ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ
Τ.Θ. 1186 Τ.Κ. 45110 Ιωάννινα
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ
Τηλ. +30-6510-98817 Fax. +30-6510-98890

ΠΑΡΟΥΣΙΑΣΗ
ΔΙΔΑΚΤΟΡΙΚΗΣ ΔΙΑΤΡΙΒΗΣ

ΗΜΕΡΟΜΗΝΙΑ: Τετάρτη, 17 Απριλίου 2013

ΩΡΑ: 15:30

ΑΙΘΟΥΣΑ: Αίθουσα Σεμιναρίων (ισόγειο Ι1-Ι2)
Κτήριο Τμήματος Πληροφορικής

ΟΜΙΛΗΤΗΣ: **Αργύρης Καλογεράτος**

Θ έ μ α

«Μέθοδοι Εξόρυξης Γνώσης από Συλλογές Εγγράφων»

Επταμελής Εξεταστική Επιτροπή

1. Αριστείδης Λύκας, Αναπλ. Καθηγητής Πανεπιστημίου Ιωαννίνων
2. Κωνσταντίνος Μπλέκας, Επικ. Καθηγητής Πανεπιστημίου Ιωαννίνων
3. Ανδρέας Σταφυλοπάτης, Καθηγητής Εθνικού Μετσόβιου Πολυτεχνείου
4. Ευαγγελία Πιτουρά, Αναπλ. Καθηγήτρια Πανεπιστημίου Ιωαννίνων
5. Παναγιώτης Τσαπάρας, Επικ. Καθηγητής Πανεπιστημίου Ιωαννίνων
6. Μιχαήλ Βαζιργιάννης, Καθηγητής Οικονομικού Πανεπιστημίου Αθηνών
7. Δημήτριος Γουνόπουλος, Καθηγητής Εθνικού Καποδιστριακού Πανεπιστημίου Αθηνών

ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ
Τ.Θ. 1186 Τ.Κ. 45110 Ιωάννινα
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ
Τηλ. +30-6510-98817 Fax. +30-6510-98890

Περίληψη

Η παρούσα διατριβή ασχολείται με το πρόβλημα της ομαδοποίησης εγγράφων (document clustering). Δοθείσης μία συλλογής εγγράφων (corpus) φυσικής γλώσσας, καταρχήν εφαρμόζεται προεπεξεργασία και εξαγωγή χαρακτηριστικών όρων (terms). Ως αποτέλεσμα κάθε έγγραφο συνήθως αναπαρίστανται με ένα διανυσματικό μοντέλο (vector space model) όπου το μη αρνητικό βάρος κάθε διάστασης περιγράφει τη σημαντικότητα του αντίστοιχου χαρακτηριστικού όρου. Οι ιδιότητες αυτού του χώρου αναπαράστασης είναι: α) η πολύ υψηλή διάσταση της τάξης των χιλιάδων χαρακτηριστικών, και β) η αραιότητα που αγγίζει το 99% (high dimensionality and sparsity). Στη διατριβή μελετώνται και αναπτύσσονται μέθοδοι αναπαράστασης και εξαγωγής πληροφορίας σχετικά με τη δομή ομάδων στη συλλογή εγγράφων (cluster structure).

Αρχικά προτείνεται ένα μοντέλο διανυσματικής αναπαράστασης εγγράφων, το οποίο, δίχως επίβλεψη, επανεξετάζει την παραδοσιακή υπόθεση ανεξαρτησίας των όρων (term independency). Για κάθε όρο του λεξικού εξάγεται το αντίστοιχο Γενικευμένο Διάνυσμα Συμφραζομένων Όρων (Global Term Context Vector) το οποίο ενσωματώνει τη συμφραζόμενη πληροφορία γύρω από τις εμφανίσεις του όρου στα έγγραφα (συν-εμφανίσεις όρων). Στη συνέχεια, κατασκευάζεται ένας σημασιολογικός πίνακας βάσει του οποίου προβάλλονται τα διανύσματα δεδομένων σε έναν πυκνότερο χώρο ίδιας διάστασης. Στο στάδιο αυτό μελετήθηκε η συμβολή του προτεινόμενου μοντέλου αναπαράστασης στην ομαδοποίηση εγγράφων.

