ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΗΛΕΚΤΡΙΚΩΝ ΒΙΟΜΗΧΑΝΙΚΩΝ ΔΙΑΤΑΞΕΩΝ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΑΠΟΦΑΣΕΩΝ

Συστήματα Προτάσεων με Πιθανοτικά Μοντέλα Θεμάτων

Κωνσταντίνος Α. Χρηστίδης

Διδακτορική Διατριβή

στο πλαίσιο του Προγράμματος Μεταπτυχιακών Σπουδών
της Σχολής Ηλεκτρολόγων Μηχανικών και Μηχανικών Η/Υ

Εθνικό Μετσόβιο Πολυτεχνείο

Ιανουάριος 2013

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΗΛΕΚΤΡΙΚΩΝ ΒΙΟΜΗΧΑΝΙΚΩΝ ΔΙΑΤΑΞΕΩΝ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΑΠΟΦΑΣΕΩΝ

Συστήματα Προτάσεων με Πιθανοτικά Μοντέλα Θεμάτων

Κωνσταντίνος Α. Χρηστίδης

Διδακτορική Διατριβή

Τριμελής Συμβουλευτική Επιτροπή:

Γρηγόριος Μέντζας, Καθηγητής Ε.Μ.Π. (επιβλέπων)

Ιωάννης Ψαρράς, Καθηγητής Ε.Μ.Π.

Δημήτριος Ασκούνης, Αν. Καθηγητής Ε.Μ.Π.

Επταμελής Εξεταστική Επιτροπή:

Γρηγόριος Μέντζας
Καθηγητής Ε.Μ.Π.

Ιωάννης Βασιλείου
Καθηγητής Ε.Μ.Π.

Ιωάννης Ψαρράς
Καθηγητής Ε.Μ.Π.

Ανδρέας Γεώργιος Σταφυλοπάτης
Καθηγητής Ε.Μ.Π.

Δημήτριος Ασκούνης
Αν. Καθηγητής Ε.Μ.Π.

Αθηνά Βακάλη
Αν. Καθηγήτρια Α.Π.Θ.

Δημήτριος Αποστόλου
Επ. Καθηγητής
Πανεπιστημίου Πειραιώς

Αθήνα, Ιανουάριος 2013

……

Κωνσταντίνος Α. Χρηστίδης

Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Η διδακτορική διατριβή τοποθετείται στον χώρο των συστημάτων αποφάσεων και ειδικότερα στην περιοχή των *συστημάτων προτάσεων* (recommender systems) όπου προτείνει την ανάπτυξη συστημάτων προτάσεων με χρήση *πιθανοτικών μοντέλων θεμάτων* (probabilistic topic models).

Στο πλαίσιο της διατριβής πραγματοποιήθηκε βιβλιογραφική μελέτη στα γνωστικά πεδία της χρήσης τεχνικών μηχανικής μάθησης και της πιθανοτικής λανθάνουσας σημασιολογικής ανάλυσης (probabilistic latent semantic analysis) για την πραγματοποίηση προτάσεων. Στην παρούσα διατριβή διερευνάται η δυνατότητα για βελτιωμένα συστήματα προτάσεων στο εσωτερικό επιχειρήσεων, κοινοτήτων και στο εμπόριο με βάση λανθάνοντα θέματα.

Παρουσιάζεται μία προσέγγιση για την ενσωμάτωση της υπάρχουσας γνώσης ενός πεδίου σε ένα σύστημα προτάσεων. Ακόμη, προτείνεται μια μεθοδολογία που εκμεταλλεύεται την εξαγωγή πιθανοτικών μοντέλων θεμάτων για την πλήρη και αποτελεσματική μοντελοποίηση της ικανότητας ενός εργαζομένου να αντιμετωπίσει ένα πρόβλημα. Περιγράφεται μια μεθοδολογία εξαγωγής προτιμήσεων για καταναλωτές σε υπεραγορές από ένα σύνολο δεδομένων με χρήση λανθανόντων θεμάτων. Τέλος, προτείνεται και αξιολογείται μια μεθοδολογία για την εκμετάλλευση του μη δομημένου κειμένου που βρίσκεται σε ηλεκτρονικές αγορές δημοπρασιών για την παραγωγή προτάσεων που απευθύνονται σε αγοραστές και πωλητές.

Τα συστήματα προτάσεων που προτείνονται αξιολογήθηκαν με ενθαρρυντικά αποτελέσματα και παρουσίασαν κοινά χαρακτηριστικά:

(1) μειώνουν τις διαστάσεις του προβλήματος και παρέχουν γρήγορα προτάσεις αφού έχει προηγηθεί η εξαγωγή των μοντέλων θεμάτων,

(2) ικανοποιούν τις απαιτήσεις των χρηστών για ακρίβεια και ανάκληση των δεδομένων που τους ενδιαφέρουν,

(3) τα θέματα που εξάγονται μπορούν να αποτελέσουν σημαντική πληροφορία για τον ιδιοκτήτη ή τον διαχειριστή του συστήματος.

Abstract

This doctoral thesis is positioned in the research area of decision support systems and specifically in recommender systems. It focuses on the design and development of *recommender systems* based on *probabilistic topic models*.

In this thesis we have thoroughly examined the literature related to utilizing machine learning techniques and probabilistic latent semantic analysis for providing recommendations. We explore the possibility to design improved recommender systems inside enterprises, communities and in electronic commerce based on latent topics.

An approach is presented for integrating existing domain knowledge in a recommender system. Additionally, a methodology is proposed for utilizing probabilistic topic models for complete and effective modeling of employee expertise on addressing problems. A methodology is described for extracting consumer preferences from datasets using topic models. Finally, we present and evaluate a methodology for utilizing the unstructured text found in electronic auction marketplaces in order to provide recommendations to buyers and sellers.

The recommender systems proposed in this thesis have displayed a number of common characteristics:

1) They reduce the dimensions of the recommendation problem and provide fast online recommendations, having trained the topic models.

2) They satisfy the user needs for accuracy and recall of all interesting objects.

3) The topics extracted can provide significant insight to the system manager or owner.

Ευχαριστίες

Η παρούσα διδακτορική διατριβή σηματοδοτεί το τέλος μιας προσπάθειας που πραγματοποιήθηκε στα πλαίσια του προγράμματος μεταπτυχιακών σπουδών της σχολής Ηλεκτρολόγων Μηχανικών και Μηχανικών Ηλεκτρονικών Υπολογιστών του Εθνικού Μετσόβιου Πολυτεχνείου.

Η συγκεκριμένη προσπάθεια δε θα μπορούσε να ολοκληρωθεί χωρίς τη μόνιμη συνδρομή του Καθηγητή κ. Γ. Μέντζα, την καθοδήγηση, την ενθάρρυνση και τη βοήθεια που μου προσέφερε. Είμαι ευγνώμων για την εμπιστοσύνη που μου έδειξε και για την ευκαιρία που μου έδωσε να δοκιμάσω τις ικανότητες και τις ιδέες μου στα πλαίσια των ερευνητικών δραστηριότητων της Μονάδας Διοίκησης Πληροφοριακών Συστημάτων.

Θα ήθελα να ευχαριστήσω τα άλλα δύο μέλη της τριμελούς συμβουλευτικής επιτροπής, τους Καθηγητές κ. Ι. Ψαρρά και κ. Δ. Ασκούνη, καθώς και την Καθηγήτρια κ. Α. Βακάλη και τους Καθηγητές κ. Α. Σταφυλοπάτη, κ. Ι. Βασιλείου, κ. Δ. Αποστόλου για την τιμή που μου έκαναν να συμμετάσχουν στην επταμελή επιτροπή εξέτασης της διατριβής.

Θέλω επίσης να ευχαριστήσω τους συναδέλφους μου Θύμιο Μπόθο, Μπάμπη Μαγγούτα, Δημήτρη Παναγιώτου, Φώτη Παρασκευόπουλο και Νίκη Παπαηλιού, καθώς και τα υπόλοιπα μέλη της ομάδας, που υπήρξαν αρωγοί και συμπαραστάτες σε αυτή την πορεία. Η συναναστροφή με τους συναδέλφους αποτέλεσε βασική πηγή υποστήριξης και έμπνευσης για την αντιμετώπιση και επίλυση των ερευνητικών προκλήσεων.

Τέλος, το πιο μεγάλο ευχαριστώ αναλογεί στους ανθρώπους που στάθηκαν δίπλα μου σε όλο το χρονικό διάστημα της συγκεκριμένης προσπάθειας. Στους γονείς μου, στην αδερφή μου και στους φίλους μου, που μου προσέφεραν αμέριστη αγάπη και ηθική υποστήριξη αυτά τα χρόνια.

Πίνακας Περιεχομένων

Κατάλογος Εικόνων

Κατάλογος Πινάκων

Σύνοψη

Η διδακτορική διατριβή τοποθετείται στον χώρο των συστημάτων αποφάσεων και ειδικότερα στην περιοχή των *συστημάτων προτάσεων* (recommender systems) όπου προτείνει την ανάπτυξη συστημάτων προτάσεων με χρήση *πιθανοτικών μοντέλων θεμάτων* (probabilistic topic models).

Καθώς ο όγκος των πληροφοριών που γίνεται διαθέσιμος μεγαλώνει ολοένα και περισσότερο στον παγκόσμιο ιστό, αλλά και γενικά στην καθημερινή ζωή, δυσχεραίνεται η δυνατότητα των ανθρώπων να εντοπίσουν την πληροφορία που τους ενδιαφέρει. Αυτή η κατάσταση κατά την οποία οι πληροφορίες που είναι διαθέσιμες στον χρήστη είναι τόσες ώστε εκείνος αδυνατεί να τις διαχειριστεί λέγεται *υπερφόρτωση πληροφορίας* (information overload). Μια απάντηση στην ανάγκη για διήθηση των διαθέσιμων πληροφοριών δίνουν τα συστήματα προτάσεων τα οποία αποτελούν τον τομέα που ασχολείται με τον σχεδιασμό και την υλοποίηση τεχνικών για την πρόβλεψη των προτιμήσεων των χρηστών. Οι προβλέψεις αυτές χρησιμοποιούνται για την υποστήριξη των χρηστών στη καθημερινότητά τους, ενώ για την πραγματοποίηση τους χρησιμοποιούνται κατά βάση δυο τεχνικές: (1) η ανάλυση του περιεχομένου των αντικειμένων που προτιμούν οι χρήστες και (2) η ανάλυση της συμπεριφοράς των χρηστών (*συνεργατική διήθηση*, collaborative filtering).

Στο πλαίσιο της διατριβής πραγματοποιήθηκε ενδελεχής βιβλιογραφική μελέτη στα γνωστικά πεδία της χρήσης τεχνικών μηχανικής μάθησης και της πιθανοτικής λανθάνουσας σημασιολογικής ανάλυσης (probabilistic latent semantic analysis) για την πραγματοποίηση προτάσεων. Οι τεχνολογίες αυτές έχουν αναδειχτεί τα τελευταία χρόνια στο πεδίο των συστημάτων προτάσεων καθώς αποτελούν μια απάντηση στην ανάγκη για μη επιβλεπόμενη εξαγωγή συμπερασμάτων που αφορούν την σημασιολογία κειμένου. Πιο συγκεκριμένα, η *λανθάνουσα κατανομή Dirichlet* αποτελεί μια αξιόπιστη προσέγγιση στην εξαγωγή λανθανόντων θεμάτων.

Στην παρούσα διατριβή διερευνάται η δυνατότητα για βελτιωμένα συστήματα προτάσεων στο εσωτερικό επιχειρήσεων, κοινοτήτων και στο εμπόριο με βάση λανθάνοντα θέματα.

Αρχικά εξετάζεται η δυνατότητα των συστημάτων προτάσεων να **ενσωματώσουν την υπάρχουσα γνώση ενός τομέα εφαρμογής**. Από τη σχετική

βιβλιογραφία προκύπτει ότι οι αντίστοιχες προσπάθειες στο παρελθόν αφορούν τεχνικές που απαιτούν τυπικές γνωσιακές δομές και διαρκείς παρεμβάσεις των χρηστών στα συστήματα διαχείρισης γνώσης, ενώ απουσιάζουν τεχνικές που περιλαμβάνουν μη δομημένες, μη τυπικές γνωσιακές δομές και κοινωνικό λογισμικό. Η ερευνητική προσπάθεια της διδακτορικής διατριβής στοχεύει στην κάλυψη του κενού που παρουσιάζεται στη βιβλιογραφία με την ενσωμάτωση μοντέλων θεμάτων και αξιολόγηση των αντίστοιχων συστημάτων προτάσεων.

Στην παρούσα διατριβή δίνεται μία προσέγγιση για την ενσωμάτωση της υπάρχουσας γνώσης ενός πεδίου σε ένα σύστημα προτάσεων. Προτείνεται ένα πλήρες σύστημα προτάσεων το οποία χρησιμοποιεί τα πιθανοτικά μοντέλα θεμάτων για να αποτυπώσει την υπάρχουσα γνώση ενός τομέα όπως αυτή διατυπώνεται σε μη-δομημένο περιεχόμενο. Επίσης, παρουσιάζεται μια μεθοδολογία για τη σύνδεση του με ελαφρού τύπου γνωσιακές δομές με μικρή τυπικότητα. Παράδειγμα αποτελεί η εταιρική γνώση στην περίπτωση ενός συστήματος προτάσεων που λειτουργεί στο εσωτερικό των επιχειρήσεων.

Η συγκεκριμένη προσέγγιση υλοποιήθηκε σε ένα σύστημα λογισμικού το οποίο υποστηρίζει τις δραστηριότητες των χρηστών στο εσωτερικό επιχειρήσεων. Η υποστήριξη αυτή αφορά τόσο την δημιουργία προτάσεων κατά την χρήση του συστήματος από τον εργαζόμενο, όσο και την υποβοήθηση της αναζήτησης με χρήση όρων. Για να πραγματοποιηθεί η υποστήριξη αυτή χρησιμοποιήθηκε η σύνδεση των μοντέλων θεμάτων με τις υπάρχουσες γνωσιακές δομές. Το σύστημα αυτό εγκαταστάθηκε και χρησιμοποιήθηκε στο εσωτερικό πέντε (5) μικρομεσαίων επιχειρήσεων στην Ευρώπη. Ακολούθησε αξιολόγηση από την πλευρά των χρηστών που οδηγεί σε ενθαρρυντικά συμπεράσματα για την χρησιμότητα του συστήματος.

Στη συνέχεια εξετάζεται η δυνατότητα για **εξαγωγή, υπολογισμό και χρήση των δεξιοτήτων των χρηστών ως βάση για την δημιουργία προτάσεων**. Οι μέθοδοι που προτείνονται στη βιβλιογραφία κυρίως αντιμετωπίζουν το πρόβλημα καταγράφοντας χωριστά ποσοτικά και ποιοτικά χαρακτηριστικά των χρηστών, ενώ αγνοούνται τα εργαλεία που χρησιμοποιούνται για τη συνεργασία.

Στην παρούσα διατριβή προτείνεται μια μεθοδολογία που εκμεταλλεύεται την εξαγωγή πιθανοτικών μοντέλων θεμάτων για την πλήρη και αποτελεσματική μοντελοποίηση της ικανότητας ενός εργαζομένου να αντιμετωπίσει ένα πρόβλημα. Η μεθοδολογία αυτή έχει δυο χαρακτηριστικά. Πρώτον, ενσωματώνει στοιχεία από πολλαπλές πηγές, δηλαδή από πολλά συνεργατικά εργαλεία στα οποία

δραστηριοποιείται ο εργαζόμενος. Δεύτερον, συνδυάζει τόσο ποιοτικά όσο και ποσοτικά χαρακτηριστικά της δραστηριότητας του εργαζομένου, λαμβάνοντας υπόψη μετρικές δραστηριότητας αλλά και το κείμενο το οποίο εκείνος παράγει.

Η μεθοδολογία αυτή ενσωματώνεται σε ένα σύστημα προτάσεων που προορίζεται για χρήση στο εσωτερικό μιας κοινότητας ανάπτυξης ελεύθερου λογισμικού ανοιχτού κώδικα. Εκεί σχηματίζονται προφίλ προγραμματιστών που περιλαμβάνουν μια εποπτική εικόνα του τομέα εξειδίκευσης (σε τι είναι ικανός) αλλά και του βαθμού δεξιότητας (πόσο ικανός είναι). Ο σχηματισμός των προφίλ βασίζεται στην εξαγωγή μοντέλων θεμάτων. Η συγκεκριμένη μεθοδολογία αξιολογήθηκε με δεδομένα τόσο από μεγάλες όσο και από μικρότερες σε μέγεθος κοινότητες με ενθαρρυντικά αποτελέσματα.

Τρίτον, στο πλαίσιο της διατριβής αναλύεται η δυνατότητα **βελτιωμένης ενσωμάτωσης της καταναλωτικής συμπεριφοράς σε ένα σύστημα προτάσεων**. Ως τώρα στη βιβλιογραφία έχουν προταθεί διάφορες μέθοδοι για την εκμετάλλευση της καταγεγραμμένης συμπεριφοράς των καταναλωτών για την παραγωγή προτάσεων.

Στην παρούσα διατριβή εξετάζεται η δυνατότητα εξαγωγής και χρήσης μοντέλων θεμάτων για την αποτύπωση των προτιμήσεων των καταναλωτών και την αποτελεσματικότερη παραγωγή προτάσεων αντικειμένων για αγορά. Για την εξαγωγή αυτή χρησιμοποιούνται δυο συναφείς τεχνικές. Πρώτον, η εξαγωγή λανθανόντων θεμάτων από ομάδες προϊόντων που εμφανίζονται μαζί συχνά στο ιστορικό των αγοραστών (λανθάνοντες χρήστες). Δεύτερον, η εξαγωγή λανθανόντων θεμάτων από ομάδες προϊόντων που εμφανίζονται συχνά μαζί σε μεμονωμένες επισκέψεις των αγοραστών (λανθάνοντα καλάθια). Τα μοντέλα λανθανόντων θεμάτων που προκύπτουν χρησιμοποιούνται για την παραγωγή προτάσεων.

Η μεθοδολογία αυτή υλοποιείται σε ένα σύστημα λογισμικού που παράγει προτάσεις προϊόντων. Παραλλαγές των τεχνικών δημιουργίας προτάσεων υλοποιήθηκαν και εφαρμόστηκαν στο σύνολο δεδομένων μιας ελληνικής υπεραγοράς. Ακολούθως, οι τεχνικές αξιολογήθηκαν θετικά, ξεπερνώντας σε επίδοση τους κανόνες συσχέτισης.

Τέλος, διερευνάται η δυνατότητα **βελτιωμένης ενσωμάτωσης του μη δομημένου περιεχομένου των περιγραφών των προϊόντων** σε ένα σύστημα προτάσεων για καταναλωτές ή πωλητές. Η βιβλιογραφική έρευνα ως τώρα

περιλαμβάνει μια ποικιλία μεθόδων για την δημιουργία προτάσεων σε ηλεκτρονικές αγορές, κυρίως με βάση τη συμπεριφορά των άλλων καταναλωτών.

Στο πλαίσιο της διδακτορικής διατριβής προτείνεται μια μεθοδολογία για την εκμετάλλευση του μη δομημένου κειμένου που βρίσκεται σε ηλεκτρονικές αγορές δημοπρασιών για την παραγωγή προτάσεων που απευθύνονται σε καταναλωτές και πωλητές. Οι προτάσεις αυτές αφορούν δυο σκέλη. Πρώτον, την πρόταση σχετικών προϊόντων με αυτό που βλέπει ο επίδοξος αγοραστής. Δεύτερον, την πρόταση συναφών αντικειμένων αλλά και σημαντικών όρων που αφορούν το προϊόν που θέλει να πουλήσει ένας πωλητής. Η δημιουργία αυτών των προτάσεων γίνεται με βάση τα πιθανοτικά μοντέλα θεμάτων που έχουν εξαχθεί από το περιεχόμενο της ηλεκτρονικής αγοράς.

Η μεθοδολογία αυτή υλοποιήθηκε σε ένα λογισμικό σύστημα προτάσεων. Το σύστημα αυτό εφαρμόστηκε σε ένα σύνολο δεδομένων που προέρχεται από ένα διεθνή ιστότοπο ηλεκτρονικών δημοπρασιών υψηλής επισκεψιμότητας. Το σύστημα που προτείνεται αξιολογήθηκε τόσο στη βάση των μοντέλων θεμάτων όσο και στη πρακτική χρησιμότητα του με ενθαρρυντικά αποτελέσματα.

Τα συστήματα προτάσεων με βάση μοντέλα θεμάτων παρουσιάζουν ποιοτικά αλλά και ποσοτικά πλεονεκτήματα σε σχέση με τα ανταγωνιστικά συστήματα. Κατά την εκπόνηση της διατριβής υλοποιήθηκαν συστήματα προτάσεων με βάση πιθανοτικά μοντέλα θεμάτων σε φάσμα εφαρμογών, με διαφοροποιημένα σύνολα δεδομένων και μεγάλο αριθμό χρηστών.

Τα συστήματα προτάσεων που προτείνονται παρουσίασαν κάποια κοινά χαρακτηριστικά που τα κάνουν να πλεονεκτούν σε σχέση με αυτά που έχουν προταθεί:

(1) μειώνουν τις απαιτούμενες διαστάσεις του προβλήματος και παρέχουν γρήγορα προτάσεις αφού έχει προηγηθεί η εξαγωγή των μοντέλων θεμάτων,

(2) ικανοποιούν τις απαιτήσεις των χρηστών για ακρίβεια και ανάκληση όλων των δεδομένων που τους ενδιαφέρουν,

(3) τα θέματα που εξάγονται μπορούν να αποτελέσουν σημαντική πληροφορία για τον ιδιοκτήτη ή τον διαχειριστή του συστήματος.

Λέξεις – Κλειδιά:

Συστήματα Προτάσεων, Πιθανοτική Λανθάνουσα Σημασιολογική Ανάλυση, Πιθανοτικά Μοντέλα Θεμάτων

Summary

This doctoral thesis is positioned in the research area of decision support systems and specifically in recommender systems. It focuses on the design and development of *recommender systems* based on *probabilistic topic models*.

As the volume of available information in the web and in everyday life increases, people find it more difficult to identify the information or the product they look for. The situation where the information available to the user is too much to handle is called *information overload*. An answer to this need for information filtering is placed by recommender systems, which constitute the field of design and implementation of techniques for predicting user preferences. These predictions are used for supporting the users in everyday life, while their algorithms are based mainly in two techniques: (1) the analysis of the content of the objects that users prefer and (2) the analysis of the user behaviour (collaborative filtering).

In this thesis we have thoroughly examined the literature related to utilizing machine learning techniques and probabilistic latent semantic analysis for providing recommendations. These technologies have lately emerged as an answer to the need for unsupervised extraction of semantics from unstructured text. More specifically, *Latent Dirichlet Allocation* is a reliable approach in latent topic extraction.

In this thesis we explore the possibility to design improved recommender systems inside enterprises, communities and in electronic commerce based on latent topics.

Initially, we examine the ability of recommender systems to **integrate existing domain knowledge**. Similar approaches in the literature are related to techniques that require formal knowledge structures and constant user interventions in knowledge management systems, while there is a lack of techniques for unstructured informal knowledge structures and social software. The approach described in this thesis is aiming on covering the gap that is found in the literature by integrating topic models and assessing the related recommender system.

In this thesis we present an approach for integrating existing domain knowledge in a recommender system. We propose a complete recommender system that utilizes probabilistic topic models for representing existing domain knowledge

as expressed in unstructured content. We also present a methodology for connecting domain knowledge with light-weight knowledge structures of low formality. An example is enterprise knowledge in the case of a recommender system that is deployed in an enterprise environment.

This approach was implemented in software that supports user activities inside enterprises. This includes not only generating suggestions for employees while they use the system, but also supporting search by extending queries. To actualize this support we have combined probabilistic topic models with existing knowledge structures. This system has been installed and used in five (5) small and medium enterprises in Europe. Users in these enterprises have evaluated the system and have provided encouraging insight on the system's usefulness.

We also consider the possibility for **extracting, calculating and utilizing the expertise of co-workers as a basis for generating recommendations**. The methods proposed in the literature mainly address the problem by recording separately quantitative and qualitative metrics of the worker activities, while ignoring the content available in text format.

In this thesis we describe a methodology for a complete and effective modelling of user suitability for handling an issue based on probabilistic topic models. This methodology has two characteristics. First, it integrates information from multiple sources; that is from the multiple tools that workers use. Second, it combines the qualitative and quantitative characteristics of worker's behaviour, by taking into account activity metrics and the text that he commits.

This methodology is integrated in a recommender system that is deployed in an open source software development community. There, developers profiles are built that contain an overview of the field of expertise (what is the developer most expert on?) and the degree of competence (how good is he?). Building developer profiles is based on the extracted topic models. This methodology has been evaluated with data not only from large but also from smaller communities with encouraging results.

Thirdly, in this thesis we explore the possibility for **improved integration of consumer behaviour in recommender systems**. In the literature a number of methods have been proposed for utilizing recorded consumer activity for generating recommendations.

In this thesis we examine the possibility of learning and using topic models for recording consumer behaviour and for providing recommendations of products for buying. For extracting topics models, we use two similar techniques. First we propose the extraction of topics from groups of products that appear together often in a user's history (latent users). Second, we propose extracting topic models from groups of products that appear often together in single consumer visits (latent baskets). The topic models that are produced are then used for providing recommendations.

This methodology is implemented in a software system for producing product recommendations. Different variations of the recommendation techniques have been implemented and applied in the dataset of a Greek super market. The techniques have been evaluated positively, surpassing association rules in accuracy.

Finally, we explore the possibility for improved **integration of unstructured content found in product descriptions into a recommender system** for buyers and sellers. The related literature contains a number of methods for generating recommendations in electronic markets, mainly based on the behaviour of consumers.

In this thesis we propose a methodology for exploiting unstructured text that can be found in electronic auction marketplaces for providing recommendations to buyers and sellers. These recommendations can take two forms. First, they can contain products related to what the prospective buyer is currently visiting. Secondly, they can include similar products and terms related to the product a seller intends to sell. These recommendations are produced based on probabilistic topic models that have been extracted from the unstructured content in the electronic marketplace.

This methodology has been implemented in recommender system software. This system has been evaluated in a dataset of a popular electronic auction marketplace. The system evaluation included not only the quality of the topics extracted but also the practical usefulness of the recommender system.

Recommender systems based on topic models display qualitative and quantitative advantages over competing systems. During the preparation of this thesis multiple recommender systems were implemented based on probabilistic topic models in a range of applications with different data sets and a large number of users.

Recommender systems that are proposed in this thesis have displayed a number of common characteristics:

4) They reduce the dimensions of the recommendation problem and provide fast online recommendations, having trained the topic models.

5) They satisfy the user needs for accuracy and recall of all interesting objects.

6) The topics extracted can provide significant insight to the system manager or owner.

Keywords:

Recommender Systems, Probabilistic Latent Semantic Analysis, Probabilistic Topic Models

1 Εισαγωγή

Η διδακτορική διατριβή τοποθετείται στον χώρο των συστημάτων αποφάσεων και ειδικότερα στην περιοχή των *συστημάτων προτάσεων* (recommender systems). Στη διατριβή διερευνάται η δυνατότητα για βελτιωμένα συστήματα προτάσεων στο εσωτερικό επιχειρήσεων, κοινοτήτων και στο εμπόριο με χρήση *πιθανοτικών μοντέλων θεμάτων* (probabilistic topic models).

Το παρόν κεφάλαιο δομείται ως εξής. Παρουσιάζουμε το ερευνητικό περιβάλλον στο οποίο κινείται η διατριβή και στη συνέχεια αναφερόμαστε συνοπτικά στις προκλήσεις που αντιμετωπίζονται. Στη συνέχεια περιγράφονται οι στόχοι της διατριβής και η συνεισφορά της στην επιστήμη. Ακολουθεί η δομή της διατριβής και, τέλος, καταγράφεται η συσχέτιση της διατριβής με ερευνητικά έργα και δημοσιεύσεις.

1.1 Ερευνητικό Περιβάλλον

Η ανθρώπινη καθημερινότητα είναι γεμάτη με πιθανές επιλογές [1]. Σε καθημερινή βάση ο καθένας βρίσκεται αντιμέτωπος με την ανάγκη λήψης αποφάσεων. Οι αποφάσεις αυτές μπορεί να αφορούν τα ρούχα που θα φορέσει, την ταινία που θα δει, τις μετοχές που θα αγοράσει, τις σελίδες που θα μελετήσει στο διαδίκτυο. Ο αριθμός των δυνατών επιλογών είναι συχνά αποτρεπτικός – δεκάδες χιλιάδες ταινίες, δεκάδες χιλιάδες βιβλία, εκατοντάδες εκατομμύρια άρθρα [2].

Όσο ο όγκος των πληροφοριών που γίνεται διαθέσιμος μεγαλώνει ολοένα και περισσότερο στον παγκόσμιο ιστό και γενικά στην καθημερινή ζωή, τόσο δυσχεραίνεται η δυνατότητα των ανθρώπων να εντοπίσουν τα αντικείμενα που τους ενδιαφέρουν. Η κατάσταση κατά την οποία οι πληροφορίες που είναι διαθέσιμες στον χρήστη είναι τόσες πολλές ώστε εκείνος αδυνατεί να τις διαχειριστεί λέγεται *υπερφόρτωση πληροφορίας* (information overload).

Μια απάντηση στην ανάγκη για διήθηση των διαθέσιμων πληροφοριών δίνουν τα συστήματα προτάσεων. Τα συστήματα προτάσεων αποτελούν τον τομέα που ασχολείται με τον σχεδιασμό και την υλοποίηση τεχνικών για την πρόβλεψη

των προτιμήσεων των χρηστών. Η περιοχή των συστημάτων προτάσεων είναι ένας διεπιστημονικός τομέας έρευνας και τεχνολογίας που προέκυψε από την ερευνητική εργασία σε περιοχές όπως η ανάκτηση πληροφορίας, οι τεχνικές προβλέψεων, η επιχειρησιακή έρευνα, η οικονομική ψυχολογία, η ανάλυση κοινωνικών δικτύων και η μοντελοποίηση επιλογών των καταναλωτών στο μάρκετινγκ.

Οι προβλέψεις των συστημάτων προτάσεων χρησιμοποιούνται για την υποστήριξη των χρηστών στη καθημερινότητά τους ενώ για την υλοποίησή τους χρησιμοποιούνται κατά βάση δυο τεχνικές: (1) η ανάλυση του περιεχομένου των αντικειμένων που προτιμούν οι χρήστες και (2) η ανάλυση της συμπεριφοράς των χρηστών (συνεργατική διήθηση, collaborative filtering).

Ιστορικά η ανάγκη για συστήματα που να υποστηρίζουν τις αποφάσεις των ανθρώπων δεν είναι καινούρια. Όμως, τα τελευταία χρόνια οι συνθήκες και οι τεχνολογίες ευνοούν τις εξελίξεις στο συγκεκριμένο ερευνητικό τομέα. Ένα σημαντικό στοιχείο που έχει αλλάξει είναι η έλευση του κοινωνικού ιστού που δίνει την δυνατότητα στους ανθρώπους να συνάπτουν κοινωνικούς δεσμούς μέσω του διαδικτύου.Ακόμη, μια σειρά από εξελίξεις έχουν επηρεάσει σημαντικά τα συστήματα προτάσεων δημιουργώντας νέες προκλήσεις για την λειτουργία τους.

Η παρούσα διατριβή επικεντρώνεται σε συγκεκριμένες προκλήσεις των συστημάτων προτάσεων.

1.2 Προκλήσεις

Οι προκλήσεις με τις οποίες ασχολείται η παρούσα διατριβή αφορούν στις ανάγκες των ανθρώπων που χειρίζονται ζητήματα της καθημερινότητας τους και ιδιαίτερα την διαχείριση του όγκου πληροφοριών και την λήψη σχετικών αποφάσεων. Η υπερφόρτωση πληροφορίας και η αδυναμία των ανθρώπων να λειτουργήσουν αποδοτικά σε ένα περιβάλλον με μεγάλο όγκο πληροφορίας δεν αποτελεί δείγμα μεμονωμένης περίπτωσης χρήσης αλλά είναι κοινός τόπος σε διαφορετικά συστήματα όπου ανταλλάσσεται πληροφορία.

Μια πρόκληση που αντιμετωπίζουν τα συστήματα διαχείρισης γνώσης στο επιχειρηματικό περιβάλλον είναι η υποστήριξη των χρηστών στην ανακάλυψη και χρήση περιεχομένου από την βάση γνώσης της επιχείρησης. Στα εταιρικά συστήματα διαχείρισης γνώσης ένας μεγάλος όγκος πληροφοριών διακινείται

καθημερινά και οι περιορισμοί στην εκμετάλλευση τους οδηγούν σε χαμένη παραγωγικότητα. Πολύ συχνά χρησιμοποιείται κάποια μορφή κατηγοριοποίησης με χρήση γνωσιακών δομών (π.χ. ταξονομιών και οντολογιών). Παρ' όλα αυτά, τα συστήματα που έχουν προταθεί στη βιβλιογραφία δεν εκμεταλλεύονται επαρκώς τις υπάρχουσες δομές για την υποστήριξη των εργαζομένων.

Ακόμη, σε ομάδες εργαζομένων που αναπτύσσουν λογισμικό παρατηρείται το φαινόμενο της απώλειας γνώσης και γίνεται εμφανής η ανάγκη για συστήματα προτάσεων. Οι προγραμματιστές αλλά και οι υπόλοιποι συμμετέχοντες έρχονται καθημερινά αντιμέτωποι με μεγάλο όγκο πληροφοριών που αλλάζουν συνεχώς, βρίσκονται σε διάσπαρτα εργαλεία και πρέπει να χρησιμοποιηθουν για την λήψη σημαντικών αποφάσεων. Τα συστήματα προτάσεων για την ανάπτυξη λογισμικού έχουν αναπτυχθεί για να υποστηρίξουν τα μέλη των κοινοτήτων στη λήψη αποφάσεων [3]. Εντούτοις, οι προγραμματιστές συχνά αποτυγχάνουν να βρουν υποστήριξη στην απάντηση ερωτήσεων όπως «Ποιος είναι ο πιο ικανός προγραμματιστής για να λύσει αυτό το πρόβλημα;» και «Με ποιο πρόβλημα πρέπει να ασχοληθώ;».

Στον τομέα του εμπορίου οι αγοραστές και οι πωλητές συχνά έχουν πολλές επιλογές. Στο εμπόριο εκτός διαδικτύου ο καταναλωτής έρχεται αντιμέτωπος με σημαντικό εύρος δυνατών επιλογών. Τα συστήματα προτάσεων μπορούν να υποστηρίξουν τους καταναλωτές στην πραγματοποίηση επιλογών. Παρόλα αυτά, τα συστήματα που χρησιμοποιούνται παρουσιάζουν κάποιες ελλείψεις. Πρώτον, η συνήθης προσέγγιση της εξαγωγής κανόνων συσχέτισης δεν παρέχει παρά περιορισμένη εποπτική δυνατότητα για τις προτιμήσεις των χρηστών και τα μοτίβα που εκείνες ακολουθούν. Δεύτερον, ένας αριθμός τεχνικών θεμάτων δεν επιτρέπει στα παραδοσιακά συστήματα προτάσεων να αναλύσουν μεγάλα σύνολα δεδομένων αγορών χωρίς προβλήματα – προβλήματα όπως η αγνόηση μεγάλων ομάδων προϊόντων, ή η δυσκολία επέκτασης.

Στο ηλεκτρονικό εμπόριο, οι συμμετέχοντες έρχονται συχνά αντιμέτωποι με μια τεράστια ποικιλία προϊόντων με διαφορετικά χαρακτηριστικά. Ο αριθμός των αντικειμένων που είναι διαθέσιμα για πώληση στις αγορές είναι εξαιρετικά μεγάλος [2]. Η πληροφορία στην περιγραφή του κάθε αντικειμένου είναι σε πολλές περιπτώσεις εκτενής και πυκνή και απαιτεί την πλήρη προσοχή του πιθανού αγοραστή. Ως αποτέλεσμα, ο αγοραστής δεν μπορεί εύκολα να βρει αυτό που ψάχνει. Επιπρόσθετα, εφόσον δεν ξέρει τον ανταγωνισμό, δεν ξέρει πόσα χρήματα μπορεί να ξοδέψει στα συγκεκριμένα αντικείμενα που βρίσκει. Από την άλλη

πλευρά, και οι πιθανοί πωλητές δυσκολεύονται να περιγράψουν και να τιμολογήσουν τα προϊόντα τους λόγω του μεγέθους της αγοράς.

1.3 Στόχοι

Τέσσερις στόχοι καλύπτονται στα πλαίσια της διατριβής.

Αρχικά, εξετάζεται η δυνατότητα των συστημάτων προτάσεων να ενσωματώσουν την υπάρχουσα γνώση ενός τομέα εφαρμογής. Από τη σχετική βιβλιογραφία προκύπτει ότι οι αντίστοιχες προσπάθειες στο παρελθόν αφορούν τεχνικές που απαιτούν τυπικές γνωσιακές δομές και διαρκείς παρεμβάσεις των χρηστών στα συστήματα διαχείρισης γνώσης, ενώ απουσιάζουν τεχνικές που περιλαμβάνουν άτυπες γνωσιακές δομές και κοινωνικό λογισμικό. Η ερευνητική προσπάθεια της διδακτορικής διατριβής στοχεύει στην κάλυψη του κενού που παρουσιάζεται στη βιβλιογραφία με τον συνδυασμό γνωσιακών δομών με μοντέλα θεμάτων και αξιολόγηση των αντίστοιχων συστημάτων προτάσεων.

Ακόμη, εξετάζεται η δυνατότητα για εξαγωγή, υπολογισμό και χρήση των δεξιοτήτων των χρηστών σε μια κοινότητα ως βάση για την δημιουργία προτάσεων. Οι μέθοδοι που προτείνονται στη βιβλιογραφία αντιμετωπίζουν το πρόβλημα κυρίως καταγράφοντας χωριστά ποσοτικά και ποιοτικά χαρακτηριστικά των εργαζομένων, ενώ αγνοούνται συχνά τα εργαλεία που χρησιμοποιούνται για τη συνεργασία των εργαζομένων. Η διατριβή επιδιώκει την κάλυψη αυτής της έλλειψης προτείνοντας ένα σύστημα προτάσεων που συνυπολογίζει ποσοτικά και ποιοτικά χαρακτηριστικά με χρήση μοντέλων θεμάτων σε πολλαπλά εργαλεία και με αξιοποίηση τους για την δημιουργία προτάσεων.

Τρίτον, στο πλαίσιο της διατριβής αναλύεται η δυνατότητα βελτιωμένης ενσωμάτωσης της καταναλωτικής συμπεριφοράς σε ένα σύστημα προτάσεων. Ως τώρα στη βιβλιογραφία έχουν προταθεί διάφορες μέθοδοι για την εκμετάλλευση της καταγεγραμμένης συμπεριφοράς των καταναλωτών για την παραγωγή προτάσεων (εξαγωγή κανόνων συνάφειας, συνεργατική διήθηση, κ.α.). Στην παρούσα διατριβή εξετάζεται η δυνατότητα εξαγωγής και χρήσης μοντέλων θεμάτων για την αποτύπωση των προτιμήσεων των καταναλωτών και την αποτελεσματικότερη παραγωγή προτάσεων αντικειμένων για αγορά.

Τέλος, διερευνάται η δυνατότητα βελτιωμένης ενσωμάτωσης του μη δομημένου περιεχομένου των περιγραφών των προϊόντων σε ένα σύστημα

προτάσεων για καταναλωτές και πωλητές. Η βιβλιογραφική έρευνα ως τώρα παρουσιάζει έναν αριθμός μεθόδων για την δημιουργία προτάσεων σε ηλεκτρονικές αγορές, κυριότερα με βάση τη συμπεριφορά των άλλων καταναλωτών. Η ερευνητική προσπάθεια που παρουσιάζεται στοχεύει στην αξιοποίηση του μη δομημένου περιεχομένου για την υποστήριξη των δραστηριοτήτων των πωλητών και των αγοραστών σε μια ηλεκτρονική αγορά δημοπρασιών.

Κατά την ανάπτυξη της διατριβής τα συγκεκριμένα ερωτήματα τίθενται σε συγκεκριμένες περιοχές χωρίς να χάνεται η γενικότητα των συμπερασμάτων. Μια επισκόπηση των στόχων της διδακτορικής διατριβής φαίνεται στον αντίστοιχο πίνακα (Πίνακας 1.1).

