

## **Διπλωματικές Εργασίες Ακαδημαϊκού Έτους 2009-2010**

Β. Μεγαλοοικονόμου, Αναπληρωτής Καθηγητής

Παρουσίαση Θεμάτων: 8 Οκτ. 2009, 15:30, Π200

### **1. 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),

Υπεύθυνη μεταπτυχιακός: Α. Σκούρα

### **2. Γεωγραφικά πληροφοριακά συστήματα και εξόρυξη χωρο-χρονικών δεδομένων**

Περιγραφή:

Ο στόχος της διπλωματικής αυτής είναι η ανάλυση χωρο-χρονικών δεδομένων από περιβαλλοντικές μελέτες, σε μορφή ArcGIS 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 (π.χ. Νομαρχίες, Δασαρχεία, κτλ) για τον αποτελεσματικό έλεγχο ακτών, παράκτιων εκτάσεων, παραποτάμιων περιοχών, κτλ. Η χρήση του λογισμικού Γεωγραφικών Πληροφοριακών Συστημάτων ArcGIS 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, Μ. Μαραγκουδάκης  
Υπεύθυνη μεταπτυχιακός : Α. Κουμπούρη

### **3. 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 για μεγάλα χρονικά διαστήματα, 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 μπορούν μερικώς ή εντελώς 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, όπως στη μαστογραφία, η τοπολογία διακλαδίσματος μιας δενδρικής δομής μπορεί να είναι μόλις ορατή ή ακόμα και απύσα από μια εικόνα, αλλά ακόμα να συμβάλλει στην υφή των γειτονικών περιοχών. Το μέγιστο βάθος της δενδρικής δομής που συλλαμβάνεται στην εικόνα μπορεί επίσης να ποικίλει, ανάλογα με τη δυνατότητα της μονάδας να εξάγει μια δομή διακλάδωσης από τα σύνθετα περίχωρά της. Οι μονάδες που προσφέρουν την απεικόνιση των πιο υψηλών επιπέδων διακλάδωσης είναι συνήθως πιο απαγορευτικές από την άποψη του κόστους, του κινδύνου υγείας, ή της άνεσης στη χρήση τους πάνω στους ασθενείς. Εναλλακτικά, οι μονάδες που μπορούν να συλλάβουν μόνο τα έμμεσα αποτελέσματα της παρουσίας μιας δομής διακλάδωσης μέσα στα σύνθετα περίχωρά της είναι ευκολότερα διαθέσιμες. Σε αυτή τη διπλωματική θα εξετασθούν θεωρητικά και πειραματικά διάφορες τεχνικές από την περιοχή του machine learning που μπορούν να εφαρμοστούν στην πιθανοτική ανίχνευση κόμβων διακλάδωσης σε εικόνες δενδρικών δομών.

Επιθυμητές γνώσεις:

Εξόρυξη γνώσης, Ανάκτηση πληροφορίας, Επεξεργασία Σημάτων, Επεξεργασία Εικόνας, Γλώσσες προγραμματισμού (C, C++, Matlab),

Υπεύθυνη μεταπτυχιακός: Α. Σκούρα

### **4. 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)

Συνεπιβλέπων: Α. Καμέας

Υπεύθυνη μεταπτυχιακός: Λ. Σερεμέτη

## **5. Κατανεμημένη ανάλυση δεδομένων και εικόνων από *clinical information repositories* και εφαρμογές στην τηλε-ιατρική**

Περιγραφή:

