιούν μοναδικά τις σχέσεις γονιού - παιδιού. Επίσης, θα δοθεί έμφαση στην εύρεση παρόμοιων δενδρικών δομών αξιοποιώντας μετρικές ομοιότητας που εφαρμόζονται για την εύρεση ομοιότητας μεταξύ ακολουθιών. Οι μεθοδολογίες που θα αναπτυχθούν θα αξιολογηθούν σε διάφορα σύνολα δεδομένων που περιέχουν δενδρικές δομές όπως ιατρικές εικόνες και δομημένα δεδομένα που παρουσιάζονται με τη μορφή δένδρου.

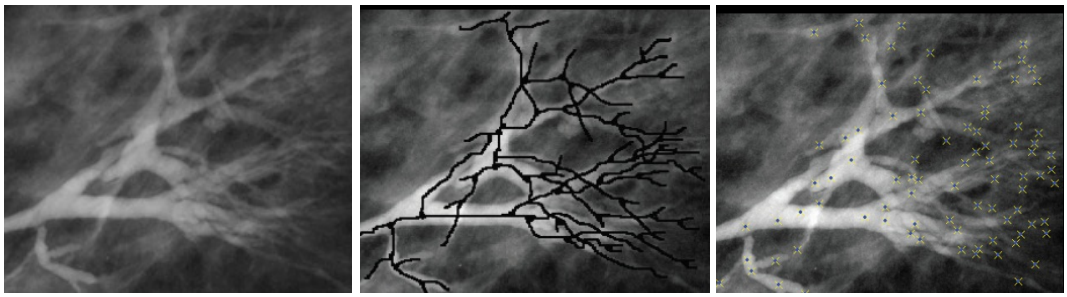
Επιθυμητές γνώσεις: Εξόρυξη γνώσης, Βιοπληροφορική, Επεξεργασία Σημάτων, Επεξεργασία Εικόνας, Γλώσσες προγραμματισμού (Matlab, C, C++).

Ενδεικτική Βιβλιογραφία:

[1] V. Megalooikonomou, M. Barnathan, D. Kontos, P. R. Bakic, A. D.A. Maidment, “A Representation and Classification Scheme for Tree-like Structures in Medical Images: Analyzing the Branching Pattern of Ductal Trees in X-ray Galactograms”, IEEE Transactions on Medical Imaging, Vol. 28, Issue 4, pp. 487-493, 2009.

2. Ανίχνευση χαρακτηριστικών σημείων σε ιατρικές εικόνες

Ο όρος "ανίχνευση χαρακτηριστικών" (feature detection) αναφέρεται στα τμήματα εκείνα μιας εικόνας στα οποία επικεντρώνει την προσοχή της η ανθρώπινη όραση, όταν πρωτοεκτίθεται σε μια στατική εικόνα. Οι κυριότερες κατηγορίες χαρακτηριστικών περιλαμβάνουν τις ακμές (edges), τις γωνίες (corners) και τις κορυφογραμμές (ridges). Η παρούσα διπλωματική θα εστιάσει στον εντοπισμό σημείων διακλάδωσης (σημεία μιας δενδρικής δομής στα οποία ξεκινά η διάσπαση ενός ιεραρχικά υψηλότερου κλάδου σε δύο ή περισσότερους ιεραρχικά χαμηλότερους κλάδους) σε ιατρικές εικόνες που απεικονίζουν δενδρικές δομές του ανθρώπινου σώματος (Εικ. 1). Η εργασία μπορεί επίσης να συνδυαστεί με την επέκταση και την εφαρμογή ήδη υλοποιημένων τεχνικών σε νέα σύνολα δεδομένων.



Εικόνα 1: Παράδειγμα εντοπισμού των σημείων διακλάδωσης: εικόνα κλινικής γαλακτογραφίας (αριστερά), ανίχνευση των κλάδων του γαλακτοφόρου δένδρου (κέντρο) και σήμανση των σημείων διακλάδωσης (δεξιά).