Πίνακας 1.1 Στόχοι της Διατριβής

	Πεδίο Εφαρμογής	**Ερευνητικός Στόχος**
Υπερφόρτωση πληροφορίας	Διαχείριση Γνώσης	Ενσωμάτωση της Υπάρχουσας Γνώσης
		Αποτύπωση και Χρήση Δεξιοτήτων
	Εμπόριο	Ενσωμάτωση Καταναλωτικής Συμπεριφοράς
		Ενσωμάτωση Μη Δομημένου Περιεχομένου

1.4 Συνεισφορά

Η συνεισφορά της διατριβής συνίσταται στα αποτελέσματα της διερεύνησης για βελτιωμένα συστήματα προτάσεων στο εσωτερικό επιχειρήσεων, κοινοτήτων και στο εμπόριο με βάση πιθανοτικά μοντέλα θεμάτων.

Α) Ενσωμάτωση υπάρχουσας γνώσης του τομέα εφαρμογής

Στην παρούσα διατριβή δίνεται μία προσέγγιση για την ενσωμάτωση της υπάρχουσας γνώσης ενός πεδίου σε ένα σύστημα προτάσεων. Προτείνεται ένα πλήρες σύστημα προτάσεων το οποία χρησιμοποιεί τα πιθανοτικά μοντέλα θεμάτων για να αποτυπώσει την υπάρχουσα γνώση ενός τομέα όπως αυτή εμφανίζεται σε μη-δομημένο περιεχόμενο. Επίσης, παρουσιάζεται μια μεθοδολογία γα τη σύνδεση του με ελαφρού τύπου γνωσιακές δομές με μικρή τυπικότητα. Ένα παράδειγμα εφαρμογής είναι η εταιρική γνώση στην περίπτωση ενός συστήματος προτάσεων που λειτουργεί στο εσωτερικό των επιχειρήσεων.

Η μεθοδολογία που περιγράφηκε υλοποιήθηκε σε ένα σύστημα λογισμικού το οποίο υποστηρίζει τις δραστηριότητες των χρηστών στο εσωτερικό επιχειρήσεων. Η υποστήριξη αφορά τόσο την δημιουργία προτάσεων κατά την χρήση του συστήματος από τον εργαζόμενο, όσο και την υποβοήθηση της αναζήτησης όρων. Στην προτεινόμενη μεθοδολογία προτείνεται η σύνδεση των μοντέλων θεμάτων με τις υπάρχουσες γνωσιακές δομές. Το αντίστοιχο σύστημα λογισμικού εγκαταστάθηκε και χρησιμοποιήθηκε στο εσωτερικό πέντε (5) μικρομεσαίων επιχειρήσεων στην Ευρώπη. Ακολούθησε αξιολόγηση από την πλευρά των χρηστών που οδηγεί σε ενθαρρυντικά συμπεράσματα για την χρησιμότητα του συστήματος.

Β) Εξαγωγή, υπολογισμός και χρήση των δεξιοτήτων των χρηστών για την δημιουργία προτάσεων

Στην παρούσα διατριβή προτείνεται μια μεθοδολογία που εκμεταλλεύεται την εξαγωγή πιθανοτικών μοντέλων θεμάτων για την πλήρη και αποτελεσματική μοντελοποίηση της ικανότητας ενός εργαζομένου να αντιμετωπίσει ένα πρόβλημα. Η μεθοδολογία αυτή έχει δυο χαρακτηριστικά. Πρώτον, ενσωματώνει στοιχεία από πολλαπλές πηγές, δηλαδή από πολλά συνεργατικά εργαλεία στα οποία δραστηριοποιείται ο εργαζόμενος. Δεύτερον, συνδυάζει τόσο ποιοτικά όσο και ποσοτικά χαρακτηριστικά της δραστηριότητας του εργαζομένου, λαμβάνοντας υπόψη μετρικές δραστηριότητας αλλά και το κείμενο το οποίο εκείνος παράγει.

Η μεθοδολογία αυτή ενσωματώνεται σε ένα σύστημα προτάσεων που προορίζεται για χρήση στο εσωτερικό μιας κοινότητας ανάπτυξης ελεύθερου λογισμικού ανοιχτού κώδικα. Εκεί σχηματίζονται προφίλ προγραμματιστών που περιλαμβάνουν μια εποπτική εικόνα του τομέα εξειδίκευσης (σε τι είναι ικανός) αλλά και του βαθμού δεξιότητας (πόσο ικανός είναι). Ο σχηματισμός των προφίλ βασίζεται στην εξαγωγή μοντέλων θεμάτων. Η συγκεκριμένη μεθοδολογία

αξιολογήθηκε με δεδομένα τόσο από μεγάλες όσο και από μικρότερες σε μέγεθος κοινότητες με ενθαρρυντικά αποτελέσματα.

Γ) Ενσωμάτωση καταναλωτικής συμπεριφοράς σε ένα σύστημα προτάσεων

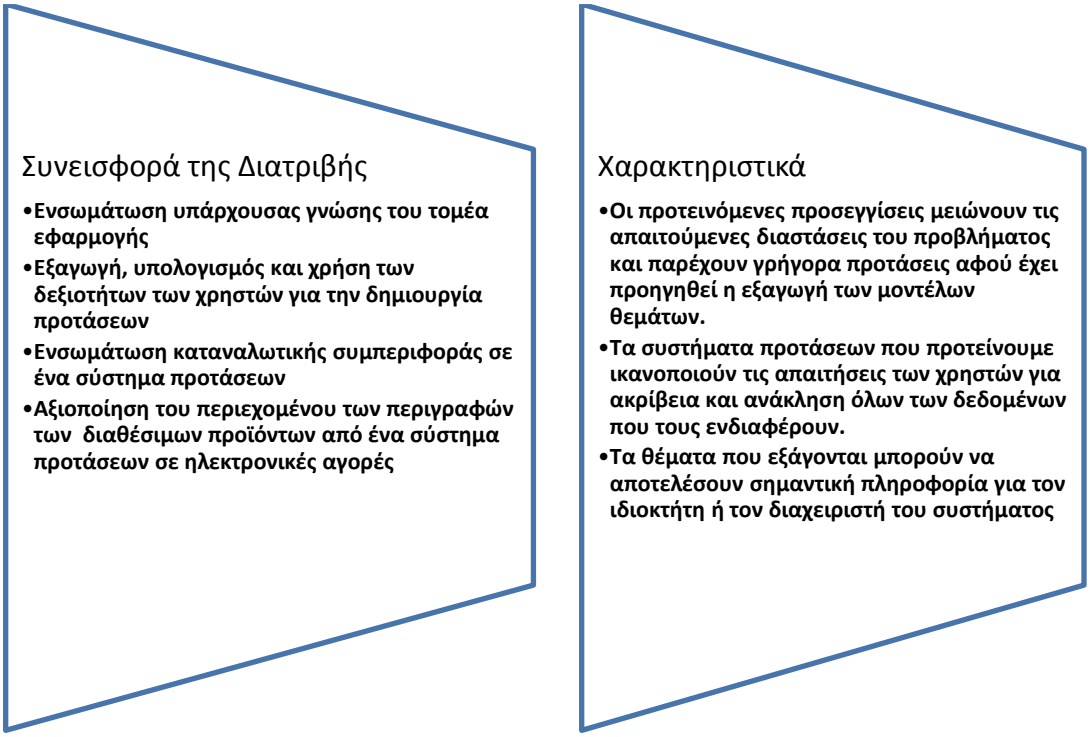
Στην διδακτορική διατριβή προτείνεται και περιγράφεται μια μεθοδολογία εξαγωγής προτιμήσεων για καταναλωτές σε υπεραγορές από ένα σύνολο δεδομένων με χρήση λανθανόντων θεμάτων. Για την εξαγωγή αυτή χρησιμοποιούνται δυο συναφείς τεχνικές. Πρώτον, η ανάλυση με χρήση λανθανόντων θεμάτων των προϊόντων που εμφανίζονται μαζί συχνά στο ιστορικό των αγοραστών (λανθάνοντες χρήστες). Δεύτερον, η ανάλυση των προϊόντων που εμφανίζονται συχνά μαζί σε μεμονωμένες επισκέψεις των αγοραστών (λανθάνοντα καλάθια). Τα εξαγόμενα λανθάνοντα θέματα στη συνέχεια χρησιμοποιούνται για την δημιουργία προτάσεων.

Η μεθοδολογία αυτή υλοποιείται σε ένα σύστημα λογισμικού που παράγει προτάσεις προϊόντων. Παραλλαγές των τεχνικών δημιουργίας προτάσεων υλοποιήθηκαν και εφαρμόστηκαν στο σύνολο δεδομένων μιας ελληνικής υπεραγοράς. Οι προτεινόμενες τεχνικές αξιολογήθηκαν θετικά, ξεπερνώντας σε επίδοση τους κανόνες συσχέτισης.

Δ) Αξιοποίηση του περιεχομένου των περιγραφών των διαθέσιμων προϊόντων από ένα σύστημα προτάσεων σε ηλεκτρονικές αγορές

Στο πλαίσιο της διδακτορικής διατριβής προτείνεται μια μεθοδολογία για την εκμετάλλευση του μη δομημένου κειμένου που βρίσκεται σε ηλεκτρονικές αγορές δημοπρασιών για την παραγωγή προτάσεων που απευθύνονται σε καταναλωτές και πωλητές. Οι προτάσεις αυτές αφορούν δυο σκέλη. Πρώτον, την πρόταση προϊόντων σχετικών με αυτό που βλέπει ο επίδοξος αγοραστής. Δεύτερον, την πρόταση συναφών αντικειμένων αλλά και σημαντικών όρων που αφορούν το προϊόν που θέλει να πουλήσει ένας πωλητής. Η δημιουργία αυτών των προτάσεων γίνεται με βάση τα πιθανοτικά μοντέλα θεμάτων που έχουν εξαχθεί από το περιεχόμενο της ηλεκτρονικής αγοράς.

Η μεθοδολογία αυτή υλοποιήθηκε σε ένα λογισμικό σύστημα προτάσεων. Το σύστημα αυτό εφαρμόστηκε σε ένα σύνολο δεδομένων που προέρχεται από ένα διεθνή ιστότοπο ηλεκτρονικών δημοπρασιών υψηλής επισκεψιμότητας. Το σύστημα που προτείνεται αξιολογήθηκε τόσο στη βάση των μοντέλων θεμάτων όσο και στη πρακτική χρησιμότητα του με ενθαρρυντικά αποτελέσματα.



Εικόνα 1.1 Συνεισφορά της Διατριβής

Κατά την εκπόνηση της διατριβής υλοποιήθηκαν συστήματα προτάσεων με βάση πιθανοτικά μοντέλα θεμάτων σε φάσμα εφαρμογών, με διαφοροποιημένα σύνολα δεδομένων και μεγάλο αριθμό χρηστών. Τα συστήματα προτάσεων που προτείνονται παρουσίασαν κοινά χαρακτηριστικά που τα κάνουν να πλεονεκτούν σε σχέση με αυτά που έχουν προταθεί στη βιβλιογραφία. Αυτά τα χαρακτηριστικά διατρέχουν τους τομείς εφαρμογής και υποστηρίζονται τόσο από την παρούσα διατριβή αλλά και από άλλες αντίστοιχες εργασίες.

Πρώτον, τα πιθανοτικά μοντέλα θεμάτων μειώνουν τις απαιτούμενες διαστάσεις του προβλήματος και παρέχουν γρήγορα προτάσεις αφού έχει προηγηθεί η εξαγωγή των μοντέλων θεμάτων. Η εξαγωγή των μοντέλων θεμάτων απαιτεί σημαντική υπολογιστική ισχύ. Παρόλα αυτά, αφού ολοκληρωθεί και με χρήση ορισμένων παραδοχών όπως θα δούμε στη συνέχεια η παραγωγή προτάσεων είναι γρήγορη.

Δεύτερον, τα πιθανοτικά μοντέλα θεμάτων ικανοποιούν τις απαιτήσεις των χρηστών για ακρίβεια και ανάκληση όλων των δεδομένων που τους ενδιαφέρουν. Αυτό παρατηρείται στα τμήματα της διατριβής όπου οι μέθοδοι αυτές αξιολογούνται από χρήστες και συγκρίνονται με εναλλακτικές όπως συνεργατική διήθηση, κανόνες συσχέτισης, κ.α.

Τρίτον, τα θέματα που εξάγονται μπορούν να αποτελέσουν σημαντική πληροφορία για τον ιδιοκτήτη ή τον διαχειριστή του συστήματος. Είτε αφορούν το περιεχόμενο, είτε αφορούν την δραστηριότητα, τα λανθάνοντα θέματα αποτελούν μια ερμηνεύσιμη αναπαράσταση του περιβάλλοντος του συστήματος προτάσεων.

Μια εποπτική εικόνα της συνεισφοράς της διατριβής μπορεί να βρεθεί στην Εικόνα 1.1.

1.5 Δομή της Διατριβής

Η παρούσα διατριβή έχει δομηθεί ως εξής.

Στο κεφάλαιο 2 παρουσιάζουμε τα συστήματα προτάσεων: την ανάγκη που οδήγησε στην ανάπτυξη τους, τις διαφορετικές κατηγορίες συστημάτων και τις μεθόδους αξιολόγησης. Στο κεφάλαιο 3 αναφερόμαστε στην τεχνολογία των πιθανοτικών μοντέλων θεμάτων, στην λανθάνουσα κατανομή Dirichlet καθώς και σε κάποιες παραλλαγές της. Στο κεφάλαιο 4 παρουσιάζουμε την πρόταση της διατριβής για συστήματα προτάσεων βασισμένα σε πιθανοτικά μοντέλα θεμάτων. Ακολουθεί στο κεφάλαιο 5 η παρουσίαση ενός συστήματος προτάσεων για εταιρικό περιβάλλον που ενσωματώνεται σε κοινωνικό λογισμικό. Στο κεφάλαιο 6 παρουσιάζεται ένα σύστημα προτάσεων για κοινότητες ανάπτυξης ελεύθερου λογισμικού και λογισμικού ανοιχτού κώδικα. Στα κεφάλαια 7 και 8 παρουσιάζουμε τις προσεγγίσεις μας για συστήματα προτάσεων στο εμπόριο και στο ηλεκτρονικό εμπόριο αντίστοιχα. Τέλος στο κεφάλαιο 9 ακολουθεί ο επίλογος, τα συμπεράσματα, οι περιορισμοί και οι κατευθύνσεις μελλοντικής έρευνας που προέκυψαν από την παρούσα διατριβή.

Στο τέλος της διατριβής περιλαμβάνεται η βιβλιογραφία που χρησιμοποιήθηκε. Με τη μορφή παραρτήματος περιλαμβάνεται ένα ευρετήριο που αναφέρει τους όρους που μεταφράστηκαν από την αγγλική γλώσσα καθώς και ένας κατάλογος των δημοσιεύσεων και τον ανακοινώσεων που πραγματοποιήθηκαν στα πλαίσια της διατριβής. Τέλος, περιλαμβάνονται τα ερωτηματολόγια που έχουν χρησιμοποιηθεί στα πλαίσια της εργασίας.

Μια εποπτική εικόνα της δομής της διατριβής μπορεί να βρεθεί στην Εικόνα 1.2.

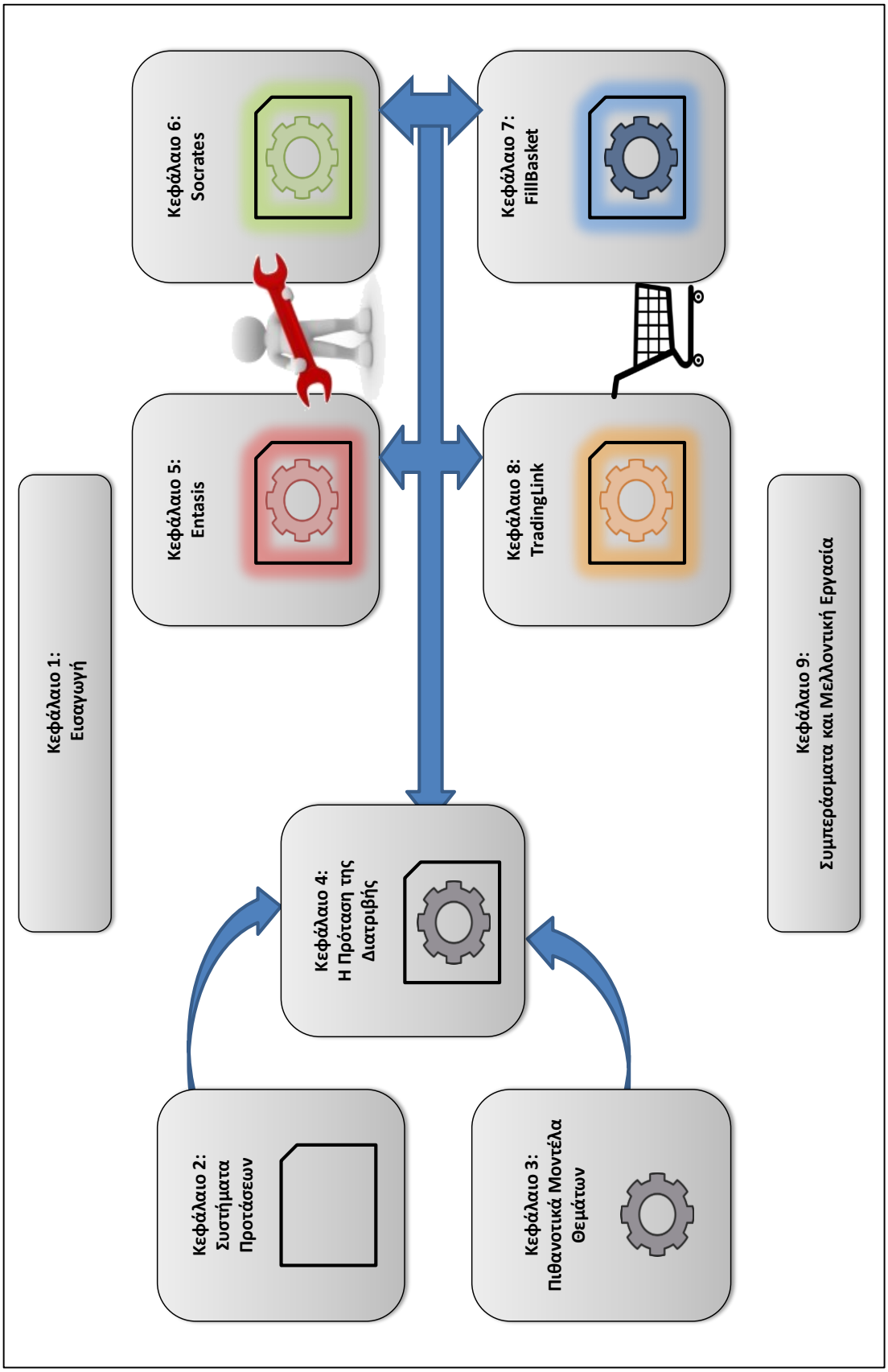
1.6 Σχέση με τις Δημοσιεύσεις

Στην παρούσα ενότητα επιχειρείται μια σύνδεση μεταξύ των δημοσιεύσεων που έχουν πραγματοποιηθεί και της εργασίας όπως χωρίζεται σε κεφάλαια. Οι δημοσιεύσεις που έγιναν στο πλαίσιο της διατριβής αναφέρονται στο τέλος του παρόντος εγγράφου σε αντίστοιχο παράρτημα (βλ. σελ 237). Οι δημοσιεύσεις σε περιοδικά αριθμούνται από [J1] ως [J4] και οι ανακοινώσεις σε συνέδρια δίνονται με τα χαρακτηριστικά [C1] ως [C8].

Στο κεφάλαιο 5 παρουσιάζεται μια προσέγγιση για την ενσωμάτωση και χρήση των γνωσιακών δομών στο εταιρικό περιβάλλον με χρήση των λανθανόντων θεμάτων. Η προτεινόμενη προσέγγιση έχει δημοσιευτεί στις ερευνητικές εργασίες [J2], [J3], [C2] και [C6]. Στο κεφάλαιο 6 παρουσιάζεται μια μεθοδολογία δημιουργίας προτάσεων για κοινότητες ανάπτυξης ελεύθερου λογισμικού και λογισμικού ανοιχτού κώδικα που συνδέεται με την αντίστοιχη ερευνητική εργασία [C3]. Στο κεφάλαιο 7 παρουσιάζεται μια προσέγγιση για την δημιουργία προτάσεων στο εμπόριο εκτός διαδικτύου που αντιστοιχεί στην εργασία [C5]. Τέλος, στο κεφάλαιο 8 παρουσιάζεται μια μεθοδολογία για την δημιουργία προτάσεων σε ηλεκτρονικές αγορές με βάση τα λανθάνοντα θέματα, όπως αυτή έχει δημοσιευτεί στις εργασίες [J1] και [C1]. Οι δημοσιευμένες μελέτες που δεν αναφέρονται έχουν συντελέσει στην διαμόρφωση του προβλήματος και της προσέγγισης της διατριβής και σχετίζονται με το κεφάλαιο 4.

1.7 Σχέση με τα Ερευνητικά Έργα

Η παρούσα διδακτορική διατριβή έχει εν μέρει υποστηριχθεί από την Ευρωπαϊκή Επιτροπή και συγκεκριμένα μέσω των ερευνητικών έργων των Τεχνολογιών της Κοινωνίας της Πληροφορίας (Information Society Technologies). Πιο συγκεκριμένα τα έργα που υποστήριξαν την ανάπτυξη της προτεινόμενης προσέγγισης ήταν το OrganiK (IST-2008-222225) και το Alert (IST-2010-258098).



Εικόνα 1.2 Δομή της Διατριβής

OrganiK (2008-2010)

Ο στόχος του ερευνητικού έργου OrganiK ήταν η έρευνα και η ανάπτυξη ενός καινοτόμου συστήματος διαχείρισης γνώσης που επιτρέπει την σημασιολογική σύνδεση μεταξύ των εταιρικών κοινωνικών εφαρμογών. Το σύστημα που προέκυψε επιτρέπει την συγκέντρωση πληροφοριών που μεταδίδονται στο εσωτερικό αλλά και στο εξωτερικό εταιριών έντασης γνώσης.

Ένα τμήμα της ερευνητικής εργασίας που εντάσσεται στην παρούσα διατριβή πραγματοποιήθηκε στα πλαίσια του OrganiK. Το τμήμα αυτό περιλαμβάνει τον σχεδιασμό και την υλοποίηση συστημάτων προτάσεων με βάση πιθανοτικά μοντέλα θεμάτων για την υποστήριξη των εργαζομένων έντασης γνώσης στο εσωτερικό των επιχειρήσεων (βλ. κεφάλαιο 5).

Alert (2010-2013)

Ο στόχος του ερευνητικού έργου ALERT είναι η ανάπτυξη μεθόδων και εργαλείων που βελτιώνουν τον συντονισμό στην ανάπτυξη ελεύθερου λογισμικού ανοιχτού κώδικα. Το έργο στοχεύει στην διατήρηση της συνεχούς επαφής των προγραμματιστών με τις δραστηριότητες της κοινότητας. Για το σκοπό αυτό χρησιμοποιούνται προσωποποιημένες ενημερώσεις πραγματικού χρόνου που εξαρτώνται από το πλαίσιο της δραστηριότητας του προγραμματιστή. Το ALERT έχει δημιουργήσει μια πλατφόρμα συνεργασίας η οποία επιτρέπει στους προγραμματιστές να αλληλεπιδρούν μεταξύ τους, τους προτείνει δράσεις και επιτρέπει την καλύτερη συνεργασία μεταξύ τους.

Ένα τμήμα της ερευνητικής εργασίας που εντάσσεται στην παρούσα διατριβή πραγματοποιήθηκε στα πλαίσια του Alert. Η εργασία αυτή οδήγησε σε μια μεθοδολογία δημιουργίας προτάσεων για προγραμματιστές που λαμβάνει υπόψη τα λανθάνοντα θέματα του περιεχομένου που δημιουργείται στην ομάδα (βλ. κεφάλαιο 6).

2 Συστήματα Προτάσεων

Η υποστήριξη των ανθρώπων στις επιλογές τους αποτελεί μια σημαντική πρόκληση καθώς ακόμη και οι απλούστερες αποφάσεις μπορεί να αποδειχτούν δύσκολες όταν απουσιάζουν οι σχετικές γνώσεις. Τα συστήματα προτάσεων αποτελούν μια απάντηση σε αυτή ακριβώς την πρόκληση.

Στο κεφάλαιο αυτό επιχειρείται να δοθεί μια γενική εικόνα των συστημάτων προτάσεων. Γίνεται μια ιστορική αναδρομή και δίνεται μια τυπική περιγραφή του προβλήματος της παραγωγής προτάσεων. Στη συνέχεια, τεκμηριώνεται η ανάγκη για συστήματα προτάσεων όπως εμφανίζεται τα τελευταία χρόνια. Ακολούθως, αναλύονται οι πιθανές ταξινομήσεις των συστημάτων προτάσεων, με βάση την τεχνολογία αλλά και την μεθοδολογία παραγωγής προτάσεων. Τέλος περιγράφουμε τις μεθόδους αξιολόγησης και τα σύνολα δεδομένων με τα οποία ελέγχεται η απόδοση των συστημάτων προτάσεων.

2.1 Εισαγωγή

Η ανθρώπινη καθημερινότητα είναι γεμάτη με πιθανές επιλογές [1]: Σε καθημερινή βάση ο κάθε άνθρωπος βρίσκεται αντιμέτωπος με την ανάγκη λήψης αποφάσεων. Οι αποφάσεις αυτές μπορεί να αφορούν τα ρούχα που θα φορέσει, την ταινία που θα δει, τις μετοχές που θα αγοράσει, τις σελίδες που θα μελετήσει στο διαδίκτυο. Ο αριθμός των δυνατών επιλογών είναι πολύ συχνά αποτρεπτικά μεγάλος – δεκάδες χιλιάδες ταινίες, δεκάδες χιλιάδες βιβλία, εκατοντάδες εκατομμύρια άρθρα [2].

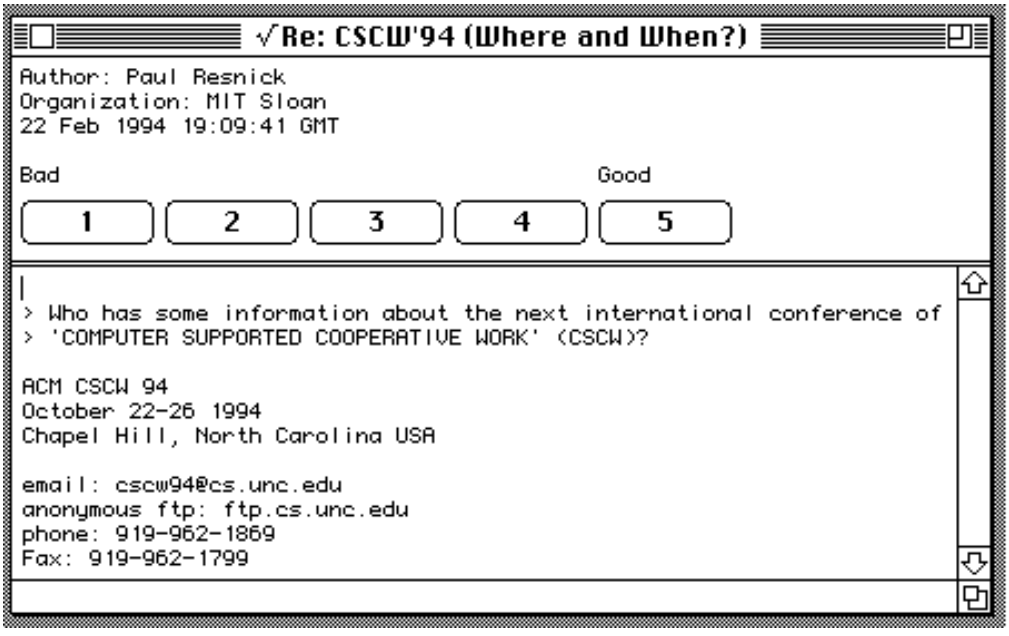
Οι χρήστες των διαφόρων συστημάτων πολύ συχνά χρειάζονται υποστήριξη στην διαδικασία επιλογής. Ακόμη και οι απλούστερες αποφάσεις μπορεί να είναι δύσκολες εάν οι χρήστες δεν διαθέτουν πρόσβαση στην γνώση που να περιγράφει τις πιθανές επιλογές αλλά και χρόνο για να αφιερώσουν στη μελέτη τους. Ιστορικά οι άνθρωποι βασίζονται στις συστάσεις και τις αναφορές από άλλους ανθρώπους, ειδικούς ή μη, ώστε να πάρουν αποφάσεις και να ανακαλύψουν νέα αντικείμενα.

Τα συστήματα προτάσεων αποτελούν μια απάντηση ακριβώς σε αυτή την πρόκληση. Η περιοχή των συστημάτων προτάσεων είναι ένας διεπιστημονικός

τομέας έρευνας και τεχνολογίας που προέκυψε από την ερευνητική εργασία σε περιοχές όπως η ανάκτηση πληροφορίας, οι τεχνικές προβλέψεων, η επιχειρησιακή έρευνα, η οικονομική ψυχολογία, η ανάλυση κοινωνικών δικτύων και η μοντελοποίηση επιλογών των καταναλωτών στο μάρκετινγκ.

2.1.1 Ιστορική Αναδρομή

Η δυνατότητα των ηλεκτρονικών υπολογιστών να παρέχουν προτάσεις έγινε εμφανής σχετικά νωρίς στην ιστορία της πληροφορικής. Το Grundy που παρουσιάστηκε το 1979 [4] ήταν ένα πρώιμο σύστημα προτάσεων που δημιουργούσε πρότυπα χρηστών με βάση μια σύντομη συνέντευξη και πληροφορίες που είχαν εισαχθεί από ειδικούς. Στις αρχές της δεκαετίας του 1990 η συνεργατική διήθηση (collaborative filtering) άρχισε να προτείνεται ως μια πιθανή λύση για την αντιμετώπιση της υπερφόρτωσης πληροφορίας (information overload). Το Tapestry [5] αποτέλεσε ένα σύστημα συνεργατικής διήθησης με ανθρώπινη παρέμβαση στο οποίο ο χρήστης πραγματοποιούσε ερωτήματα σε μια περιοχή με βάση τη δραστηριότητα άλλων χρηστών, για παράδειγμα «Ποια μηνύματα έχουν προωθηθεί από τον συγκεκριμένο χρήστη;».



Εικόνα 2.1 Το Σύστημα Προτάσεων GroupLens

Τα επόμενα χρόνια εμφανίστηκαν τα αυτοματοποιημένα συστήματα συνεργατικής διήθησης που εντοπίζουν αυτόματα απόψεις άλλων χρηστών και τις συνοψίζουν σε χρήσιμες προτάσεις. Το GroupLens [6], η διεπαφή του οποίου

φαίνεται στην Εικόνα 2.1, χρησιμοποίησε τεχνικές συνεργατικής διήθησης για να εντοπίσει άρθρα τα οποία μπορεί να είναι χρήσιμα σε έναν δεδομένο χρήστη. Οι χρήστες απλώς εισήγαγαν βαθμολογίες ή παρείχαν κάποιου τύπου ανάδραση και στη συνέχεια το σύστημα συνδύαζε τις συγκεκριμένες δραστηριότητες των χρηστών ώστε να παράγει εξατομικευμένες προτάσεις. Παρ' όλα αυτά, οι χρήστες δεν αποκτούν ελεύθερη πρόσβαση σε πληροφορίες που αφορούν τις προτιμήσεις των υπόλοιπων χρηστών.

Ακολούθησαν αρκετά συστήματα συνεργατικής διήθησης, όπως το Ringo το οποίο πραγματοποιούσε προσωποποιημένες προτάσεις για άλμπουμ μουσικής και καλλιτέχνες [7]. Στα πλαίσια μιας άλλης μελέτης [8] παρουσιάζεται ένα σύστημα προτάσεων για εικονικές κοινότητες καθώς και η εφαρμογή του για πραγματοποίηση προτάσεων βίντεο. Σε μια διαφορετική προσέγγιση [9], οι συγγραφείς χρησιμοποιούν την *ανάλυση κύριων συνιστωσών* (principal component analysis) ώστε να πραγματοποιήσουν μείωση διαστάσεων για την ομαδοποίηση των χρηστών και τον γρήγορο υπολογισμό των προτάσεων. Το πεδίο εφαρμογής του συγκεκριμένου συστήματος είναι το Jester, ένα online σύστημα πρότασης ανεκδότων. Εκτός των συστημάτων παροχής περιεχομένου, τα συστήματα προτάσεων έδειξαν την χρησιμότητά τους και στο πεδίο του μάρκετινγκ όπου αυξάνουν τις πωλήσεις και βελτιώνουν την εμπειρία των καταναλωτών [10],[11].

Στο τέλος της δεκαετίας του 1990 άρχισαν να εμφανίζονται τα πρώτα εμπορικά συστήματα προτάσεων. Το πρώτο, και ενδεχομένως το γνωστότερο, παράδειγμα συστήματος προτάσεων ενσωματώθηκε στην εταιρία Amazon.com [12]. Οι προτάσεις λάμβαναν την μορφή «χρήστες που είδαν το αντικείμενο που βλέπετε, είδαν επίσης το…». Στη συνέχεια έκαναν την εμφάνιση τους πολλοί ακόμη εμπορικοί τόποι ηλεκτρονικού εμπορίου που υλοποίησαν συστήματα προτάσεων με βάση τη συνεργατική διήθηση. Ακολούθως προτάθηκαν διάφορες μέθοδοι δημιουργίας προτάσεων οι οποίες επίσης λάμβαναν υπόψη το περιεχόμενο των αντικειμένων αλλά και άλλα στοιχεία των προϊόντων. Επίσης προέκυψαν υβριδικά συστήματα προτάσεων στα οποία ώριμοι αλγόριθμοι προτάσεων συνδυάζονται ώστε να σχηματιστούν συστήματα που αποδίδουν καλύτερα από τον κάθε αλγόριθμο που τα αποτελεί.

Το 2006 αποτέλεσε σταθμό στην περιοχή των συστημάτων προτάσεων καθώς η εταιρία πωλήσεων και ενοικιάσεων βιντεοταινιών Netflix προκήρυξε έναν διαγωνισμό ώστε να βελτιώσει τον αλγόριθμο προτάσεων που χρησιμοποιούσε. Ο στόχος του διαγωνισμού ήταν η βελτίωση του υπάρχοντος αλγορίθμου της εταιρίας

κατά 10%, και ο νικητής ανταμείφθηκε με ένα μεγάλο χρηματικό ποσό αποδεικνύοντας το ενδιαφέρον των εμπορικών εταιριών για τον τομέα των συστημάτων προτάσεων [13].

Η βασική αιτία των προβλημάτων που αντιμετωπίζουν τα συστήματα προτάσεων είναι η εγγενής αδυναμία των ανθρώπων να διαβάσουν όλα τα διαθέσιμα βιβλία, να καταναλώσουν όλα τα διαθέσιμα προϊόντα, να παρακολουθήσουν όλες τις διαθέσιμες ταινίες και να ακούσουν όλη τη διαθέσιμη μουσική κατά τη διάρκεια της ζωής τους. Τα συστήματα προτάσεων δανείζονται τεχνικές από την στατιστική και την τεχνητή νοημοσύνη ώστε να προβλέψουν τις προτιμήσεις των χρηστών και να τους υποστηρίξουν στις επιλογές τους. Τα συστήματα προτάσεων έχουν εξελιχθεί σε ένα σημαντικό τομέα έρευνας τις τελευταίες δύο δεκαετίες, γεγονός που αποτυπώνεται τόσο στο ακαδημαϊκό όσο και στο επιχειρηματικό περιβάλλον. Η πρόοδος αυτή είναι περισσότερο εμφανής αν αναλογιστεί κανείς την πορεία από τις πρώτες προσπάθειες για την εφαρμογή συνεργατικής διήθησης έως τα σημερινά πολύπλοκα συστήματα.

2.1.2 Τυπική Περιγραφή Προβλήματος

Τα συστήματα προτάσεων στην πιο γενική τους μορφή, αντιμετωπίζουν το πρόβλημα της εκτίμησης της χρησιμότητας ή της αξιολόγησης αντικειμένων τα οποία ο χρήστης δεν έχει δει ακόμη. Για την επίλυση του συγκεκριμένου προβλήματος, μια σειρά από διαφορετικούς τύπους συστημάτων προτάσεων έχουν προταθεί: συστήματα βασισμένα στο περιεχόμενο και στις προτιμήσεις των χρηστών, αλλά και συστήματα βασισμένα στην μνήμη ή σε μοντέλα [14].

Η αναζήτηση και τα συστήματα προτάσεων αποτελούν, κατά μια έννοια, δυο όψεις του ίδιου νομίσματος που χρησιμοποιούν παρόμοιες τεχνολογίες. Οι Belkin και Croft [15] μετά από εξέταση των δυο τομέων κατέληξαν στο γεγονός ότι υπάρχει μικρή διαφορά μεταξύ της ανάκτησης πληροφορίας και της διήθησης της σε ένα αφαιρετικό επίπεδο, καθώς ο στόχος αλλά και το πλαίσιο λειτουργίας τους είναι παρόμοιο.

Η αναζήτηση αφορά την εύρεση περιεχομένου (συνήθως εγγράφων) μη δομημένης μορφής (συνήθως κειμένου) όπου ικανοποιεί μια ανάγκη για πληροφορία που αφορά μεγάλες συλλογές δεδομένων που συνήθως έχουν αποθηκευτεί σε υπολογιστές [16]. Η αναζήτηση αφορά οποιαδήποτε μορφή διεπαφής που επιτρέπει την πραγματοποίηση μιας σειράς από εργασίες ανάκτησης

πληροφορίας από τον χρήστη. Τα τελευταία πενήντα χρόνια η αναζήτηση έχει εξελιχθεί από μία αναζήτηση δεδομένων, σε αναζήτηση πληροφορίας, σε αναζήτηση με χρήση σύνταξης και τέλος σε σημασιολογική αναζήτηση [17]. Μια τυπική εργασία αναζήτησης είναι αυτή που χρησιμοποιεί ερωτήματα (queries) αλλά η αναζήτηση επίσης περιλαμβάνει την πλοήγηση ανάμεσα σε έγγραφα αλλά και το φιλτράρισμά τους κατά τη διάρκεια της πλοήγησης, συμπίπτοντας έτσι με τα συστήματα προτάσεων. Όσο το περιεχόμενο γίνεται πιο εύκολα διαθέσιμο, τόσο καταβάλλεται περισσότερη προσπάθεια ώστε οι αλγόριθμοι να μπορούν να αξιολογήσουν τη συνάφεια των αποτελεσμάτων καθώς και τη σημαντικότητα της πηγής, ιδιαίτερα στον ιστό. Τελευταία, η σημασιολογική αναζήτηση έχει κερδίσει έδαφος στο συνδυασμό της δομημένης και της μη δομημένης πληροφορίας για την βέλτιστη ανάκτηση αποτελεσμάτων για τις αναζητήσεις των τελικών χρηστών.

Στη γενική του μορφή το πρόβλημα της δημιουργίας προτάσεων μπορεί να οριστεί ως το πρόβλημα της εκτίμησης αξιολογήσεων αντικειμένων τα οποία δεν έχει δει (ή καταναλώσει) ακόμη ένας συγκεκριμένος χρήστης. Για να πραγματοποιηθεί αυτή η εκτίμηση, τα συστήματα προτάσεων μπορούν να λάβουν υπόψη όλες τις πληροφορίες που είναι διαθέσιμες σε εκείνο το χρονικό σημείο (παρελθούσες αξιολογήσεις χρηστών, χαρακτηριστικά των αντικειμένων, αξιολογήσεις από άλλους χρήστες, χαρακτηριστικά των χρηστών, κ.α.). Αφότου πραγματοποιηθεί η εκτίμηση της αξιολόγησης των αντικειμένων το σύστημα προτάσεων μπορεί να πραγματοποιήσει συγκεκριμένες προτάσεις στον χρήστη.