Οι περισσότερες τεχνικές εξόρυξης γνώσης έχουν προταθεί για centralized συστήματα βάσεων δεδομένων (όπου τα δεδομένα είναι συγκεντρωμένα σε ένα κεντρικό σύστημα). Η διπλωματική αυτή έχει σαν αντικείμενο τη μελέτη τεχνικών εξόρυξης γνώσης σε κατανεμημένα συστήματα βάσεων δεδομένων όπου τα δεδομένα είναι αποθηκευμένα σε διάφορες επιμέρους βάσεις δεδομένων συνήθως γεωγραφικά διαχωρισμένες αλλά συνδεδεμένες μεταξύ τους και όπου τα transactions είναι όχι μόνο local αλλά και global. Ένα τέτοιο κατανεμημένο σύστημα μπορεί να χρησιμοποιηθεί για την ανάλυση μεγάλου όγκου δεδομένων και εικόνων που βρίσκονται σε διάφορα repositories με σκοπό την υποστήριξη λήψης αποφάσεων. Θα εξετασθούν θεωρητικά και πειραματικά διάφορες τεχνικές εξόρυξης γνώσης που μπορούν να εφαρμοστούν κάτω από διάφορους περιορισμούς που υπάρχουν στο σύστημα οσον αφορά την επικοινωνία μεταξύ των sites (bandwidth limitations), την υπολογιστική τους ισχύ, κ.α. Θα εξετασθεί επίσης η εφαρμογή αυτού του συστήματος σε τηλε-ιατρική με σκοπό τη βελτίωση πρόσβασης στην υγεία και στην περίθαλψη.

Επιθυμητές γνώσεις:

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

## **6. Μελέτη και υλοποίηση τεχνικών εξόρυξης δεδομένων χρονοσειρών**

Περιγραφή:

Αντικείμενο αυτής της εργασίας είναι η μελέτη των τεχνικών εξόρυξης γνώσης από δεδομένα που εξελίσσονται χρονικά. Σκοπός αυτών των τεχνικών είναι η ανακάλυψη ομοιοτήτων, η πρόβλεψη μελλοντικών τιμών, η ανακάλυψη ακολουθιακών προτύπων (sequential patterns) ή κανόνων συσχετίσεων ακολουθιών (sequence association rules), η εύρεση ομάδων (clustering) ή η ταξινόμηση τους (classification), η δημιουργία περιλήψεων (summarization), κ.λ.π. Η αναπαράσταση των χρονοσειρών είναι πολύ σημαντική για την περαιτέρω εφαρμογή των τεχνικών εξόρυξης. Στα πλαίσια αυτής της διπλωματικής θα μελετηθούν αυτές οι τεχνικές και θα υλοποιηθούν κάποιες από 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, Βάσεις δεδομένων, Επεξεργασία Σημάτων  
Γλώσσες προγραμματισμού (C, C++, Matlab)

## **7. Μελέτη και ανάπτυξη τεχνικών εντοπισμού συσχετίσεων βασισμένων σε Bayesian Networks**

Περιγραφή:

Αντικείμενο αυτής της διπλωματικής εργασίας είναι α) η μελέτη τεχνικών εξόρυξης συσχετίσεων από μεγάλες βάσεις δεδομένων και β) η υλοποίηση κάποιων από αυτών των τεχνικών για τον εντοπισμό συσχετίσεων σε οικονομικά και ιατρικά δεδομένα. Πιο συγκεκριμένα οι τεχνικές που θα μελετηθούν και θα υλοποιηθούν θα είναι βασισμένες σε Bayesian Networks.

Επιθυμητές γνώσεις:

Εξόρυξη γνώσης, Ανάκτηση πληροφορίας, Βάσεις δεδομένων  
Γλώσσες προγραμματισμού (C, C++, Matlab)

## **8. Ανάλυση χρονοσειρών που συλλέγονται από κατανομημένα συστήματα και δίκτυα αισθητήρων**

Περιγραφή:

Στα πλαίσια αυτής της διπλωματικής θα μελετηθούν τεχνικές ανάλυσης μεγάλου όγκου χρονοσειρών που συλλέγονται από δίκτυα αισθητήρων με σκοπό την εξαγωγή συμπερασμάτων σχετικά με τη λειτουργία μιας υποδομής, όπως για παράδειγμα της κίνησης των τροχοφόρων σε έναν αυτοκινητόδρομο. Το πρόβλημα αυτό μπορεί να μελετηθεί offline ή online (ανάλυση stream data). Πέρα από τη μελέτη θα υπάρχει υλοποίηση κάποιων τεχνικών και εφαρμογή τους σε πραγματικά δεδομένα με σκοπό την εύρεση προτύπων καθώς και συσχετίσεων μεταξύ των δεδομένων διαφορετικών αισθητήρων.