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

Ενδεικτική Βιβλιογραφία:

[1] Angeliki Skoura, Tatyana Nuzhnaya, Predrag R. Bakic, Vasilis Megalooikonomou: Detecting and Localizing Tree Nodes in Anatomic Structures of Branching Topology. ICIAR 2013: 485-493

3. Κατηγοριοποίηση ιατρικών εικόνων που απεικονίζουν δενδρικές δομές του ανθρώπινου σώματος

Στο ανθρώπινο σώμα απαντώνται αρκετά όργανα με τοπολογία δένδρου. Χαρακτηριστικά παραδείγματα αποτελούν το αγγειακό δίκτυο, το βρογχικό δένδρο, το νευρικό σύστημα και το γαλακτοφόρο δίκτυο των μαστών. Η παρούσα διπλωματική εστιάζει στην εξαγωγή χαρακτηριστικών που μοντελοποιούν τις τοπολογιών δένδρων. Η ανάλυση των εν λόγω τοπολογιών έχει απώτερο στόχο την εύρεση νέων συσχετίσεων μεταξύ μορφολογίας και λειτουργικότητας των μελετώμενων οργάνων με κύρια εφαρμογή τη διάκριση μεταξύ φυσιολογικών και παθολογικών καταστάσεων. Η εργασία μπορεί επίσης να συνδυαστεί με την επέκταση και την εφαρμογή ήδη υλοποιημένων τεχνικών που έχουν αναπτυχθεί για το σκοπό αυτό.

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

Ενδεικτική Βιβλιογραφία:

[1] Angeliki Skoura, Michael Barnathan, Vasileios Megalooikonomou: Classification of Ductal Tree Structures in Galactograms. ISBI 2009: 1015-1018.

4. Γραφοθεωρητικές βάσεις δεδομένων

Τα τελευταία χρόνια έχουν κάνει δυναμικά την εμφάνισή τους στο προσκήνιο οι γραφοθεωρητικές βάσεις δεδομένων [1] ως βασική κατηγορία των noSQL βάσεων δεδομένων, οι οποίες δεν ακολουθούν την παραδοσιακή θεώρηση, δόμηση, και διαχείριση των δεδομένων υπό μορφή πινάκων. Αξίζει να σημειωθεί πως η ερμηνεία του όρου διαφέρει, καθώς σύμφωνα με μια άποψη σημαίνει no SQL, ενώ σύμφωνα με άλλη θεώρηση σημαίνει not only SQL. Σημαντικές γραφοθεωρητικές βάσεις είναι μεταξύ άλλων η neo4j της ομώνυμης εταιρείας, η FlockDB του twitter και η Spatial and Graph και η NoSQL Database της Oracle.

Σήμερα οι γραφοθεωρητικές βάσεις δεδομένων βρίσκονται στο επίκεντρο έντονης ακαδημαϊκής έρευνας. Επιπλέον οι γραφοθεωρητικές βάσεις δεδομένων βρίσκουν σημαντικές εφαρμογές στα κοινωνικά μέσα δικτυώσεως, στον σημασιολογικό ιστό (semantic Web) μέσω της αναπαραστάσεως οντολογιών, στα γραφικά, στην υπολογιστική όραση, και στην βιοϊατρική μέσω της αναπαραστάσεως πρωτεϊνικών αλληλεξαρτήσεων. Τα κυριότερα χαρακτηριστικά τους είναι ο εγγενής και εν δυνάμει παραλληλισμός πράξεων, η έμφαση στις σχέσεις και στις ιδιότητες των υπό αναπαράσταση αντικειμένων, και η ενδεχόμενη μερική μόνον συμμόρφωση με τις απαιτήσεις ACID. Στόχος της διπλωματικής εργασίας είναι η συγκριτική μελέτη των δυνατοτήτων τουλάχιστον τριών γραφοθεωρητικών βάσεων δεδομένων.

Επιθυμητές γνώσεις: Αλγόριθμοι, Διακριτά μαθηματικά, Βάσεις δεδομένων I, Βάσεις δεδομένων II, Αλγόριθμοι και συνδυαστική βελτιστοποίηση, Παράλληλη επεξεργασία.

Βιβλιογραφία:

[1] Ian Robinson, Jim Webber, Emil Eifrem, "Graph databases", O'Reilly media 2013
[2] Malewicz et al "Pregel: a system for large-scale graph processing", SIGMOD 2010

Συνεπιβλέπων: Ε. Γαλλόπουλος.