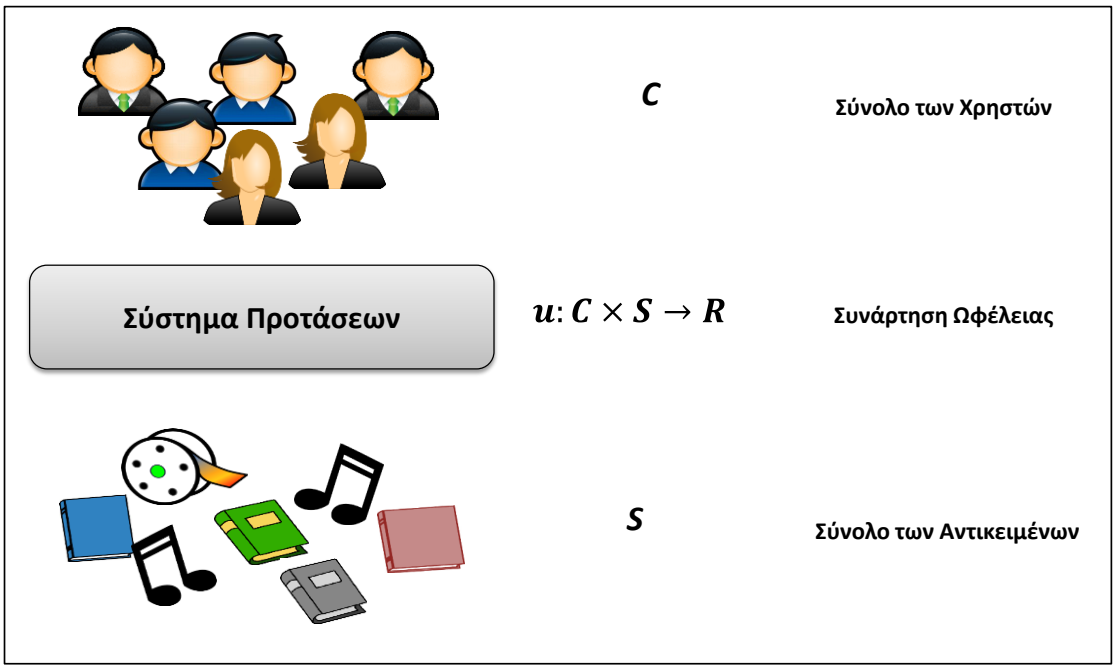
Μια τυπική περιγραφή του προβλήματος παροχής προτάσεων είναι η παρακάτω [14]: έστω ***C*** είναι το σύνολο όλων των χρηστών και ***S*** ένα σύνολο από όλα τα πιθανά αντικείμενα που μπορεί να προταθούν, όπως βιβλία, μουσική, ταινίες, άνθρωποι ή εστιατόρια. Ο χώρος των πιθανών αντικειμένων σε μερικές εφαρμογές είναι εξαιρετικά μεγάλος – για παράδειγμα στην πρόταση μουσικής ή βιβλίων. Επίσης ο χώρος των χρηστών μπορεί να είναι μεγάλος, ανάλογα με την περίπτωση. Ως ***U*** ορίζεται η συνάρτηση ωφέλειας η οποία χρησιμοποιείται για να εκτιμήσει την χρησιμότητα του αντικείμενου ***s*** για τον χρήστη ***c***, όπως στην εξίσωση (2.1) όπου το ***R*** είναι ένα πλήρως ταξινομημένο σύνολο.

$$u: C \times S \rightarrow R \qquad (2.1)$$

Για την δημιουργία προτάσεων, για κάθε χρήστη ***c*** στο σύνολο ***C*** επιλέγουμε έναν αριθμό αντικειμένων ***S*** που μεγιστοποιούν την ωφέλεια του χρήστη (2.2).

$$\forall c \in C, \acute{s}_c = argmax u_{s \in S}(c, s) \qquad (2.2)$$

Μια εποπτική παρουσίαση του προβλήματος της παροχής προτάσεων φαίνεται στην Εικόνα 2.2.



Εικόνα 2.2 Σύστημα Προτάσεων

2.2 Ανάγκη για Προτάσεις

Ιστορικά η ανάγκη για συστήματα που να υποστηρίζουν τις αποφάσεις των ανθρώπων δεν είναι καινούρια. Όπως αναφέρθηκε προηγουμένως, η ανθρώπινη καθημερινότητα είναι γεμάτη με πιθανές επιλογές που μπορεί να αφορούν τα ρούχα που θα φορέσει, την ταινία που θα δει, τις μετοχές που θα αγοράσει, τις σελίδες που θα μελετήσει στο διαδίκτυο. Ιστορικά οι άνθρωποι βασίζονται στις συστάσεις και τις αναφορές από άλλους ανθρώπους, ειδικούς ή μη, ώστε να πραγματοποιήσουν αποφάσεις και να ανακαλύψουν νέα αντικείμενα. Όμως, οι τεχνολογικές εξελίξεις έχουν κάνει τον αριθμό των δυνατών επιλογών αποτρεπτικά μεγάλο – δεκάδες χιλιάδες ταινίες, δεκάδες χιλιάδες βιβλία, εκατοντάδες εκατομμύρια άρθρα [2], δημιουργώντας υπερφόρτωση πληροφορίας (information overload).

Τα τελευταία χρόνια οι συνθήκες και οι τεχνολογίες έχουν επιτείνει και διαφοροποιήσει το πρόβλημα της υπερφόρτωσης πληροφορίας. Ένα σημαντικό στοιχείο που έχει αλλάξει είναι η έλευση του κοινωνικού ιστού που ενθαρρύνει την

δημιουργία περιεχομένου ενώ δίνει τη δυνατότητα στους ανθρώπους να συνάπτουν κοινωνικούς δεσμούς μέσω του διαδικτύου.

2.2.1 Ο Κοινωνικός Ιστός

Κατά τη δημιουργία του ο παγκόσμιος ιστός ήταν μια συλλογή ημι-στατικών σελίδων συνδεδεμένων μεταξύ τους που δεχόντουσαν ανεξάρτητες επισκέψεις χρηστών. Το υπερκείμενο (hypertext) προτάθηκε το 1990 από τον Tim Berners-Lee και τον Robert Cailliau ως ένα ενιαίο περιβάλλον χρήστη μέσω του οποίου θα μπορεί να πραγματοποιείται πλοήγηση σε κόμβους με διαφορετικό περιεχόμενο [18]. Μετά την κατάρρευση πολλών εταιριών υψηλής τεχνολογίας το 2001, διατυπώθηκε η άποψη από τον Dale Dougherty ότι οι εταιρίες που επιβίωσαν από την κρίση είχαν κοινά χαρακτηριστικά που αφορούσαν την χρήση του παγκόσμιου ιστού ως ένα κοινωνικό και διαδραστικό εργαλείο [19].

Ο κοινωνικός ιστός ορίζεται ως το σύνολο των κοινωνικών δεσμών μεταξύ ανθρώπων στα πλαίσια του παγκοσμίου ιστού [20]. Κατά τα τελευταία χρόνια ο παγκόσμιος ιστός έχει μεταλλαχθεί ριζικά: κάποιοι ιστότοποι αλλά και το δικτυακό λογισμικό έχει επανασχεδιαστεί και αναπτυχθεί ώστε να υποστηρίξει τις κοινωνικές σχέσεις επιτρέποντας αλληλεπίδραση μεταξύ ανθρώπων [21]. Αυτές οι αλληλεπιδράσεις στο περιβάλλον του ιστού μπορούν να πάρουν μια σειρά από μορφές: συνεργασία, ψώνια, παιχνίδια, εκπαίδευση κ.α.. Τα κοινωνικά μέσα και οι ιστότοποι κοινωνικής δικτύωσης έχουν επιτείνει αυτή την τάση. Σταδιακά, και με δεδομένη την συνδεσιμότητα παντού αυτές οι αλληλεπιδράσεις δεν περιορίζονται στον προσωπικό υπολογιστή αλλά επεκτείνονται στα κινητά τηλέφωνα και σε υπολογιστές - ταμπλέτες.

Ένας αριθμός εφαρμογών έχει προκύψει που υποστηρίζουν την απευθείας συνεργασία στον κοινωνικό ιστό: άμεσα μηνύματα (instant messaging), δημόσιες συζητήσεις, ιστολόγια, σελίδες wiki, μικρο-ιστολόγια και κοινωνικοί σελιδοδείκτες. Τα άμεσα μηνύματα αποτελούν μια από τις παλιότερες υπηρεσίες του παγκόσμιου ιστού που επιτρέπουν την επικοινωνία των χρηστών του δικτύου σε πραγματικό χρόνο. Οι δημόσιες συζητήσεις παρέχουν χώρο για επικοινωνία για γενικά ή και πιο εξειδικευμένα θέματα. Τα ιστολόγια είναι ιστότοποι όπου κάποιοι μπορούν να εκθέσουν δημόσια τις απόψεις τους και να δεχτούν κριτική με τη μορφή σχολίων. Οι σελίδες wiki είναι εργαλεία για την συνεργατική συγγραφή κειμένου που επιτρέπουν την επεξεργασία από οποιονδήποτε του έχει δοθεί άδεια. Τα μικρο-

ιστολόγια επιτρέπουν την δημοσίευση σύντομων μηνυμάτων που περιγράφουν την τρέχουσα κατάσταση του χρήστη, ενώ οι κοινωνικοί σελιδοδείκτες επιτρέπουν το μοίρασμα ενδιαφερόντων συνδέσμων μεταξύ των χρηστών που έχουν παρόμοια ενδιαφέροντα.

Εκτός από τα επιμέρους εργαλεία, οι ιστότοποι κοινωνικής δικτύωσης και κοινωνικών μέσων έχουν γίνει δημοφιλείς τελευταία. Ξεκινώντας από το MySpace[1], και αργότερα τα Facebook[2], LinkedIn[3] και Twitter[4] αποκτούν ολοένα και περισσότερους χρήστες. Επιπλέον, ένας αριθμός άλλων ιστοτόπων αποκτά σταδιακά κοινωνικές λειτουργίες, κάτι που επηρεάζει και την συμπεριφορά του κοινού. Οι παρεχόμενες υπηρεσίες επιτρέπουν στους ανθρώπους και στους οργανισμούς να έρχονται εύκολα και άμεσα σε επαφή μέσω των σελίδων τους. Εκατοντάδες εκατομμύρια χρηστών χρησιμοποιούν τέτοιους ιστότοπους ώστε να μένουν σε επαφή με την οικογένεια, τους φίλους και τους συναδέλφους τους, να γνωρίζουν καινούργιους ανθρώπους και να μοιραστούν περιεχόμενο, όπως φωτογραφίες, βίντεο, σελιδοδείκτες και δημοσιεύσεις ιστολογίων.

Μια ακόμη θεμελιώδης αλλαγή στα πλαίσια της τεχνολογίας, έχει επηρεάσει ριζικά τον κοινωνικό ιστό – τόσο την έκταση του όσο και την χρήση του. Αυτή η αλλαγή έχει να κάνει με την διάδοση των κινητών συσκευών που παρέχουν πρόσβαση στο διαδίκτυο. Οι περισσότεροι δικτυακοί τόποι κοινωνικής δικτύωσης και κοινωνικών μέσων πλέον υποστηρίζουν και προωθούν την χρήση τους από φορητές συσκευές και έξυπνα τηλέφωνα. Οι χρήστες ενθαρρύνονται να δημοσιεύσουν περιεχόμενο, να μοιραστούν την κατάστασή τους με τους φίλους τους και να λάβουν τις ενημερώσεις του περιεχομένου χρησιμοποιώντας αντίστοιχες πλατφόρμες στο κινητό τους τηλέφωνο. Αυτό επιτρέπει την διατήρηση της επαφής με τους ιστότοπους, και μέσω αυτών με τους φίλους τους, σε όλη τη διάρκεια της ημέρας – ακόμα και όταν είναι μακριά από έναν προσωπικό υπολογιστή. Αυτή η εξέλιξη σηματοδοτεί ριζικές αλλαγές σε κοινωνικές εφαρμογές, επιτρέποντας την εισαγωγή διαστάσεων όπως ο χρόνος, ο τόπος ή το πλαίσιο (context) στο οποίο πραγματοποιούνται οι δραστηριότητες. Παραδείγματα τέτοιων εφαρμογών αποτελούν η επαυξημένη πραγματικότητα (augmented reality), οι υπηρεσίες που βασίζονται στην τοποθεσία, τα μαζικά παιχνίδια ρόλων μέσω

[1] http://www.myspace.com
[2] http://www.facebook.com
[3] http://www.linkedin.com
[4] http://www.twitter.com

δικτύου με πολλούς παίκτες (massive multi-player online role playing games) και άλλα.

Ο κοινωνικός ιστός εξελίσσεται παράλληλα με τις τεχνολογικές αλλαγές και επεκτείνεται σε ολοένα και μεγαλύτερα τμήματα της κοινωνίας. Οι αλλαγές στον κοινωνικό ιστό επηρεάζουν ένα μεγάλο κομμάτι του πληθυσμού, καθώς το 2011 για παράδειγμα υπολογίζεται ότι περίπου το 32% του παγκόσμιου πληθυσμού[5] έχει πρόσβαση στο διαδίκτυο. Ο πληθυσμός του τόπου κοινωνικής δικτύωσης Facebook αγγίζει τα 900 εκατομμύρια[6] – αποτελώντας, θεωρητικά, τη τρίτη μεγαλύτερη χώρα στον κόσμο. Ο αριθμός των ιστολογίων αγγίζει τα 181 εκατομμύρια στο τέλος του 2011 [7] ενώ οι χρήστες των μικρο-ιστολογίων του Twitter φτάνουν τα 140 εκατομμύρια.[8].

Οι αλλαγές στον τρόπο που χρησιμοποιείται ο παγκόσμιος ιστός, μέσω υπολογιστών και φορητών συσκευών, έχουν επηρεάσει τον τομέα των συστημάτων προτάσεων με δυο τρόπους. Πρώτον, η έλευση του κοινωνικού ιστού ευνόησε την δημιουργία μεγάλου όγκου περιεχομένου από τους χρήστες. Ο μεγάλος όγκος περιεχομένου καθιστά δύσκολη την εύρεση ενδιαφέρουσας πληροφορίας κάθε τύπου από τους χρήστες. Δεύτερον, ο τρόπος που οι άνθρωποι αλληλεπιδρούν σε ένα online κοινωνικό περιβάλλον δημιουργεί νέες προκλήσεις αλλά και παρέχει νέες πληροφορίες για τα συστήματα προτάσεων.

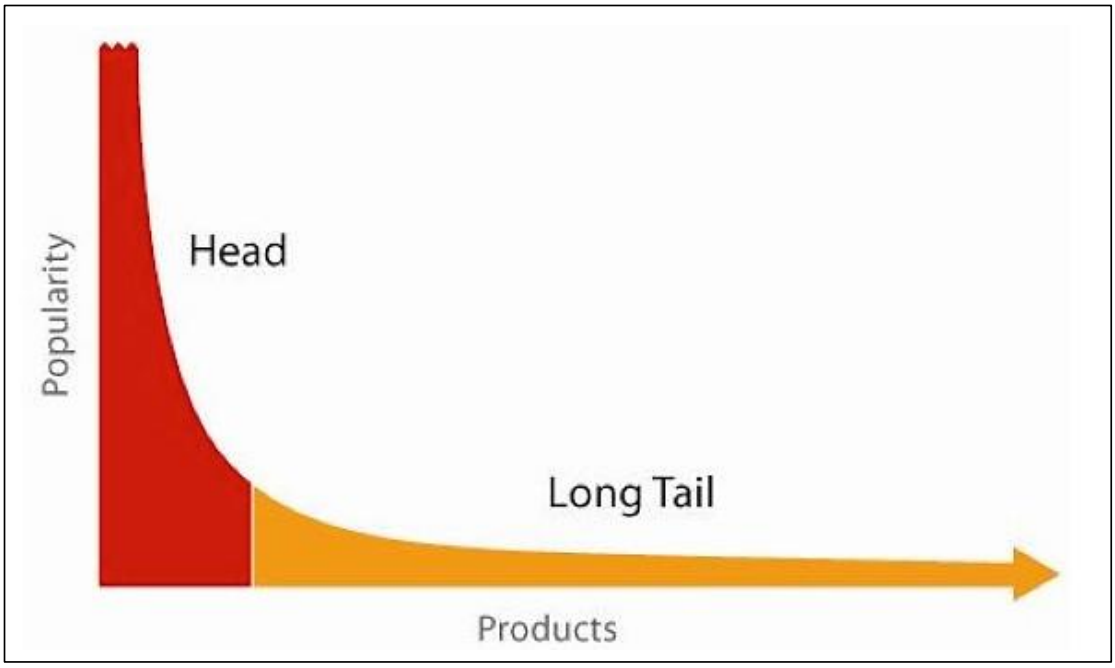
Στον κοινωνικό ιστό, ο καταναλωτής περιεχομένου έχει γίνει και παραγωγός του περιεχομένου. Ο χρήστης μπορεί πλέον να δημιουργήσει και να παρέχει το δικό του περιεχόμενο με τη μορφή κειμένου, εικόνων, μουσικής, ήχου και βίντεο. Ερασιτέχνες σε τομείς όπως η δημοσιογραφία, φωτογραφία, παραγωγή βίντεο, μουσική, ανάπτυξη λογισμικού, σχέδια κ.α. έχουν την ευκαιρία να δημιουργήσουν το περιεχόμενό τους και να το παρέχουν ελεύθερα μέσω του διαδικτύου. Αυτή η έκρηξη στη δημιουργία δημιουργεί ένα πρόβλημα στην κατανάλωση του περιεχόμενου: οι άνθρωποι δεν γνωρίζουν ποια από τα εκατομμύρια αντικείμενα πληροφορίας στον ιστό θα τους ενδιέφεραν.

Η κατάσταση αυτή οδηγεί ένα μεγάλο μέρος του περιεχομένου να παραμένει αθέατο και μη προσβάσιμο για τους περισσότερους χρήστες. Η

[5]http://royal.pingdom.com/2012/04/19/world-internet-population-has-doubled-in-the-last-5-years/
[6]http://blogs.wsj.com/digits/2012/04/23/facebook-passes-the-900-million-monthly-users-barrier/
[7] http://www.dazeinfo.com/2012/03/10/number-of-blogs-up-from-35-million-in-2006-to-181-million-by-the-end-of-2011/
[8] http://blog.twitter.com/2012/03/twitter-turns-six.html

πρόσβαση σε αυτό όμως είναι επιθυμητή από πολλούς χρήστες οι οποίοι ενδεχομένως να δυσκολευτούν να το εντοπίσουν. Το συγκεκριμένο φαινόμενο που σχετίζεται άμεσα με τον κοινωνικό ιστό ονομάζεται φαινόμενο της «μακράς ουράς» (long-tail) και απεικονίζεται και στην Εικόνα 2.3 ([22]). Το καινούριο κοινωνικό και οικονομικό μοντέλο που περιγράφει ο εμπνευστής του όρου [22] επιτρέπει στους καταναλωτές να πραγματοποιήσουν απεριόριστες επιλογές. Θεωρεί επίσης ότι οι καταναλωτές προτιμούν να περιπλανηθούν σε βάθος στον ιστό για να πραγματοποιήσουν τις επιλογές τους, αναζητώντας τίτλους μουσικής ή ταινίες που απομακρύνονται από τους πιο δημοφιλείς και κινούνται σε αυτό που ονομάζει μακρά ουρά του περιεχομένου.

Αποτελεί στόχο ενός συστήματος προτάσεων να δώσει στο χρήστη πρόσβαση σε αντικείμενα που μπορεί να βρει ενδιαφέροντα. Το φαινόμενο της μακράς ουράς τονίζει αυτή την ανάγκη καθώς το σύστημα δεν μπορεί να αρκεστεί στις πλέον δημοφιλείς επιλογές σε μια κοινότητα. Αντιθέτως το φαινόμενο αυτό συμβαδίζει με την ανάγκη για ολοένα και μεγαλύτερη εξατομίκευση του προσφερόμενου περιεχομένου.



Εικόνα 2.3 Το Φαινόμενο «Μακράς Ουράς» στο Περιεχόμενο

Όπως αναφέρθηκε προηγουμένως, η άνοδος του κοινωνικού ιστού έχει επιδράσει βαθιά στην σχεδίαση, λειτουργία αλλά και στόχευση των συστημάτων προτάσεων. Η σχεδίαση των συστημάτων προτάσεων πρέπει να αλλάξει ώστε να

λειτουργήσει σε ένα κοινωνικό περιβάλλον πολλαπλών πηγών δεδομένων, με ενδεχόμενη προστασία της ιδιωτικής ζωής. Τα δεδομένα που χρησιμοποιούνται σε ένα τέτοιο περιβάλλον δεν είναι μόνο προτιμήσεις που συσχετίζουν χρήστες με αντικείμενα, αλλά περιλαμβάνουν και κοινωνικούς δεσμούς μεταξύ των ανθρώπων. Οι περιγραφές των αντικειμένων και των χρηστών περιλαμβάνουν περιεχόμενο που δημιουργείται από τους χρήστες σε πολλαπλές μορφές. Η λειτουργία των συστημάτων προτάσεων πρέπει να αλλάξει ώστε να είναι εφικτό να συμπεριλάβει τις πολλαπλές μορφές δεδομένων και να παρέχει προτάσεις. Σε αυτή την κατεύθυνση, προχωρημένες τεχνικές μηχανικής μάθησης και αλγόριθμοι εξόρυξης δεδομένων έχουν χρησιμοποιηθεί. Τέλος ο στόχος των συστημάτων προτάσεων πρέπει να αλλάξει ώστε να καλύπτει τις ανάγκες των χρηστών των κοινωνικών μέσων. Ο χρήστης πλέον δε χρειάζεται μόνο μια πρόβλεψη για την επόμενη ταινία που θα παρακολουθήσει, αλλά μια πρόταση ανθρώπων ή ομάδων για συνεργασία, δραστηριότητας, κ.α.

Η επίδραση των κοινωνικών μέσων, και των ιστοσελίδων κοινωνικής δικτύωσης στα συστήματα προτάσεων είναι τόσο εμφανής ώστε ορισμένοι ερευνητές χρησιμοποιούν τον όρο «κοινωνικά συστήματα προτάσεων» όταν αναφέρονται σε συστήματα που λειτουργούν στα μέσα κοινωνικής δικτύωσης [23]. Ωστόσο, σε αυτή την διατριβή δεν χρησιμοποιείται κάποιος ειδικός όρος για να σηματοδοτήσει αυτή τη διαφορά.

2.3 Εξελίξεις στα Συστήματα Προτάσεων

Ο τομέας των συστημάτων προτάσεων είναι εξαιρετικά ενεργός καθώς αφορά την επίλυση προβλημάτων που σε μεγάλο βαθμό απαντώνται σε εφαρμογές της καθημερινής ζωής. Για την βελτίωση των συστημάτων προτάσεων, οι ερευνητές αλλά και οι επαγγελματίες του χώρου έχουν σχεδιάσει πολύπλοκες μεθόδους προτάσεων, έχουν επεξεργαστεί βελτιωμένα μοντέλα αναπαράστασης της συμπεριφοράς του χρήστη αλλά και του περιεχομένου των αντικειμένων, χρησιμοποίησαν πολυκριτηριακές μεθόδους αξιολογήσεων και έχουν ενσωματώσει πληροφορίες που αφορούν το πλαίσιο της πρότασης (context).

Παραδείγματα τέτοιων εφαρμογών αποτελούν οι προτάσεις βιβλίων, μουσικής, ταινιών, νέων αντικειμένων και προϊόντων σε ένα εμπορικό περιβάλλον. Άλλα συστήματα προτάσεων μπορούν να χρησιμοποιηθούν για να προτείνουν πληροφοριακούς πόρους, εικόνες ή επισημειώσεις στο εταιρικό και στο προσωπικό

περιβάλλον. Τέλος μια ακόμη ομάδα εφαρμογών που ανήκει στην οικογένεια των συστημάτων προτάσεων περιλαμβάνει την πρόταση ομάδων για συνεργασία, την πρόταση πακέτων διακοπών και προτάσεις πραγματικού χρόνου σε κινητά τηλέφωνα και άλλες συσκευές. Οι εξελίξεις των τελευταίων ετών έχουν δημιουργήσει προκλήσεις που σχετίζονται με τα συστήματα προτάσεων. Κάποιες από τις προκλήσεις που περιγράφουμε στην παρούσα ενότητα είναι το πρόβλημα της ψυχρής έναρξης (cold start), της επίγνωσης του πλαισίου (context awareness), της ετερογένειας, της ποικιλίας (diversity), και της εύνοιας τυχαίων ανακαλύψεων (serendipity).

Ένα σημαντικό πρόβλημα που αφορά τα συστήματα προτάσεων, όπως και άλλα συστήματα μοντελοποίησης χρηστών από δεδομένα, είναι το πρόβλημα της ψυχρής έναρξης (cold start problem) [24]. Το πρόβλημα αυτό αφορά στην παροχή προτάσεων σε χρήστες οι οποίοι μόλις έχουν εισέλθει στο σύστημα και δεν έχουν παρουσιάσει κάποια συμπεριφορά η οποία να μπορεί να αναλυθεί. Επιπρόσθετα, το πρόβλημα αυτό περιλαμβάνει τα αντικείμενα τα οποία μόλις έχουν εισαχθεί στο σύστημα και κανένας από τους καταναλωτές δεν τα γνωρίζει – οπότε δεν μπορεί και να τα προτείνει. Έτσι το σύστημα δεν μπορεί να εξάγει κάποια συμπεράσματα ή κάποιες προβλέψεις για την μελλοντική συμπεριφορά των χρηστών εφόσον δεν έχει αρκετές πληροφορίες από το παρελθόν.

Μια ακόμη πρόκληση για τον σχεδιασμό των συστημάτων προτάσεων αφορά στο πλαίσιο (context) στο οποίο γίνεται η πρόταση. Οι πληροφορίες αυτές περιγράφουν συμπληρωματικά στοιχεία (τα «συμφραζόμενα») στην επιλογή των χρηστών και η αξία τους για την παραγωγή προτάσεων εκτιμάται τόσο από τους ερευνητές όσο και από επαγγελματίες. Τα συστήματα προτάσεων με χρήση πλαισίου γίνονται ολοένα και πιο δημοφιλή στην σχετική έρευνα και βιβλιογραφία [25]. Ένας τρόπος να προσεγγιστεί το πλαίσιο των προτάσεων είναι η προσπάθεια για αποκάλυψη των μακροπρόθεσμων ενδιαφερόντων των χρηστών, αλλά και των βραχυπρόθεσμων αναγκών και της σημασιολογικής γνώσης των χρηστών [26]. Σε μια πρόσφατη προσπάθεια για ταξινόμηση τέτοιων συστημάτων, εφαρμογές αναζήτησης πληροφορίας και ταξιδιωτικοί οδηγοί ξεχώρισαν ως οι πλέον συνηθισμένες μέχρι σήμερα [25]. Κάποιες εφαρμογές έχουν εστιάσει στην αναζήτηση μέσω κινητών συσκευών [27], [28] και βασίζονται σε μοντέλα εκ των υστέρων διήθησης (post-filtering) που ενσωματώνουν το πλαίσιο των χρονικών και χωρικών δεδομένων με τα δεδομένα προτιμήσεων ώστε να υποστηρίξουν την εξερεύνηση δια μέσω της αναζήτησης. Για τους online ταξιδιωτικούς οδηγούς, τα

συστήματα προτάσεων προσαρμόζουν το περιεχόμενο που παρουσιάζεται θεωρώντας στοιχεία όπως ο χρόνος, το κλίμα, η τοποθεσία ή η θερμοκρασία για πρόταση εστιατορίων [29] και φυσικές, κοινωνικές και τοπικές πληροφορίες για την πρόταση σημείων ενδιαφέροντος (place-of-interest, POI).

Η ετερογένεια στα συστήματα προτάσεων αναφέρεται στην πολλαπλότητα των τύπων και των πηγών του περιεχομένου που πρέπει να ληφθούν υπόψη καθώς και στην εφαρμογή πολλαπλών αλγορίθμων προτάσεων [30]. Πιο συγκεκριμένα, στον κοινωνικό ιστό, η πληροφορία που συνδέεται με κάποιους χρήστες και αντίστοιχα αντικείμενα προέρχεται από πολλαπλές πηγές καθώς ο χρήστης μπορεί να διαθέτει λογαριασμούς σε διάφορα κοινωνικά μέσα (σε Facebook, LinkedIn, Twitter, MySpace, κ.λπ.). Αυτοί οι πολλαπλοί λογαριασμοί πρέπει να εντοπιστούν και η πληροφορία που αφορά τις δραστηριότητες του χρήστη πρέπει να συγχωνευτεί στην κατεύθυνση των καλύτερων προτάσεων. Επιπρόσθετα ο τύπος του περιεχομένου μπορεί να είναι σχεδόν οτιδήποτε: κείμενο, εικόνες, βίντεο, ήχος ή μουσική, όπως επίσης και άλλοι άνθρωποι ως πιθανοί φίλοι ή συνεργάτες. Η ανάλυση τέτοιων πολλαπλών τύπων περιεχομένου από διαφορετικές πηγές μπορεί να οδηγήσει σε καλύτερα ενημερωμένες προτάσεις, για παράδειγμα προτάσεις ταινιών με βάση τις μουσικές προτιμήσεις. Τέλος, καθώς διαφορετικοί αλγόριθμοι και προσεγγίσεις προσφέρουν καλύτερα αποτελέσματα σε κάθε σύνολο δεδομένων, είναι σημαντικό να δούμε πως αυτοί οι αλγόριθμοι θα συνδυαστούν για την παραγωγή προτάσεων.

Τέλος, υπάρχει ένα χαρακτηριστικό των συστημάτων προτάσεων που δεν είναι εύκολα αντιληπτό αν τα προσεγγίσουμε ως συστήματα ανάκτησης πληροφορίας. Αυτό το χαρακτηριστικό αναφέρεται με τους όρους ποικιλότητα και εύνοια τυχαίων ανακαλύψεων. Η ποικιλότητα (diversity) αφορά την ιδιότητα της κάθε πρότασης να είναι διαφορετική σε σχέση με τις προηγούμενες προτιμήσεις του χρήστη ή σε σχέση με τις υπόλοιπες προτάσεις. Η εύνοια τυχαίων ανακαλύψεων (serendipity) στο πλαίσιο των συστημάτων προτάσεων αφορά την πιθανότητα ευτυχών ανακαλύψεων σε ένα νέο πλαίσιο ή σε μια καινούρια περιοχή ενώ επίσης απαντάται με τον όρο νεωτερισμός (novelty). Αυτά τα χαρακτηριστικά κινούνται πέρα από γνωστές μετρικές και αντικατοπτρίζουν νέες προκλήσεις καθώς έρχονται σε αντίθεση με κάποιες από τις προϋποθέσεις για ομοιότητα που θεμελιώνουν την χρησιμότητα των συστημάτων προτάσεων. Τα συγκεκριμένα χαρακτηριστικά είναι σημαντικά για την αξιολόγηση όχι μόνο της ακρίβειας των προτάσεων αλλά και της ποιότητας των προτάσεων όπως αυτή γίνεται αντιληπτή

από τον χρήστη – αποτυπώνοντας τον βαθμό που το σύστημα προτάσεων υποστήριξε την εξερεύνηση νέων περιοχών περιεχομένου και την ανακάλυψη ενδιαφερουσών πληροφοριών.

Για να αντιμετωπιστούν οι προκλήσεις που παρουσιάστηκαν σε αυτό το κεφάλαιο τα συστήματα προτάσεων εκμεταλλεύονται τόσο τις προτιμήσεις των χρηστών όσο και το περιεχόμενο των διαφόρων αντικειμένων. Για τον σκοπό αυτό έχει προταθεί η εφαρμογή τεχνικών και αλγορίθμων που προέρχονται από διάφορους τομείς, από την μηχανική μάθηση και την εξόρυξη δεδομένων μέχρι τις εθνολογικές μελέτες και τους ελέγχους A/B στο μάρκετινγκ.

2.4 Ταξινόμηση Συστημάτων Προτάσεων

2.4.1 Μεθοδολογία Προτάσεων

Τα συστήματα προτάσεων ακολουθούν μια βασική ταξινόμηση που αφορά στον τρόπο με τον οποίο παρέχουν προτάσεις στον τελικό χρήστη. Χωρίζονται σε εκείνα που βασίζονται στο περιεχόμενο και σε εκείνα που αναλύουν τη συμπεριφορά των χρηστών (συνεργατική διήθηση).

Στα συστήματα που βασίζονται στο περιεχόμενο, το περιεχόμενο κάθε αντικειμένου αναλύεται και συσχετίζεται με χρήστες με βάση τις προηγούμενες προτιμήσεις τους. Με αυτή τη διαδικασία προβλέπεται η μελλοντική γνώμη κάποιου χρήστη για ένα δεδομένο αντικείμενο. Ένα σύστημα προτάσεων βασισμένο στο περιεχόμενο, στην γενικευμένη μορφή του, μπορεί να περιγραφεί σαν ένα σύστημα επίλυσης του προβλήματος ταξινόμησης κειμένου. Στο συγκεκριμένο πρόβλημα το περιεχόμενο των νέων αντικειμένων αντιστοιχίζεται σε κλάσεις πληροφορίας και ακολούθως σε προφίλ χρηστών.

Τα συστήματα προτάσεων τα οποία βασίζονται στο περιεχόμενο αναλύουν τις περιγραφές των αντικειμένων ώστε να εντοπιστούν εκείνα που μπορεί να ενδιαφέρουν κάθε χρήστη [31]. Τα στοιχεία που διαφοροποιούν τα συστήματα αυτά είναι αφενός ο τρόπος αναπαράστασης των αντικειμένων και των χρηστών, και αφετέρου οι αλγόριθμοι και οι τεχνικές για την πραγματοποίηση προτάσεων.

Συχνά τα αντικείμενα που μπορεί να προταθούν στον χρήστη αποθηκεύονται σε σχεσιακές βάσεις δεδομένων, και πιο συγκεκριμένα σε πίνακες που κάθε γραμμή αντιστοιχεί σε ένα αντικείμενο και κάθε στήλη περιλαμβάνει ένα διαφορετικό

χαρακτηριστικό του. Κάθε αντικείμενο έχει χαρακτηριστικά που το ξεχωρίζουν από τα υπόλοιπα, και αυτός είναι και ο συνηθισμένος τρόπος αποθήκευσης δομημένης πληροφορίας. Σε περιπτώσεις που χρησιμοποιούνται τεχνολογίες σημασιολογικού ιστού για την αναπαράσταση της γνώσης (όπως ταξονομίες ή οντολογίες) η τελική αποθήκευση των δεδομένων μπορεί να γίνει σε γλώσσες σημασιολογικού ιστού όπως RDF[9] ή OWL[10].

Κάποιες φορές το περιεχόμενο που έχουμε στη διάθεσή μας για κάθε αντικείμενο περιλαμβάνει ένα κείμενο που το περιγράφει. Αυτή η πληροφορία είναι μη-δομημένη και μπορεί να αποθηκευτεί ως έχει. Όμως, για να γίνει πιο χρήσιμη για το σύστημα προτάσεων μπορεί να μετατραπεί σε μια δομημένη αναπαράσταση λέξεων και τιμών. Το πρώτο βήμα της μετατροπής αφορά την αντικατάσταση της λέξης από το λήμμα το οποίο την παράγει [32]. Στη συνέχεια, υπολογίζεται η τιμή που συσχετίζει το αντικείμενο με το λήμμα. Η τιμή αυτή είτε είναι δυαδική (0 ή 1, ανάλογα με το αν το λήμμα βρίσκεται ή όχι στην περιγραφή του αντικειμένου), είτε είναι η τιμή TF*IDF [33] (το βάρος της λέξης) ώστε το αντικείμενο να μπορεί να απεικονιστεί στο διανυσματικό χώρο των λέξεων.

Για κάθε χρήστη, τα στοιχεία των αντικειμένων που βρίσκονται στο ιστορικό του συναθροίζονται ώστε να σχηματίσουν το προφίλ του. Το προφίλ του χρήστη μπορεί να συμπληρωθεί με προσθήκες από τον ίδιο το χρήστη οι οποίες περιγράφουν τα ενδιαφέροντά του. Στη γενική της μορφή, η πραγματοποίηση προτάσεων αποτελεί μια προσπάθεια επίλυσης ενός προβλήματος κατηγοριοποίησης, όπου προσπαθούμε να εκπαιδεύσουμε έναν μηχανισμό ώστε να προβλέπει ποια αντικείμενα θα αρέσουν στο χρήστη και ποια όχι.

Ένας αριθμός μεθόδων που χρησιμοποιούν το περιεχόμενο έχει προταθεί στη βιβλιογραφία. Η τεχνική RIPPER [34], παρόμοια με εκείνη των δένδρων αποφάσεων, έχει αποδειχθεί ότι παρουσιάζει επιτυχία στην ταξινόμηση ηλεκτρονικών μηνυμάτων. Η τεχνική των k-πλησιέστερων γειτόνων έχει χρησιμοποιηθεί από συστήματα όπως το «The Daily Learner» [35] για τον εντοπισμό των ενδιαφερόντων του χρήστη. Γραμμικοί ταξινομητές (linear classifiers) και μηχανές διανυσμάτων υποστήριξης (support vector machines) έχουν προταθεί για αντίστοιχα προβλήματα. Επίσης αφελής ταξινομητής Bayes (naïve bayes classifier) φαίνεται ότι αποδίδει ικανοποιητικά σε εφαρμογές προτάσεων όπως φαίνεται στο Syskill & Webert [36].

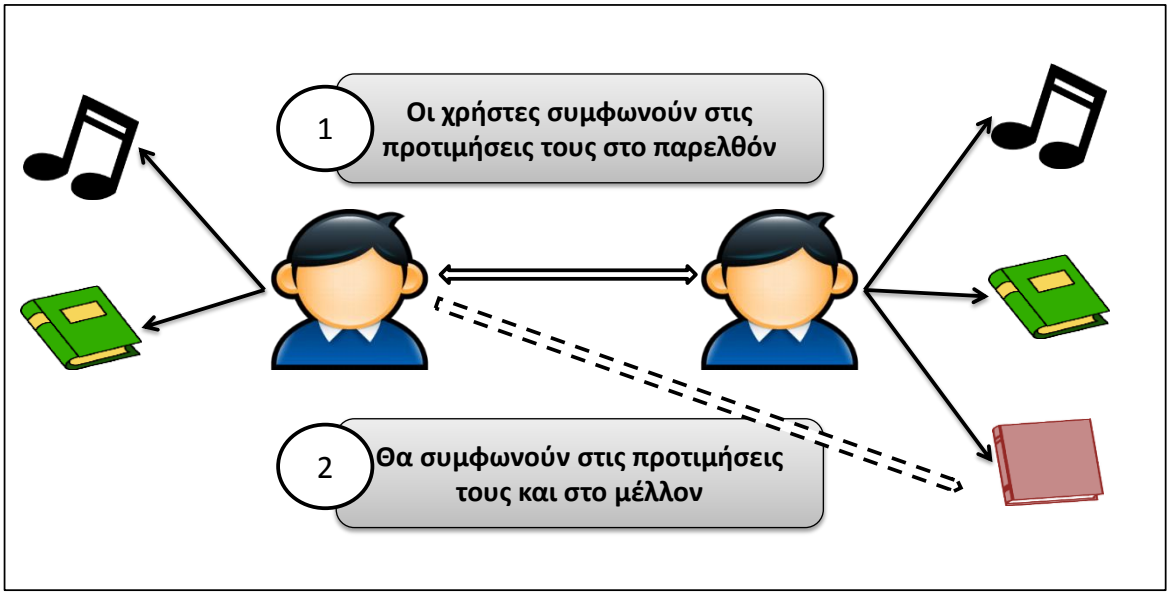
[9] http://www.w3.org/RDF/
[10] http://www.w3.org/OWL/

Στην τεχνική της συνεργατικής διήθησης η συνολική συμπεριφορά των χρηστών λαμβάνεται υπόψη. Για τη λειτουργία τους θεωρείται ως αρχική υπόθεση ότι οι χρήστες που εμφανίζουν παρόμοια συμπεριφορά στο παρελθόν θα συνεχίσουν να συμπεριφέρονται με παρόμοιο τρόπο και στο μέλλον. Στη γενική της μορφή η τεχνική αυτή αναζητά χρήστες που μοιράζονται το ίδιο μοτίβο συμπεριφοράς με τον ενεργό χρήστη.

Πιο συγκεκριμένα η τεχνική της συνεργατικής διήθησης βασίζεται στις γνωστές δραστηριότητες των χρηστών για να υπολογίσει μια πρόταση για τον ενεργό χρήστη [37]. Η πλειονότητα των αλγορίθμων συνεργατικής διήθησης που χρησιμοποιούνται σήμερα, παράγουν αρχικά προβλέψεις των προτιμήσεων του χρήστη και στη συνέχεια πραγματοποιούν προτάσεις ταξινομώντας τα υποψήφια αντικείμενα. Αυτή η στρατηγική που εφαρμόζεται είναι ανάλογη με εκείνη της ανάκτησης πληροφορίας όπου υπολογίζονται οι βαθμολογίες συνάφειας του κάθε αποτελέσματος με το ερώτημα και στη συνέχεια παρέχονται τα αποτελέσματα με την μεγαλύτερη βαθμολογία.

Παρακάτω βλέπουμε διαφορετικές μεθόδους που έχουν προταθεί για την υποστήριξη προτάσεων μέσω συνεργατικής διήθησης. Πρώτον, βλέπουμε μεθόδους αναφοράς χωρίς εξατομίκευση αποτελεσμάτων. Δεύτερον, βλέπουμε την ομάδα των πλέον διαδεδομένων αλγορίθμων προτάσεων Κ πλησιέστερων γειτόνων, τόσο με βάση τα αντικείμενα όσο και με βάση τους χρήστες. Στο σημείο αυτό γίνεται μια αναφορά και στις μετρικές συνάφειας. Τέλος, αναφέρουμε συνοπτικά μεθόδους μείωσης διαστάσεων και πιθανοτικές μεθόδους, στις οποίες συγκαταλέγονται τα μοντέλα θεμάτων.