Ύστερα, παρουσιάζεται η μέθοδος ομαδοποίησης εγγράφων *k*-συνθετικών πρωτοτύπων (*k*-synthetic prototypes). Η μέθοδος βασίζεται στον σφαιρικό αλγόριθμο *k*-μέσων (spherical *k*-means) με την πρωτοτυπία ότι εισάγει τους συνθετικούς αντιπροσώπους για τις ομάδες. Η προτεινόμενη αυξητική προσέγγιση για τον υπολογισμό ενός συνθετικού αντιπροσώπου χρησιμοποιεί τα *K* αντικείμενα που βρίσκονται εγγύτερα στο ενδιάμεσο αντικείμενο μίας ομάδας (medoid). Η ενδιαφέρουσα ιδιότητα αυτής της προσέγγισης είναι ότι ευνοεί την αναπαράσταση της κυρίαρχης κλάσης δεδομένων σε μία ομάδα επιτρέποντας με αυτό τον τρόπο την αποφυγή λύσεων τοπικών ελαχίστων λόγω κακής αρχικοποίησης. Στην πειραματική μελέτη συγκρίνουμε την προτεινόμενη μέθοδο με μία σειρά από ευρέως χρησιμοποιούμενους αλγόριθμους ομαδοποίησης.

Στη συνέχεια, μελετώνται αλγόριθμοι αυξητικής ομαδοποίησης (incremental clustering), οι οποίοι εισάγουν την *k*+1 ομάδα δεδομένων βασιζόμενοι στη λύση *k* ομάδων. Αρχικά παρουσιάζεται ένα γενικό πλαίσιο (clustering framework) που εισάγει τη μερική ενημέρωση της *k*+1 λύσης κατά την εισαγωγή μίας νέας ομάδας (partial update). Το πλαίσιο αυτό καλύπτει γνωστούς αυξητικούς αλγορίθμους, όπως ο διαμεριστικός *k*-μέσων (bisecting *k*-means), ο γενικευμένος *k*-μέσων (global *k*-means), και διάφορες παραλλαγές τους. Προτείνεται, δε, ο γενικευμένος αλγόριθμος *k*-συνθετικών πρωτοτύπων (global *k*-synthetic prototypes) ο οποίος συγκρίνεται πειραματικά με υπάρχουσες αυξητικές προσεγγίσεις επιδεικνύοντας καλύτερα αποτελέσματα ομαδοποίησης σε συλλογές εγγράφων.

Η τελευταία ενότητα της διατριβής αφορά το πρόβλημα εκτίμησης του αριθμού των ομάδων σε ένα σύνολο δεδομένων. Για την προσέγγιση του προβλήματος προτείνεται το κριτήριο dip-dist το οποίο θεωρεί κάθε αντικείμενο της υπό εξέταση ομάδας ως ‘παρατηρητή’ και εφαρμόζει ένα στατιστικό τεστ μονοτροπικότητας (unimodality dip-test) στην κατανομή των αποστάσεων μεταξύ του παρατηρητή και των υπολοίπων αντικειμένων της ομάδας. Στη συνέχεια, περιγράφεται η αυξητική μέθοδος dip-μέσων (dip-means) της οποίας η μοναδική υπόθεση είναι η μονοτροπικότητα κάθε ομάδας. Τα πλεονεκτήματα της προτεινόμενης προσέγγισης είναι ότι το στατιστικό τεστ εφαρμόζεται σε 1Δ κατανομές, ενώ θα μπορούσε να χρησιμοποιηθεί και σε μεθόδους που βασίζονται στον πίνακα ομοιότητας (kernel-based methods), όπου δεν απαιτούνται τα πραγματικά διανύσματα δεδομένων.