5. Μέθοδοι Ανάλυσης και Διαχείρισης Δεδομένων σε Πραγματικό Χρόνο

Τα τελευταία χρόνια έχουν εμφανιστεί πολλές εφαρμογές που απαιτούν την διαχείριση και την επεξεργασία ροών δεδομένων (data streams). Χαρακτηριστικά παραδείγματα αποτελούν τα δίκτυα αισθητήρων, τα κοινωνικά δίκτυα και γενικότερα, το σύνολο των εφαργομών που περιλαμβάνουν δεδομένα τα οποία δημιουργούνται με μεγάλους ρυθμούς και είναι απαραίτητη η εξαγωγή συμπερασμάτων σε πραγματικό χρόνο. Η εξόρυξη γνώσης από ροές δεδομένων απαιτεί την χρήση ιδιαίτερα αποδοτικών αλγορίθμων, ικανών να ανταπεξέλθουν σε ιδιαίτερα υψηλούς ρυθμούς δεδομένων. Παράλληλα με την εξόρυξη γνώσης από ροές έχει αναπτυχθεί ένα σύνολο εργαλείων διαχείρισης ροών δεδομένων (Data Stream Management Systems), τα οποία προσφέρουν μηχανισμούς που υποστηρίζουν την επεξεργασία σε πραγματικό χρόνο.

Στην παρούσα διπλωματική εργασία θα κληθείτε να μελετήσετε υπάρχουσες τεχνικές εξώρυξης γνώσης κατάλληλες για ροές δεδομένων που αφορούν βασικά προβλήματα όπως εξαγωγή χαρακτηριστικών, συσταδοποίηση και κατηγοριοποίηση. Επίσης, θα πρέπει να

πραγματοποιήσετε υλοποίηση κάποιων τεχνικών πάνω σε κάποιο υπάρχον σύστημα διαχείρισης ροών δεδομένων.

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

Ενδεικτική Βιβλιογραφία:

[1] C.C. Aggarwal, J. Han, J. Wang, P. Yu, A framework for clustering evolving data streams, in: Proceedings of the 29th International Conference on Very Large Data Bases, Berlin, Germany, 2003

[2] Chen, Y.. Density-based clustering for real-time stream data. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2007

[3] Yasushi Sakurai, Spiros Papadimitriou, and Christos Faloutsos. Braid: stream mining through group lag correlations. In Proceedings of the 2005 ACM SIGMOD international conference on Management of data, SIGMOD'05, pages 599–610, New York, NY, USA, 2005. ACM.

[4] Yunyue Zhu and Dennis Shasha. Statstream: statistical monitoring of thousands of data streams in real time. In Proceedings of the 28th international conference on Very Large Data Bases, VLDB'02, pages 358–369. VLDBEndowment, 2002.

[5] http://wis.cs.ucla.edu/wis/stream-mill/index.php

6. Συστήματα Διαχείρισης Ροών Δεδομένων

Τα συστήματα διαχείρισης ροών δεδομένων (Data Stream Management Systems - DSMSs) έχουν εμφανιστεί τα τελευταία χρόνια με σκοπό την επίλυση του προβλήματος της οργάνωσης ροών δεδομένων σε εφαρμογές που απαιτούν την εξαγωγή αποτελεσμάτων σε πραγματικό χρόνο. Σε αντίθεση με τα παραδοσιακά συστήματα διαχείρισης βάσεων δεδομένων, όπου τα δεδομένα είναι στατικά ή ο ρυθμός ανανέωσης είναι σχετικά μικρός, τα συστήματα διαχείρισης ροών δεδομένων παρέχουν την δυνατότητα συνεχούς εκτέλεσης ερωτημάτων (continuous queries) πάνω σε χρονικά τμήματα των ροών δεδομένων που ορίζονται με τελεστές χρονικών παραθύρων.

Στην παρούσα εργασία θα κληθείτε να μελετήσετε τα προβλήματα που αναδύονται στα συστήματα διαχείρισης ροών δεδομένων, όπως η χρονοδρομολόγιση (query scheduling), η βελτιστοποίηση ερωτημάτων (query optimization) και η απόρριψη φορτίου (load shedding). Επίσης, θα πρέπει να μελετήσετε και να συγκρίνετε υπάρχοντα συστήματα διαχείρισης ροών δεδομένων, με στόχο την εξαγωγή συμπερασμάτων σχετικά με την αποδοτικότητά τους, την ευελιξία και τις δυνατότητες που προσφέρουν, στα πλαίσια μιας εφαρμογής που θα αναπτυχθεί κατα τη διάρκεια της διπλωματικής εργασίας.

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

Ενδεικτική Βιβλιογραφία:

[1] Daniel J. Abadi, Don Carney, Ugur Çetintemel, Mitch Cherniack, Christian Convey, Sangdon Lee, Michael Stonebraker, Nesime Tatbul, and Stan Zdonik. Aurora: a new model and architecture for data stream management. The VLDB Journal, 12:120–139, 2003.