Εκτός των αλγορίθμων που υποστηρίζουν την παραγωγή εξατομικευμένων προτάσεων, υπάρχουν και μέθοδοι που παρέχουν μη εξατομικευμένες προτάσεις. Οι συγκεκριμένες μέθοδοι μπορούν να χρησιμοποιηθούν ως βάση αναφοράς ή για την προ-επεξεργασία και την κανονικοποίηση των δεδομένων. Ένας τρόπος πραγματοποίησης τέτοιων προβλέψεων είναι ο υπολογισμός της μέσης βαθμολογίας με χρήση όλων των βαθμολογιών στο σύστημα. Μια βελτιωμένη εκδοχή μπορεί να συγκεντρώνει την μέση τιμή των βαθμολογιών για το συγκεκριμένο αντικείμενο.



Εικόνα 2.4 Μέθοδος Πλησιέστερων Γειτόνων

Η πρώτη μεθοδολογία αυτόματων εξατομικευμένων προβλέψεων είναι η συνεργατική διήθηση με βάση τον χρήστη. Στη συγκεκριμένη μεθοδολογία επιχειρούμε να εντοπίσουμε χρήστες που στο παρελθόν είχαν παρόμοια συμπεριφορά με τον ενεργό χρήστη και χρησιμοποιούμε τις προτιμήσεις τους για να προβλέψουμε τις προτιμήσεις του ενεργού χρήστη. Για τον εντοπισμό των χρηστών με παρόμοια συμπεριφορά χρησιμοποιούμε την μέθοδο των πλησιέστερων γειτόνων (βλ. Εικόνα 2.4).

Μια παρουσίαση της γενικής μορφής της συνεργατικής διήθησης γίνεται στην Εικόνα 2.5. Με βάση έναν πίνακα βαθμολογιών ***R*** μπορούμε να υπολογίσουμε την ομοιότητα μεταξύ δυο χρηστών ***s*** (βλ. εξίσωση (2.3)).

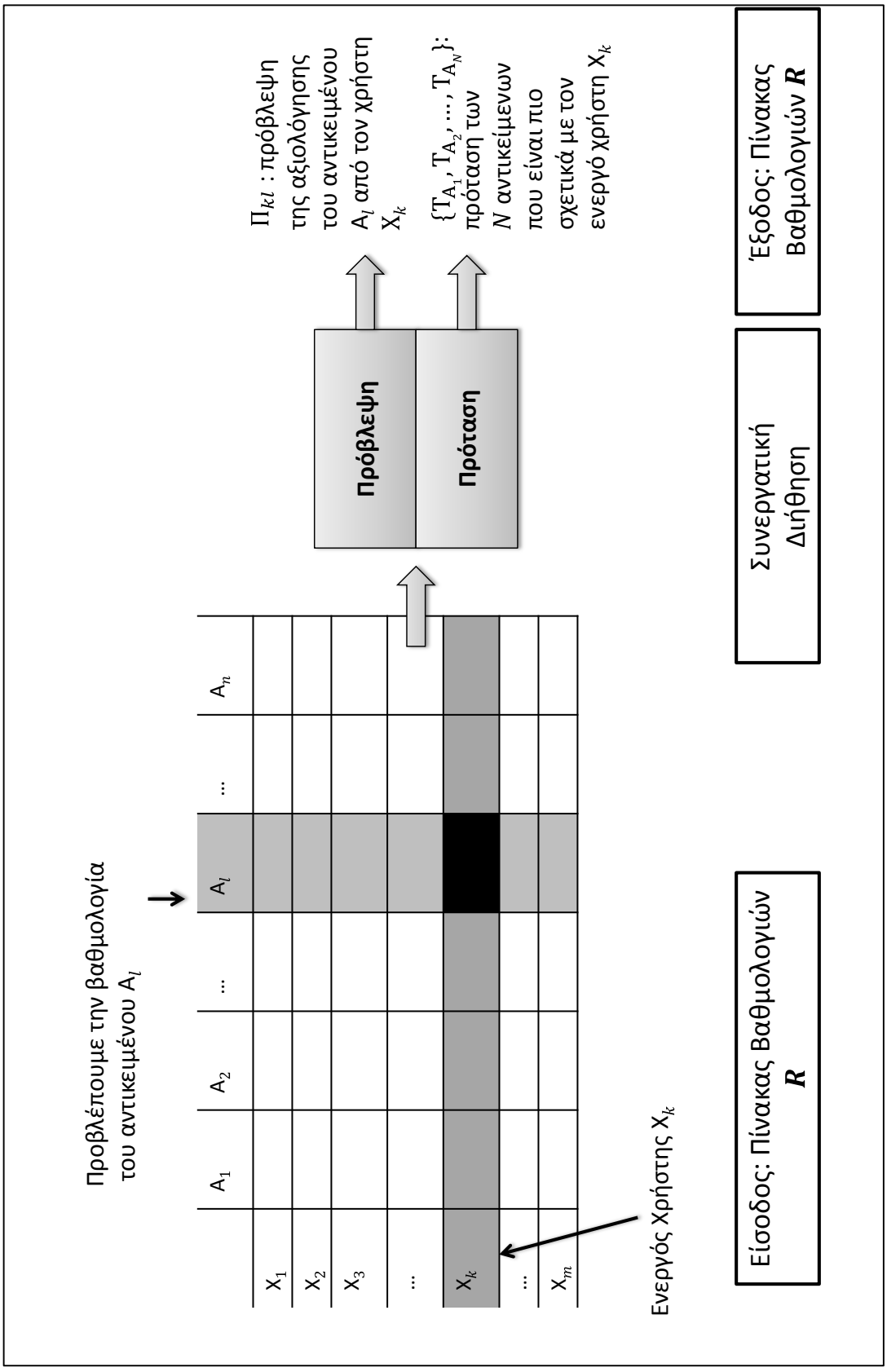
$$s: UxU \rightarrow R \qquad (2.3)$$

Ενώ η συνεργατική διήθηση με βάση τους χρήστες παρουσιάζει ακριβή αποτελέσματα, παρουσιάζει προβλήματα επεκτασιμότητας όσο περισσότεροι χρήστες χρησιμοποιούν το σύστημα (με υπολογιστικό κόστος τουλάχιστον O([U]). Μια πιο επεκτάσιμη μορφή του αλγορίθμου είναι η συνεργατική διήθηση με βάση τα αντικείμενα.

Η συνεργατική διήθηση με βάση τα αντικείμενα αποτελεί την πλέον διαδομένη μορφή συνεργατικής διήθησης που χρησιμοποιείται για εμπορικούς σκοπούς σήμερα. Η συγκεκριμένη μέθοδος [38] υπολογίζει την ομοιότητα μεταξύ των αντικειμένων την οποία χρησιμοποιεί στην συνέχεια για την παραγωγή

προτάσεων. Αν δυο αντικείμενα αξιολογούνται με τον ίδιο τρόπο από τους χρήστες τότε είναι συναφή και οι χρήστες προβλέπεται ότι θα τα αξιολογήσουν με παρόμοιο τρόπο. Η γενικότερη δομή είναι παρόμοια με την πρόταση με βάση το περιεχόμενο, με τη διαφορά ότι δεν χρησιμοποιεί δεδομένα που να περιγράφουν τα αντικείμενα αλλά τις προτιμήσεις των χρηστών. Οι μεγαλύτερες βελτιώσεις που επιτυγχάνει η συγκεκριμένη μέθοδος αφορούν την δυνατότητα της να ανταπεξέρχεται σε σύνολα δεδομένων όπου οι χρήστες είναι πολύ περισσότεροι από τα αντικείμενα αλλά και την δυνατότητα να προϋπολογίζει την ομοιότητα μεταξύ των αντικειμένων.

Για τον υπολογισμό της ομοιότητας έχουν προταθεί διάφορες μετρικές. Η απλούστερη μετρική ομοιότητας είναι η Ευκλείδεια απόσταση στην οποία αντιμετωπίζουμε τα αντικείμενα και τους χρήστες ως διανύσματα. Στην περίπτωση αυτή η μεγαλύτερη απόσταση σημαίνει μικρότερη ομοιότητα. Επιπλέον, πολύ συχνά χρησιμοποιείται η ομοιότητα συνημίτονου η οποία είναι απλώς το συνημίτονο μεταξύ των δυο διανυσμάτων. Ο υπολογισμός της μετρικής αυτής είναι γρήγορος και μπορεί να προβλέψει με ακρίβεια την ομοιότητα μεταξύ δυο διανυσμάτων. Επίσης έχουν προταθεί η συσχέτιση Pearson για την παραγωγή προτάσεων με συνεργατική διήθηση αλλά και η ομοιότητα Tanimoto για σύνολα δεδομένων όπου ο χρήστης δεν αξιολογεί αλλά απλώς επιλέγει να αγοράσει (ή όχι) κάποια προϊόντα.



Εικόνα 2.5 Γενική Μορφή της Συνεργατικής Διήθησης

Στις προσεγγίσεις της συνεργατικής διήθησης πολύ συχνά μεταφέρουμε τις αξιολογήσεις των χρηστών σε διανύσματα. Αυτά τα διανύσματα όμως είναι εξαιρετικά μεγάλων διαστάσεων για να καλύψουν όλα τα πιθανά αντικείμενα. Επίσης από τα συγκεκριμένα διανύσματα απουσιάζουν ορισμένες τιμές, για αντικείμενα που δεν έχουν αξιολογηθεί από τους χρήστες. Για το σκοπό αυτό έχει διερευνηθεί η δυνατότητα μείωσης διαστάσεων των διανυσμάτων. Συγκεκριμένα, έχει προταθεί η εφαρμογή λανθάνουσας σημασιολογικής ανάλυσης για τον περιορισμό των διαστάσεων του προβλήματος της παραγωγής προτάσεων. Οι μεθοδολογίες που ενσωματώνουν την λανθάνουσα σημασιολογική ανάλυση έχουν σχεδιαστεί και αξιολογηθεί με σημαντική επιτυχία και περιγράφονται σε μεγαλύτερη ανάλυση στο κεφάλαιο 4.

Εκτός από τις προαναφερθείσες μεθόδους, έχουν προταθεί και άλλοι τύποι συστημάτων προτάσεων. Ένας από αυτούς είναι οι μέθοδοι μηχανικής γνώσης (knowledge engineering) όπου με τη συνεισφορά ανθρώπων, εντοπίζονται, μοντελοποιούνται και χρησιμοποιούνται οι παράγοντες που επηρεάζουν τις προτιμήσεις των χρηστών [39]. Ο Burke [40] παρέχει μια επισκόπηση των ειδών των υβριδικών συστημάτων προτάσεων – συστημάτων που συνδυάζουν διάφορες τεχνικές σε ένα σύστημα.

2.4.2 Τεχνολογία Προτάσεων

Μια άλλη ταξινόμηση των συστημάτων προτάσεων μπορεί να γίνει με βάση τον τρόπο που το σύστημα αποθηκεύει και επεξεργάζεται τις πληροφορίες που συγκεντρώνονται ώστε να πραγματοποιηθούν προβλέψεις των αξιολογήσεων. Αυτή η ταξινόμηση σε υψηλό επίπεδο πραγματοποιείται μεταξύ των αλγορίθμων με βάση τη μνήμη (memory- based) και αλγορίθμων βασισμένων σε μοντέλα (model-based) [14].

Οι αλγόριθμοι που βασίζονται στη μνήμη, πραγματοποιούν εκτιμήσεις για μελλοντικές αξιολογήσεις με βάση όλες τις προηγούμενες αξιολογήσεις που έχουν γίνει στο παρελθόν. Για να πραγματοποιηθεί αυτό, διατηρούν ένα ιστορικό όλων των γνωστών αξιολογήσεων των χρηστών για όλα τα αντικείμενα. Με βάση αυτά τα δεδομένα πραγματοποιούνται οι προβλέψεις για μια νέα αξιολόγηση: για παράδειγμα, η τιμή μιας άγνωστης αξιολόγησης $r_{c,s}$ από τον χρήστη c για το αντικείμενο s μπορεί να υπολογιστεί ως μια συνάθροιση των αξιολογήσεων κάποιων άλλων χρηστών (συνήθως των πιο συναφών) για το ίδιο αντικείμενο s.

Αυτή η συνάθροιση μπορεί να είναι απλώς η μέση τιμή, ή ένας σταθμισμένος μέσος. Για να πραγματοποιηθούν οι αναγκαίοι υπολογισμοί, συνήθως ορίζεται μια μετρική ομοιότητας μεταξύ χρηστών. Αυτό αποτελεί ένα μέτρο της απόστασης μεταξύ των χρηστών και χρησιμοποιείται ως βάρος στη συνάθροιση. Οι συγκεκριμένοι υπολογισμοί αποτελούν ευρετικές προσεγγίσεις που μοντελοποιούν την διαφορά μεταξύ των χρηστών και απλοποιούν τις προβλέψεις των εκτιμήσεων.

Οι μέθοδοι που βασίζονται στην μνήμη είναι προγενέστερες και αυτή τη στιγμή χρησιμοποιούνται σε πολλά εμπορικά συστήματα καθώς όχι μόνο είναι αποτελεσματικές αλλά είναι και εύκολο να υλοποιηθούν. Χαρακτηριστικά παραδείγματα τέτοιων συστημάτων αποτελούν συστήματα συνεργατικής διήθησης (collaborative filtering) που βασίζονται στη χρήση των πλησιέστερων γειτόνων και χρησιμοποιούν είτε τα αντικείμενα είτε τους χρήστες σαν μέτρο σύγκρισης για να παρέχουν προτάσεις. Στα πλεονεκτήματα των μεθόδων με βάση την μνήμη περιλαμβάνεται η ικανότητα να εξηγηθούν τα αποτελέσματα στους χρήστες, η δυνατότητα για επέκταση του συστήματος και το γεγονός ότι νέα δεδομένα μπορούν να προστεθούν στο σύστημα. Από την άλλη πλευρά, οι συγκεκριμένες μέθοδοι έχουν και κάποια μειονεκτήματα: η απόδοση μειώνεται όταν οι προτιμήσεις που εισάγονται στο σύστημα είναι πάρα πολλές ενώ το σύστημα δεν μπορεί να παρέχει προτάσεις σε νέους χρήστες ή για εντελώς καινούρια αντικείμενα που κανείς δεν έχει αξιολογήσει.

Οι μέθοδοι που βασίζονται σε μοντέλα μετατρέπουν τις προτιμήσεις των χρηστών σε ένα περιγραφικό μοντέλο χρηστών, αντικειμένων και αξιολογήσεων. Οι προτάσεις παράγονται από το μοντέλο ενώ όλες οι προηγούμενες αξιολογήσεις μπορούν να παραμεριστούν καθώς δε χρησιμοποιούνται. Τα μοντέλα συνήθως αναπτύσσονται χρησιμοποιώντας αλγορίθμους από την εξόρυξη δεδομένων και από την μηχανική μάθηση που αποσκοπούν στην εξαγωγή μοτίβων από τις πρότερες προτιμήσεις των χρηστών. Ένα γενικό παράδειγμα της προσέγγισης αυτής είναι η χρήση τεχνικών ταξινόμησης (classification) ή ομαδοποίησης (clustering) για την ταυτοποίηση του χρήστη ενώ ο αριθμός των παραμέτρων περιορίζεται με χρήση τεχνικών μείωσης διαστάσεων. Παραδείγματα αλγορίθμων που χρησιμοποιούνται σε τέτοιες προσεγγίσεις είναι τα μπεϋζιανά δίκτυα (bayesian networks), ομαδοποίηση, λανθάνουσα σημασιολογική ανάλυση (latent semantic alanysis) πολλαπλός πολλαπλασιαστικός παράγοντας (multiple multiplicative factor), λανθάνουσα κατανομή dirichlet (latent dirichlet allocation) και οι διαδικασίες απόφασης markov (markov decision processes).

Τα συστήματα που βασίζονται σε μοντέλα επιτρέπουν μια συνολική προσέγγιση στο πρόβλημα των προτάσεων. Στοχεύουν στην αποκάλυψη παραγόντων που επηρεάζουν τις προτιμήσεις των χρηστών. Αυτές οι προσεγγίσεις παρουσιάζουν ορισμένα πλεονεκτήματα σε σύγκριση με εκείνες που βασίζονται στη μνήμη. Οι προσεγγίσεις που βασίζονται σε μοντέλα μπορούν να χειριστούν αραιά σύνολα δεδομένων με μεγαλύτερη επιτυχία. Επίσης, παρέχουν μια διαισθητική εξήγηση για τις εκτιμήσεις, ενώ ορισμένες φορές αποκαλύπτουν λανθάνοντες παράγοντες στις προτιμήσεις των χρηστών. Τα κύρια μειονεκτήματα αυτής της προσέγγισης είναι το υπολογιστικό κόστος για τη δημιουργία του μοντέλου, καθώς και η πιθανή απώλεια πολύτιμων πληροφοριών λόγω της μείωσης των διαστάσεων.

	Μέθοδοι Βασισμένες στη Μνήμη	**Μέθοδοι Βασισμένες σε Μοντέλα**
Με Βάση το Περιεχόμενο	• TF-IDF • Ομαδοποίηση	• Μπεϋζιανοί Ταξινομητές • Ομαδοποίηση • Δένδρα αποφάσεων • Νευρωνικά Δίκτυα • Πιθανοτικά Μοντέλα
Με Βάση τις Προτιμήσεις	• Εύρεση Πλησιέστερων Γειτόνων • Ομαδοποίηση • Θεωρία Γράφων	• Μπεϋζιανά Δίκτυα • Ομαδοποίηση • Γραμμική Παλινδρόμηση • Νευρωνικά Δίκτυα • Πιθανοτικά Μοντέλα
Υβριδικά Συστήματα Προτάσεων	• Γραμμικός Συνδυασμός • Ψηφοφορίες • Η μία μέθοδος χρησιμοποιείται ως ευρετική	• Η μία μέθοδος χρησιμοποιείται ως ευρετική • Ενοποιημένο Μοντέλο

Πίνακας 2.1 Προσεγγίσεις και Μέθοδοι στα Συστήματα Προτάσεων

2.4.3 Επισκόπηση Εφαρμογών

Ο Πίνακας 2.1 περιέχει μια επισκόπηση της αντιστοίχισης μεταξύ των αλγορίθμων που μπορούν να χρησιμοποιηθούν και των μεθόδων και των προσεγγίσεων συστημάτων προτάσεων. Πρέπει να σημειωθεί ότι αυτή η αντιστοίχιση είναι ενδεικτική, καθώς και ότι ο παρών πίνακας αποτελεί τροποποίηση του πίνακα που έχει παρουσιαστεί σε αντίστοιχη μελέτη [14].

2.5 Αξιολόγηση Συστημάτων Προτάσεων

Η αξιολόγηση των συστημάτων προτάσεων παρουσιάζει ομοιότητες τόσο με τις αξιολογήσεις πληροφοριακών συστημάτων όσο και με την αξιολόγηση συστημάτων μηχανικής μάθησης. Παρακάτω παρουσιάζονται μέθοδοι αξιολόγησης, εκτός σύνδεσης, online και με τη χρήση ερωτηματολογίων [41]. Επίσης παρουσιάζονται τα χαρακτηριστικά των συνόλων δεδομένων που χρησιμοποιούνται συχνότερα στις αξιολογήσεις των συστημάτων προτάσεων.

2.5.1 Μέθοδοι Αξιολόγησης

Για την αξιολόγηση των συστημάτων προτάσεων έχουν προταθεί διάφορες μέθοδοι μεταξύ των οποίων η αξιολόγηση εκτός σύνδεσης (offline), οι αξιολογήσεις από χρήστες και η online αξιολόγηση.

Ο στόχος της αξιολόγησης εκτός σύνδεσης είναι να εξεταστούν οι αλγόριθμοι οι οποίοι χρησιμοποιούνται στα συστήματα προτάσεων. Ο σχεδιασμός της αξιολόγησης πρέπει να γίνεται έτσι ώστε τα δεδομένα που χρησιμοποιούνται εκτός σύνδεσης να ταιριάζουν με τα δεδομένα που θα αντιμετωπίσει το σύστημα αν εγκατασταθεί στην εφαρμογή για την οποία προορίζεται. Πιο συγκεκριμένα πρέπει να βεβαιωθούμε ότι το σύστημα δεν ευνοεί κάποια συγκεκριμένη κατανομή χρηστών, αντικειμένων και αξιολογήσεων. Για παράδειγμα, πρέπει να αποφευχθεί η προεπεξεργασία των δεδομένων στο βαθμό που αποκλείει περιπτώσεις που εμφανίζονται στην πραγματικότητα. Για τη μείωση των δεδομένων μπορεί να χρησιμοποιείται η τυχαία επιλογή, αν αυτό είναι απαραίτητο.

Η βασική δομή της αξιολόγησης εκτός σύνδεσης βασίζεται στον διαχωρισμό των δεδομένων σε δεδομένα εκπαίδευσης και αξιολόγησης. Ως είσοδο το σύστημα δέχεται ένα σύνολο δεδομένων που περιέχει αξιολογήσεις χρηστών σε μια χρονική

περίοδο αλλά και το περιεχόμενο των αντικειμένων. Οι χρήστες στα σύνολα δεδομένων χωρίζονται σε δύο μέρη, στο σύνολο εκπαίδευσης και στο σύνολο αξιολόγησης. Χρησιμοποιώντας το σύνολο εκπαίδευσης, εκπαιδεύεται ένα σύστημα προτάσεων. Στη συνέχεια αξιολογείται η επίδοση του σε ένα σύνολο αξιολόγησης.

Λαμβάνοντας διαφορετικά τμήματα του συνόλου δεδομένων ως σύνολο εκπαίδευσης και αξιολόγησης αντίστοιχα μπορούμε να πραγματοποιήσουμε μια διασταυρωμένη επικύρωση των αποτελεσμάτων με k τμήματα. (k-fold cross-validation). Για να επιτύχουν μεγαλύτερη ομοιότητα με τις πραγματικές συνθήκες στις οποίες πραγματοποιούνται προτάσεις οι ερευνητές πολλές φορές χρησιμοποιούν χρονικά δεδομένα. Όταν στο σύνολο δεδομένων περιέχονται χρονικές πληροφορίες αυτές χρησιμοποιούνται για τον διαχωρισμό σε σύνολα εκπαίδευσης και αξιολόγησης. Ο συγκεκριμένος διαχωρισμός επιβάλλει στο σύστημα προτάσεων να χρησιμοποιήσει μόνο δεδομένα που έχουν ανακληθεί σε προηγούμενο χρόνο για το σχηματισμό μιας πρότασης.

Μια εναλλακτική μέθοδος αξιολόγησης των συστημάτων προτάσεων αφορά την αξιολόγηση από χρήστες. Για το λόγο αυτό απαιτείται από τους χρήστες να χρησιμοποιήσουν το σύστημα προτάσεων. Κατά τη διάρκεια της χρήσης πρέπει να συγκεντρωθούν κάθε τύπου πληροφορίες (ποιες ενέργειες ολοκληρώθηκαν, σε πόσο χρόνο, με πόση ακρίβεια, κ.α.). Τέλος, αφού αλληλεπιδράσουν με το σύστημα οι χρήστες καλούνται να απαντήσουν ερωτήσεις για να καλύψουν στοιχεία τα οποία δεν καλύπτονται από την καταγραφή των δραστηριοτήτων τους.

Για την αξιολόγηση από χρήστες μπορούν να χρησιμοποιηθούν και ερωτηματολόγια. Η χρονική στιγμή της συμπλήρωσης των ερωτηματολογίων μπορεί να τοποθετηθεί πριν, κατά τη διάρκεια και μετά από τη χρήση του συστήματος.

Η online αξιολόγηση αποτελεί τον πλέον ρεαλιστικό τρόπο αξιολόγησης των συστημάτων προτάσεων. Στην αξιολόγηση αυτή ο σχεδιαστής του συστήματος μπορεί να μετρήσει ποσοτικά τις δραστηριότητες των χρηστών. Έτσι μπορεί να μετρηθεί στην πράξη ο βαθμός στον οποίο αυτές επηρεάζονται από την χρήση διαφορετικών στρατηγικών στα συστήματα προτάσεων . Μια τέτοια μεθοδολογία είναι η A/B αξιολόγηση (A/B test).

2.5.2 Σύνολα Δεδομένων

Τα συστήματα προτάσεων αναπτύσσονται στα πλαίσια συγκεκριμένων εφαρμογών, και αντίστοιχα οι αξιολογήσεις τους γίνονται με βάση συγκεκριμένα και διαφορετικά κάθε φορά σύνολα δεδομένων. Στη βιβλιογραφία εμφανίζονται γενικά σύνολα δεδομένων που έχουν γίνει διαθέσιμα στο κοινό. Αποτελούν μια βάση μέτρησης για την αποτελεσματικότητα των μεθόδων παραγωγής προτάσεων, ιδίως για την σύγκριση διαφορετικών προσεγγίσεων. Επίσης μπορούν να χρησιμοποιηθούν στα πλαίσια μιας προκαταρκτικής αξιολόγησης των τεχνικών προτάσεων.

Κάποια σύνολα δεδομένων έχουν χρησιμοποιηθεί αρκετά στη βιβλιογραφία. Το σύνολο δεδομένων του MovieLens περιλαμβάνει αξιολογήσεις ταινιών και έχει γίνει διαθέσιμο σε τρεις διαφορετικές εκδόσεις με διαφορετικά μεγέθη[11]. Ένα δεύτερο σύνολο δεδομένων που έχει χρησιμοποιηθεί ευρέως είναι το σύνολο δεδομένων που έγινε διαθέσιμο το 2006 στα πλαίσια του διαγωνισμού της Netflix και αποσύρθηκε στα τέλη του 2009. Τέλος, το σύνολο δεδομένων του Epinions[12] έχει χρησιμοποιηθεί εκτενώς για την αξιολόγηση αλγορίθμων που αφορούν κοινωνικά δίκτυα και σχέσεις εμπιστοσύνης.

Για να αποτυπωθεί μια πιο ρεαλιστική εικόνα της αξιολόγησης του αλγορίθμου παραγωγής προτάσεων είναι σκόπιμο να χρησιμοποιούνται σύνολα δεδομένων που προέρχονται από τον τομέα εφαρμογής.

2.6 Συμπεράσματα

Τα συστήματα προτάσεων αφορούν έναν τομέα που σχετίζεται τόσο με την μηχανική μάθηση και με την εξόρυξη δεδομένων, όσο και με το μάρκετινγκ, την ψυχολογία και την επιχειρησιακή και οργανωσιακή έρευνα. Η ανάγκη για συστήματα προτάσεων έχει διαμορφωθεί και μεγεθυνθεί κατά την τελευταία χρονική περίοδο επηρεασμένη από την ανάπτυξη του κοινωνικού ιστού.

Στο κεφάλαιο αυτό έχουμε καταγράψει την ιστορική πορεία του κλάδου των συστημάτων προτάσεων καθώς και τις προκλήσεις που δημιούργησαν την ανάγκη για τα συγκεκριμένα συστήματα. Επίσης έχουμε παρουσιάσει τις ταξινομήσεις των

[11] www.grouplens.org/node/73
[12] http://www.epinions.com/?sb=1

συστημάτων με βάση την πηγή δεδομένων που χρησιμοποιούν αλλά και με βάση την μεθοδολογία παραγωγής προτάσεων. Τέλος περιγράφουμε διαφορετικές μεθοδολογίες για την αξιολόγηση των συστημάτων προτάσεων καθώς και τα σύνολα δεδομένων.

Για την επέκταση και βελτίωση των προτάσεων που παράγονται έχει προταθεί η τεχνολογία των πιθανοτικών μοντέλων θεμάτων. Οι βασικές αρχές της συγκεκριμένης τεχνολογίας παρουσιάζονται στο επόμενο κεφάλαιο.

3 Πιθανοτικά Μοντέλα Θεμάτων

Στο συγκεκριμένο κεφάλαιο παρουσιάζουμε την τεχνολογία των πιθανοτικών μοντέλων θεμάτων ως τρόπο μοντελοποίησης συνόλων διακριτών δεδομένων και πιο συγκεκριμένα συλλογών κειμένων.

Παρουσιάζουμε μια σύντομη επισκόπηση και σύγκριση μεταξύ των διαφόρων μεθοδολογιών ανάκτησης πληροφορίας. Στη συνέχεια παρουσιάζουμε τα πιθανοτικά μοντέλα θεμάτων και την πιο διαδεδομένη μορφή τους, την λανθάνουσα κατανομή Dirichlet και τον αντίστοιχο αλγόριθμο εύρεσης θεμάτων. Τέλος παρουσιάζουμε τις πιο σημαντικές παραλλαγές των πιθανοτικών μοντέλων θεμάτων.

3.1 Εισαγωγή

Ένα σημαντικό πρόβλημα που αντιμετωπίζεται στο χώρο της ανάκτησης πληροφορίας (information retrieval) είναι η μοντελοποίηση συλλογών διακριτών δεδομένων και πιο συγκεκριμένα συλλογών κειμένων. Στόχος της έρευνας στο συγκεκριμένο πεδίο είναι η εξαγωγή περιγραφών των μελών της συλλογής που να επιτρέπουν αποτελεσματική επεξεργασία διατηρώντας τις πληροφορίες που είναι απαραίτητες για εργασίες όπως πλοήγηση, αναζήτηση εγγράφων, περίληψη κειμένου και υπολογισμός συνάφειας.

Για την αντιμετώπιση του συγκεκριμένου προβλήματος έχουν προταθεί μια σειρά από προσεγγίσεις που παρουσιάζουν διαφορετική ακρίβεια αλλά και διαφορετικούς βαθμούς εξάπλωσης σε πρακτικές εφαρμογές [17].

Για την ανάκτηση πληροφορίας το πρώτο και απλούστερο μοντέλο που έχει προταθεί είναι το λογικό μοντέλο το οποίο βασίζεται στην ύπαρξη ή όχι ενός χαρακτηριστικού σε ένα αντικείμενο. Το διανυσματικό μοντέλο μετατρέπει κάθε κείμενο σε ένα διάνυσμα πραγματικών αριθμών και εφαρμόστηκε με επιτυχία στις σύγχρονες μηχανές αναζήτησης στο Διαδίκτυο. Επίσης, για την αναπαράσταση και επισημείωση των εγγράφων έχει προταθεί η χρήση διαφόρων γνωσιακών δομών, κυρίως από το χώρο του σημασιολογικού ιστού. Τέλος, στα πλαίσια του διανυσματικού μοντέλου έχει προταθεί η λανθάνουσα σημασιολογική ανάλυση ως

μια προσέγγιση για τον σχηματισμό διανυσματικών αναπαραστάσεων του κειμένου που αντανακλούν το σημασιολογικό τους περιεχόμενο.

3.1.1 Ορολογία

Κάποιοι όροι είναι κοινοί στις μεθόδους ανάκτησης πληροφορίας και χρησιμοποιούνται σε όλες τις μεθοδολογίες που περιγράφονται στο παρόν κεφάλαιο.

Η λέξη «***έγγραφο***» περιγράφει τα αντικείμενα τα οποία αναζητούμε, έχουν διαφορετικά χαρακτηριστικά και συνήθως περιέχουν κείμενο. Η λέξη «***όρος***» αφορά στην πιθανή λέξη που μπορεί να βρεθεί σε κάποιο έγγραφο και να επηρεάζει την σημασιολογική του περιγραφή. Το σύνολο των όρων που μπορούν να βρεθούν σε μια συλλογή κειμένων είναι το «**λεξιλόγιο**» της συλλογής. Τέλος η λέξη «***ερώτημα***» αφορά στην έκφραση που χρησιμοποιεί ο χρήστης ενός συστήματος ανάκλησης πληροφοριών για να περιγράψει τα αντικείμενα που θέλει να του επιστραφούν.

3.1.2 Λογικό Μοντέλο

Το λογικό μοντέλο [17] είναι ένα απλό μοντέλο ανάκτησης πληροφορίας που βασίζεται στην θεωρία συνόλων και στην άλγεβρα Boole.

Στο μοντέλο θεωρούμε ότι οι διάφοροι όροι μπορούν είτε να είναι παρόντες είτε να απουσιάζουν από ένα συγκεκριμένο έγγραφο. Για να περιγράψουμε ένα έγγραφο χρησιμοποιούμε μια σειρά από βάρη τα οποία το συνδέουν με όλους τους πιθανούς όρους στο λεξιλόγιο. Τα βάρη που συνδέουν τα έγγραφα με τους αντίστοιχους όρους είναι δυαδικά και παίρνουν είτε την τιμή 0 είτε την τιμή 1. Ένα ερώτημα διατυπώνεται ως μια συλλογή όρων οι οποίοι συνδέονται με διάφορους τελεστές Boole. Το ερώτημα αυτό αποτελεί στην ουσία μια έκφραση Boole και η απάντηση σε αυτό το ερώτημα είναι η απάντηση του συστήματος στην αναζήτηση που κάνει ο χρήστης. Το λογικό μοντέλο μπορεί να προβλέψει αν ένα έγγραφο είναι σχετικό ή όχι με το ερώτημα, ενώ δεν μπορεί κάποιο έγγραφο να είναι εν μέρει σχετικό.

Τα βασικά πλεονεκτήματα του μοντέλου σε σχέση με ανταγωνιστικές προσεγγίσεις είναι ο υψηλός βαθμός τυπικότητας των ερωτημάτων αλλά και η απλότητά του. Στα μειονεκτήματά του συγκαταλέγεται η πιθανότητα το μοντέλο να επιστρέψει πολύ λίγα ή πάρα πολλά έγγραφα ως απάντηση σε κάποιο ερώτημα.

Επίσης πολύ συχνά η ανάγκη για κάποια πληροφορία δεν μπορεί να μεταφραστεί εύκολα σε μια έκφραση Boole.

3.1.3 Απλό Διανυσματικό Μοντέλο

Στο διανυσματικό μοντέλο [42] αντιμετωπίζονται οι περιορισμοί που τίθενται από το λογικό μοντέλο και προτείνεται ένα πλαίσιο που επιτρέπει την μερική συσχέτιση εγγράφων με συγκεκριμένα ερωτήματα. Αυτό γίνεται εφικτό καθώς χρησιμοποιούνται βάρη για τη συσχέτιση μεταξύ εγγράφων και όρων. Τα συγκεκριμένα βάρη παίρνουν τιμές πραγματικών αριθμών στο διάστημα από 0 έως 1. Έτσι, ένα σύστημα ανάκτησης πληροφορίας που βασίζεται στο διανυσματικό μοντέλο επιστρέφει μια λίστα από αποτελέσματα ταξινομημένη με βάση το βαθμό που αυτά σχετίζονται με το ερώτημα που έχει διατυπωθεί.

Η επεξεργασία των εγγράφων από ένα διανυσματικό μοντέλο μπορεί να χωριστεί σε δυο στάδια. Το πρώτο στάδιο είναι το στάδιο της ευρετηρίασης (indexing) των εγγράφων ενώ στο δεύτερο στάδιο υπολογίζονται τα βάρη των όρων σε κάθε έγγραφο.

Κατά την ευρετηρίαση των εγγράφων ακολουθείται μια διαδικασία καταγραφής των όρων που απαρτίζουν το κάθε έγγραφο [33]. Σε αυτή τη διαδικασία αφαιρούνται λέξεις που δεν περιγράφουν το περιεχόμενο (συνήθεις λέξεις, stopwords). Για τον εντοπισμό τέτοιων λέξεων χρησιμοποιείται είτε η συχνότητά τους είτε μια προκαθορισμένη λίστα με λέξεις που δεν προσφέρουν σημασιολογική αξία. Επιπρόσθετα, σε κάποια συστήματα οι λέξεις αντικαθίστανται από τα λήμματα από τα οποία προέρχονται (εύρεση λήμματος).

Για να υπολογίσουμε το βάρος κάθε όρου σε κάθε έγγραφο υπολογίζουμε συνήθως δυο μετρικές που αφορούν τον συγκεκριμένο όρο. Πρώτον, υπολογίζεται η συχνότητα με την οποία εμφανίζεται ο όρος στο παρόν έγγραφο. Η συγκεκριμένη συχνότητα μπορεί να κανονικοποιηθεί με βάση το μήκος του κειμένου. Δεύτερον, υπολογίζουμε την αντίστροφη συχνότητα κειμένων, που αποτελεί το αντίστροφο της συχνότητας με την οποία συναντούμε τον συγκεκριμένο όρο σε ολόκληρη την συλλογή κειμένων. Οι δυο αυτές μετρικές πολλαπλασιάζονται μεταξύ τους για να δώσουν το συγκεκριμένο βάρος του όρου. Το υπολογιζόμενο βάρος ευνοεί τους όρους που εμφανίζονται με μεγάλη συχνότητα στο τρέχον έγγραφο αλλά δεν βρίσκονται σε πολλά έγγραφα στην συλλογή που χρησιμοποιούμε. Τα βάρη των λέξεων ενός εγγράφου αποτελούν την διανυσματική περιγραφή του.

Εφόσον προηγηθούν οι συγκεκριμένοι υπολογισμοί, το σύστημα είναι έτοιμο να απαντήσει σε ερωτήματα που μπορεί να θέσει ο χρήστης. Για το σχηματισμό της απάντησης υπολογίζεται η ομοιότητα μεταξύ του ερωτήματος που διατυπώνει ο χρήστης και των εγγράφων που αποτελούν την συλλογή. Πολύ συχνά επιλέγονται συναρτήσεις συνάφειας όπως το συνημίτονο, που χρησιμοποιούν την γωνία που σχηματίζεται ανάμεσα στο ερώτημα και στα έγγραφα.

Το μοντέλο που περιγράφηκε αναπαριστά και επιλύει το πρόβλημα σε έναν διανυσματικό χώρο, για αυτό λέγεται και μοντέλο διανυσματικού χώρου (vector space model).

3.1.4 Γνωσιακές Δομές

Οντολογίες – Γνωσιακές Δομές

Η σημασιολογία ως όρος απαντάται σε διάφορα πεδία όπως στη γλωσσολογία, στην τυπική λογική, στις γλώσσες προγραμματισμού και στη σημειολογία. Στα πεδία που σχετίζονται με την πληροφορική, ο όρος αναφέρεται στην ανάπτυξη μοντέλων, μεθόδων και εργαλείων για την εκφορά νοήματος και την μετάδοση πληροφορίας μεταξύ ανθρώπων ενώ περιλαμβάνει την κατανόηση και την επεξεργασία νοήματος από μηχανές. Ένα αρκετά διαδεδομένο μοντέλο που χρησιμοποιείται για το συγκεκριμένο σκοπό είναι οι οντολογίες ή και γενικότερα οι γνωσιακές δομές.

Η οντολογία ως γνωσιακή δομή περιγράφεται συνηθέστερα ως «ένας τυπικός, ρητός ορισμός μιας κοινής σύλληψης». Η σύλληψη αφορά ένα αφηρημένο μοντέλο που περιγράφει το πώς οι άνθρωποι σκέφτονται αναφορικά με αντικείμενα στον πραγματικό κόσμο και συνήθως για μια συγκεκριμένη περιοχή του κόσμου. Ο ρητός ορισμός προϋποθέτει ότι οι έννοιες και οι σχέσεις στο αφηρημένο μοντέλο ονοματίζονται και ορίζονται με σαφήνεια [43].

Οι γνωσιακές δομές μπορούν να επιτρέψουν την ανάπτυξη ενός καινούριου τύπου αλληλεπιδράσεων. Μια διάδοχη μορφή του παγκόσμιου ιστού μπορεί να αποτελέσει ο σημασιολογικός ιστός ο οποίος θα αλλάξει τον τρόπο με τον οποίο παράγεται και μοιράζεται η γνώση [44].

Ο ιστός είχε σχεδιαστεί αρχικά ως χώρος πληροφοριών. Στόχος του ήταν να αποτελέσει ένα χρήσιμο εργαλείο για την επικοινωνία μεταξύ ανθρώπων αλλά και να επιτρέπει σε μηχανές να λαμβάνουν μέρος και να βοηθούν αυτή την

επικοινωνία. Ένα σημαντικό εμπόδιο σε αυτό τον στόχο αποτελεί το γεγονός το μεγαλύτερο μέρος των πληροφοριών που είναι διαθέσιμες στον ιστό έχουν δημιουργηθεί αποκλειστικά για κατανάλωση από ανθρώπους. Οι υπολογιστές είναι καλύτεροι στο να διαχειρίζονται δεδομένα που έχουν δομηθεί με προσοχή και σχεδιαστεί με ακρίβεια. Ακόμη και στην περίπτωση μιας δομημένης και ορισμένης βάσης δεδομένων, η σημασιολογία των δεδομένων που περιέχονται σε αυτήν δεν είναι προφανής σε έναν πράκτορα λογισμικού (software agent) που τα διαβάζει. Στο μέλλον περισσότερη πληροφορία στον ιστό μετατρέπεται σε μια μορφή που οι μηχανές όχι μόνο θα μπορούν να επεξεργαστούν αλλά και να κατανοήσουν.