Επιθυμητές γνώσεις:

Εξόρυξη γνώσης, Ανάκτηση πληροφορίας, Βάσεις δεδομένων, Καταναμημένα συστήματα, Επεξεργασία Σημάτων, Γλώσσες προγραμματισμού (C, C++, Matlab)

### **9. 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, Δ. Χριστοδουλάκης  
Υπεύθυνοι μεταπτυχιακοί: Σ. Στάμου, Α. Κοζανίδης

### **10. Ανάλυση του γράφου της Wikipedia για την αυτόματη δημιουργία ιεραρχικών ευρετηρίων τύπων οντοτήτων**

Περιγραφή:

Κάθε άρθρο της Wikipedia περιέχει την περιγραφή μιας οντότητας και παρέχει για αυτήν ένα σύνολο κατηγοριών στις οποίες ανήκει. Παρόλο που οι οντότητες είναι διασυνδεδεμένες με χρήση σχέσεων redirect, disambiguation, reference, κτλ, οι αντίστοιχες κατηγορίες τους δεν εμφανίζουν την ανάλογη δομή. Σκοπός της παρούσας διπλωματικής είναι η μελέτη τόσο της δομής του XML corpus της Wikipedia όσο και των σχέσεων μεταξύ των οντοτήτων που περιέχει,

προκειμένου να οδηγηθούμε στον σχεδιασμό και την υλοποίηση εργαλείων με την βοήθεια των οποίων θα πραγματοποιηθεί η οργάνωση των οντοτήτων της Wikipedia σε πολλαπλές ιεραρχίες, για παράδειγμα θέματος, γεωγραφικού προσανατολισμού, τύπου δεδομένων κα.

Ζητούμενο της εργασίας είναι η υλοποίηση των αντίστοιχων ιεραρχικών ευρετηρίων για το XML corpus της Wikipedia.

Επιθυμητές Γνώσεις: Βάσεις Δεδομένων, Δομές Δεδομένων, Γλωσσική Τεχνολογία, Ανάκτηση Πληροφορίας, Τεχνολογίες Διαδίκτυου. Επιπλέον γνώσεις: εξοικείωση με τεχνικές αντικειμενοστραφούς προγραμματισμού (Python ή C# ή Java) , XML και HTML.

Συνεπιβλέποντες: Β. Μεγαλοοικονόμου, Δ. Χριστοδουλάκης  
Υπεύθυνη μεταπτυχιακός : Π. Τζέκου

## ***11. Εξόρυξη γνώσης από συστήματα μεγάλων βάσεων δεδομένων χρησιμοποιώντας διάσπαση πολυδιάστατων πινάκων***

Περιγραφή:

Δεδομένης μιας μεγάλης συλλογής από εικόνες που μεταβάλλεται με το χρόνο πως μπορεί κάποιος να βρεί πρότυπα και συσχετίσεις. Παρόμοια, δεδομένης μιας ροής από δεδομένα που τρέχουν με συνεχή ρυθμό και σε μεγάλες ποσότητες σε δίκτυα υπολογιστών πως μπορεί κάποιος να ανιχνεύσει ανωμαλίες, εισβολές, ή πιθανές ανεπάρκειες? Πολλά τέτοια προβλήματα εξόρυξης δεδομένων μπορούν να αντιμετωπιστούν χρησιμοποιώντας διάσπαση πολυδιάστατων πινάκων. Αυτοί οι πολυδιάστατοι πίνακες αντιστοιχούν στα DataCubes της εξόρυξης δεδομένων. Αρκετή δουλειά έχει ήδη γίνει σε δυσδιάστατους πίνακες. Σκοπός αυτής της διπλωματικής είναι η μελέτη της υπάρχουσας 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 γνώσης, Ανάκτηση πληροφορίας, Βάσεις δεδομένων, Γραμμική Άλγεβρα  
Γλώσσες προγραμματισμού (C, C++, 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, είτε με κατάθεση στην ταχυδρομική θυρίδα του διδάσκοντα στο Β' κτίριο.

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