[2] Lewis Girod, Kyle Jamieson, Yuan Mei, Ryan Newton, Stanislav Rost, Arvind Thiagarajan, Hari Balakrishnan, and Samuel Madden. Wavescope: a signal-oriented data stream management system. In Proceedings of the 4th international conference on Embedded networked sensor systems, SenSys '06, pages 421–422, New York, NY, USA, 2006. ACM.

[3] Lukasz Golab and M. Tamer Özsu. Issues in data stream management. SIGMOD Rec., 32:5–14, June 2003.

[4] Jiang, Qingchun and Chakravarthy, Sharma. Scheduling Strategies for Processing Continuous Queries over Streams. Key Technologies for Data Management, Lecture Notes in Computer Science, 2004.

7. Ανάλυση οικονομικών δεδομένων με χρήση τεχνικών εξόρυξης

Ο διαθέσιμος όγκος οικονομικών δεδομένων σήμερα είναι τεράστιος και έχει δημιουργήσει την ανάγκη για ανάλυση και επεξεργασία αυτών των δεδομένων ώστε να μπορούν να μετατραπούν σε χρήσιμες πληροφορίες και να μας βοηθήσουν στη λήψη αποφάσεων. Οι τεχνικές εξόρυξης δεδομένων σε συνδυασμό με τις στατιστικές μεθόδους αποτελούν σπουδαία εργαλεία για την ανάλυση αυτών των δεδομένων. Ένας τομέας που παρουσιάζει μεγάλο ενδιαφέρον, λόγω του όγκου των πληροφοριών που συσσωρεύει καθημερινά, είναι το χρηματιστήριο. Στα πλαίσια αυτής της διπλωματικής θα γίνει αρχικά μια βιβλιογραφική ανασκόπηση των τεχνικών ανάλυσης που έχουν προταθεί για χρηματιστηριακά δεδομένα. Έπειτα η εργασία αυτή θα εστιάσει στην ανάλυση χρηματιστηριακών δεδομένων με τεχνικές εξόρυξης όπως η συσταδοποίηση, η κατηγοριοποίηση και η πρόβλεψη. Κάποιες απο αυτές τις τεχνικές θα αξιολογηθούν και θα εφαρμοστούν σε πραγματικά δεδομένα από τον ημερήσιο δείκτη S&P500 (Standard and Poor's 500).

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

Ενδεικτική Βιβλιογραφία:

[1] Chi-Jie Lu, Tian-Shyug Lee, Chih-Chou Chiu, Financial Time Series Forecasting Using Independent Component Analysis And Support Vector Regression, Decision Support Systems, Volume 47 Issue 2, May, 2009, Pages 115-125.

[2] Kyoung-Jae Kim, Financial Time Series Forecasting Using Support Vector Machines, Neurocomputing 55, pp. 307-319, 2003.

[3] Α.Μαζαράκης, Πρόβλεψη Χρηματιστηριακών Μεγεθών με Τεχνικές Εξόρυξης Δεδομένων, Μεταπτυχιακή Εργασία, Τμήμα Εφαρμοσμένης Πληροφορικής, Πανεπιστήμιο Μακεδονίας, 2007.

8. Δημιουργία Αντιστοιχίσεων μεταξύ Ετερογενών Οντολογιών

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, Βάσεις δεδομένων, Εξόρυξη γνώσης, Ανάκτηση πληροφορίας, Τεχνολογίες Διαδικτύου, 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.

Συνεπιβλέπων: Α. Καμέας (ΕΑΠ)

9. Εφαρμογή Τεχνικών Εξόρυξης σε Πολυδιάστατα Αιματολογικά Δεδομένα

Η ανάλυση των αιματολογικών δεδομένων είναι μια αρκετά πολύπλοκη διαδικασία. Η κυτταρομετρία ροής, μια μέθοδος ανάλυσης αιματολογικών δεδομένων χρησιμοποιείται για την ταυτόχρονη μέτρηση και ανάλυση πολλαπλών φυσικών ή/και χημικών χαρακτηριστικών μικροσκοπικών σωματιδίων, συνήθως κυττάρων. Σημαντική τεχνολογική πρόοδος στα υλικό/πειραματικά όργανα και την ανάπτυξη φθοριζόντων 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 μετριούνται 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 χρονοβώρα, 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 (2n, 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.

Συνεπιβλέπων: Ευγενία Βερίγου (Ιατρική Σχολή)

10. Μελέτη ιδιοτήτων μεγάλων πραγματικών γραφημάτων

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 στιγμιοτύπου 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.