Η υποστήριξη ανάγνωσης και κατανόησης εγγράφων από μηχανές δε προϋποθέτει την ύπαρξη μίας μαγικής τεχνητής νοημοσύνης που επιτρέπει στις μηχανές να αντιλαμβάνονται το κείμενο που γράφουν οι άνθρωποι. Στηρίζεται αποκλειστικά στην ικανότητα ενός υπολογιστή να επιλύσει σαφώς καθορισμένα προβλήματα με την εκτέλεση ορισμένων λειτουργιών σε ορισμένα γεγονότα. Έτσι, αντί να ζητάμε από τις μηχανές να κατανοήσουν τη γλώσσα των ανθρώπων, η τεχνολογία των γνωσιακών δομών απαιτεί από τους ανθρώπους να κάνουν κάποια επιπλέον προσπάθεια στις διατυπώσεις τους – για την οποία θα ανταμειφθούν με επιπρόσθετες λειτουργίες.

Ανοιχτά Συνδεδεμένα Δεδομένα (Linked Open Data)

Τα συνδεδεμένα δεδομένα αφορούν την χρήση του ιστού για την δημιουργία δεσμών μεταξύ δεδομένων από διαφορετικές πηγές [45].

Τα συνδεδεμένα δεδομένα αναφέρονται σε δεδομένα που δημοσιεύονται στο διαδίκτυο με τέτοιο τρόπο ώστε να είναι αναγνώσιμα από μηχανές. Τα συγκεκριμένα δεδομένα έχουν κάποια κοινά χαρακτηριστικά: το νόημά τους ορίζεται ρητά και μπορούν να συνδεθούν με εξωτερικά σύνολα δεδομένων. Ενώ οι βασικές μονάδες του ιστού είναι έγγραφα που συνδέονται με υπερσυνδέσεις (σελίδες HTML), τα συνδεδεμένα δεδομένα βασίζονται σε έγγραφα που περιέχουν στοιχεία σε μορφή RDF (resource description framework) [46]. Ωστόσο, αντί απλώς να συνδέουν αυτά τα έγγραφα, τα συνδεδεμένα δεδομένα χρησιμοποιούν το πλαίσιο RDF για να εκφράσουν τύπους προτάσεων που συνδέουν αυθαίρετα πράγματα στον πραγματικό κόσμο.

Το αποτέλεσμα της εμφάνισης και ανάπτυξης των συνδεδεμένων δεδομένων ονομάζεται ιστός δεδομένων (web of data) και μπορεί να περιγραφεί ως ένα δίκτυο

από αντικείμενα στον κόσμο που περιγράφονται από δεδομένα στον ιστό. Ο Berners-Lee [47] έχει παρουσιάσει μια σειρά από «κανόνες» για τη δημοσίευση δεδομένων στον ιστό δεδομένων:

- Κάθε αντικείμενο αντιστοιχίζεται σε ένα ενιαίο αναγνωριστικό πόρου (URI, uniform resource identifier)
- Για κάθε αντικείμενο χρησιμοποιούνται URI μορφής HTTP ώστε οι χρήστες να μπορούν να διαβάζουν τι σημαίνουν τα ονόματα των αντικειμένων.
- Κάθε URI παρέχει χρήσιμες πληροφορίες χρησιμοποιώντας τα αντίστοιχα πρότυπα (RDF, SPARQL)
- Πρέπει να περιλαμβάνονται συνδέσεις με άλλα URI ώστε οι χρήστες να μπορούν να ανακαλύψουν περισσότερα πράγματα που έχουν γίνει γνωστά ως συνδεδεμένα δεδομένα.

Σύμφωνα με κάποιες μελέτες [47] ακριβώς όπως τα παραδοσιακά προγράμματα περιήγησης στο Web επιτρέπουν στους χρήστες να πλοηγούνται μεταξύ σελίδων χρησιμοποιώντας συνδέσμους, περιηγητές συνδεδεμένων δεδομένων θα επιτρέπουν στους χρήστες να πλοηγούνται μεταξύ των πηγών δεδομένων μέσω συνδέσμων που εκφράζονται με χρήση RDF.

Παρά τα πλεονεκτήματα που υπόσχονται οι γνωσιακές δομές και οι αντίστοιχες τεχνολογίες σημασιολογικού ιστού, ο σχεδιασμός και η υλοποίηση τους ενδέχεται να αντιμετωπίσει προβλήματα. Η σημασιολογία στις τυπικές γνωσιακές δομές σημασιολογικού ιστού εισάγεται στον ίδιο κώδικα που υλοποιεί τις αντίστοιχες υπηρεσίες. Η εισαγωγή γίνεται σύμφωνα με τις προδιαγραφές που έχουν διατυπωθεί στις αντίστοιχες οντολογίες και στην ανεπίσημη τεκμηρίωση. Αντίθετα, η σημασιολογία που αφορά την ερμηνεία των φυσικών γλωσσών είναι ενσωματωμένη στην ανθρώπινη γνωσιακή και πολιτιστική συμπεριφορά, όπου η γλωσσική έκφραση προκαλεί αντίστοιχες απαντήσεις και αλλαγές στη γνώση. Λόγω του τεράστιου πλαισίου γνώσης και πολιτισμού που έχει γίνει διαθέσιμο στον ιστό, η γλωσσική έκφραση μπορεί να είναι αρκετά αμφίσημη και όμως να γίνεται τελικά κατανοητή σωστά [48], ενώ οι γνωσιακές δομές πιθανόν να αποτύχουν σε αυτό τον τομέα.

3.1.5 Λανθάνουσα Σημασιολογική Ανάλυση

Η λανθάνουσα σημασιολογική ανάλυση (latent semantic analysis, LSA) αποτελεί μια τεχνική για το σχηματισμό διανυσματικών αναπαραστάσεων του κειμένου που αντανακλούν το σημασιολογικό τους περιεχόμενο [49].

Μια συνήθης μορφή της λανθάνουσας σημασιολογικής ανάλυσης επεκτείνει το διανυσματικό μοντέλο (βλ. ενότητα 3.1.3) με τη χρήση ανάλυσης ιδιοτιμών (singular value decomposition, SVD) για τον ανακαθορισμό των δεδομένων. Η υπόθεση που στηρίζει την διαδικασία είναι ότι υπάρχει ένα σύνολο λανθανόντων μεταβλητών που καλύπτουν τα διάφορα νοήματα που μπορούν να εκφραστούν σε μια συγκεκριμένη γλώσσα. Οι μεταβλητές αυτές θεωρούνται ανεξάρτητες. Για την καταγραφή των νοημάτων σε μια συλλογή κειμένων πολύ συχνά θεωρούμε ότι μερικές εκατοντάδες μεταβλητών είναι αρκετές.

Αν και η λανθάνουσα σημασιολογική ανάλυση εμφανίζει διάφορες παραλλαγές, κάποια από τα πιο συνηθισμένα βήματα είναι τα παρακάτω [50]:

- Συλλέγεται ένας αριθμός από κείμενα τα οποία χωρίζονται σε έγγραφα με βάση τον τρόπο που έχουν γραφτεί (π.χ. παραγράφους).
- Στη συνέχεια σχηματίζεται ένας πίνακας εμφάνισης λέξεων στα έγγραφα. Κάθε κελί στον πίνακα καταγράφει τον αριθμό των εμφανίσεων του συγκεκριμένου όρου στο έγγραφο. Ο συγκεκριμένος πίνακας αποτελεί μια συλλογή διανυσματικών απεικονίσεων των εγγράφων.
- Οι τιμές της συχνότητας των λέξεων που βρίσκονται σε κάθε κελί μπορούν να κανονικοποιηθούν ώστε να περιοριστεί η σημασία των συνηθισμένων λέξεων (π.χ., με την αντίστροφη συχνότητα κειμένων).
- Για τον περιορισμό των διαστάσεων εφαρμόζεται μια μέθοδος (για παράδειγμα, *ανάλυση ιδιοτιμών*) με μια παράμετρο που ελέγχει τον αριθμό τον διαστάσεων. Η μέθοδος αυτή οδηγεί στην εξαγωγή λανθανουσών μεταβλητών. Στη συνέχεια, για τους σκοπούς της ανάκτησης πληροφορίας, οι διανυσματικές περιγραφές των εγγράφων εκφράζονται με βάση τις λανθάνουσες μεταβλητές που εξάχθηκαν.

Τα ερωτήματα μετατρέπονται στο αντίστοιχο διανυσματικό χώρο και χρησιμοποιούνται για την ανάκτηση των σχετικών εγγράφων με χρήση συναρτήσεων συνάφειας (π.χ. συνάρτηση συνημίτονου). Τα πλησιέστερα έγγραφα

επιστρέφονται σε μια ταξινομημένη λίστα, όπως και στο απλό διανυσματικό μοντέλο.

Η μέθοδος της λανθάνουσας σημασιολογικής ανάλυσης μπορεί να βοηθήσει στον περιορισμό των διαστάσεων του προβλήματος της ανάκτησης πληροφορίας. Επίσης μπορεί να υποστηρίξει την ύπαρξη συνωνύμων αλλά και μεταφορών στα πλαίσια της γλώσσας. Ένα από τα μειονεκτήματά της αφορά τον περιορισμό της σε περιεχόμενο κειμένου. Επίσης σημαντικό είναι το ζήτημα του μεγέθους της συλλογής εγγράφων και του αντίστοιχου υπολογιστικού κόστους. Τέλος, όπως όλες οι μέθοδοι που αγνοούν το συντακτικό της γλώσσας η μέθοδος αυτή δεν μπορεί να εξάγει συμπεράσματα από την σειρά των λέξεων και να αντιληφθεί τις αρνήσεις.

3.2 Πιθανοτικά Μοντέλα Θεμάτων

3.2.1 Πιθανοτική Λανθάνουσα Σημασιολογική Ανάλυση (pLSA)

Ένα σημαντικό βήμα στην κατεύθυνση των πιθανοτικών μοντέλων θεμάτων έγινε από τον Hofmann [51] που παρουσίασε την πιθανοτική λανθάνουσα σημασιολογική ανάλυση (pLSA). Η προσέγγιση αυτή μοντελοποιεί κάθε λέξη ενός κειμένου ως το αποτέλεσμα δειγματοληψίας από ένα μοντέλο ανάμειξης όπου τα αναμεμειγμένα στοιχεία είναι τυχαίες πολυωνυμικές συναρτήσεις και ονομάζονται θέματα. Έτσι κάθε λέξη δημιουργείται από ένα απλό θέμα, και διαφορετικές λέξεις σε ένα κείμενο μπορεί να δημιουργηθούν από διαφορετικά θέματα.

Η συγκεκριμένη προσέγγιση αποτέλεσε ένα σημαντικό βήμα στα μοντέλα θεμάτων. Παρ όλα αυτά δεν είναι πλήρης καθώς δεν προβλέπει την ύπαρξη ενός αντίστοιχου πιθανοτικού μοντέλου. Στο pLSA, κάθε κείμενο αναπαριστάται από μια λίστα αριθμών που αντιστοιχούν στις αναλογίες των θεμάτων και δεν υπάρχει κάποιο γενετικό πιθανοτικό μοντέλο για την παραγωγή των συγκεκριμένων αριθμών. Αυτό επιφέρει αρκετά προβλήματα: (1) ο αριθμός των παραμέτρων στο μοντέλο αυξάνεται γραμμικά με το μέγεθος της συλλογής δεδομένων, κάτι που οδηγεί σε σοβαρά προβλήματα υπερπροσαρμογής και (2) δεν είναι ξεκάθαρο πως εξάγεται η πιθανοτική κατανομή για ένα κείμενο εκτός του εκπαιδευτικού συνόλου δεδομένων.

Ένα γενετικό μοντέλο (generative model) [52] έχει ως στόχο την εκμάθηση της από κοινού κατανομής πιθανοτήτων $p(x, y)$ με βάση τις εισόδους x και τις

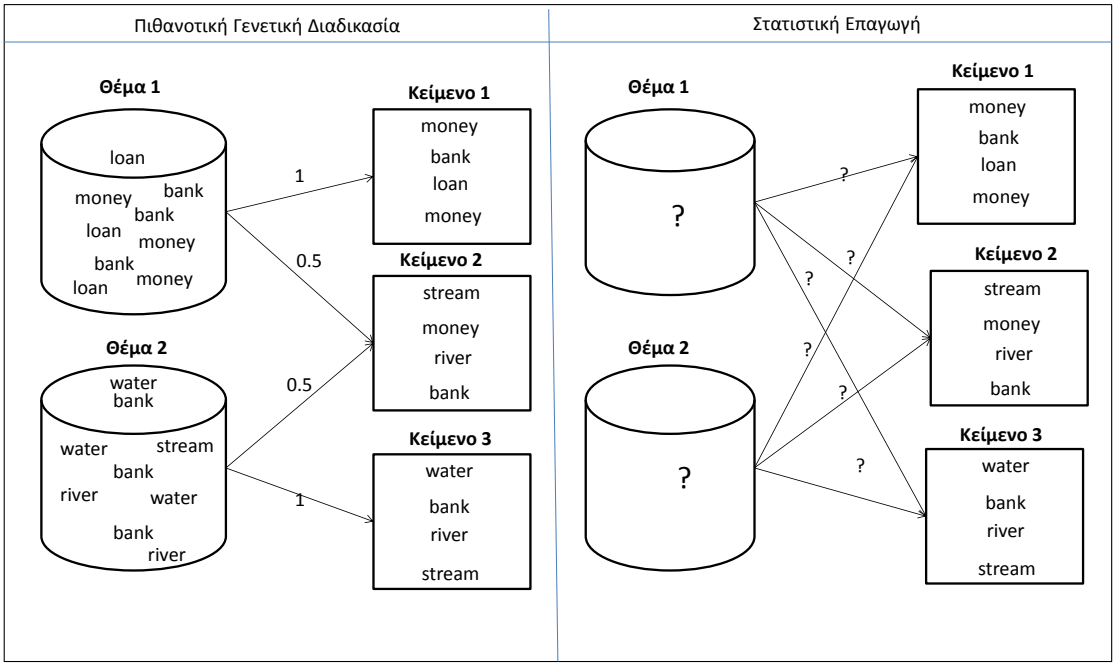
γνωστές εξόδους y του φαινομένου που αναλύεται. Στη συνέχεια χρησιμοποιείται Μπεϋζιανή στατιστική για τον υπολογισμό της κατανομής πιθανοτήτων του y με δεδομένο το x, $p(y|x)$, και εκτιμάται το y στην συγκεκριμένη περίπτωση. Αντίθετα, τα διαχωριστικά μοντέλα (discriminative models) μοντελοποιούν απευθείας την εκ των υστέρων κατανομή $p(y|x)$.

Στη γενική του μορφή ένα γενετικό μοντέλο εγγράφων αποτελείται από πιθανοτικούς κανόνες που περιγράφουν τον τρόπο με τον οποίο έγγραφα που περιέχουν κείμενο μπορεί να δημιουργηθούν μέσω τυχαίων λανθανουσών μεταβλητών. Στόχος της δημιουργίας ενός γενετικού μοντέλου είναι η εύρεση του βέλτιστου συνόλου λανθανουσών μεταβλητών που μπορεί να εξηγήσει τα παρατηρούμενα δεδομένα με βάση την υπόθεση ότι τα παρατηρούμενα έγγραφα δημιουργήθηκαν από το μοντέλο.

Η Εικόνα 3.1, που προέρχεται από αντίστοιχη μελέτη [53], είναι μια παρουσίαση των πιθανοτικών μοντέλων θεμάτων με δύο τρόπους: ως γενετικό μοντέλο και ως πρόβλημα στατιστικής επαγωγής. Στα αριστερά, η γενετική διαδικασία παρουσιάζεται με δύο θέματα. Τα Θέματα 1 και 2 σχετίζονται με την οικονομία και με τα ποτάμια αντίστοιχα. Διαφορετικά κείμενα μπορούν να παραχθούν καθώς το μοντέλο επιλέγει λέξεις από ένα θέμα ανάλογα του βάρους που έχει. Για παράδειγμα, τα κείμενα 1 και 3 σχηματίζονται με δειγματοληψία μόνο από το θέμα 1 και το θέμα 2, ενώ το κείμενο 2 δημιουργήθηκε από μείξη και των δύο θεμάτων.

Ο τρόπος με τον οποίο ορίστηκε το μοντέλο επιτρέπει στις λέξεις να προέρχονται από περισσότερο του ενός θέματα. Το χαρακτηριστικό αυτό επιτρέπει στο μοντέλο να υποστηρίζει την πολυσημία, όπου ο ίδιος όρος μπορεί να έχει πολλές σημασίες ταυτόχρονα. Έτσι, για παράδειγμα η λέξη bank υπάρχει και στα δύο θέματα που βλέπουμε στην Εικόνα 3.1 καθώς σημαίνει τόσο τράπεζα όσο και όχθη.

Το δεξί κομμάτι που περιέχεται στην Εικόνα 3.1 απεικονίζει το πρόβλημα στατιστικής επαγωγής. Με δεδομένες τις παρατηρούμενες λέξεις σε ένα σύνολο δεδομένων επιχειρούμε να υπολογίσουμε το μοντέλο θεμάτων που είναι πιο πιθανό να έχει δημιουργήσει τα δεδομένα. Αυτό συμπεριλαμβάνει τον υπολογισμό της πιθανοτικής κατανομής των λέξεων που συσχετίζονται με κάθε θέμα, των θεμάτων για κάθε κείμενο και την κατανομή του θέματος που είναι υπεύθυνο για την δημιουργία κάθε λέξης.



Εικόνα 3.1 Πιθανοτικά Μοντέλα Θεμάτων

3.2.2 Λανθάνουσα Κατανομή Dirichlet (Latent Dirichlet Allocation)

Εδώ περιγράφουμε τον αλγόριθμο λανθάνουσας κατανομής Dirichlet (Latent Dirichlet Allocation, LDA) ως ένα γενετικό πιθανοτικό μοντέλο. Η LDA βασίζεται σε μια ιεραρχική πιθανοτική δομή τριών επιπέδων όπου κάθε αντικείμενο της συλλογής μοντελοποιείται ως ένα μείγμα από ένα σύνολο πιθανοτήτων. Η μέθοδος αυτή έχει προκύψει ως επέκταση της πιθανοτικής λανθάνουσας σημασιολογικής ανάλυσης (pLSA).

Στην λανθάνουσα κατανομή Dirichlet η λέξη αποτελεί τη βασική μονάδα των διακριτών δεδομένων, η οποία ορίζεται ως ένα αντικείμενο που προέρχεται από ένα λεξιλόγιο όρων *{1,....,V}*. Κάθε όρος απεικονίζεται ως διάνυσμα όπου ένα και μόνο στοιχείο είναι ίσο με 1 και όλα τα υπόλοιπα είναι ίσα με 0 (αντίστοιχα και με το απλό διανυσματικό μοντέλο, βλ. ενότητα 3.1.3). Έτσι ο ν-οστός όρος απεικονίζεται στο λεξιλόγιο ως ένα διάνυσμα ***V*** διαστάσεων ***w*** όπου $w^v=1$ και $w^u=0$ για $u \neq v$.Με βάση την λέξη ορίζονται τα κείμενα και οι συλλογές κειμένων. Ένα κείμενο είναι μια σειρά από ***N*** λέξεις ορισμένο ως $\boldsymbol{w}=(w_1,w_2,....,w_N)$ όπου το $\boldsymbol{w_n}$ είναι η n-οστή λέξη του κειμένου. Μια συλλογή κειμένων είναι ένα σύνολο από ***M*** κείμενα ορισμένα ως D={$\mathbf{w_1,w_2,....w_M}$}.

Σκοπός της μεθόδου είναι η δημιουργία ενός πιθανοτικού μοντέλου που έχει δύο χαρακτηριστικά. Πρώτον το μοντέλο δίνει υψηλή πιθανότητα εμφάνισης στα κείμενα που υπάρχουν στο σύνολο δεδομένων. Δεύτερον το μοντέλο ευνοεί την εμφάνιση νέων κείμενων, παρόμοιων από άποψη σύνθεσης. Η κεντρική ιδέα της μεθόδου είναι ότι τα κείμενα αναπαρίστανται από τυχαίες αναμείξεις λανθανόντων θεμάτων, όπου κάθε θέμα χαρακτηρίζεται από μία κατανομή λέξεων.

Η λανθάνουσα κατανομή Dirichlet υιοθετεί την παρακάτω γενετική διαδικασία για την δημιουργία κάθε κειμένου **w** σε ένα σώμα **D**, όπου ~ ισοδυναμεί με δειγματοληψία:

- Επιλογή N ~ Poisson(ξ)
- Επιλογή θ ~ Dir(α)
- Για κάθε μια από τις N λέξεις w_n:
    - Επιλογή θέματος z_n ~ Multinomial(θ)
    - Επιλογή μιας λέξης w_n από $p(w_n|z_n,\beta)$, μια πολυωνυμική κατανομή πιθανοτήτων όρων με δεδομένο το θέμα z_n

Το μοντέλο της μεθόδου στηρίζεται σε δυο βασικές υποθέσεις. Πρώτον, η διάσταση ***k*** της Dirichlet κατανομής και της μεταβλητής θέματος ***z*** που αντιστοιχεί στον αριθμό των λανθανόντων θεμάτων θεωρείται γνωστή και σταθερή. Δεύτερον, οι πιθανότητες της εμφάνισης των λέξεων με δεδομένο ένα θέμα θεωρούμε ότι προέρχονται από ένα πίνακα ***β*** διαστάσεων $K \times V$ όπου $\beta_{ij} = p(w^j = 1 \mid z^j = 1)$. Τέλος η κατανομή Poisson επηρεάζει μόνο το μήκος των εγγράφων και στη θέση της μπορούν να χρησιμοποιηθούν άλλες κατανομές.

Μία τυχαία μεταβλητή ***ϑ*** με κατανομή Dirichlet η οποία έχει ***k*** διαστάσεις μπορεί να πάρει τιμές σε ένα simplex ***k-1*** διαστάσεων. Ένα διάνυσμα ***k*** διαστάσεων ***ϑ*** βρίσκεται στο ίδιο simplex αν ***ϑi*** ≥ *0,* $\sum_{i=1}^{k} \theta_i = 1$. Η συνάρτηση πυκνότητας πιθανότητας της μεταβλητής ***ϑ*** στο συγκεκριμένο simplex εμφανίζεται στην εξίσωση (3.1).

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k} a_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_{1.}^{\alpha_1 - 1} \dots . \theta_k^{\alpha_k - 1} \qquad (3.1)$$

Στην εξίσωση (3.1) η παράμετρος **α** είναι ένα διάνυσμα **k** διαστάσεων με στοιχεία $\alpha_i > 1$ και η *Γ(x)* είναι η συνάρτηση Γάμμα.

Η Dirichlet είναι μια κατανομή στο simplex που ανήκει στην εκθετική οικογένεια και είναι συζυγής με την πολυωνυμική κατανομή. Με δεδομένες τις υπερπαραμέτρους ***α*** και ***β***, η από κοινού κατανομή μιας κατανομής θεμάτων ***ϑ***, ενός συνόλου από ***N*** θέματα ***z***, και ενός συνόλου από ***N*** λέξεις ***w*** δίνεται από την εξίσωση (3.2). Η εξίσωση αυτή αφορά ένα κείμενο της συλλογής και το $p(z_n|\vartheta)$ είναι το ϑ_i για ένα μοναδικό ***i*** έτσι ώστε $z^i_n=1$.

$$p(\theta, \boldsymbol{z}, \boldsymbol{w}|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta), \qquad (3.2)$$

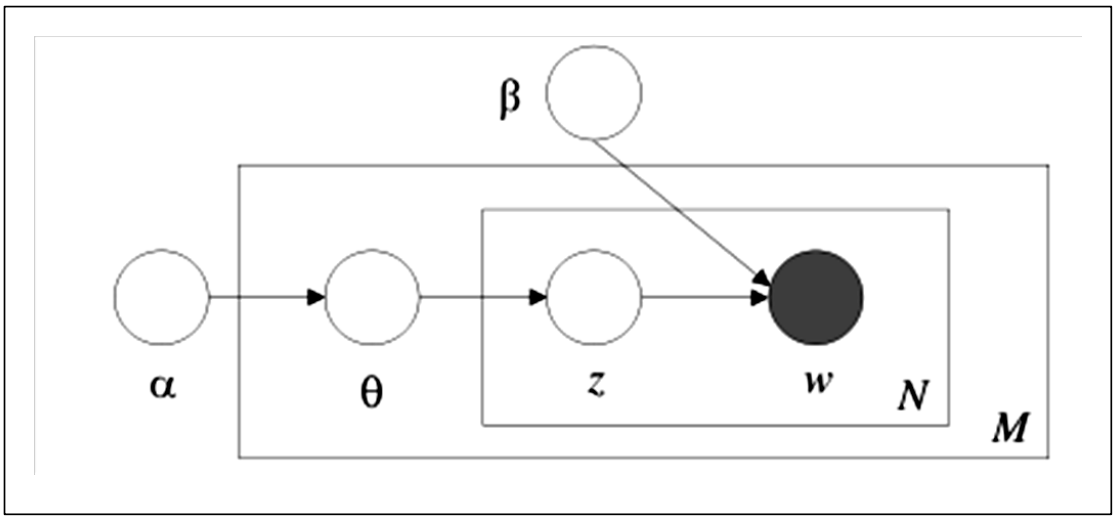
Ολοκληρώνοντας ως προς ***ϑ*** και αθροίζοντας ως προς **z**, παίρνουμε την οριακή κατανομή πιθανοτήτων για ένα κείμενο (βλ. εξίσωση (3.3))

$$p(\boldsymbol{w}|\alpha, \beta) = \int p(\theta|\alpha) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n \mid \theta) p(w_n \mid z_n, \beta)\right) d\theta \qquad (3.3)$$

Τέλος, πολλαπλασιάζοντας τις οριακές κατανομές πιθανοτήτων όλων των κειμένων μπορούμε να έχουμε την συνολική πιθανότητα μιας συλλογής κειμένων (εξίσωση (3.4)).

$$p(\boldsymbol{D}|\alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha) \left(\prod_{n=1}^{N_d} \sum_{Z_{dn}} p(z_{dn} \mid \theta) p(w_{dn} \mid z_{dn}, \beta)\right) d\theta_d$$

(3.4)



Εικόνα 3.2 Γραφικό Μοντέλο Απεικόνισης του LDA.

Το μοντέλο της LDA απεικονίζεται ως πιθανοτικό γραφικό μοντέλο με τρία επίπεδα στην Εικόνα 3.2. Τα πλαίσια απεικονίζουν τις επαναλήψεις: το εξωτερικό πλαίσιο απεικονίζει τα κείμενα της συλλογής ενώ το εσωτερικό πλαίσιο απεικονίζει την επαναληπτική επιλογή των θεμάτων και των λέξεων μέσα σε ένα κείμενο. Οι παράμετροι $\boldsymbol{\alpha}$ και $\boldsymbol{\beta}$ είναι οι υπερπαράμετροι επιπέδου συνόλου δεδομένων που αφορούν ολόκληρη τη συλλογή κειμένων κι έχουν οριστεί στην διαδικασία ανάλυσης του σώματος. Οι μεταβλητές $\boldsymbol{\vartheta_d}$ είναι μεταβλητές επιπέδου κειμένου που αφορούν ένα κείμενο. Τέλος, οι μεταβλητές $\mathbf{z_{dn}}$ και $\mathbf{w_{dn}}$ είναι μεταβλητές επιπέδου λέξεων και αφορούν κάθε λέξη κάθε κειμένου.

Αντίστοιχες προσεγγίσεις έχουν μελετηθεί στην μοντελοποίηση Μπεϋζιανής στατιστικής όπου αναφέρονται ως ιεραρχικά μοντέλα ή ως υπό συνθήκες ανεξάρτητα ιεραρχικά μοντέλα. Επίσης περιγράφονται ως παραμετρικά εμπειρικά Μπεϋζιανά μοντέλα.

3.2.3 Αλγόριθμος Εξαγωγής Θεμάτων

Οι κύριες μεταβλητές που μας ενδιαφέρουν σε ένα μοντέλο είναι οι μεταβλητές $\boldsymbol{\varphi}$ που συνδέουν θέματα με λέξεις και οι κατανομές θεμάτων $\boldsymbol{\vartheta}$ για κάθε κείμενο. Σε μελέτες [51] χρησιμοποιείται ο αλγόριθμος μεγιστοποίησης αναμονής (expectation-maximization) για να υπολογιστούν τα $\boldsymbol{\varphi}$ και $\boldsymbol{\vartheta}$. Αυτή η προσέγγιση παρουσιάζει προβλήματα σχετικά με τα τοπικά μέγιστα στην συνάρτηση πιθανοφάνειας, κάτι που οδήγησε στην εύρεση καλύτερων αλγόριθμων όπως στο [54].

Αντί για τον απευθείας υπολογισμό των κατανομών λέξεων για κάθε θέμα $\boldsymbol{\varphi}$ και των θεματικών κατανομών $\boldsymbol{\vartheta}$ για κάθε κείμενο, μια άλλη προσέγγιση είναι να υπολογίσουμε κατευθείαν την εκ των υστέρων πιθανοτική κατανομή ως προς $\boldsymbol{z}$, δεδομένου των παρατηρούμενων λέξεων **w,** καθώς αφαιρούμε από την διατύπωση τα $\boldsymbol{\varphi}$ και $\boldsymbol{\vartheta}$.

Κάθε $\boldsymbol{z_i}$ έχει μια ακέραια τιμή στο διάστημα *[1....T]* που αντιστοιχεί στο θέμα στο οποίο αντιστοιχίζεται η λέξη $\boldsymbol{i}$. Επειδή πολλές συλλογές κειμένων περιέχουν εκατομμύρια τέτοιων δειγμάτων, ο υπολογισμός της εκ των υστέρων κατανομής ως προς το $\boldsymbol{z}$ απαιτεί αποτελεσματικές διαδικασίες υπολογισμού.

Περιγράφουμε τον αλγόριθμο που χρησιμοποιεί δειγματοληψία Gibbs, μια μορφή αλυσίδων Markov Monte Carlo (MCMC), που μπορούν να εφαρμοστούν και

να εκτιμήσουν τα στοιχεία του συνόλου των θεμάτων για ένα μεγάλο σύνολο δεδομένων.

Ο όρος αλυσίδα Markov Monte Carlo αναφέρεται σε ένα σύνολο επαναληπτικών τεχνικών που έχουν σχεδιαστεί να εκτιμούν τιμές για πολύπλοκες κατανομές. Η δειγματοληψία Gibbs, μια συγκεκριμένη μορφή των MCMC, χρησιμοποιείται για να εκτιμήσει τιμές που ανήκουν σε μια πολυδιάστατη κατανομή. Η μεθοδολογία αυτή συλλέγει υποσύνολα μικρότερης διάστασης από τις μεταβλητές της ζητούμενης κατανομής, όπου κάθε υποσύνολο είναι δεσμευμένο ως προς τις τιμές των άλλων. Η διαδικασία γίνεται διαδοχικά και δεν σταματάει αν δεν υπολογίσει τις τιμές που να ανήκουν στην κατανομή που ψάχνουμε. Η δειγματοληψία Gibbs που θα περιγράψουμε, δεν παρέχει τους απευθείας υπολογισμούς των φ και ϑ αλλά τα φ και ϑ μπορούν να υπολογιστούν χρησιμοποιώντας τους εκ των υστέρων υπολογισμούς του z.

Δειγματοληψία Gibbs

Αναπαριστούμε την συλλογή κειμένων με ένα σύνολο όρων $\mathbf{w_i}$ και δεικτών κειμένων $\mathbf{d_i}$, για κάθε λέξη i. Η δειγματοληψία Gibbs θεωρεί γνωστή κάθε λέξη που βρίσκεται στην συλλογή κειμένων και υπολογίζει την πιθανότητα να πάρουμε την παρούσα λέξη από κάθε θέμα, δεδομένων των υπόλοιπων συνδέσεων λέξεων και θεμάτων. Γράφουμε αυτή την δεσμευμένη κατανομή ως $P(z_i=j|z_{-i}, w_i, d_i,.)$, όπου το $z_i=j$ αναφέρεται στην σύνδεση της λέξης i με το θέμα j, *το* z_{-i} αναφέρεται στα θέματα από τα οποία παράχθηκαν οι άλλες λέξεις, και το **'.'** αναφέρεται σε όλες τις άλλες γνωστές πληροφορίες όπως όλες οι άλλες λέξεις και κείμενα w_{-i} και d_{-i} και οι υπερπαράμετροι α και β.

Οι Griffiths και Steyvers [55] έδειξαν πως αυτή η κατανομή μπορεί να υπολογιστεί όπως στη σχέση (3.5) όπου $\mathbf{C^{WT}}$ και $\mathbf{C^{DT}}$ είναι πίνακες των μετρήσεων με διάσταση **W x T** και **D x T** αντίστοιχα.

$$P(z_i = j \mid z_{-i}, w_i, d_i, \cdot) \propto \frac{C^{WT}_{w_i j} + \beta}{\sum_{w=1}^{W} C^{WT}_{wj} + W\beta} \frac{C^{DT}_{d_i j} + \alpha}{\sum_{t=1}^{T} C^{DT}_{d_i j} + T\alpha} \qquad (3.5)$$

Το στοιχείο $\mathbf{C^{WT}_{wj}}$ είναι ο αριθμός των φορών που η λέξη w έχει συνδεθεί με το θέμα j, μη συμπεριλαμβανομένης της παρούσας τιμής i, και το $\mathbf{C^{DT}_{dj}}$ είναι ο αριθμός των συνδέσεων του θέματος j με κάποια λέξη στο έγγραφο **d**, μη συμπεριλαμβάνοντας την παρούσα τιμή i. Η εξίσωση (3.5) δίνει την μη

κανονικοποιημένη κατανομή, ενώ η πραγματική κατανομή σύνδεσης μιας λέξης με ένα θέμα ***j*** υπολογίζεται διαιρώντας την ποσότητα της εξίσωσης (3.5) για το θέμα ***t*** με το άθροισμα όλων των θεμάτων.

Οι παράγοντες που επηρεάζουν την επιλογή του θέματος που θα συνδεθεί με μια συγκεκριμένη λέξη γίνονται εμφανείς εξετάζοντας τα δύο μέρη της εξίσωσης (3.5). Το αριστερό μέρος της εξίσωσης είναι η πιθανότητα σύνδεσης την λέξης ***w*** με το θέμα ***j*** και το δεξί κομμάτι είναι η πιθανότητα εμφάνισης του θέματος ***j*** στη κατανομή θεμάτων του κειμένου ***d***. Καθόσον συνδέονται πολλές λέξεις με ένα θέμα ***j***, θα αυξηθεί η πιθανότητα να συνδεθεί ένας συγκεκριμένος όρος με το θέμα ***j***. Την ίδια στιγμή, αν το θέμα ***j*** έχει χρησιμοποιηθεί πολλές φορές σε ένα κείμενο, θα αυξηθεί η πιθανότητα κάθε λέξη του κειμένου να συνδεθεί με το θέμα ***j***. Συνεπώς, η σύνδεση των λέξεων με τα θέματα εξαρτάται από πιθανότητα μια λέξη να ανήκει στο θέμα, καθώς επίσης και από το πόσο κυρίαρχο είναι ένα θέμα σε ένα κείμενο.

Η δειγματοληψία Gibbs αρχίζει συνδέοντας τις λέξεις σε ένα τυχαίο θέμα [1....T]. Για κάθε λέξη που συνδέεται, τα περιεχόμενα των πινάκων C^{WT} και C^{DT} που αφορούν την παρούσα σύνδεση θέματος μειώνονται κατά 1. Μετά, ένα νέο θέμα λαμβάνεται δειγματοληπτικά από την κατανομή της εξίσωσης (3.5) και οι πίνακες C^{WT} και C^{DT} ενημερώνονται με την νέα σύνδεση λέξης και θέματος. Κάθε εφαρμογή της δειγματοληψίας Gibbs περιλαμβάνει μια σειρά συνδέσεων θεμάτων με όλες τις N λέξεις του σώματος. Καθ' όλη την αρχική φάση της δειγματοληψίας (burn-in phase), οι εκτιμήσεις των συνδέσεων απορρίπτονται καθώς δεν βρίσκονται κοντά στον υπολογισμό της εκ των υστέρων κατανομής. Μετά την περίοδο αυτή οι συνδέσεις θεμάτων και λέξεων συνυπολογίζονται στην κατανομή. Για να πάρουμε ένα αντιπροσωπευτικό σύνολο τιμών από την κατανομή, ένας αριθμός από συνδέσεις σώζονται ανά τακτά χρονικά διαστήματα, για να αποφευχθούν οι συσχετίσεις μεταξύ των δειγματοληψιών.

Υπολογίζοντας τα φ και θ.

Ο αλγόριθμος δειγματοληψίας υπολογίζει απευθείας το ***z*** για κάθε λέξη. Όμως, πολλές εφαρμογές του μοντέλου απαιτούν τους υπολογισμούς του ϑ και του φ των κατανομών λέξεων θεμάτων και των κατανομών θεμάτων κειμένων αντίστοιχα. Αυτές μπορούν να ληφθούν από τους τύπους στην εξίσωση (3.6).

$$\phi'^{(j)}_i = \frac{C^{WT}_{ij} + \beta}{\sum_{k=1}^{W} C^{WT}_{kj} + W\beta} \theta'^{(d)}_j \frac{C^{DT}_{d\,j} + \alpha}{\sum_{k=1}^{T} C^{DT}_{d_i k} + T\alpha} \qquad (3.6)$$

Οι συγκεκριμένες τιμές αντιστοιχούν στις προβλεπόμενες κατανομές της δειγματοληψίας μια λέξης ***i*** από το θέμα ***j***, και της δειγματοληψίας ενός νέου δείγματος σε ένα κείμενο ***d*** από το θέμα ***j***. *Αποτελούν* επίσης τους εκ των υστέρων μέσους όρους αυτών τον ποσοτήτων σε ένα συγκεκριμένο δείγμα.

Ανταλλαξιμότητα των Θεμάτων

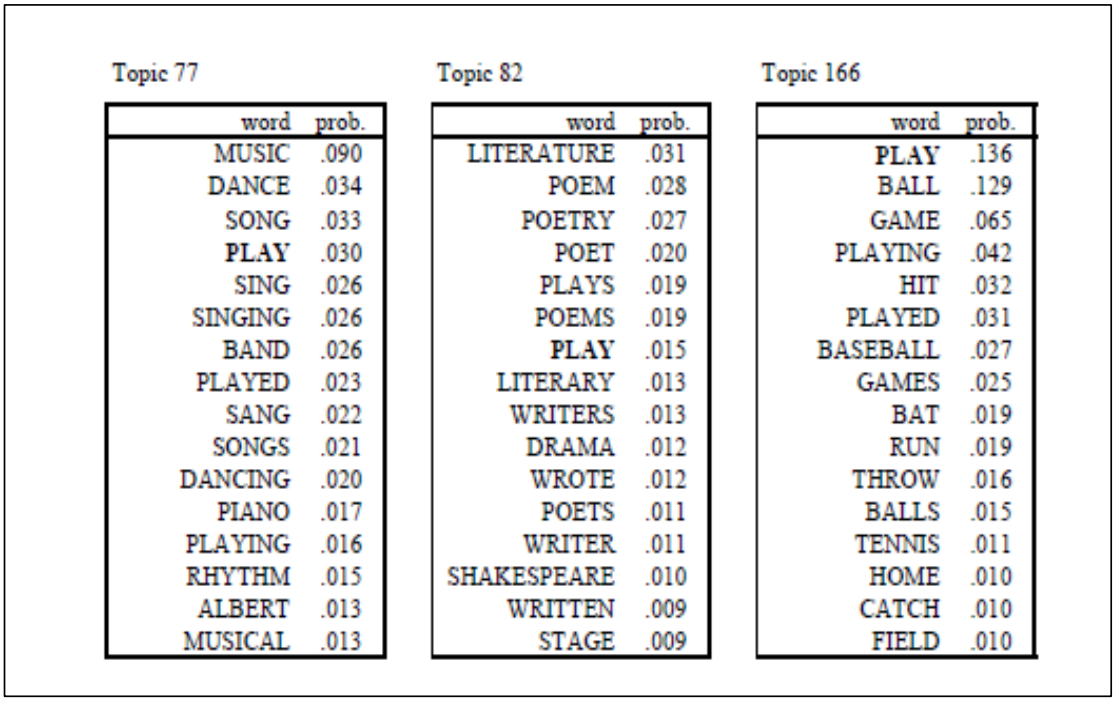
Στην λανθάνουσα κατανομή Dirichlet δεν υπάρχει κάποια εκ των προτέρων ορισμένη σειρά των θεμάτων. Το θέμα j σε ένα δείγμα δεν συνδέεται με κανέναν τρόπο με το αντίστοιχο θέμα j σε ένα άλλο δείγμα ασχέτως αν τα δείγματα έρχονται από την ίδια η διαφορετική αλυσίδα Markov. Συνεπώς δεν μπορούμε να χρησιμοποιήσουμε διαφορετικά δείγματα για να εξάγουμε τον μέσο όρο κάποιας τιμής που αφορά ένα συγκεκριμένο θέμα. Μόνον όταν η σειρά των θεμάτων δεν επηρεάζει τη ζητούμενη μετρική μπορούμε να πραγματοποιήσουμε πολλαπλές δειγματοληψίες.