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

[4] 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.

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

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] 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.

[2] 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.

12. Βιοπληροφορική - Ανάλυση γονιδιακών δεδομένων

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 ελεχθεί η ομοιότητα των αντίστοιχων γονιδιακών λειτουργιών. Ο υπολογισμόςς 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.

13. Βιοπληροφιρική: Ανάπτυξη Εργαλείων Πρωτεομικής Ανάλυσης και Οπτικοποίησης Αποτελεσμάτων

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 φυσιοπαθολογικές 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. Tsagkrasoulis, 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.

Συνεπιβλέπων: Σ. Κοσσίδα (ΙΙΒΕΑΑ, Ακαδημία Αθηνών)

14. Μέθοδοι ανάλυσης υφής και εφαρμογή τους σε εικόνες

Η ανάλυση υφής αποτελεί μια από τις σημαντικότερες τεχνικές ανάλυσης εικόνων για την εξαγωγή χρήσιμης πληροφορίας. Αρκετές μέθοδοι έχουν παρουσιαστεί στη διεθνή βιβλιογραφία οι οποίες αποσκοπούν στη βελτίωση της ικανότητας ανίχνευσης περιοχών ειδικού ενδιαφέροντος σε εικόνες αλλά και στην υποβοήθηση της αξιολόγησης των περιοχών αυτών 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.

15. ***Τεχνικές διαχείρισης και αποδοτικής ανάκτησης πολυδιάστατων ακολουθιών***

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. Τεχνικές ανάλυσης δεδομένων απο τον ανθρώπινο εγκέφαλο

Αντικείμενο αυτής της εργασίας είναι η μελέτη τεχνικών για την ανάλυση δεδομένων που προέρχονται από συστήματα απεικόνησης 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: Κ. Μπερμπερίδης, 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. Devuyst, 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. Εξόρυξη γνώσης από πολυδιάστατα δεδομένα χρησιμοποιώντας διάσπαση τανυστών (πολυδιάστατων πινάκων)

Δεδομένης μιας μεγάλης συλλογής πολυδιάστατων δεδομένων (μέσα στις διαστάσεις είναι και αυτές του χρόνου και του χώρου) πως μπορεί κάποιος να βρεί πρότυπα και συσχετίσεις; Παρόμοια, δεδομένης μιας ροής από δεδομένα που τρέχουν με συνεχή ρυθμό και σε μεγάλες ποσότητες πως μπορεί κάποιος να ανιχνεύσει ανωμαλίες, προβλήματα, κ.α.; Πολλά τέτοια θέματα εξόρυξης δεδομένων μπορούν να αντιμετωπιστούν χρησιμοποιώντας διάσπαση τανυστών, δηλ. πολυδιάστατων πινάκων. Αυτοί οι πολυδιάστατοι πίνακες αντιστοιχούν στα DataCubes της εξόρυξης δεδομένων. Αρκετή δουλειά έχει ήδη γίνει σε δυσδιάστατους πίνακες (μητρώα). Σκοπός αυτής της διπλωματικής είναι η μελέτη της υπαρχουσας βιβλιογραφίας σε πολυδιάστατους πίνακες, η σχεδίαση αλγορίθμων για διάσπαση τέτοιων πινάκων που θα μπορούν να δουλέψουν με μεγάλους όγκους δεδομένων, και η εφαρμογή αυτών των αλγορίθμων σε διάφορα δεδομένα.

Επιθυμητές γνώσεις:

Εξόρυξη δεδομένων και αλγόριθμοι μάθησης, Ανάκτηση πληροφορίας, Γραμμική Άλγεβρα, Επιστημικός Υπολογισμός Ι, Επιστημονικός Υπολογισμός ΙΙ, Γλώσσες προγραμματισμού (C, C++, Matlab, Python)

Ενδεικτική Βιβλιογραφία:

[1] M. Barnathan, V. Megalooikonomou, C. Faloutsos, F.B. Mohamed, S. Faro, "TWave: High-Order Analysis of Spatiotemporal Data", In Proceedings of the 14th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Hyderabad, India, June, 21-24, 2010, Advances in Knowledge Discovery and Data Mining, Lecture Notes in Computer Science, 2010, Volume 6118/2010, pp. 246-253.