Πολυσημία των Θεμάτων

Σε ένα νέο κείμενο έχοντας παρατηρήσει μόνο την λέξη *"play",* θα υπάρχει αβεβαιότητα ανάμεσα στα θέματα από τα οποία προέρχεται η συγκεκριμένη λέξη. Η αβεβαιότητα μπορεί να μειωθεί παρατηρώντας άλλες λέξεις του κειμένου. Η συγκεκριμένη διαδικασία μπορεί να περιγραφεί στα πλαίσια της επαναληπτικής δειγματοληψίας όπου η σύνδεση κάθε λέξης με ένα θέμα εξαρτάται από τα θέματα των άλλων λέξεων του κειμένου.

Στην Εικόνα 3.4 [53], βλέπουμε τρία κείμενα από τη συλλογή κειμένων TASA[13] που χρησιμοποιούν την λέξη *"play"*σε τρεις διαφορετικές εκδοχές. Τα θέματα τα οποία μπορεί να τις έχουν παράγει φαίνονται στην Εικόνα 3.3. Οι εκθέτες των λέξεων στην Εικόνα 3.4 δείχνουν τις συνδέσεις των λέξεων με τα θέματα. Οι γκρίζες λέξεις είναι συνήθεις λέξεις ή λέξεις χαμηλής συχνότητας και δεν έχουν χρησιμοποιηθεί στην ανάλυση. Η διαδικασία δειγματοληψίας των λιγότερο αβέβαιων λέξεων ισχυροποιεί ένα συγκεκριμένο θέμα στο κείμενο. Όταν μια λέξη έχει αβεβαιότητα, η κατανομή θέματος που αναπτύσσεται για το κείμενο είναι ο βασικός παράγοντας για τον προσδιορισμό του θέματος της λέξης.

[13] Στη συγκεκριμένη μελέτη το σύνολο δεδομένων TASA έχει παραχωρηθεί από την Touchstone Applied Science Associates και τον Tom Landauer.



Topic 77

word	prob.
MUSIC	.090
DANCE	.034
SONG	.033
PLAY	.030
SING	.026
SINGING	.026
BAND	.026
PLAYED	.023
SANG	.022
SONGS	.021
DANCING	.020
PIANO	.017
PLAYING	.016
RHYTHM	.015
ALBERT	.013
MUSICAL	.013

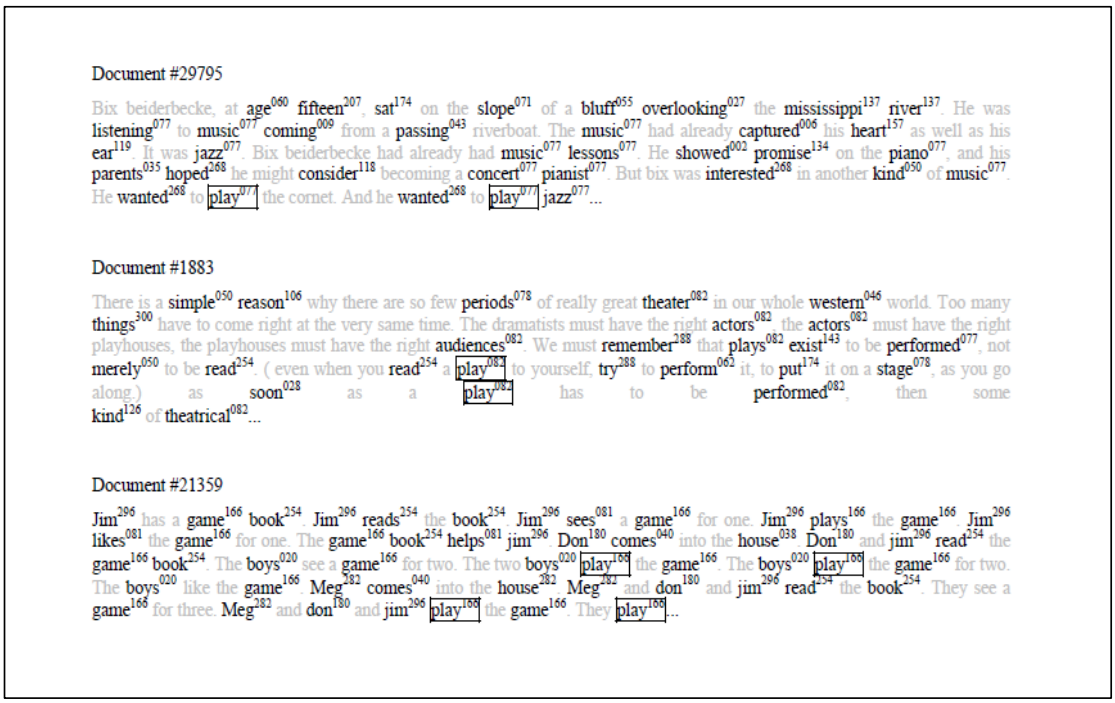
Topic 82

word	prob.
LITERATURE	.031
POEM	.028
POETRY	.027
POET	.020
PLAYS	.019
POEMS	.019
PLAY	.015
LITERARY	.013
WRITERS	.013
DRAMA	.012
WROTE	.012
POETS	.011
WRITER	.011
SHAKESPEARE	.010
WRITTEN	.009
STAGE	.009

Topic 166

word	prob.
PLAY	.136
BALL	.129
GAME	.065
PLAYING	.042
HIT	.032
PLAYED	.031
BASEBALL	.027
GAMES	.025
BAT	.019
RUN	.019
THROW	.016
BALLS	.015
TENNIS	.011
HOME	.010
CATCH	.010
FIELD	.010

Εικόνα 3.3 Πολυσημία σε Θέματα



Document #29795

Bix beiderbecke, at age[060] fifteen[207], sat[174] on the slope[071] of a bluff[055] overlooking[027] the mississippi[137] river[137]. He was listening[077] to music[077] coming[009] from a passing[043] riverboat. The music[077] had already captured[006] his heart[157] as well as his ear[119]. It was jazz[077]. Bix beiderbecke had already had music[077] lessons[077]. He showed[002] promise[134] on the piano[077], and his parents[035] hoped[268] he might consider[118] becoming a concert[077] pianist[077]. But bix was interested[268] in another kind[050] of music[077]. He wanted[268] to play[077] the cornet. And he wanted[268] to play[077] jazz[077]...

Document #1883

There is a simple[050] reason[106] why there are so few periods[078] of really great theater[082] in our whole western[046] world. Too many things[300] have to come right at the very same time. The dramatists must have the right actors[082], the actors[082] must have the right playhouses, the playhouses must have the right audiences[082]. We must remember[288] that plays[082] exist[143] to be performed[077], not merely[050] to be read[254]. (even when you read[254] a play[082] to yourself, try[288] to perform[062] it, to put[174] it on a stage[078], as you go along.) as soon[028] as a play[082] has to be performed[082], then some kind[126] of theatrical[082]...

Document #21359

Jim[296] has a game[166] book[254]. Jim[296] reads[254] the book[254]. Jim[296] sees[081] a game[166] for one. Jim[296] plays[166] the game[166]. Jim[296] likes[081] the game[166] for one. The game[166] book[254] helps[081] jim[296] Don[180] comes[040] into the house[038] Don[180] and jim[296] read[254] the game[166] book[254]. The boys[020] see a game[166] for two. The two boys[020] play[166] the game[166]. The boys[020] play[166] the game[166] for two. The boys[020] like the game[166]. Meg[282] comes[040] into the house[282]. Meg[282] and don[180] and jim[296] read[254] the book[254]. They see a game[166] for three. Meg[282] and don[180] and jim[296] play[166] the game[166]. They play[166]...

Εικόνα 3.4 Έγγραφα με την Λέξη Play

Τα πιθανοτικά μοντέλα θεμάτων δεν σχετίζονται αποκλειστικά με την ανάλυση συλλογών κειμένων αλλά έχουν και άλλες εφαρμογές όπως η συνεργατική διήθηση, η ανάκτηση εικόνων βασισμένη σε κείμενο και η βιοπληροφορική. Οι

απαιτήσεις από το μοντέλο θεμάτων διαφοροποιούνται ανάλογα με τις εφαρμογές. Η επόμενη ενότητα περιλαμβάνει συνοπτικές περιγραφές των σημαντικότερων παραλλαγών των πιθανοτικών μοντέλων θεμάτων.

3.3 Παραλλαγές των Μοντέλων Θεμάτων

Σε αυτή την ενότητα παρουσιάζονται συνοπτικά ορισμένες παραλλαγές των μοντέλων θεμάτων που αφορούν την σύνδεση των θεμάτων με ετικέτες, την εισαγωγή συσχετίσεων μεταξύ θεμάτων, τον εντοπισμό θεμάτων που αλλάζουν με τον χρόνο και την εισαγωγή ιεραρχιών.

3.3.1 Μοντέλο Θεμάτων με Ετικέτες (L-LDA)

Η περιγραφή του μοντέλου βασίζεται στην εργασία των Ramage, Hall, Nallapati και Manning [56].

Εισαγωγή

Τα διάφορα κείμενα που βρίσκονται στον ιστό, όπως αυτά δημοσιεύονται σε ειδησεογραφικούς ιστοτόπους και σε ιστοτόπους κοινωνικών δικτύων, πολύ συχνά χαρακτηρίζονται με επισημειώσεις από ανθρώπους. Οι συλλογές αυτές δεδομένων δείχνουν πολύ συχνά ότι ένα κείμενο σχετίζεται με περισσότερα από ένα θέματα – για παράδειγμα, ένα άρθρο για την νομοθεσία των μεταφορών αφορά τόσο τις μεταφορές όσο και την πολιτική, χωρίς να υπάρχει αναγκαστικά κάποια ιεραρχική δομή μεταξύ των όρων. Ακόμη, μια σελίδα που αποθηκεύεται σε σελίδες κοινωνικών συνδέσμων μπορεί να συνδεθεί με διαφορετικούς όρους όπως τέχνη, φυσική, ομορφιά, κ.α.

Ωστόσο, δεν συνδέονται όλες οι επισημειώσεις με ολόκληρο το κείμενο – και αυτό το γεγονός δημιουργεί νέες προκλήσεις για την ανάκτηση πληροφορίας και την ανάλυση επισημειωμένων δεδομένων. Για παράδειγμα, οι χρήστες που χρησιμοποιούν έναν συγκεκριμένο όρο για να βρουν κάποια έγγραφα, μπορεί να προτιμούν να δουν περιλήψεις που εστιάζουν στο κομμάτι του κειμένου που σχετίζεται με τον όρο που χρησιμοποιούν στην αναζήτηση. Ακόμη, όταν ένας χρήστης διαβάζει ένα συγκεκριμένο κείμενο, μια αντίστοιχη διεπαφή χρήστη μπορεί να παρέχει κάποια οπτικοποίηση της συσχέτισης τμημάτων του κειμένου με

συγκεκριμένες λέξεις - κλειδιά βοηθώντας τον χρήστη να εντοπίσει γρήγορα την πληροφορία που ψάχνει.

Μια απλή προσέγγιση στις συγκεκριμένες προκλήσεις μπορεί να βρεθεί σε μοντέλα που συνδέουν διαφορετικές λέξεις του κειμένου με αντίστοιχες ετικέτες. Για παράδειγμα, στο άρθρο που αφορά το νομοσχέδιο για τις μεταφορές ο όρος «*οδός*» αφορά τις μεταφορές ενώ ο όρος «*υπουργός*» την πολιτική. Χρησιμοποιώντας αυτή τη γνώση μπορούν να εντοπιστούν τα πιο σχετικά τμήματα του κειμένου με κάθε λέξη.

Εδώ αναζητούμε μια προσέγγιση για να εξάγουμε την εκ των υστέρων κατανομή κάθε λέξης ενός κειμένου με βάση τις επισημειώσεις του. Η λανθάνουσα κατανομή Dirichlet (LDA) αποτελεί μια πολλά υποσχόμενη προσέγγιση για την συσχέτιση τμημάτων του κειμένου με κάποιες επισημειώσεις. Το στοιχείο της μεθόδου που σχετίζεται με τις επισημειώσεις είναι ότι κάθε λέξη προέρχεται από ένα λανθάνον θέμα.

Αν και η λανθάνουσα κατανομή Dirichlet μπορεί να προβλέψει την μοντελοποίηση πολλαπλών θεμάτων ανά έγγραφο, δεν ενδείκνυται για συλλογές κειμένων με πολλαπλές επισημειώσεις καθώς δεν ενσωματώνει τις επισημειώσεις στη λογική της. Πιο συγκεκριμένα, η μέθοδος μπορεί να καταλήξει σε θέματα κάποια από τα οποία δεν ερμηνεύονται εύκολα από τους ανθρώπους. Τα θέματα αυτά δεν μπορούν εύκολα να συνδεθούν με διεπαφές χρήστη – ακόμη κι αν υπάρχει χρόνος και διάθεση των χρηστών να συμμετέχουν στην διαδικασία.

Προσέγγιση

Εφόσον οι επισημειώσεις που προέρχονται από ανθρώπους έχουν νόημα για εκείνους που τις χρησιμοποίησαν, μια λύση στο πρόβλημα μπορεί να είναι να χρησιμοποιούνται οι επισημειώσεις των χρηστών ως ετικέτες με βάση τις οποίες προήλθαν οι λέξεις αντί για τα λανθάνοντα και μη ερμηνεύσιμα θέματα. Με βάση αυτή τη σκέψη, έχει προταθεί η λανθάνουσα κατανομή Dirichlet με Ετικέτες (Labelled Latent Dirichlet Allocation, L-LDA), ένα γενετικό μοντέλο για σύνολα δεδομένων με πολλαπλές ετικέτες το οποίο συνδυάζει την επιβλεπόμενη μέθοδο της προσθήκης ετικετών ως επισημειώσεις με την μη – επιβλεπόμενη μέθοδο των πιθανοτικών μοντέλων θεμάτων.

Σε αντίθεση με την κλασική λανθάνουσα κατανομή Dirichlet η προτεινόμενη προσέγγιση συνδέει κάθε ετικέτα με ένα και μόνο λανθάνον θέμα.

Η L-LDA είναι ένα πιθανοτικό γραφικό μοντέλο που περιγράφει μια διαδικασία για την δημιουργία ενός συνόλου εγγράφων χαρακτηρισμένου με ετικέτες (επισημειώσεις). Όπως και στη λανθάνουσα κατανομή Dirichlet, κάθε έγγραφο μοντελοποιείται ως μια ανάμειξη λανθανόντων θεμάτων και κάθε λέξη προέρχεται από ένα θέμα. Σε αντίθεση όμως με την LDA, η προτεινόμενη μέθοδος ενσωματώνει επίβλεψη καθώς περιορίζει το μοντέλο θεμάτων να χρησιμοποιήσει μόνο τα θέματα που αντιστοιχούν στις επισημειώσεις που έχουν συνδεθεί με το συγκεκριμένο έγγραφο.

Έστω ότι κάθε έγγραφο ***d*** εκπροσωπείται από δυο στοιχεία: μια λίστα δεικτών σε λέξεις $w(d) = (w_1, w_2, .., w_{Nd})$ και μια λίστα δυαδικών τιμών που δείχνουν την σύνδεση ή όχι με τις πιθανές ετικέτες $\Lambda(d) = (l_1, l_2, .., l_K)$. Κάθε λέξη μπορεί να πάρει την τιμή ενός όρου που βρίσκεται στο λεξιλόγιο ***V***, $w_i \in \{1,...,V\}$, και κάθε δείκτης επισημείωσης μπορεί να πάρει τιμή 0 ή 1, $k \in \{0,1\}$. Εδώ το ***Nd*** είναι το μήκος του εγγράφου, ***V*** είναι το μέγεθος του λεξιλογίου και ***K*** ο συνολικός αριθμός των μοναδικών ετικετών στο σύνολο δεδομένων. Ορίζουμε τον αριθμό των θεμάτων στην λανθάνουσα κατανομή Dirichlet με ετικέτες να είναι ακριβώς ίσος με τον αριθμό των διαφορετικών ετικετών που χρησιμοποιούνται στη συλλογή εγγράφων.

Η γενετική διαδικασία για τον αλγόριθμο μπορεί να βρεθεί στο αντίστοιχο σχήμα (Εικόνα 3.5). Τα βήματα 1 και 2 ταυτίζονται με την λανθάνουσα κατανομή Dirichlet καθώς αφορούν την (τυχαία) επιλογή κατανομών θεμάτων από μια κατανομή Dirichlet (όπως στην αντίστοιχη εργασία [54]). Στη συνέχεια, η μέθοδος LDA επιλέγει πολυωνυμικές κατανομές που μπορεί να περιέχουν όλα τα πιθανά θέματα για κάθε έγγραφο από μια εκ των προτέρων συνάρτηση Dirichlet ***α***. Στη συγκεκριμένη προσέγγιση όμως, περιορίζουμε το ***ϑ(d)*** ώστε να περιλαμβάνει μόνο στα θέματα που αντιστοιχούν στις ετικέτες ***Λ(d)*** του τρέχοντος εγγράφου. Καθώς οι συνδέσεις λέξεων με θέματα γίνονται με βάση την συγκεκριμένη κατανομή, αυτός ο περιορισμός είναι αρκετός για να περιορίσουμε τα θέματα του κειμένου στις συγκεκριμένες ετικέτες.

1 Για κάθε θέμα $k \in \{1,...,K\}$
2 Επιλογή $\beta_k = (\beta_{k,1},...,\beta_{k,V})^T \sim \text{Dir}(\bullet|\eta)$
3 Για κάθε έγγραφο d
4 Για κάθε θέμα $k \in \{1,...,K\}$
5 Επιλογή $\Lambda^{(d)} \in \{0,1\} \sim \text{Bernoulli}(\bullet|\Phi_k)$
6 Επιλογή $\alpha^{(d)} = L^{(d)} \times \alpha$
7 Επιλογή $\theta^{(d)} = (\theta_{l,1},...,\theta_{l,Md})^T \sim \text{Dir}(\bullet|\theta^{(d)})$
8 Για κάθε i στο $\{1,...,N_d\}$:
9 Επιλογή $z_i \in \{\lambda_1^{(d)},...,\lambda_{Md}^{(d)}1\} \sim \text{Mult}(\bullet|\theta^{(d)})$
10 Επιλογή $w_i \in \{1,...,V\} \sim \text{Mult}(\bullet|\beta_{zi})$

Εικόνα 3.5 Γενετική Διαδικασία της L-LDA

Για να εισάγουμε τον περιορισμό που περιγράψαμε στην γενετική διαδικασία, χρησιμοποιούμε μια συνάρτηση Bernoulli για κάθε θέμα ώστε να επιλέξουμε τις ετικέτες του μοντέλου με βάση μια εκ τον προτέρων πιθανότητα ***Φk*** (βλ. βήμα 5). Στη συνέχεια ορίζουμε το διάνυσμα των ετικετών του εγγράφου να είναι $\lambda^{(d)} = \{k|\Lambda_k^{(d)} = 1\}$. Αυτή η διαδικασία μας επιτρέπει να ορίσουμε έναν πίνακα προβολής $L^{(d)}$ με διαστάσεις M_d *X K* για κάθε έγγραφο ***d,*** όπου $M_d = |\lambda^{(d)}|$ ως εξής: Για κάθε γραμμή $i \in \{1,...,M_d\}$ και στήλη $j \in \{1,...,K\}$ χρησιμοποιούμε την εξίσωση (3.7).

$$L_{ij}^{(d)} = \begin{cases} 1 \; \alpha\nu \; \lambda_i^{(d)} = j \\ \alpha\lambda\lambda\iota\acute{\omega}\varsigma \end{cases} \qquad (3.7)$$

Με άλλα λόγια, η i-οστή γραμμή του πίνακα $\boldsymbol{L^{(d)}}$ έχει την τιμή 1 στην στήλη ***j*** αν και μόνον αν η i-οστή ετικέτα του εγγράφου $\boldsymbol{\lambda_i^{(d)}}$ ισούται με το θέμα ***j***. Ο πίνακας $\boldsymbol{L^{(d)}}$ χρησιμοποιείται για να προβληθεί το διάνυσμα παραμέτρων της εκ των προτέρων συνάρτησης Dirichlet των θεμάτων σε ένα διάνυσμα μικρότερης διάστασης. Η διαδικασία αυτή φαίνεται στην εξίσωση (3.8).

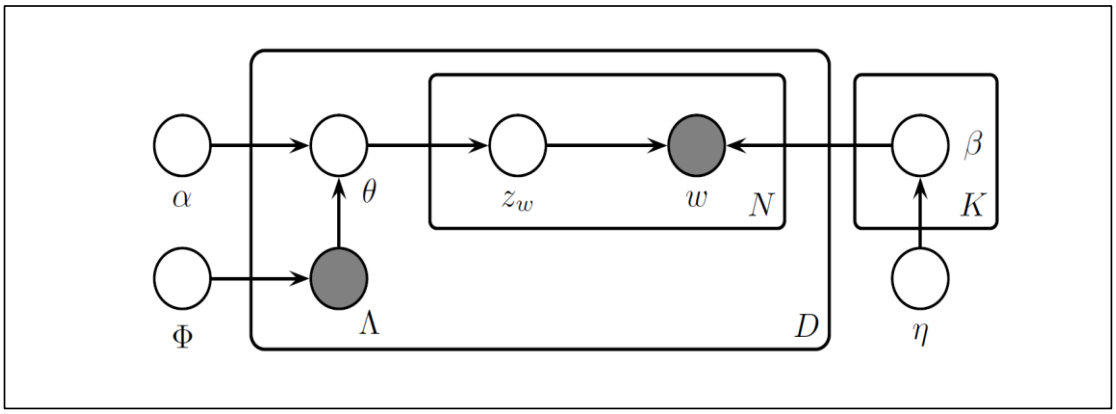
$$\alpha^{(d)} = L^{(d)} \times \boldsymbol{\alpha} = (\alpha_{\lambda 1(d)}, \dots, \alpha_{\lambda Md(d)})^T \qquad (3.8)$$

Οι διαστάσεις του προβαλλόμενου διανύσματος αντιστοιχούν στα θέματα που αναπαρίστανται από τις ετικέτες του εγγράφου.

Για παράδειγμα, αν υποθέσουμε ότι ***K***=4 και οι ετικέτες του εγγράφου ***d*** δίνονται από το $\Lambda^{(d)} = \{0,1,1,0\}$ αυτό σημαίνει ότι $\lambda^{(d)} = \{2,3\}$, και αντίστοιχα το διάνυσμα ***L(d)*** ισούται με $\begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$. Στη συνέχεια το διάνυσμα $\boldsymbol{\vartheta^{(d)}}$ επιλέγεται από μια κατανομή Dirichlet με παραμέτρους $\alpha^{(d)} = L^{(d)} \times \alpha = (\alpha_2, \alpha_3)^T$ που σημαίνει ότι περιορίζεται στα θέματα 2 και 3. Αυτό ικανοποιεί τον περιορισμό που είχαμε θέσει ώστε το έγγραφο να συνδέεται μόνο με τα θέματα τα οποία αντιστοιχούν στις ετικέτες τους.

Το βήμα 6 που αφορά την προβολή είναι το μόνο ντετερμινιστικό βήμα της μεθόδου. Τα υπόλοιπα βήματα από το 7 έως το 10 ταυτίζονται με εκείνα της κλασικής μεθόδου LDA.

Το μοντέλο της L-LDA που προτείνεται παρουσιάζεται και στην Εικόνα 3.6. Εκεί είναι εμφανής η εξάρτηση του $\boldsymbol{\vartheta}$ τόσο από το $\boldsymbol{\alpha}$ αλλά και το $\boldsymbol{\Lambda}$ με τις ακμές που τα συνδέουν.



Εικόνα 3.6 Το Μοντέλο της Λανθάνουσας Κατανομής Dirichlet με Ετικέτες

Αξιολόγηση

Στην αντίστοιχη εργασία, αποδεικνύεται από τους συγγραφείς ότι η προτεινόμενη μέθοδος L-LDA μπορεί να συνεισφέρει στην μοντελοποίηση συλλογών εγγράφων που περιέχουν επισημειώσεις. Τα αποτελέσματα της μεθόδου είναι πιο εύκολα ερμηνεύσιμα συγκριτικά με εκείνα της τυπικής μορφής της λανθάνουσας κατανομής Dirichlet. Επίσης αποδεικνύεται ότι αποδίδει καλύτερα στην εξαγωγή επισημειώσεων από άγνωστο κείμενο σε σχέση με μηχανές διανυσμάτων υποστήριξης (support vector machines)

3.3.2 Μοντέλο Συσχετισμένων Θεμάτων (CTM)

Η περιγραφή του μοντέλου βασίζεται στην εργασία των Blei και Lafferty [57].

Εισαγωγή

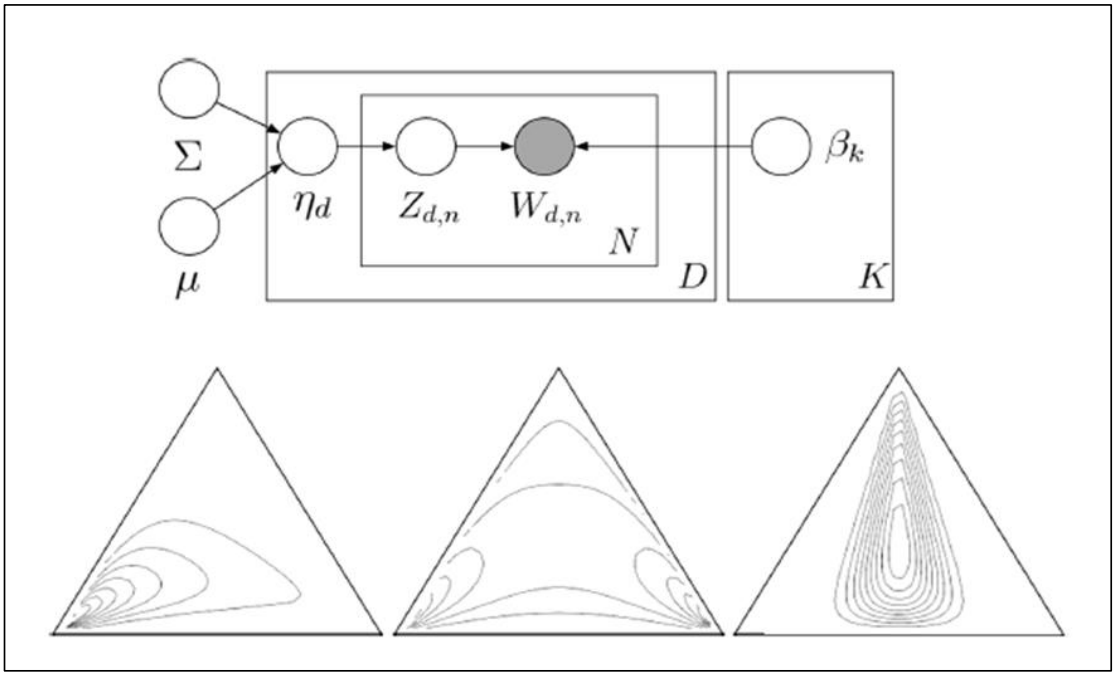
Η προσέγγιση της λανθάνουσας κατανομής Dirichlet (LDA), μπορεί να αποδειχθεί ένα χρήσιμο εργαλείο για την στατιστική ανάλυση συλλογών κειμένων και άλλων διακριτών δεδομένων.

Η LDA εμφανίζει περιορισμούς καθώς δεν έχει την δυνατότητα να συσχετίζει θέματα μεταξύ τους. Αυτή η συσχέτιση υπάρχει στις συλλογές πραγματικών δεδομένων όπου, για παράδειγμα, ένα κείμενο για την γενετική είναι περισσότερο πιθανό να αναφέρεται και στην ιατρική πάρα στην αστρονομία. Ο περιορισμός αυτός προέρχεται από την χρήση της κατανομής Dirichlet για την μοντελοποίηση της ποικιλίας στις διάφορες κατανομές θεμάτων. Σε αυτό το κεφάλαιο θα δούμε το μοντέλο συσχετισμένων θεμάτων (correlated topic model, CTM), όπου οι κατανομές θεμάτων εμφανίζουν συσχέτιση μεταξύ των θεμάτων και αυτό γίνεται με χρήση μίας λογιστικής κανονικής κατανομής (logistic normal distribution). Καθώς η λογιστική κανονική κατανομή δεν είναι συζυγής με την πολυωνυμική, απαιτείται μια ξεχωριστή διαδικασία για την εκτίμηση των κατανομών των θεμάτων από σύνολα δεδομένων.

Προσέγγιση

Το CTM είναι μια ιεραρχική γενετική προσέγγιση που περιγράφει συλλογές κειμένων. Η προσέγγιση αυτή χρησιμοποιεί μοντέλο ανάμειξης για την μοντελοποίηση των λέξεων και οι αναλογίες που χρησιμοποιούνται είναι τυχαίες μεταβλητές.

Ένα κρίσιμο στοιχείο της προτεινόμενης προσέγγισης είναι η λογιστική κανονική κατανομή (logistic normal distribution [58]). Η κατανομή αυτή ορίζεται στο simplex και επιτρέπει την μεταβλητότητα μεταξύ των συνιστωσών της χρησιμοποιώντας μια συνάρτηση κανονικής κατανομής για πολλές μεταβλητές. Η λογιστική κανονική κατανομή χρησιμοποιείται για να εκφράσει συσχετίσεις μεταξύ των μερών της τυχαίας μεταβλητής χρησιμοποιώντας τον πίνακα συνδιακύμανσης της κανονικής κατανομής. Εδώ χρησιμοποιείται για να υποστηρίξει ένα ιεραρχικό μοντέλο και την λανθάνουσα σύνθεση των θεμάτων που σχετίζονται με κάθε έγγραφο.



Εικόνα 3.7 Μοντέλο Συσχετισμένων Θεμάτων

Στο μοντέλο CTM θεωρούμε ότι ένα κείμενο ***N*** λέξεων παράγεται από μια γενετική διαδικασία. Η διαδικασία αυτή φαίνεται στην Εικόνα 3.8. Ο υπολογισμός γίνεται με δεδομένα τα θέματα $\boldsymbol{\beta_{1:K}}$, ένα διάνυσμα μέσων *K* διαστάσεων **μ**, και έναν *K x K* πίνακα συνδιακύμανσης **Σ.**

1 Επιλογή $\eta_d|\{\mu,\Sigma\} \sim N(\mu,\Sigma)$.
2 Για κάθε $\eta \in \{1,...,N_d\}$:
3 Επιλογή της κατανομής θεμάτων $Z_{d,n}|\ \eta_d$ από $Mult(f(\eta_d))$
4 Επιλογή της λέξης $w_{d,n}|\{z_{d,n},\beta_{1:K}\}$ από $Mult(\beta_{zd,n})$

Εικόνα 3.8 Γενετική Διαδικασία Μοντέλου Συσχετισμένων Θεμάτων

Η γενετική διαδικασία του προτεινόμενου μοντέλου είναι σχεδόν όμοια με την γενετική διαδικασία της λανθάνουσας κατανομής Dirichlet με τη διαφορά ότι οι αναλογίες των θεμάτων επιλέγονται από την λογιστική κανονική κατανομή και όχι από την Dirichlet.

Το μοντέλο CTM έχει την δυνατότητα να εκφράσει περισσότερα στοιχεία από την λανθάνουσα κατανομή Dirichlet. Η υπόθεση που κάνουμε χρησιμοποιώντας την LDA είναι ότι κάθε θέμα είναι ανεξάρτητο από κάποιο άλλο. Αυτή η υπόθεση όμως δεν ισχύει στην πραγματικότητα. Ο πίνακας συνδιακύμανσης που χρησιμοποιείται στο προτεινόμενο μοντέλο εισάγει την δυνατότητα τέτοιων συσχετίσεων. Η συγκεκριμένη δομή που προέρχεται από το προτεινόμενο μοντέλο μπορεί να χρησιμοποιηθεί για την εξερεύνηση, κατανόηση και πλοήγηση ενός

εκτενούς συνόλου δεδομένων. Επίσης, η μοντελοποίηση των συσχετίσεων μπορεί να παράγει πιο ακριβείς κατανομές για να χρησιμοποιηθούν σε αντίστοιχες εφαρμογές.

Στην Εικόνα 3.7 παρουσιάζεται το μοντέλο συσχετισμένων θεμάτων. Στην κορυφή εμφανίζεται το γραφικό μοντέλο του CTM. Η λογιστική κανονική κατανομή μπορεί να απεικονίσει συσχετίσεις μεταξύ θεμάτων που είναι αδύνατο να υποστηρίξει μία απλή Dirichlet. Στο κάτω μέρος εμφανίζονται παραδείγματα των πυκνοτήτων της λογιστικής κανονικής κατανομής σε ένα simplex 2 διαστάσεων.

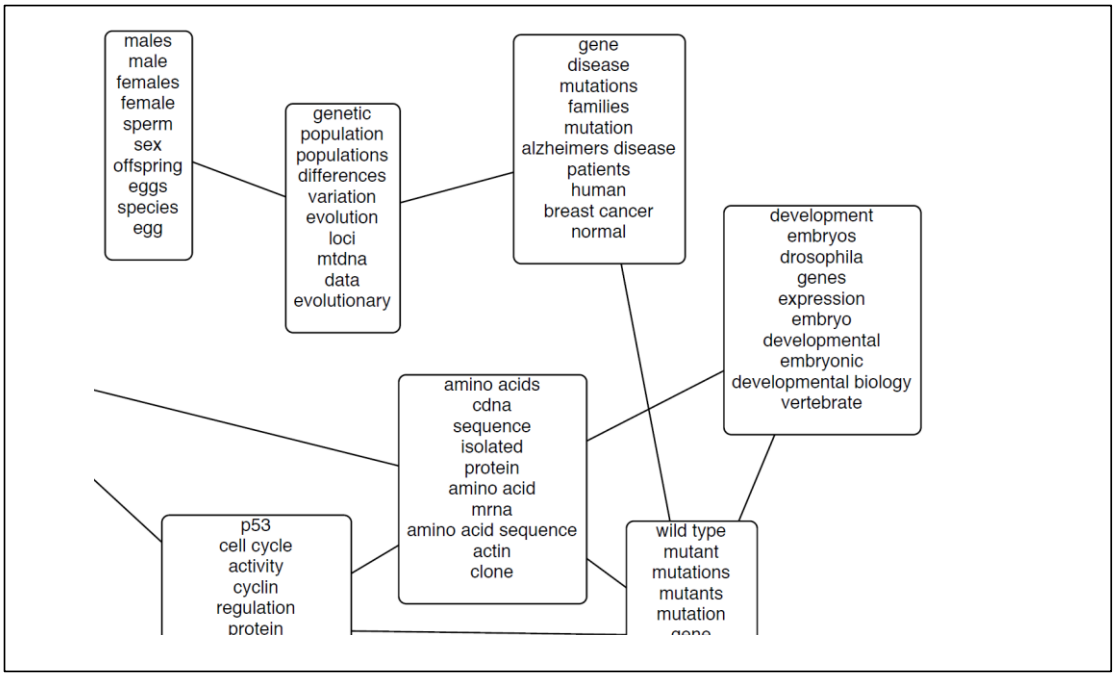
Ένα μειονέκτημα χρήσης της CTM είναι ότι η λογιστική κανονική κατανομή δεν είναι συζυγής με την πολυωνυμική και αυτό δυσκολεύει τον υπολογισμό της εκ των υστέρων κατανομής. Δύο υπολογιστικά προβλήματα πρέπει να λυθούν όταν χρησιμοποιούμε το CTM για να αναλύσουμε δεδομένα.

Πρώτον, με δεδομένες μία συλλογή θεμάτων και μια κατανομή θεματικών αναλογιών $\{\boldsymbol{\beta}_{1:K}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ πρέπει να υπολογίσουμε την εκ των υστέρων κατανομή των κρυφών μεταβλητών δεσμευμένων ως προς τα κείμενα που παρατηρούμε $p(\boldsymbol{\eta}, \boldsymbol{z} | \boldsymbol{w}, \boldsymbol{\beta}_{1:K}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$. Αυτός ο υπολογισμός δεν μπορεί να γίνει, αλλά μπορούν να γίνουν εκτιμήσεις. Για το σκοπό αυτό χρησιμοποιούμε την μέθοδο προσέγγισης μέσου πεδίου (mean field variational method) για να εκτιμήσουμε την εκ των υστέρων κατανομή, κάτι που μας επιτρέπει να αναλύσουμε αποδοτικά μεγάλες συλλογές δεδομένων κάτω από σχετικά πολύπλοκες προϋποθέσεις.

Δεύτερον, δεδομένης μια συλλογής κειμένων $\{\mathbf{w}_1, \ldots, \mathbf{w}_D\}$ πρέπει να επαναλάβουμε τον αλγόριθμο για να έχουμε την μέγιστη πιθανοφάνεια αλλά και να υπολογίσουμε την λογιστική κανονική κατανομή με βάση τις υποθέσεις του CTM. Για το σκοπό αυτό προτείνεται η χρήση μιας μορφής του αλγορίθμου μεγιστοποίησης αναμονής (variational expectation maximization).

Περισσότερα στοιχεία για την αντιμετώπιση των συγκεκριμένων υπολογιστικών ζητημάτων μπορούν να βρεθούν στην αντίστοιχη εργασία.

Αξιολόγηση



Εικόνα 3.9 Γράφος Συσχετισμένων Θεμάτων

Για τον αξιολόγηση του προτεινόμενου μοντέλου οι συγγραφείς της εργασίας εξήγαγαν ένα μοντέλο 100 συσχετισμένων θεμάτων με βάση 16.351 άρθρα του περιοδικού *Science* που δημοσιεύτηκαν από το 1990 μέχρι το 1999. Το αποτέλεσμα της διαδικασίας είναι ένας γράφος που περιέχει τα λανθάνοντα θέματα και τις συνδέσεις μεταξύ τους. Ένα μέρος του αποτελέσματος φαίνεται στην Εικόνα 3.9.

Επίσης αξιολογήθηκε η απόδοση του CTM σε σχέση με το LDA χρησιμοποιώντας μια μικρότερη συλλογή άρθρων. Χρησιμοποιώντας διασταυρωμένη επικύρωση με 10 επαναλήψεις, υπολογίστηκε η πιθανοφάνεια των δεδομένων που είχαν κρατηθεί με βάση τα υπόλοιπα δεδομένα. Η μεγαλύτερη πιθανότητα εμφάνισης σημαίνει ότι υπάρχει κι ένα καλύτερο μοντέλο. Το μοντέλο συσχετισμένων θεμάτων φαίνεται να υπερτερεί και να υποστηρίζει περισσότερα θέματα. Τέλος αξιολογήθηκε η απόδοση του μοντέλου στην πρόβλεψη του υπόλοιπου κειμένου ενός εγγράφου εάν γνωρίζουμε ένα τμήμα του.

Τα αποτελέσματα της συνολικής αξιολόγησης έδειξαν ότι το προτεινόμενο μοντέλο παρουσίασε συγκεκριμένα και μετρήσιμα πλεονεκτήματα.

3.3.3 Δυναμικό Μοντέλο Θεμάτων (DTM)

Η περιγραφή του δυναμικού μοντέλου θεμάτων βασίζεται στην εργασία των Blei και Lafferty [59].

Εισαγωγή

Σε ένα μοντέλο θεμάτων που ισχύει η αρχή της ανταλλαξιμότητας κάθε θέμα μπορεί να αντικατασταθεί από κάποιο άλλο σε επόμενη δειγματοληψία. Οι αναλογίες ανάμειξης επιλέγονται στην τύχη για κάθε έγγραφο αλλά τα θέματα είναι κοινά σε ολόκληρο το μοντέλο. Σε αυτά τα μοντέλα οι λέξεις είναι ανταλλάξιμες η μία με την άλλη και αυτή η απλοποίηση βοηθά τον εντοπισμό των εννοιών που επηρεάζουν την δημιουργία ενός κειμένου.