18. Σύστημα αναγνώρισης φυσικής δραστηριότητας από δεδομένα 3D επιταχυνσιόμετρων με εφαρμογές ιατρικής πρόληψης/αποκατάστασης

Ένα από τα μεγαλύτερα προβλήματα στην παρακολούθηση ιατρικών περιπτώσεων είναι η ικανότητα των θεραπευτών να γνωρίζουν τα πραγματικά επίπεδα φυσικής δραστηριότητας των ασθενών τους. Για να αποκτηθούν αξιόπιστες πληροφορίες σχετικά με την πρόδο της θεραπείας ενός ασθενή, υπάρχουν ειδικές συσκευές (πολλαπλά επιταχυνσιόμετρα) που μπορούν να δράσουν ως εργαλεία παροχής δεδομένων τα οποία είναι ανώτερα από λ.χ. ερωτηματολόγια που συμπληρώνονται από τους ασθενείς.

Σκοπός αυτής της διπλωματικής εργασίας είναι η δημιουργία ενός συστήματος σχεδιασμένο για 3D επιταχυνσιόμετρα που τοποθετούνται στο ισχύο, με σκοπό την παρακολούθηση της φυσικής δραστηριότητας ενός ασθενή. Τα δεδομένα που θα αναλυθούν είναι απο την βάση δεδομένων USC-HAD που περιλαμβάνει βασικές κινήσεις όπως περπάτημα, τρέξιμο, κάθισμα, ύπνο, κ.α.. Στα δεδομένα θα εφαρμοστούν βασικές τεχνικές προεπεξεργασίας, τμηματοποίησης, εξαγωγής χαρακτηριστικών, μείωσης της διαστατικότητας των δεδομένων και κατηγοριοποίησης.

Επιθυμητές γνώσεις: Επεξεργασία Σημάτων, Γλώσσες προγραμματισμού (Java, Python, Matlab), Εξόρυξη Δεδομένων και Αλγόριθμοι Μάθησης

Ενδεικτική Βιβλιογραφία:

[1] M. Zhang, A. A. Sawchuk, USC-HAD: A Daily Activity Dataset for Ubiquitous Activity Recognition Using Wearable Sensors, ACM UbiComp'12, Sept. 5-8, 2012.

[2] 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.

[3] 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.

[4] 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.

[5] Alberto G. Bonomi (2011) Physical Activity Recognition Using a Wearable Accelerometer, in Sensing Emotions, Philips Research Book Series, 2011, Volume 12, 41-51.

19. Μέθοδοι ανάλυσης σεισμολογικών δεδομένων και εφαρμογές

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, Python)

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/

Συνεπιβλέπων: Ε. Σώκος (Τμήμα Γεωλογίας)

20. Ανάλυση φυσικής δραστηριότητας και κοινωνικής συμπεριφοράς χρησιμοποιώντας τεχνικές εξόρυξης δεδομένων

Η φυσική δραστηριότητα και συμπεριφορά αλλάζει με το χρόνο. Διάφοροι παράγοντες, όπως π.χ., κοινωνικοί, οικονομικοί, επαγγελματικοί, παράγοντες υγείας, οικογένειας μπορούν να επηρεάσουν τις δραστηριότητές μας και την συμπεριφορά μας. Διάφορες κινητές συσκευές που χρησιμοπούμε καθημερινά ή απλά φέρουμε μαζί μας για μεγάλα χρονικά διαστήματα διαθέτουν

διάφορους αισθητήρες που μπορούν να καταγράψουν στοιχεία τα οποία όταν αναλυθούν να μας δώσουν περισσότερες πληροφορίες για τις δραστηριότητες μας. Οι δραστηριότητες αυτές μπορούν να αναπαρασταθούν σαν χρονικά εξελισσόμενα γραφήματα. Σκοπός της διπλωματικής αυτής εργασίας πέρα απο την μελέτη της υπάρχουσας βιβλιογραφίας σε πραγματικά δυναμικά γραφήματα, είναι η ανάλυση υπαρχόντων δεδομένων φυσικής δραστηριότητας και κοινωνικής συμπεριφοράς με τεχνικές εξόρυξης που βασίζονται σε εργαλεία από την γραμμική και πλειογραμμική άλγεβρα και τη θεωρία γραφημάτων, με σκοπό την εύρεση και μελέτη δυναμικών ιδιοτήτων τέτοιων γραφημάτων, την ανίχνευση ανωμαλιών (anomaly detection) και την εύρεση προτύπων (pattern discovery).

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

Συνεπιβλέπων: Ε. Γαλλόπουλος

21. Εξαγωγή χαρακτηριστικών σε ΗΕΓ για εντοπισμό επιληπτικής κρίσης

Το πρόβλημα της ανίχνευσης επιληπτικής κρίσης μπορεί να αντιμετωπιστεί ως ένα πρόβλημα ταξινόμησης, στο οποίο πρώτα εξάγονται χαρακτηριστικά από καταγεγραμμένα δεδομένα, όπως ηλεκτροεγκεφαλογράφημα (ΗΕΓ), και στη συνέχεια εισάγονται τα χαρακτηριστικά αυτά σε εκπαιδευμένους ταξινομητές. Τυπικά χαρακτηριστικά που προέρχονται από σήματα ΗΕΓ περιλαμβάνουν την κυρίαρχη φασματική κορυφή, αναλογία ισχύος, το εύρος ζώνης των κυρίαρχων φασματικών αιχμών, μη γραμμική ενέργεια, φασματική εντροπία, το μήκος της γραμμής, κ.α.. Ο σκοπός της εργασίας αυτής είναι να εξαχθούν διάφορα ενδιαφέροντα χαρακτηριστικά που έχουν χρησιμοποιηθεί στη βιβλιογραφία για εντοπισμό επιληπτικής κρίσης σε ΗΕΓ. Ο υπολογισμός των χαρακτηριστικών μπορεί είτε να γίνει απευθείας σε MATLAB, είτε να αναπτυχθεί κατάλληλο interface που θα φορτώνει διαθέσιμα προγράμματα που υπολογίζουν τα χαρακτηριστικά αυτά.

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

Βιβλιογραφία:

[1] P. McSharry, T. He, L. Smith, et al., "Linear and non-linear methods for automatic seizure detection in scalp electro-encephalogram recordings,"Med Biol Eng Comput, vol. 40, pp.447–461, 2002.

[2] B.R. Greene, et al., "Combination of EEG and ECG for improved automatic neonatal seizure detection," Clinical Neurophysiology, vol. 118, pp. 1348–1359, 2007.

[3] C.A. Teixeira et al., "EPILAB: A software package for studies on the prediction of epileptic seizures," Journal of Neuroscience Methods, vol. 200 pp. 257– 271, 2011.

Συνεπιβλέποντες: Κ. Μπερμπερίδης, Ε. Ζαχαράκη

22. Ανάλυση Χωρο-χρονικών Δεδομένων χρησιμοποιώντας Γεωγραφικά Πληροφοριακά Συστήματα

Ο στόχος της διπλωματικής αυτής είναι η ανάλυση χωρο-χρονικών δεδομένων και η εξόρυξη γνώσης από αυτά, είτε για πρόβλεψη μελλοντικών τιμών ή για συσταδοποίηση ομοίων παρατηρήσεων είτε για ανακάλυψη προτύπων. 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)

23. Ανίχνευση εστιών εγκεφαλικής βλάβης σε δεδομένα MRI

Στην εργασία αυτή θα αναπτυχθεί μέθοδος για αυτόματη ανίχνευση αλλοιώσεων ιστού του εγκεφάλου που οφείλονται σε αγγειακή νόσο ή εγκεφαλικά επεισόδια. Η ανίχνευση θα γίνει σε δεδομένα μαγνητικής τομογραφίας (MRI) και θα βασιστεί στη χρήση μοντέλου που περιγράφει την απεικονιστική πληροφορία φυσιολογικού εγκεφάλου (χωρίς βλάβη). Το στατιστικό μοντέλο θα δημιουργηθεί χρησιμοποιώντας MRI δεδομένα ενός πληθυσμού υγιών ατόμων. Η μελέτη θα βασιστεί σε προηγούμενη έρευνα η οποία εφάρμοσε πολυ-παραμετρική ανάλυση μοντελοποιώντας τη μορφολογία του εγκεφάλου στο σύνολό της [1] ή voxel-based ανάλυση μοντελοποιώντας την τιμή φωτεινότητας κάθε ιστού [2].

Επιθυμητές γνώσεις: Εξόρυξη γνώσης, Μηχανική μάθηση, Επεξεργασία Εικόνας, Γλώσσες προγραμματισμού (C, C++, Python, Matlab).

Ενδεικτική Βιβλιογραφία:

[1] E.I. Zacharaki, A. Bezerianos, “Abnormality segmentation in brain images via distributed estimation,” IEEE Transaction on Information Technology in Biomedicine, vol. 16, no. 3, pp. 330-338, 2012.