Όμως για μερικές συλλογές κειμένων αυτή η υπόθεση της ανταλλαξιμότητας δεν ευσταθεί. Για παράδειγμα, τα επιστημονικά περιοδικά, το ηλεκτρονικό ταχυδρομείο ή τα άρθρα σε μέσα ενημέρωσης αφορούν περιεχόμενο που εξελίσσεται. Τα θέματα σε μια συλλογή εγγράφων εξελίσσονται στη διάρκεια του χρόνου και η μοντελοποίηση της εξέλιξης μπορεί να ευνοήσει την επίλυση προβλημάτων.

Προσέγγιση

Ενώ η κλασσική μοντελοποίηση χρονοσειρών επικεντρώνεται στην ανάλυση συνεχών δεδομένων, τα μοντέλα θεμάτων σχεδιάστηκαν για διακριτά δεδομένα. Η προσέγγισή μας αφορά την χρήση μοντέλων χώρου κατάστασης (state space models) στον χώρο των παραμέτρων των πολυωνυμικών μεταβλητών θεμάτων, καθώς επίσης και των παραμέτρων που ορίζουν τις λογιστικές κανονικές κατανομές που χρησιμοποιούνται στις αναλογίες θεμάτων κάθε εγγράφου.

Στο δυναμικό μοντέλο θεμάτων, θεωρούμε ότι τα δεδομένα χωρίζονται σε περιόδους χρόνου, για παράδειγμα ανά έτος. Μοντελοποιούμε τα κείμενα κάθε περιόδου με ένα μοντέλο ***K*** θεμάτων, όπου τα θέματα που αφορούν την περίοδο ***t*** εξελίσσονται με βάση τα θέματα που αφορούν την περίοδο ***t-1***.

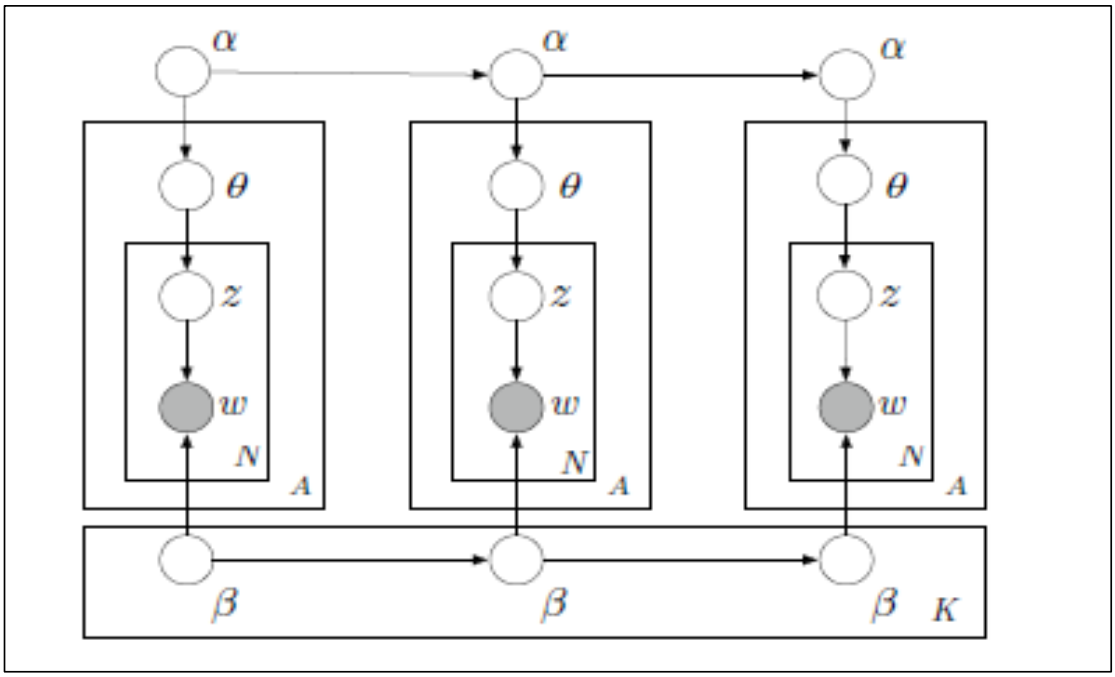
Για ένα μοντέλο **K** θεμάτων με V όρους, ορίζουμε ως $\boldsymbol{\beta}_{t,k}$ ένα διάνυσμα V διαστάσεων των φυσικών παραμέτρων του θέματος ***k*** στην περίοδο ***t***. Η συνηθισμένη απεικόνιση μια πολυωνυμικής κατανομής γίνεται με την παραμετροποίηση του μέσου της. Αν ορίσουμε την παράμετρο μέσω της

πολυωνυμικής κατανομής *V* διαστάσεων ως **π**, το *i*-οστό στοιχείο της φυσικής παραμέτρου δίνεται με βάση το $\beta_i = log(\pi_i/\pi_V)$.

Στις τυπικές εφαρμογές γλωσσικής μοντελοποίησης, οι Dirichlet κατανομές χρησιμοποιούνται για να μοντελοποιήσουν την αβεβαιότητα των κατανομών πάνω στις λέξεις. Όμως η Dirichlet κατανομή δεν μπορεί να υποστηρίξει την διαδοχική μοντελοποίηση. Αντί αυτής, προτείνεται η σύνδεση των φυσικών παραμέτρων κάθε θέματος $\boldsymbol{\beta_{t,k}}$ με ένα μοντέλο χώρου κατάστασης που εξελίσσεται με Γκαουσιανό θόρυβο (gaussian noise). Η απλούστερη εκδοχή ενός τέτοιου μοντέλου φαίνεται στην εξίσωση (3.9).

$$\beta_{t,k}|\beta_{t-1,k} \sim N(\beta_{t-1,k}, \sigma^2 I) \qquad (3.9)$$

Η προσέγγιση περιλαμβάνει την μοντελοποίηση σειρών σύνθετων τυχαίων μεταβλητών συνδέοντας Γκαουσιανές κατανομές με ένα δυναμικό μοντέλο και αντιστοιχίζοντας τις προκύπτουσες τιμές στο simplex.



Εικόνα 3.10 Γραφική Απεικόνιση του Δυναμικού Μοντέλου Θεμάτων

Στην λανθάνουσα κατανομή Dirichlet, οι αναλογίες μεταξύ των θεμάτων επιλέγονται από μία κατανομή Dirichlet. Στο δυναμικό μοντέλο θεμάτων, χρησιμοποιείται μια λογιστική κανονική κατανομή με μέσο $\boldsymbol{\alpha}$ για να εκφραστεί η αβεβαιότητα στις αναλογίες των θεμάτων. Η εξέλιξη μεταξύ των μοντέλων λαμβάνει την δυναμική μορφή που περιγράφεται στην εξίσωση (3.10).

$$\alpha_t|\alpha_{t-1} \sim N(\alpha_{t-1,}\delta^2 I) \quad (3.10)$$

Συσχετίζοντας τα θέματα με κατανομές θεμάτων, έχει κατ' επέκταση σχηματιστεί μια αλυσίδα εξελισσόμενων μοντέλων θεμάτων. Η γενετική διαδικασία για περίοδο ***t*** για ένα σύνολο χρονικά τοποθετημένων δεδομένων είναι όπως φαίνεται στην Εικόνα 3.11.

1 Επιλογή θεμάτων $\beta_{t,k}|\beta_{t-1,k} \sim N(\beta_{t-1,k,}\sigma^2 I)$
2 Επιλογή $\alpha_t|\alpha_{t-1} \sim N(\alpha_{t-1,}\delta^2 I)$
3 Για κάθε έγγραφο:
4 Επιλογή $\eta \sim N(\alpha_t,\alpha^2 I)$
5 Για κάθε λέξη
6 Επιλογή $Z \sim Mult(\pi(\eta))$
7 Επιλογή $W_{t,d,n} \sim Mult(\pi(\beta_t,z))$

Εικόνα 3.11 Γενετική Διαδικασία Δυναμικού Μοντέλου Θεμάτων

Σε αυτή τη διαδικασία το π συνδέει τις πολυωνυμικές φυσικές παραμέτρους με τις παραμέτρους μέσου (βλ. εξίσωση (3.11)).

$$\pi(\beta_{k,t})_w = \frac{exp(\beta_{k,t,w})}{\sum_w exp(\beta_{k,t,w})} \quad (3.11)$$

Στην Εικόνα 3.10 παρουσιάζεται το γραφικό μοντέλο του δυναμικού μοντέλου θεμάτων για τρεις χρονικές περιόδους. Οι παράμετροι κάθε θέματος $\boldsymbol{\beta_{t,k}}$ εξελίσσονται καθώς περνάει ο χρόνος, μαζί με τις παραμέτρους μέσου $\boldsymbol{\alpha_t}$ και την λογιστική κανονική κατανομή των αναλογιών θεμάτων. Αν τα οριζόντια βέλη του σχήματος διαγραφούν καταστρέφεται η δυναμικότητα με βάση τον χρόνο και το μοντέλο μετατρέπεται σε ένα σύνολο ανεξάρτητων μοντέλων θεμάτων. Με την δυναμικότητα του χρόνου, το *k*-οστό θέμα την περίοδο ***t*** προκύπτει ως εξέλιξη *από* το k-οστό θέμα της περιόδου ***t-1***.

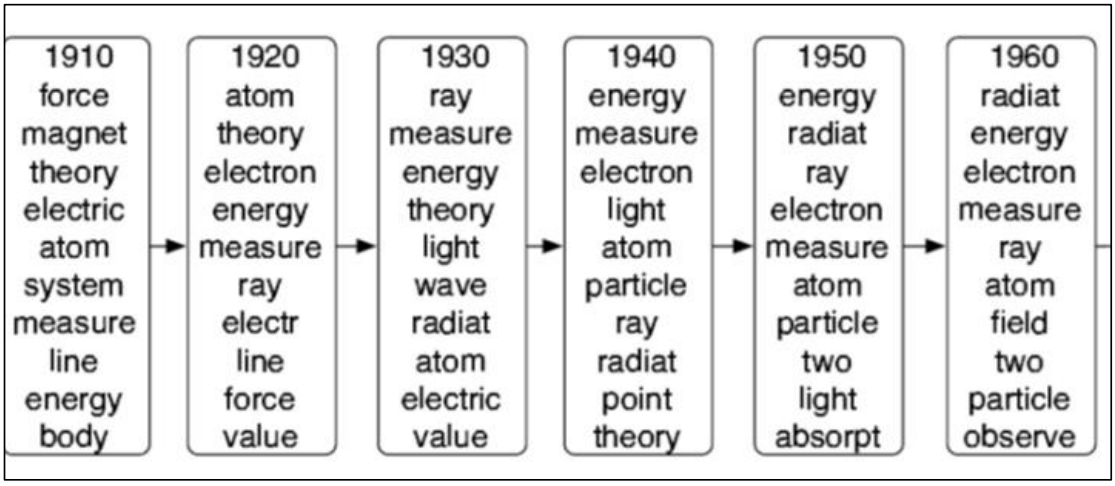
Στη συνέχεια οι συγγραφείς εστιάζουν σε ένα μοντέλο που αποτελείται από K δυναμικά μοντέλα που εξελίσσονται όπως στην Εικόνα 3.10, και όπου η αναλογία θεμάτων προέρχεται σταθερά από μία Dirichlet κατανομή. Τα τεχνικά ζητήματα που συσχετίζονται με την μοντελοποίηση των θεματικών αναλογιών στις χρονοσειρές ταυτίζονται με εκείνα των θεμάτων που συνδέονται μεταξύ τους στην πορεία του χρόνου.

Η εργασία με χρονοσειρές πάνω σε φυσικές παραμέτρους μας οδηγεί στην χρήση Γκαουσιανών μοντέλων για την δυναμική του χρόνου. Όμως, εφόσον η Γκαουσιανή κατανομή και η φυσικές παράμετροι δεν είναι συζυγείς δεν μπορούμε να υπολογίσουμε τις εκ των υστέρων κατανομές. Έτσι οι συγγραφείς επιλύουν το

πρόβλημα προτείνοντας μεταβολικές μεθόδους (variational methods) για την πραγματοποίηση εκτιμήσεων. Περισσότερες λεπτομέρειες για την επίλυση των προβλημάτων μπορούν να βρεθούν στην αντίστοιχη ερευνητική εργασία [59].

Αξιολόγηση

Οι συγγραφείς χρησιμοποίησαν την μέθοδο για να αναλύσουν ένα σύνολο 30.000 άρθρων από το περιοδικό Science, 250 από κάθε χρόνο μεταξύ του 1881 και του 1999. Η συλλογή εγγράφων περιέχει περίπου 7,5 εκατομμύρια λέξεις, ενώ το λεξιλόγιο περιέχει περίπου 15.955. Για την ανάλυση χρησιμοποιήθηκε ένα δυναμικό μοντέλο 20 θεμάτων.



Εικόνα 3.12 Παράδειγμα του Αποτελέσματος Δυναμικού Μοντέλου Θεμάτων

Το μοντέλο μπορεί να εξάγει επιστημονικά θέματα κι επίσης μπορεί να χρησιμοποιηθεί για να παρατηρήσουμε την εξέλιξη των λέξεων που χρησιμοποιούμε στα συγκεκριμένα θέματα. Ένα τέτοιο παράδειγμα εμφανίζεται στην Εικόνα 3.12 όπου φαίνεται η εξέλιξη ενός θέματος από το έτος 1910 έως το 1960. Για την ποσοτική αξιολόγηση του μοντέλου έχει ελεγχθεί η δυνατότητα του μοντέλου να προβλέψει την θεματική κατανομή του περιοδικού με βάση τα δεδομένα των προηγούμενων ετών. Το δυναμικό μοντέλο θεμάτων αποδίδει ικανοποιητικά τόσο στην ποιοτική όσο και στην ποσοτική αξιολόγηση.

3.3.4 Δυναμικό Μοντέλο Θεμάτων Συνεχούς Χρόνου (cDTM)

Η περιγραφή του δυναμικού μοντέλου θεμάτων συνεχούς χρόνου βασίζεται στην εργασία των Wang, Blei και Heckerman [60].

Εισαγωγή

Τα τελευταία χρόνια τα μοντέλα θεμάτων χρησιμοποιούνται όλο και περισσότερο για την ανάλυση συλλογών κειμένων.

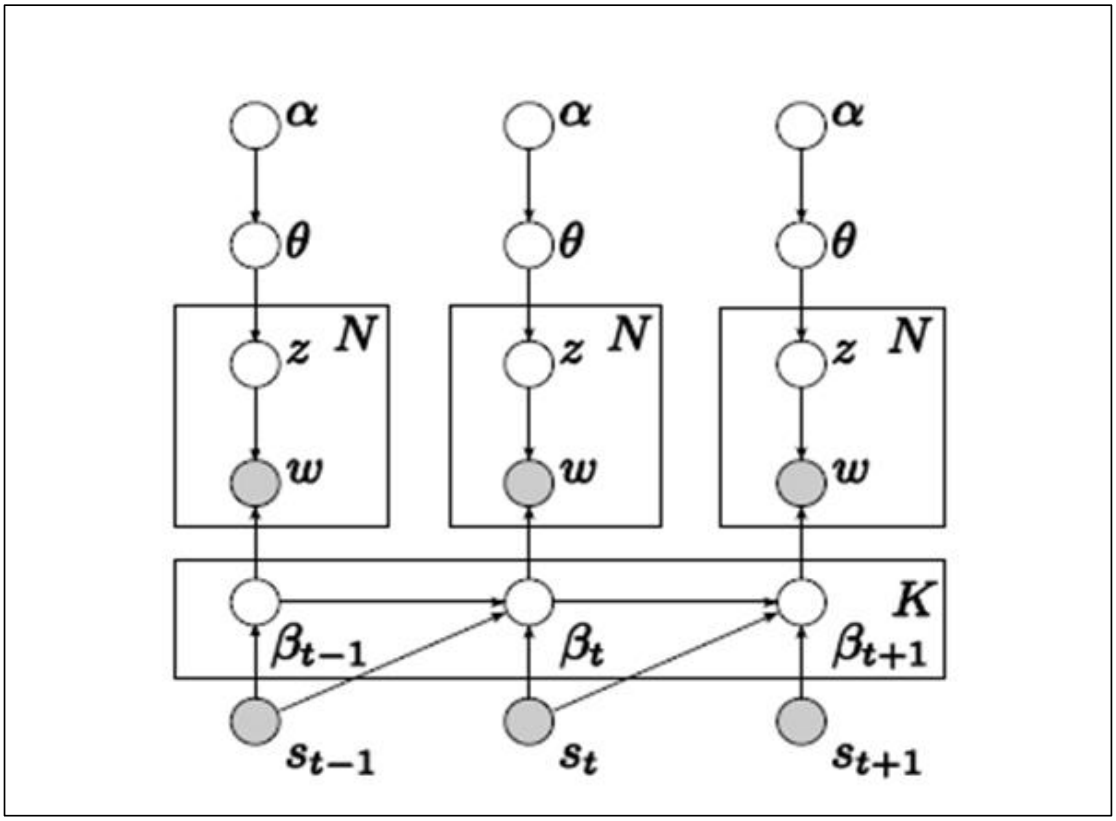
Για να αντιμετωπιστεί το θέμα της εξέλιξης των θεμάτων στη διάρκεια του χρόνου έχει προταθεί το δυναμικό μοντέλο θεμάτων (βλ. ενότητα 3.3.3) το οποίο αφορά διακριτές μονάδες χρόνου. Το δυναμικό μοντέλο Θεμάτων χρησιμοποιεί ένα μοντέλο χώρου κατάστασης στις φυσικές παραμέτρους των πολυωνυμικών κατανομών που απεικονίζουν τα θέματα. Αυτό απαιτεί ο χρόνος να διαμερίζεται σε περιόδους, και μέσα σε κάθε περίοδο η λανθάνουσα κατανομή Dirichlet χρησιμοποιείται για να αναλυθούν τα κείμενα. Παρόλο που το δυναμικό μοντέλο Θεμάτων είναι ένα ισχυρό μοντέλο, η επιλογή της διαμέρισης επηρεάζει αρνητικά τις απαιτήσεις μνήμης και την υπολογιστική πολυπλοκότητα της εκ των υστέρων επαγωγής. Αυτό καθορίζει σε μεγάλο βαθμό την ανάλυση, στην οποία θα προσαρμόσουμε το μοντέλο.

Για να λύσουμε το πρόβλημα της διαμέρισης, θεωρούμε τον χρόνο ως μια συνεχή μεταβλητή. Το δυναμικό μοντέλο θεμάτων συνεχούς χρόνου (cDTM) που προτείνεται αντικαθιστά το διακριτό μοντέλο κατάστασης χώρου του DTM με την γενίκευση του, την κίνηση Brown. Το cDTM γενικεύει το DTM περιγράφοντας μόνο τον βαθμό ανάλυσής με τον οποίο καταγράφεται ο χρόνος δημοσίευσης των κειμένων.

Το cDTM μοντέλο εισάγει πολύ περισσότερες λανθάνουσες μεταβλητές από το DTM. Όμως αυτό το φαινομενικά πολύπλοκο μοντέλο είναι απλούστερο και πιο αποτελεσματικό Τα cDTM και DTM δεν είναι τα μόνα θεματικά μοντέλα που λαμβάνουν τον χρόνο υπόψη τους. Τόσο τα θέματα μοντέλων χρόνου (TOT) [61] όσο και τα δυναμικά μοντέλα ανάμειξης (DMM) [62] περιλαμβάνουν την διάσταση του χρόνου στην ανάλυση των κειμένων.

Προσέγγιση

Σε μία συλλογή κειμένων που έχουν αντιστοιχιστεί σε χρονικές στιγμές επιχειρείται να μοντελοποιηθούν τα λανθάνοντα θέματα όπως αλλάζουν κατά την εξέλιξη της συλλογής. Σε δεδομένα ειδήσεων, για παράδειγμα, ένα απλό θέμα θα αλλάξει καθώς τα άρθρα που σχετίζονται με αυτό εξελίσσονται. Το δυναμικό θεματικό μοντέλο διακριτού χρόνου (DTM) στηρίζεται στα ανταλλάξιμα μοντέλα θεμάτων. Στο DTM, τα κείμενα χωρίζονται σε ομάδες τοποθετημένες με χρονική σειρά, και τα θέματα κάθε περιόδου εξελίσσονται από τα θέματα της προηγούμενης περιόδου. Τα κείμενα όμως στο εσωτερικό μίας ομάδας θεωρούνται ανταλλάξιμα.



Εικόνα 3.13 Γραφικό Μοντέλο cDTM

Ένα μειονέκτημα του DTM που έχει ήδη προταθεί είναι ότι ο χρόνος είναι διαμερισμένος. Αν ο βαθμός ανάλυσης που έχει επιλεγεί είναι αρκετά μικρός, τότε η υπόθεση ότι τα κείμενα μέσα σε ένα χρονικό βήμα είναι ανταλλάξιμα δεν θα είναι αληθινή. Αν η ανάλυση είναι αρκετά μεγάλη, τότε ο αριθμός των μεταβολικών παραμέτρων θα μεγαλώνει δραματικά καθώς προστίθενται περισσότερα χρονικά σημεία. Η επιλογή του βαθμού ανάλυσης στον οποίο γίνεται η διαμέριση πρέπει να

βασίζεται στις υποθέσεις σχετικά με τα δεδομένα. Όμως, οι υπολογιστικές απαιτήσεις μπορεί να εμποδίσουν την ανάλυση στη κατάλληλη κλίμακα χρόνου.

Στο δυναμικό μοντέλο θεμάτων συνεχούς χρόνου απεικονίζουμε τα θέματα στην φυσική τους παραμετροποίηση, αλλά θα χρησιμοποιήσουμε την κίνηση Brown για να μοντελοποιήσουμε την εξέλιξη στη διάρκεια του χρόνου. Θεωρούμε τα *i,j (j > i > 0)* να είναι δύο αυθαίρετοι δείκτες χρόνου, τα s_i και s_j είναι οι αντίστοιχες χρονοσφραγίδες, και $\Delta_{sj,si}$ να είναι ο χρόνος μεταξύ τους. Σε ένα cDTM μοντέλο K θεμάτων, η κατανομή της k-οστής $(1 \leq k \leq K)$ παραμέτρου θέματος στον όρο w εμφανίζεται στην εξίσωση (3.12). Στην εξίσωση (3.12) η διακύμανση αυξάνεται γραμμικά σε σχέση με την καθυστέρηση. Η συγκεκριμένη εξίσωση αποτελεί βασικό συστατικό της γενετικής διαδικασίας του προτεινόμενου μοντέλου.

$$\beta_{0,k,w} \sim N(m, v_0)$$
$$\beta_{j,k,w} | \beta_{i,k,w,s} \sim N(\beta_{i,k,w,v} \Delta_{sj,si}) \qquad (3.12)$$

Όταν j = i + 1, γράφουμε $\Delta_{sj,si}$ σαν Δ_{sj} για συντομία.

1. Για κάθε θέμα k, $1 \leq k \leq K$
   α. Επιλογή $\beta_{0,k} \sim N(m, v_0 I)$
2. Για κάθε έγγραφο d_t τη χρονική στιγμή S_t $(t>0)$:
   α. Για κάθε θέμα k, $1 \leq k \leq K$,
      i. Επιλογή από το μοντέλο κίνησης Brown
         $\beta_{t,k} | \beta_{t-1,k,s} \sim N(\beta_{t-1,k,v} \Delta_{st} I)$
   β. Επιλογή $\theta_t \sim Dir(\alpha)$
   γ. Για κάθε λέξη
      i. Επιλογή $z_{t,n} \sim Mult(\theta_t)$
      ii. Επιλογή $w_{t,n} \sim Mult(\pi(\beta_{t,zt,n}))$

Εικόνα 3.14 Γενετική Διαδικασία cDTM

Όπως και στο διακριτό δυναμικό μοντέλο θεμάτων, το π συνδέει τις πολυωνυμικές φυσικές παραμέτρους, που στην περίπτωση μας είναι απεριόριστες, με τις παραμέτρους μέσου (βλ. εξίσωση (3.11)). Το προτεινόμενο μοντέλο απεικονίζεται στην Εικόνα 3.13. Η εξέλιξη των θεματικών παραμέτρων $\boldsymbol{\beta_t}$ εξαρτάται από την κίνηση Brown. Η μεταβλητή s_t είναι το παρατηρούμενο στιγμιότυπο του κειμένου $\boldsymbol{d_t}$.

Το δυναμικό μοντέλο θεμάτων συνεχούς χρόνου μπορεί να ερμηνευτεί και ως μία γενίκευση του DTM. Και στα δύο μοντέλα η πιθανότητα εμφάνισης ενός όρου παρουσιάζει διακύμανση κατά τη διάρκεια ενός χρονικού διαστήματος μεταξύ δυο παρατηρήσεων.

Στο DTM το διάστημα μεταξύ δυο παρατηρήσεων χωρίζεται σε όμοια διακριτά κομμάτια. Σε κάθε χρονική στιγμή υπάρχει η πλήρης απεικόνιση των θεμάτων και των όρων που τα αποτελούν. Έτσι οδηγούμαστε σε μεγάλες απαιτήσεις για μνήμη για τον υπολογισμό του μοντέλου ακόμα και αν οι παρατηρήσεις είναι αραιά διαμερισμένες σε όλο τον χρόνο. Στο δυναμικό μοντέλο θεμάτων συνεχούς χρόνου, όμως, η διακύμανση είναι μια συνάρτηση της καθυστέρησης μεταξύ των παρατηρήσεων, και οι πιθανότητες στα διακριτά βήματα αυτών των παρατηρήσεων δεν χρειάζεται να ληφθούν υπόψη. Ένα DTM μπορεί να παραχθεί από ένα cDTM θεωρώντας μόνο κάποιες χρονοσφραγίδες των κειμένων στον επιθυμητό βαθμό ανάλυσης.

Η γενετική διαδικασία του cDTM περιγράφεται στην Εικόνα 3.14. Η αναλυτική αντιμετώπιση των προβλημάτων της στατιστικής επαγωγής μπορεί να βρεθεί στην αντίστοιχη εργασία [60].

Αξιολόγηση

Οι συγγραφείς της αντίστοιχης εργασίας έχουν αξιολογήσει το προτεινόμενο μοντέλο σε δύο συλλογές δεδομένων που προέρχονται από ειδησεογραφικές πηγές. Το πρώτο σύνολο δεδομένων προέρχεται από το TREC AP corpus [63] και περιέχει ειδήσεις από την 1/5/1988 ως την 30/6/1988. Τα δεδομένα αυτά αφορούν τις εκλογές και περιλαμβάνονται σε 1.342 έγγραφα. Δεύτερον, τα Election 8 δεδομένα αποτελούν περιλήψεις άρθρων από το γνωστό ιστότοπο Digg[14] και αφορούν τις εκλογές του 2008. Στο πρώτο σύνολο δεδομένων στα άρθρα καταγράφεται ο χρόνος συγγραφής με ακρίβεια ώρας ενώ στο δεύτερο με ακρίβεια μέρας.

Στα δεδομένα αυτά τα μοντέλα αξιολογήθηκαν για την ικανότητα τους να προβλέψουν τις λέξεις και την χρονική στιγμή των κειμένων και για την γενικότερη ποιότητα των εξαγόμενων θεμάτων. Τα αποτελέσματα είναι ενθαρρυντικά.

3.3.5 Ιεραρχικό Μοντέλο Θεμάτων (hLDA)

Η περιγραφή του ιεραρχικού μοντέλου βασίζεται στην εργασία των Blei, Griffiths, Jordan και Tenenbaum [64].

[14] http://digg.com

Εισαγωγή

Τα πολύπλοκα πιθανοτικά μοντέλα κυριαρχούν όλο και περισσότερο σε περιοχές όπως της βιοπληροφορικής, ανάκτησης πληροφορίας και όρασης μηχανών. Ένα σημαντικό παράδειγμα των σημερινών προκλήσεων είναι το πρόβλημα εκμάθησης ιεραρχίας θεμάτων από τα δεδομένα. Δεδομένης μιας συλλογής από κείμενα το καθένα από τα οποία περιέχει λέξεις, επιθυμούμε να ανακαλύψουμε χρήσιμες δομές ή "θέματα" σε κείμενα και να οργανώσουμε αυτά τα θέματα σύμφωνα με μια ιεραρχία. Η προτεινόμενη προσέγγιση αποτελεί μια στατιστική τεχνική για την κατασκευή μιας ιεραρχίας που μπορεί να επεκτείνεται και να αλλάζει καθώς συσσωρεύονται όλο και περισσότερα δεδομένα.

Προσεγγίζουμε αυτό το πρόβλημα επιλογής μοντέλων προσδιορίζοντας ένα γενικό πιθανοτικό μοντέλο ιεραρχικών δομών και χρησιμοποιώντας την Μπεϋζιανή οπτική στο πρόβλημα εκμάθησης αυτών των δομών από τα δεδομένα. Έτσι οι ιεραρχίες μας είναι τυχαίες μεταβλητές. Οι συγκεκριμένες μεταβλητές ορίζονται μέσω διαδικασιών, με βάση με έναν αλγόριθμο που κατασκευάζει την ιεραρχία χρησιμοποιώντας ως είσοδο τα δεδομένα. Το πιθανοτικό αντικείμενο που υποστηρίζει αυτή την προσέγγιση είναι μια κατανομή διαμερίσεων ακέραιων αριθμών γνωστή ως διαδικασία Chinese Restaurant (CRP). Η διαδικασία Chinese Restaurant επεκτείνεται σε μια ιεραρχία διαμερίσεων και χρησιμοποιείται για την απεικόνιση των εκ των προτέρων και των εκ των υστέρων κατανομών σε ιεραρχίες θεμάτων.

Υπάρχουν πολλές πιθανές προσεγγίσεις στην μοντελοποίηση ιεραρχιών θεμάτων. Στην τρέχουσα προσέγγιση, κάθε κόμβος στην ιεραρχία συσχετίζεται με ένα θέμα, όπου ένα θέμα είναι μια κατανομή από λέξεις. Ένα κείμενο παράγεται επιλέγοντας ένα μονοπάτι από την ρίζα της ιεραρχίας ως ένα φύλλο, συνεχώς επιλέγοντας θέματα καθώς το μονοπάτι διασχίζει την ιεραρχία, και επιλέγοντας λέξεις με βάση τα θέματα. Έτσι η οργάνωση των θεμάτων σε μια ιεραρχία στοχεύει στο να συλλάβει την χρήση των θεμάτων στο σύνολο δεδομένων.

Αυτή η προσέγγιση διαφέρει από μοντέλα ιεραρχιών θεμάτων που στηρίζονται στην υπόθεση ότι οι κόμβοι-γονείς είναι συσχετισμένοι με τους κόμβους παιδιά. Εδώ δεν γίνεται κάποια τέτοια υπόθεση.

Προσέγγιση

Περιγράφοντας την προσέγγιση του ιεραρχικού πιθανοτικού μοντέλου και πιο συγκεκριμένα της μεθόδου ιεραρχικής λανθάνουσας κατανομής Dirichlet, ξεκινάμε με μια σύντομη περιγραφή της διαδικασίας Chinese Restaurant και στην συνέχεια θα δείξουμε πως αυτή η διαδικασία μπορεί να υποστηρίξει το προτεινόμενο μοντέλο.

Η διαδικασία Chinese Restaurant (CRP) είναι μια κατανομή διαμερίσεων ακεραίων που προκύπτει αν υποθέσουμε μια διαδικασία στην οποία ***M*** πελάτες κάθονται σε ένα εστιατόριο με άπειρο αριθμό τραπεζιών. Η βασική διαδικασία ορίζεται ως εξής: Ο πρώτος πελάτης κάθεται στο πρώτο τραπέζι και ο m-οστός πελάτης κάθεται σε ένα τραπέζι σύμφωνα με την κατανομή που εμφανίζεται στην εξίσωση (3.13), όπου το $\boldsymbol{m_i}$ είναι ο αριθμός των προηγούμενων πελατών στο τραπέζι ***i***, και ***γ*** μια παράμετρος.

$$p(previously\ occupied\ table\ i) = \frac{m_i}{\gamma+m-1}$$
$$p(the\ next\ unoccupied\ table) = \frac{\gamma}{\gamma+m-1} \qquad (3.13)$$

Αφού καθίσουν ***M*** πελάτες, η διαδικασία αυτή παράγει μια διαμέριση ***M*** αντικειμένων. Αυτή η διαδικασία δίνει την ίδια δομή διαμερίσεων όπως αν πραγματοποιούσαμε επιλογές από μια διαδικασία Dirichlet. Όμως, η CRP επιτρέπει την διατύπωση πολλών παραλλαγών του βασικού κανόνα της εξίσωσης (3.13), συμπεριλαμβανομένης και μιας επιλογής του ***γ*** που εξαρτάται από τα δεδομένα και την παρούσα διαμέριση. Αυτή η ευελιξία της διαδικασίας αποδεικνύεται χρήσιμη για την δημιουργία ιεραρχιών θεμάτων.

Η διαδικασία Chinese Restaurant έχει χρησιμοποιηθεί για να απεικονίσει την αβεβαιότητα στον αριθμό στοιχείων που συμμετέχουν σε ένα μοντέλο ανάμειξης. Η διαδικασία μπορεί να χρησιμοποιηθεί καθώς μπορούμε να σχηματίσουμε μια ένα-προς-ένα σχέση μεταξύ των τραπεζιών και των αναμεμειγμένων στοιχείων και μια ένα-προς-πολλά σχέση ανάμεσα στα αναμεμειγμένα στοιχεία και στα δεδομένα. Στα μοντέλα που θα θεωρήσουμε, όμως, κάθε σημείο δεδομένων είναι συσχετισμένο με πολλαπλά αναμεμειγμένα στοιχεία που βρίσκονται σε ένα μονοπάτι μιας ιεραρχίας.

Η φωλιασμένη διαδικασία Chinese Restaurant (nested Chinese restaurant process) μπορεί να οριστεί υποθέτοντας το εξής σενάριο. Υποθέτουμε ότι υπάρχει ένας άπειρος αριθμός εστιατορίων με άπειρα τραπέζια σε μια πόλη. Ένα εστιατόριο ορίζεται να είναι το εστιατόριο-ρίζα, και πάνω σε καθένα από τα άπειρα τραπέζια

του, υπάρχει μια κάρτα με το όνομα ενός άλλου εστιατόριου. Σε κάθε ένα από τα τραπέζια αυτών των εστιατορίων υπάρχουν κάρτες που αναφέρονται σε άλλα εστιατόρια, και αυτή η δομή επαναλαμβάνεται απείρως. Κάθε εστιατόριο αναφέρεται ακριβώς μια φορά. Έτσι τα εστιατόρια της πόλης είναι οργανωμένα σε ένα δέντρο με άπειρα κλαδιά. Σημειώστε ότι κάθε εστιατόριο τοποθετείται σε ένα μόνο επίπεδο αυτού του δέντρου.

Για τον σχηματισμό των μονοπατιών ακολουθούμε το εξής παράδειγμα. Ένας τουρίστας φτάνει στην πόλη. Το πρώτο βράδυ, πηγαίνει στο εστιατόριο-ρίζα και διαλέγει ένα τραπέζι σύμφωνα με την εξίσωση (3.13). Το δεύτερο βράδυ πηγαίνει στο εστιατόριο που ορίστηκε από το τραπέζι της πρώτης νύχτας και διαλέγει ένα άλλο τραπέζι, πάλι από την εξίσωση (3.13). Επαναλαμβάνει την διαδικασία για ***L*** ημέρες. Στο τέλος της εκδρομής, ο τουρίστας έχει επισκεφτεί ***L*** εστιατόρια που συγκροτούν ένα μονοπάτι μέχρι το *L*-οστό επίπεδο του δέντρου. Μετά από τις διακοπές ***L*** ημερών ***M*** τουριστών, μια συλλογή μονοπατιών σχηματίζει ένα υποδένδρο ***L*** επιπέδων.

Αυτή η κατανομή μπορεί να χρησιμοποιηθεί για να μοντελοποιήσει ιεραρχίες θεμάτων. Όπως η κλασσική CRP μπορεί να χρησιμοποιηθεί για να εκφράσει την αβεβαιότητα του αριθμών των στοιχείων σε μοντέλα ανάμειξης, έτσι η φωλιασμένη CRP μπορεί να χρησιμοποιηθεί για να εκφράσει την αβεβαιότητα πιθανών δένδρων *L*-επιπέδων.

Ας θεωρήσουμε ένα σύνολο δεδομένων που αποτελείται από μία συλλογή κειμένων. Κάθε κείμενο είναι μια συλλογή λέξεων, όπου μια λέξη είναι όρος ενός λεξιλογίου. Υποθέτουμε επίσης ότι οι λέξεις σε ένα κείμενο δημιουργούνται σύμφωνα με ένα μοντέλο ανάμειξης όπου οι αναλογίες των θεμάτων είναι τυχαίες ανά κείμενο. Θεωρούμε μια πολυωνυμική μεταβλητή θεμάτων ***z*** και ένα συσχετισμένο σύνολο κατανομών λέξεων ***p(w| z, β),*** όπου ***β*** είναι μια παράμετρος. Αυτά τα θέματα είναι τα βασικά αναμεμειγμένα στοιχεία στο μοντέλο μας. Οι αναλογίες είναι συσχετισμένες με στοιχεία που δηλώνονται από ένα διάνυσμα ***ϑ***. Αν υποθέσουμε (προσωρινά) ότι υπάρχουν ***K*** θέματα σε ένα σώμα, το ***z*** επεκτείνεται σε *K* τιμές και το ***ϑ*** ένα διάνυσμα *K* διαστάσεων. Η κατανομή είναι $p(w|\vartheta) = \sum_{i=1}^{K} \theta_i p(w \mid z = i, \beta_i)$ που είναι μια τυχαία κατανομή αφού το ϑ είναι τυχαίο.

Μια γενική γενετική πιθανοτική διαδικασία δύο επιπέδων για την δημιουργία ενός κειμένου έχει δύο στάδια: (1) Επιλογή ενός διανύσματος *K* διαστάσεων ***ϑ*** από μια κατανομή **p(*ϑ*|*α*)** που θα ελέγχει τις αναλογίες των θεμάτων

τα έγγραφα άνω το $\boldsymbol{\alpha}$ είναι μία παράμετρος κοινή για ολόκληρη τη συλλογή εγγράφων. (2) Επιλογή λέξεων από την κατανομή ανάμειξης $\mathbf{p(w|\vartheta)}$ για την επιλεγμένη τιμή του $\boldsymbol{\vartheta}$. Αν η κατανομή $\boldsymbol{p(\vartheta|\alpha)}$ είναι μια κατανομή Dirichlet τότε η διαδικασία είναι η λανθάνουσα κατανομή Dirichlet.

Το μοντέλο αυτό μπορεί να επεκταθεί ώστε να καλύψει την τοποθέτηση θεμάτων σε μια ιεραρχία. Προς στιγμήν, ας υποθέσουμε ότι έχουμε δεδομένο ένα δέντρο ύψους $\boldsymbol{L}$ και κάθε κόμβος είναι συνδεδεμένος με ένα θέμα. Ένα κείμενο παράγεται ως εξής: (1) επιλογή ενός μονοπατιού από την ρίζα του δέντρου ως ένα φύλλο, (2) επιλογή ενός διανύσματος αναλογιών θεμάτων $\boldsymbol{\vartheta}$ από μια κατανομή Dirichlet $\boldsymbol{L}$ διαστάσεων, (3) δημιουργία λέξεων στο κείμενο από ένα μείγμα $\boldsymbol{L}$ θεμάτων με αναμεμειγμένες αναλογίες $\boldsymbol{\vartheta}$.