[2] E.I. Zacharaki, G. Erus, A. Bezerianos, C. Davatzikos, “Fuzzy multi-channel clustering with individualized spatial priors for segmenting brain lesions and infarcts,” 2nd Artificial Intelligence Applications in Biomedicine Workshop (AIAB 2012), 27-30 September 2012, Halkidiki, Greece.

[3] http://www.ia.unc.edu/MSseg/papers.php

Συνεπιβλέπων: Ε. Ζαχαράκη

24. Μελέτη ηλεκτροκαρδιογραφήματος και μοντελοποίηση

Η διπλωματική εργασία αφορά την μελέτη καρδιακών σημάτων (ECG) και την επεξεργασία τους με χρήση signal processing και data mining τεχνικών για την αυτόματη εξαγωγή προτύπων από τα σήματα αυτά. Τα δεδομένα που θα χρησιμοποιηθούν είναι αυτά του physionet (http://www.physionet.org).

Επιθυμητά Προσόντα: Εξόρυξη γνώσης, βασικές γνώσεις signal processing, βασικές γνώσεις matlab.

25. Αναγνώριση στρες σε οδηγούς αυτοκινήτων

Η διπλωματική εργασία αφορά την μελέτη εγκεφαλικών (EEG) και καρδιακών (ECG) σημάτων και την επεξεργασία τους με χρήση signal processing και data mining τεχνικών για την μοντελοποίηση του stress κατά την διάρκεια της οδήγησης. Τα δεδομένα που θα χρησιμοποιηθούν είναι αυτά του: http://www.physionet.org/physiobank/database/drivedb/

Επιθυμητά Προσόντα: Εξόρυξη γνώσης, βασικές γνώσεις signal processing, βασικές γνώσεις matlab.

26. Ανάλυση άπνιας, μοντελοποίηση και ανίχνευση

Η διπλωματική εργασία αφορά την μελέτη καρδιακών σημάτων (ECG) και την επεξεργασία τους με χρήση signal processing και data mining τεχνικών για την μελέτη της άπνιας. Τα δεδομένα που θα χρησιμοποιηθούν είναι αυτά του:
http://www.physionet.org/physiobank/database/apnea-ecg/

Επιθυμητά Προσόντα: Εξόρυξη γνώσης, βασικές γνώσεις signal processing, βασικές γνώσεις matlab.

27. Ανάλυση των σταδίων του ύπνου, μοντελοποίηση και ανίχνευση χρησιμοποιώντας δεδομένα από Polysomnography

Η διπλωματική εργασία αφορά την μελέτη φυσιολογικών σημάτων (εγκεφαλικών, καρδιακών) και την επεξεργασία τους με χρήση signal processing και data mining τεχνικών για την μελέτη της δομής του ύπνου. Τα δεδομένα που θα χρησιμοποιηθούν είναι αυτά του:
http://www.physionet.org/physiobank/database/slpdb/

Επιθυμητά Προσόντα: Εξόρυξη γνώσης, βασικές γνώσεις signal processing, βασικές γνώσεις matlab.

28. Ανάλυση και μοντελοποίηση των επιπτώσεων της γήρανσης και ασθενιών στο βάδισμα

Η διπλωματική εργασία αφορά την μελέτη φυσιολογικών σημάτων (εγκεφαλικών, καρδιακών) και την επεξεργασία τους με χρήση signal processing και data mining τεχνικών για την μελέτη των επιπτώσεων που προκαλεί η γήρανση και τύποι ασθενιών στο βάδισμα. Τα δεδομένα που θα χρησιμοποιηθούν είναι αυτά του: http://www.physionet.org/physiobank/database/gaitdb/

Επιθυμητά Προσόντα: Εξόρυξη γνώσης, βασικές γνώσεις signal processing, βασικές γνώσεις matlab.

29. Ανάλυση και μοντελοποίηση της ανθρώπινης συναισθηματικής κατάστασης

Η διπλωματική εργασία αφορά την μελέτη φυσιολογικών σημάτων (εγκεφαλικών - EEG) και την επεξεργασία τους με χρήση signal processing και data mining τεχνικών για την μελέτη της συναισθηματικής κατάστασης του ανθρώπου. Τα δεδομένα που θα χρησιμοποιηθούν είναι αυτά του: http://www.eecs.qmul.ac.uk/mmv/datasets/deap/

Επιθυμητά Προσόντα: Εξόρυξη γνώσης, βασικές γνώσεις signal processing, βασικές γνώσεις matlab.

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).

Αιτήσεις με email στην ηλεκτρονική διεύθυνση 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).