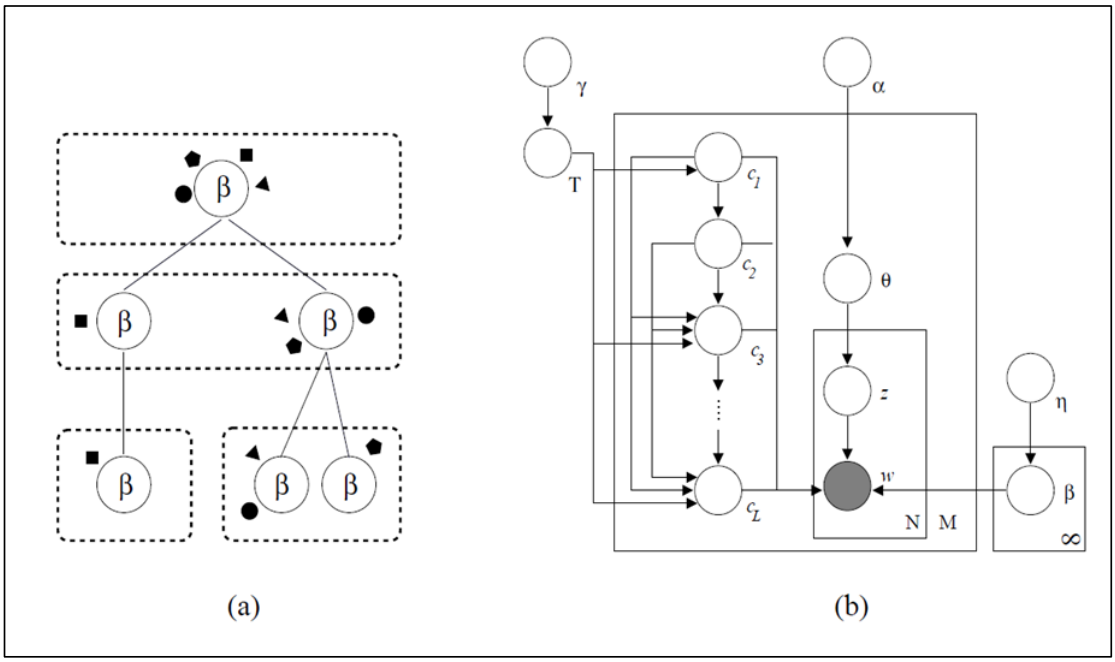
Τέλος, θα χρησιμοποιηθεί η φωλιασμένη μορφή της διαδικασίας Chinese Restaurant (nested CRP) για να υποστηριχθούν δέντρα με μη καθορισμένη δομή. Συσχετίζουμε επίσης μια εκ των προτέρων κατανομή με κάθε θέμα $\boldsymbol{\beta_i}$, καθένα από τα οποία συσχετίζεται με ένα εστιατόριο στο άπειρο δέντρο. Ένα κείμενο δημιουργείται επιλέγοντας πρώτα ένα μονοπάτι $\boldsymbol{L}$ επιπέδων και μετά παίρνοντας λέξεις από $\boldsymbol{L}$ θέματα που συσχετίζονται με τα εστιατόρια από τα οποία περνά το μονοπάτι. Σημειώστε ότι όλα τα κείμενα έχουν λέξεις που παράγονται από το θέμα που συσχετίζεται με το εστιατόριο-ρίζα.

1 Έστω ότι το c_1 είναι το εστιατόριο - ρίζα
2 Για κάθε επίπεδο $l \in \{2,...,L\}$:
(α) Επιλογή ενός τραπεζιού από το εστιατόριο c_{l-1} με βάση την αντίστοιχη εξίσωση της διαδικασίας. Έστω ότι το c_l είναι το εστιατόριο στο οποίο αναφέρεται το συγκεκριμένο τραπέζι
3 Επιλογή του διανύσματος αναλογιών θεμάτων θ με L διαστάσεις από την Dir(α)
4 Για κάθε λέξη $n \in \{1,...,N\}$:
(α) Επιλογή $z \in \{1,...,L\} \sim \text{Mult}(\theta)$
(β) Επιλογή w_n από την κατανομή του θέματος που σχετίζεται με το εστιατόριο c_z

Εικόνα 3.15 Γενετική Διαδικασία Ιεραρχικού Μοντέλου Θεμάτων

Στην Εικόνα 3.16 παρουσιάζεται η φωλιασμένη διαδικασία Chinese Restaurant και το Γραφικό Μοντέλο του Ιεραρχικού Μοντέλου Θεμάτων. Στο αριστερό κομμάτι, (a), εμφανίζονται τα μονοπάτια που έχουν ακολουθήσει τέσσερις τουρίστες (τρίγωνο, τετράγωνο, πεντάγωνο και κύκλος) στο άπειρο δέντρο των εστιατορίων με $\boldsymbol{L=3}$. Οι γραμμές συνδέουν τα εστιατόρια που το ένα αναφέρει το άλλο στα τραπέζια του. Τα μονοπάτια των τουριστών διαγράφουν ένα συγκεκριμένο υποδέντρο, ένα δείγμα που έχει ληφθεί από την φωλιασμένη CRP. Στο δεξί κομμάτι,

(b), απεικονίζεται το γραφικό μοντέλο που προτείνεται στο οποίο καθένα από τα άπειρα $\boldsymbol{\beta}$ αντιστοιχεί σε ένα από τα εστιατόρια. Ο κόμβος $\boldsymbol{T}$ αναφέρεται σε μια συλλογή από έναν άπειρο αριθμό μονοπατιών $\boldsymbol{L}$ επιπέδων και προκύπτει από την φωλιασμένη CRP. Με δεδομένο τον $\boldsymbol{T}$ προκύπτουν οι μεταβλητές $\boldsymbol{c_{m,l}}$. Αν δεν γνωρίζουμε τον $\boldsymbol{T}$, η κατανομή των $\boldsymbol{c_{m,l}}$ ορίζεται από την φωλιασμένη CRP, με βάση τα $\boldsymbol{c_{q,l}}$ για $q < m$.



Εικόνα 3.16 Ιεραρχικό Μοντέλο Θεμάτων

Αν υποθέσουμε ότι μας έχει δοθεί ένα σύνολο M κειμένων, $\mathbf{w_1,\ldots,w_M}$, η εκ των υστέρων κατανομή των $\boldsymbol{c}$ αντιστοιχεί ουσιαστικά, σε μια εκ των υστέρων κατανομή των πρώτων **M** μονοπατιών στο $\boldsymbol{T}$. Αν θεωρήσουμε ένα νέο κείμενο $\mathbf{w_{M+1}}$, η εκ των υστέρων κατανομή του μονοπατιού που το έχει παράγει θα εξαρτάται από το άγνωστο $\boldsymbol{T}$. Ακολούθως τα νέα κείμενα θα εξαρτώνται επίσης από το αρχικό σύνολο δεδομένων και από όλα τα άλλα καινούρια κειμένα που είχαν παρατηρηθεί πριν από αυτά. Μέσω της φωλιασμένης διαδικασίας Chinese Restaurant, κάθε νέο κείμενο μπορεί να επιλέξει ένα προηγούμενο εστιατόριο, που ήταν μέχρι στιγμής μη επισκέψιμο, σε οποιοδήποτε από τα επίπεδα του δέντρου.

Οι συγγραφείς περιγράφουν πως μπορεί να εξαχθεί το μοντέλο με βάση μια συλλογή παρατηρούμενων κειμένων. Περιγράφουν με ακρίβεια την χρήση της δειγματοληψίας Gibbs για την δειγματοληψία της εκ των υστέρων κατανομής της φωλιασμένης διαδικασίας Chinese Restaurant και των σχετικών κατανομών θεμάτων [64].

Αξιολόγηση

Κατά την αξιολόγηση του μοντέλου αποδεικνύεται η δυνατότητα του να εξάγει ιεραρχίες χρησιμοποιώντας συνθετικά δεδομένα από ένα σχετικά περιορισμένο λεξιλόγιο. Ο αλγόριθμος εφαρμόστηκε σε ένα σύνολο 100 εγγράφων που περιέχουν 1.000 λέξεις. Τα έγγραφα έχουν δημιουργηθεί από μια ιεραρχία τριών θεμάτων. Η ιεραρχία ανακτήθηκε με σημαντική επιτυχία από την προτεινόμενη προσέγγιση. Επίσης, η προτεινόμενη προσέγγιση συγκρίθηκε με ανταγωνιστικές προσεγγίσεις επιδεικνύοντας μεγαλύτερη αποτελεσματικότητα υπό προϋποθέσεις.

Τέλος η μέθοδος αξιολογήθηκε σε πραγματικά δεδομένα και εκτίμησε μια ιεραρχία τριών επιπέδων. Τα αποτελέσματα της αξιολόγησης είναι ενθαρρυντικά δείχνοντας ότι πρόκειται για ένα ευέλικτο γενικό μοντέλο ιεραρχιών θεμάτων που μπορεί να ανταποκριθεί σε εξελισσόμενες συλλογές δεδομένων.

3.4 Συμπεράσματα

Στις προηγούμενες παραγράφους παρουσιάστηκαν επιλεκτικά κάποιες μέθοδοι ανάκτησης πληροφορίας αλλά και η ιστορική εξέλιξη που οδήγησε στα πιθανοτικά μοντέλα θεμάτων. Επίσης παρουσιάσαμε τον ορισμό των μοντέλων θεμάτων και της πιο διαδεδομένης παραλλαγής τους, της λανθάνουσας κατανομής Dirichlet. Τέλος είδαμε κάποιες από τις βασικές παραλλαγές του αλγορίθμου.

Τα πιθανοτικά μοντέλα θεμάτων αποτελούν ένα σημαντικό τεχνολογικό εργαλείο στο οποίο μπορούν να βασιστούν διάφορες μεθοδολογίες διαχείρισης, ανάκτησης και διήθησης πληροφορίας. Η διατριβή εστιάζει στην υλοποίηση συστημάτων προτάσεων με βάση τα πιθανοτικά μοντέλα θεμάτων. Στο επόμενο κεφάλαιο καταγράφεται η πρόταση της διδακτορικής διατριβής, τα ερευνητικά ερωτήματα και ο τρόπος που τα αντιμετωπίζουμε.

4 Η Πρόταση της Διατριβής

Στα προηγούμενα κεφάλαια είδαμε την δημιουργία, την εξέλιξη και την εξάπλωση των συστημάτων προτάσεων. Επίσης, παρουσιάστηκε η τεχνολογία των πιθανοτικών μοντέλων θεμάτων και ο τρόπος που σχετίζεται με τις υπάρχουσες τεχνολογίες για την ανάλυση περιεχομένου.

Στο παρόν κεφάλαιο, καταγράφεται με βάση το συγκεκριμένο πλαίσιο το πρόβλημα που αντιμετωπίζεται στη διατριβή και οι άξονες που το περιγράφουν. Ακολουθεί η διατύπωση των ερευνητικών ερωτημάτων, η συνεισφορά της διατριβής και μια συνοπτική αναφορά στο υπόλοιπο της διατριβής.

4.1 Διαμόρφωση του Προβλήματος

Στα κεφάλαια 2 και 3 περιλαμβάνεται μια συνοπτική παρουσίαση της βιβλιογραφίας που συνδέεται με τα συστήματα προτάσεων και με τις τεχνικές που υλοποιούν πιθανοτικά μοντέλα θεμάτων. Σε αυτή την ενότητα περιγράφουμε ερευνητικές εργασίες που συσχετίζουν την εφαρμογή πιθανοτικών μοντέλων θεμάτων και συναφών τεχνολογιών με την σχεδίαση συστημάτων προτάσεων. Επίσης γίνεται μια απόπειρα να διαμορφωθεί το ερευνητικό πρόβλημα της διατριβής σε σχέση με την έρευνα που έχει ήδη πραγματοποιηθεί στην περιοχή.

Τα συστήματα προτάσεων όπως περιγράφονται στη βιβλιογραφία αποτελούν ώριμες τεχνολογικές λύσεις οι οποίες αντιμετωπίζουν σε πολλές περιπτώσεις με επιτυχία το πρόβλημα της υπερφόρτωσης πληροφορίας. Εντούτοις, ένας αριθμός ελλείψεων και προβλημάτων παραμένουν να λυθούν από τους ερευνητές και τους επαγγελματίες του χώρου. Η χρήση των πιθανοτικών μοντέλων θεμάτων μπορεί να βοηθήσει στην αντιμετώπιση κάποιων από τα ζητήματα που αφορούν τα συστήματα προτάσεων, καθώς αποτελούν μια αξιόπιστη προσέγγιση για την εξαγωγή λανθάνουσας σημασιολογίας από περιεχόμενο.

Στη συνέχεια περιγράφουμε δύο στοιχεία ταξινόμησης των συστημάτων προτάσεων, τους τομείς εφαρμογής και τον τύπο των χρησιμοποιούμενων δεδομένων.

4.1.1 Τομείς Εφαρμογής

Τα συστήματα προτάσεων που βασίζονται σε πιθανοτικά μοντέλα θεμάτων καλύπτουν ένα εύρος εφαρμογών.

Η διαχείριση γνώσης στο εργασιακό περιβάλλον αποτελεί έναν γενικό τομέα εφαρμογής των συστημάτων προτάσεων που μπορεί να πάρει διαφορετικές μορφές, όπως στο εταιρικό κοινωνικό λογισμικό ή σε μια ομάδα που συνεργάζεται για την ανάπτυξη λογισμικού. Με την εμφάνιση του εταιρικού κοινωνικού λογισμικού, οι προκλήσεις που αφορούν τα συστήματα προτάσεων γίνονται ακόμη πιο έντονες λόγω της αύξησης του διαθέσιμου περιεχομένου στον οργανισμό [65]. Αντίστοιχα, οι μεγάλες ομάδες προγραμματιστών που συνεργάζονται για την ανάπτυξη λογισμικού συχνά χρειάζονται υποστήριξη για την λήψη αποφάσεων από αντίστοιχα συστήματα [66].

Ένας διαφορετικός τομέας εφαρμογής είναι ο κοινωνικός ιστός και τα κοινωνικά δίκτυα. Στον κοινωνικό ιστό, που αποτελείται από κοινωνικούς δεσμούς μεταξύ ανθρώπων στα πλαίσια του παγκοσμίου ιστού [20], τα συστήματα προτάσεων επιχειρούν να συνδέσουν αντικείμενα με χρήστες. Κάποιες προσεγγίσεις βασίζονται στη χρήση πιθανοτικών μοντέλων θεμάτων ([67], [68]).

Τέλος έχουν σχεδιαστεί εφαρμογές σε διαφορετικούς τομείς όπως ειδησεογραφικοί ιστότοποι, πληροφορίες τοποθεσίας και λίστες αναπαραγωγής μουσικής (π.χ. [69], [70], [71]).

4.1.2 Τύπος Δεδομένων

Η τεχνολογία των μοντέλων θεμάτων έχει τη δυνατότητα να χρησιμοποιηθεί για την ανάλυση δεδομένων. Οι συγκεκριμένες εφαρμογές ταξινομούνται σε τρεις κατηγορίες: εκείνες που αναλύουν τις περιγραφές των αντικειμένων, εκείνες που αναλύουν την συμπεριφορά των χρηστών και εκείνες που χρησιμοποιούν έναν συνδυασμό των δυο.

4.1.2.1 Περιεχόμενο

Αρκετές εφαρμογές συστημάτων αποφάσεων με πιθανοτικά μοντέλα θεμάτων αναλύουν το περιεχόμενο που έχει παραχθεί για την πραγματοποίηση προβλέψεων. Οι συγκεκριμένες εφαρμογές περιλαμβάνουν τον κοινωνικό ιστό

(ιστολόγια, άρθρα Wikipedia, επισημειώσεις εγγράφων και μουσικής) αλλά και το εργασιακό περιβάλλον (εταιρικά έγγραφα).

Τα λανθάνοντα θέματα έχουν χρησιμοποιηθεί για την εξόρυξη πληροφοριών από ιστολόγια και για την πρόταση σχετικών επισημειώσεων για συγκεκριμένες δημοσιεύσεις [67]. Σε αντίστοιχη μελέτη έχουν επιλεγεί 4.625 άρθρα από την Wikipedia (Wikipedia Selection for Schools) στα οποία εφαρμόζεται η λανθάνουσα κατανομή Dirichlet ώστε να εξαχθεί ένα μοντέλο θεμάτων [72]. Στη συνέχεια, οι κατανομές που εξάγονται από το συγκεκριμένο μοντέλο χρησιμοποιούνται για την δημιουργία προτάσεων με βάση το τρέχον άρθρο. Τα αποτελέσματα δείχνουν την αυξημένη συνάφεια των άρθρων που παράγονται από το σύστημα προτάσεων.

Στη βιβλιογραφία έχουν προταθεί μεθοδολογίες για την πρόταση επισημειώσεων (tags) στους χρήστες [73],[74]. Σε μια μεθοδολογιά τέτοιου τύπου έχει γίνει μια αξιολόγηση εκτός σύνδεσης στο σύνολο δεδομένων του Bibsonomy[15] καθώς και μια online αξιολόγηση με χρήστες που αποδεικνύει την αποτελεσματικότητα της μεθόδου [74]. Επίσης έχει παρατηρηθεί ότι η χρήση πιθανοτικών μοντέλων θεμάτων παρουσιάζει αυξημένη ακρίβεια σε σχέση με την εξαγωγή κανόνων συσχέτισης [73].

Στον τομέα της μουσικής, με βάση τις τρέχουσες προσεγγίσεις για την πρόταση λιστών αναπαραγωγής έχει παρουσιαστεί μια μετρική σύγκρισης που χρησιμοποιεί τις επισημειώσεις και τα μοντέλα θεμάτων για την σύγκριση λιστών αναπαραγωγής [71].

Τέλος, στο εταιρικό περιβάλλον έχει προταθεί η χρήση των λανθανόντων θεμάτων ώστε να εντοπιστούν και να οπτικοποιηθούν λανθάνουσες κοινότητες με παρεμφερή ενδιαφέροντα [75].

4.1.2.2 Συμπεριφορά Χρηστών

Κάποιες εφαρμογές συστημάτων αποφάσεων με πιθανοτικά μοντέλα θεμάτων αναλύουν την συμπεριφορά των χρηστών. Για την εκμετάλλευση του συγκεκριμένου τύπου δεδομένων έχουν προταθεί αντίστοιχα πλαίσια ενώ οι συγκεκριμένες εφαρμογές περιλαμβάνουν την ανάλυση αξιολογήσεων αντικειμένων και ταινιών, καθώς επίσης και προτάσεις για συμμετοχή σε κοινότητες και δρομολόγια ταξιδιών.

[15] http://www.bibsonomy.org/

Ο Hofmann περιγράφει ένα γενικότερο πλαίσιο συνεργατικής διήθησης βασισμένο σε μοντέλα θεμάτων [76]. Οι αλγόριθμοι του πλαισίου χρησιμοποιούν στατιστικά μοντέλα και εισάγουν μεταβλητές λανθανουσών κλάσεων σε ένα μοντέλο ανάμειξης ώστε να ανακαλύψουν κοινότητες χρηστών και πρότυπα προφίλ ενδιαφερόντων.

Μέχρι τώρα, το κυρίαρχο υπόδειγμα για την πραγματοποίηση συνεργατικής διήθησης ήταν ο αλγόριθμος των πλησιέστερων γειτόνων και οι διάφορες παραλλαγές του. Οι συνηθέστερες παραλλαγές παρουσιάζουν κατά τον συγγραφέα τέσσερις βασικές ελλείψεις: (α) Η ακρίβεια των προβλέψεων πολλές φορές δεν είναι μεγάλη. (β) Δεν κατασκευάζεται κάποιο στατιστικό μοντέλο, και συνεπώς το σύστημα δεν μαθαίνει κάτι από τα υπάρχοντα δεδομένα και δεν παρουσιάζεται κάποια εποπτική εικόνα της δραστηριότητας των χρηστών. (γ) Οι μέθοδοι που βασίζονται στην μνήμη δεν είναι επεκτάσιμες χωρίς κόστος όσο αφορά τις απαιτήσεις τους σε μνήμη και υπολογιστικό χρόνο, και (δ) είναι δύσκολο να προσαρμοστούν ώστε να μεγιστοποιήσουν κάποια συνάρτηση ωφέλειας η οποία σχετίζεται με τη συγκεκριμένη δραστηριότητα. Στη συγκεκριμένη μελέτη ο συγγραφέας προτείνει μια γενικότερη προσέγγιση βασισμένη σε στατιστικά μοντέλα η οποία: (α) επιτυγχάνει μεγαλύτερη ακρίβεια στις προβλέψεις, (β) συμπιέζει τα δεδομένα σε ένα συμπαγές στατιστικό μοντέλο που εντοπίζει κοινότητες χρηστών, (γ) επιτρέπει την πραγματοποίηση προβλέψεων σε σταθερό χρόνο και (δ) δίνει στον διαχειριστή ή τον σχεδιαστή του συστήματος μεγαλύτερη ευελιξία στην περιγραφή του στόχου της εφαρμογής.

Το πρόβλημα της συνεργατικής διήθησης έχει διατυπωθεί εκ νέου σε ένα γενετικό (generative) πιθανοτικό πλαίσιο [77]. Στο συγκεκριμένο πλαίσιο κάθε αξιολόγηση αντικειμένου από χρήστη χρησιμοποιείται ως ένδειξη για αξιολογήσεις που δεν γνωρίζουμε ακόμη. Η τελική βαθμολογία εκτιμάται συνδυάζοντας προβλέψεις από τρεις πηγές: προβλέψεις με βάση τις αξιολογήσεις του ίδιου αντικειμένου από διαφορετικούς χρήστες, προβλέψεις με βάση τις αξιολογήσεις διαφορετικών αντικειμένων από τον ίδιο χρήστη και τέλος προβλέψεις με βάση τις αξιολογήσεις που έχουν κάνει διαφορετικοί αλλά συναφείς χρήστες σε διαφορετικά αλλά συναφή αντικείμενα. Λόγω της άντλησης δεδομένων από τρεις πηγές, το μοντέλο είναι πιο σταθερό σε περιπτώσεις αραιής πληροφορίας. Η πειραματική αξιολόγηση του μοντέλου καταδεικνύει την σταθερότητα και την βελτιωμένη απόδοση του σε σχέση με ανταγωνιστικές μεθόδους.

Ακόμη, έχει περιγραφεί μια εφαρμογή που σχετίζεται με την μελέτη των επιλογών των χρηστών που αφορούν τις ταινίες [78]. Πιο συγκεκριμένα παρουσιάζεται ένα πιθανοτικό μοντέλο που περιγράφει δυο πράξεις που αφορούν την διαδικασία της επιλογής ταινιών: (1) την πράξη με την οποία οι χρήστες επιλέγουν ποια αντικείμενα θα αξιολογήσουν και (2) την πράξη με την οποία αξιολογούν τα συγκεκριμένα αντικείμενα. Χρησιμοποιούν σύνολα δεδομένων τα οποία έχουν εξαχθεί από τη δημοφιλή υπηρεσία Netflix[16] ώστε να αποδείξουν ότι το μοντέλο αυτό μπορεί να παρέχει ακριβείς προβλέψεις για τις αξιολογήσεις ταινιών.

Επίσης, έχει εξεταστεί η ικανότητα της λανθάνουσας κατανομής Dirichlet να παράγει προτάσεις για τη συμμετοχή ανθρώπων σε κοινότητες και η αποτελεσματικότητά της έχει συγκριθεί με τη εξαγωγή κανόνων συσχέτισης [68]. Συγκρίνοντας τις δυο προσεγγίσεις, γίνεται μια πειραματική αξιολόγηση με βάση το σύνολο δεδομένων του ιστοτόπου κοινωνικής δικτύωσης Orkut[17]. Η εμπειρική αξιολόγηση με χρήση των k-κορυφαίων προβλέψεων δείχνει ότι η λανθάνουσα κατανομή Dirichlet είναι πιο αποτελεσματική όταν προτείνουμε μια σχετικά μεγάλη λίστα με ομάδες.

Σε μια παρόμοια μελέτη [70] προτείνεται ένα σύστημα προτάσεων για πιθανά δρομολόγια ταξιδιών. Βάση για την παραγωγή τέτοιων προτάσεων έχει αποτελέσει το ιστορικό δημοσίευσης φωτογραφιών από τουρίστες στο δημοφιλή ιστότοπο Flickr [18]. Οι προτάσεις παράγονται από ένα πιθανοτικό μοντέλο συμπεριφοράς που συνυπολογίζει την πιθανότητα ένας φωτογράφος να επισκεφτεί ένα τοπόσημο (έναν σημαντικό και αναγνωρίσιμο τόπο σε μια περιοχή). Η προτεινόμενη προσέγγιση χρησιμοποιεί πιθανοτικά μοντέλα θεμάτων και μοντέλα Markov για να ενσωματώσει τόσο τη συμπεριφορά των χρηστών όσο και τη γεωγραφική πληροφορία. Η αποτελεσματικότητα της προτεινόμενης προσέγγισης αξιολογείται σε ένα σύνολο πραγματικών δεδομένων.

4.1.2.3 Συνδυασμός Περιεχομένου και Συμπεριφοράς Χρηστών

Ένας σημαντικός αριθμός από εφαρμογές συστημάτων αποφάσεων με πιθανοτικά μοντέλα θεμάτων συνδυάζει την συμπεριφορά των χρηστών με το περιεχόμενο των αντικειμένων. Αντίστοιχες εφαρμογές περιλαμβάνουν τη χρήση περιγραφών και αξιολογήσεων αντικειμένων, την ανάλυση ειδησεογραφικών

[16] http://www.netflix.com
[17] http://www.orkut.com
[18] http://www.flickr.com

ιστοτόπων, την ανάλυση επιστημονικών και εταιρικών άρθρων, και τέλος την ανάλυση πηγαίου κώδικα και της επίλυσης προβλημάτων λογισμικού.

Οι Popescul, Unger, Pennock και Lawrence [79] επεκτείνουν την προηγούμενη εργασία του Hofmann [76] και παρουσιάζουν ένα πιθανοτικό πλαίσιο για ένα ενοποιημένο σύστημα προτάσεων με βάση το περιεχόμενο και τις προτιμήσεις.

Περιγράφουν ένα μοντέλο όψεων (aspect model) τριών δρόμων που χρησιμοποιεί χρήστες, αντικείμενα και περιεχόμενο αντικειμένων. Στη συνεισφορά της μελέτης περιλαμβάνεται η αυτόματη εξαγωγή του σχετικού βάρους μεταξύ των συνεργατικών προτάσεων και των προτάσεων περιεχομένου το οποίο μπορεί να προκύπτει από τα δεδομένα και δε χρειάζεται να ρυθμίζεται παραμετρικά. Το μοντέλο που παρουσιάζεται υποθέτει ότι οι χρήστες ενδιαφέρονται για λανθάνοντα θέματα τα οποία με τη σειρά τους παράγουν τόσο αντικείμενα όσο και περιεχόμενο αντικειμένων. Για την εκπαίδευση του μοντέλου χρησιμοποιείται η τεχνική της μεγιστοποίησης αναμονής (expectation maximization).

Στη βιβλιογραφία έχει επίσης προταθεί η μέθοδος fLDA [80], μια μέθοδος παραγοντοποίησης πινάκων η οποία μπορεί να χρησιμοποιηθεί για την πρόβλεψη βαθμολογιών σε ένα σύστημα προτάσεων. Η μέθοδος αυτή προορίζεται για τις περιπτώσεις εφαρμογής όπου είναι κατάλληλη μια προσέγγιση συνόλου λέξεων (bag-of-words) για την περιγραφή των αντικειμένων. Η προτεινόμενη μέθοδος κανονικοποιεί ταυτόχρονα τους παράγοντες που αφορούν τους χρήστες και τα αντικείμενα, χρησιμοποιώντας τις περιγραφές των χρηστών και των αντικειμένων αντίστοιχα. Πιο συγκεκριμένα, κάθε λέξη στην περιγραφή ενός αντικειμένου συσχετίζεται με έναν λανθάνοντα παράγοντα (το θέμα της λέξης). Τα θέματα ενός αντικειμένου εξάγονται λαμβάνοντας την μέση τιμή των θεμάτων όλων των λέξεων που το περιγράφουν. Στη συνέχεια, η βαθμολογία ενός χρήστη για ένα συγκεκριμένο αντικείμενο χρησιμοποιείται ως η συνάφεια του χρήστη με το αντικείμενο και τα θέματα τα οποία το συγκεκριμένο αντικείμενο αφορά. Στη συνέχεια πραγματοποιείται επιβλεπόμενη μάθηση της συνάφειας του χρήστη με τα θέματα (παράγοντες χρήστη) και της συνάφειας του αντικειμένου με τα θέματα (παράγοντες αντικειμένου). Για την αποφυγή υπερπροσαρμογής, οι παράγοντες των χρηστών και των αντικειμένων κανονικοποιούνται με βάση Γκαουσιανή γραμμική αναδρομή και λανθάνουσα κατανομή Dirichlet (LDA) αντίστοιχα. Το συγκεκριμένο μοντέλο αποδεικνύεται ότι είναι ακριβές, ερμηνεύσιμο και μπορεί να χειριστεί σενάρια που παρουσιάζεται ψυχρή έναρξη (cold start).

Επίσης, έχει διερευνηθεί η δυνατότητα των συστημάτων προτάσεων να ασχοληθούν με εξαιρετικά μεγάλα σύνολα δεδομένων, της τάξεως των εκατομμυρίων χρηστών και αντικειμένων, που παρουσιάζουν δυναμική συμπεριφορά (τα αντικείμενα αλλάζουν συνεχώς) [69]. Ειδικότερα, η συγκεκριμένη μελέτη ασχολείται με τους ειδησεογραφικούς ιστοτόπους και τη δυνατότητα των συστημάτων προτάσεων να παρέχουν προτάσεις σε αυτούς. Πιο συγκεκριμένα, διερευνούν την δυνατότητα τους να παράγουν προτάσεις για το Google News[19] χρησιμοποιώντας για το σκοπό αυτό τρεις πηγές: συνεργατική διήθηση με ομαδοποίηση, πιθανοτική λανθάνουσα σημασιολογική ανάλυση και μετρικές αθροισμάτων επισκέψεων. Ένα γραμμικό μοντέλο χρησιμοποιείται για τον συνδυασμό των παραπάνω αλγορίθμων. Η προτεινόμενη προσέγγιση παρουσιάζει πλεονεκτήματα όπως ανεξαρτησία από περιεχόμενο, ανεξαρτησία από τον τομέα εφαρμογής και προσαρμοστικότητα σε διαφορετικές εφαρμογές.

Μια αντίστοιχη προσέγγιση έχει προταθεί για την πραγματοποίηση προτάσεων επιστημονικών άρθρων [81]. Οι ερευνητές πολύ συχνά εγγράφονται σε online κοινότητες όπου μοιράζονται παραπομπές σε επιστημονικά άρθρα. Η εργασία αυτή εκμεταλλεύεται αυτή τη δραστηριότητα για να προτείνει επιστημονικά άρθρα σε χρήστες κοινοτήτων. Συγκεκριμένα συνδυάζει την συνεργατική διήθηση με τα πιθανοτικά μοντέλα θεμάτων και παρέχει μια ερμηνεύσιμη δομή λανθανόντων θεμάτων για χρήστες και αντικείμενα. Η συγκεκριμένη προσέγγιση ξεπερνά σε αποτελεσματικότητα την απλή συνεργατική διήθηση σε μεγάλα σύνολα δεδομένων.

Στη βιβλιογραφία έχει περιγραφεί η χρήση λανθανόντων θεμάτων για να αποτυπώσουν τα ενδιαφέροντα των εργαζομένων αλλά και για να προτείνονται σχετικά έγγραφα [65]. Το προτεινόμενο σύστημα ανιχνεύει τις συμπεριφορές των χρηστών στο εταιρικό περιβάλλον και τα θέματα που τους ενδιαφέρουν. Τα θέματα στη συνέχεια χρησιμοποιούνται για την ανάκτηση συναφών εγγράφων τα οποία σχετίζονται με τα θέματα για τα οποία ενδιαφέρεται ο χρήστης. Η πειραματική αξιολόγηση δείχνει την δυνατότητα της μεθόδου να ανακτήσει συναφή άρθρα με μεγάλη ακρίβεια.

Τέλος, έχει περιγραφεί μια προσέγγιση με το όνομα DRETOM (Developer REcommendation based on TOpic Models) που προτείνει προγραμματιστές για να επιλύσουν προβλήματα λογισμικού σε μια κοινότητα [66]. Η συγκεκριμένη

[19] http://news.google.com

προσέγγιση μοντελοποιεί τα ενδιαφέροντα και την εμπειρία των προγραμματιστών στην επίλυση προβλημάτων με βάση το ιστορικό τους. Για κάθε πρόβλημα που ανακύπτει, το σύστημα προτείνει μια ταξινομημένη λίστα προγραμματιστών οι οποίοι μπορούν να συμμετέχουν και να συνεισφέρουν στην επίλυση του συγκεκριμένου προβλήματος. Η λίστα αυτή συντίθεται με βάση τα ενδιαφέροντα και την εμπειρία των προγραμματιστών στην επίλυση παρομοίων προβλημάτων. Η πειραματική αξιολόγηση της προσέγγισης στις κοινότητες του Eclipse και του Mozilla Firefox δείχνουν υψηλό ποσοστό ανάκλησης.

Μια εποπτική εικόνα της σχετικής βιβλιογραφίας παρέχεται στον αντίστοιχο πίνακα (Πίνακας 4.1).

Πίνακας 4.1 Εφαρμογές με Βάση τον Τύπο Δεδομένων

Τύπος Δεδομένων	**Παραδείγματα Εφαρμογής Από Τη Βιβλιογραφία**
Περιεχόμενο	Ιστολόγια [67], Άρθρα Wikipedia [72],Επισημειώσεις [73],[74], Επισημειώσεις Μουσικής [71], Εταιρικά Έγγραφα [75]
Συμπεριφορά Χρηστών	Γενικότερο Πλαίσιο Συνεργατικής Διήθησης [76], Αξιολογήσεις Αντικειμένων [77], Επιλογή και Αξιολόγηση Ταινιών [78], Συμμετοχή σε Κοινότητες [68], Δρομολόγια Ταξιδιών [70]
Συνδυασμός Αναλύσεων	Γενικότερο Πλαίσιο [79], Περιγραφές και Αξιολογήσεις Αντικειμένων [80], Άρθρα Ειδήσεων και Επισκέψεις [69], Επιστημονικά Άρθρα και Επισκέψεις [81], Εταιρικά Έγγραφα και Δραστηριότητες Χρηστών [65], Πηγαίος Κώδικας και Επίλυση Προβλημάτων [66]

4.1.3 Άξονες της Έρευνας

Συνολικά η συνεισφορά των πιθανοτικών μοντέλων θεμάτων στα συστήματα προτάσεων έχει υπάρξει σημαντική. Τα πιθανοτικά μοντέλα θεμάτων μπορούν να παρέχουν σημαντική ακρίβεια στην παραγωγή προτάσεων η οποία συχνά είναι καλύτερη από ανταγωνιστικές μεθόδους (συνεργατική διήθηση και κανόνες συσχέτισης). Οι προτάσεις αυτές μπορούν να παραχθούν από μια συμπιεσμένη μορφή των δεδομένων (που είναι το εξαγόμενο μοντέλο θεμάτων) σε σταθερό

χρόνο. Επίσης τα πιθανοτικά μοντέλα θεμάτων μπορούν να χρησιμοποιηθούν για το συνδυασμό πληροφοριών από πολλαπλές πηγές με διαφορετικούς τρόπους δημιουργώντας ένα αξιόπιστο σύστημα που μπορεί να αντιμετωπίσει σενάρια ψυχρής έναρξης. Τέλος, τα πιθανοτικά μοντέλα θεμάτων μπορούν να χρησιμοποιηθούν συμπληρωματικά από τον σχεδιαστή ή τον διαχειριστή του συστήματος προτάσεων για την εξαγωγή πρακτικών συμπερασμάτων και την απόκτηση μιας εποπτικής εικόνας της συμπεριφοράς των χρηστών.

Παρά τη συνεισφορά των πιθανοτικών μοντέλων θεμάτων στο τεχνολογικό υπόβαθρο των συστημάτων προτάσεων, παραμένουν προκλήσεις που αφορούν την σχεδίαση και τη λειτουργία τους.

Η συγκεκριμένη διατριβή τοποθετείται σε δύο άξονες. Ο πρώτος άξονας αφορά στην δυνατότητα εφαρμογής πιθανοτικών μοντέλων θεμάτων σε δυο τύπους δεδομένων: στο περιεχόμενο των συνεισφορών των χρηστών και στη συμπεριφορά των χρηστών. Ο δεύτερος άξονας αφορά στην γενικότητα της προτεινόμενης προσέγγισης και την δυνατότητα εφαρμογής αντίστοιχων μεθοδολογιών σε σημαντικό εύρος τομέων.

Πρώτος Άξονας: Τύπος Δεδομένων

Η μια επιλογή για την σχεδίαση και υλοποίηση ενός συστήματος προτάσεων περιλαμβάνει την εκμετάλλευση του περιεχομένου για την παραγωγή προτάσεων. Το περιεχόμενο αυτό μπορεί να είναι τμήμα των αντικειμένων ή των περιγραφών των αντικειμένων ενώ έχει τη μορφή κειμένου ή επισημειώσεων. Αυτό το περιεχόμενο μπορεί να αναλυθεί και να συνδεθεί με συγκεκριμένους χρήστες, δημιουργώντας σχετικά προφίλ χρηστών. Συγκρίνοντας το περιεχόμενο των προφίλ των χρηστών και των αντικειμένων πραγματοποιούνται προτάσεις νέων αντικειμένων.

Εναλλακτικά ως βάση για τη παραγωγή προτάσεων μπορεί να χρησιμοποιηθεί η συμπεριφορά των χρηστών του συστήματος. Ανάλογα με το τομέα εφαρμογής που μελετάται κάθε φορά, η συμπεριφορά μπορεί να περιλαμβάνει διαφορετικές δραστηριότητες. Σε ένα σύστημα πρότασης ενημερωτικών άρθρων η δραστηριότητα μπορεί να αφορά την επίσκεψη σε σελίδες με άλλα άρθρα, ενώ για μια ηλεκτρονική αγορά προϊόντων περιλαμβάνει τις προηγούμενες αγορές του κάθε καταναλωτή. Αυτές οι δραστηριότητες αθροίζονται

για όλους τους χρήστες του συστήματος, αναλύονται και χρησιμοποιούνται για την παραγωγή προτάσεων.

Δεύτερος Άξονας: Τομέας Εφαρμογής

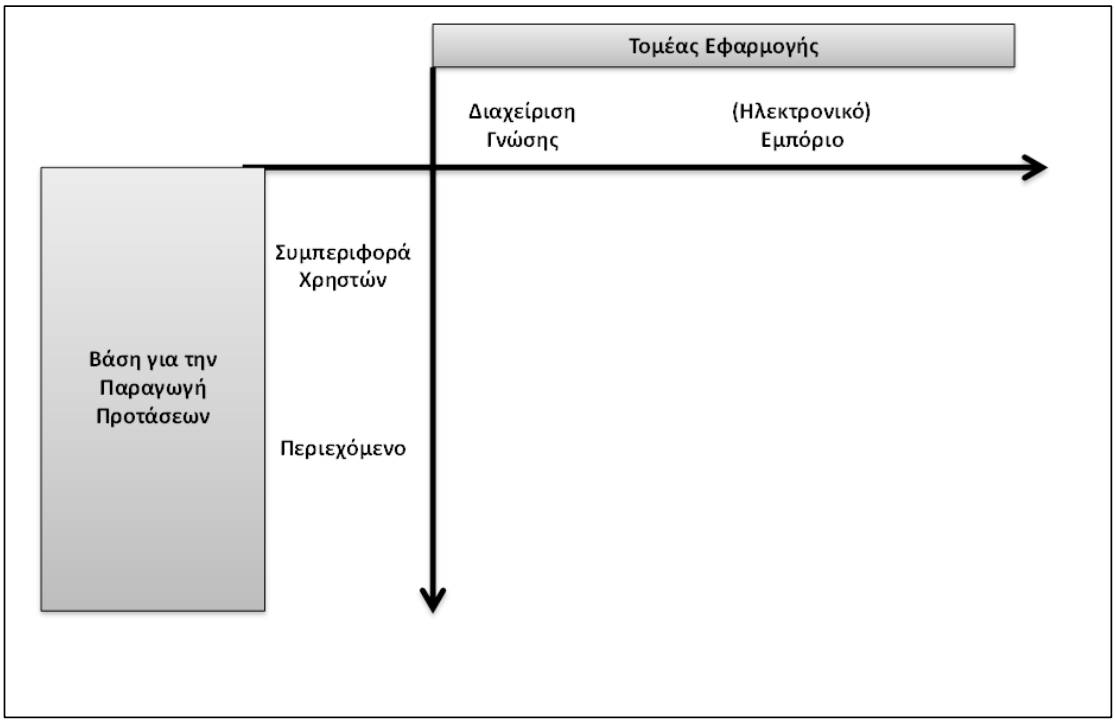
Τα συστήματα που περιγράφονται μπορούν να εφαρμοστούν σε διαφορετικούς τομείς. Στην παρούσα διατριβή εξετάζονται δυο κατηγορίες δραστηριοτήτων που θεωρούμε ότι έχουν αντιπροσωπευτικό χαρακτήρα για το πρόβλημα της υπερφόρτωσης πληροφορίας.

Πρώτον εξετάζουμε την διαχείριση γνώσης όπως αυτή πραγματοποιείται στο εσωτερικό επιχειρήσεων και κοινοτήτων.

Στις επιχειρήσεις έντασης γνώσης η δημιουργία, η επεξεργασία, ο διαμοιρασμός και η επαναχρησιμοποίηση των πληροφοριών έχει κεντρική σημασία. Τελευταία, και με τις αλλαγές που επιτελούνται στο επίπεδο του παγκόσμιου ιστού, ένας νέος τύπος επιχειρήσεων έχει αρχίσει να αναδεικνύεται – επιχειρήσεις που χρησιμοποιούν τεχνολογίες του κοινωνικού ιστού ώστε να συντονίσουν τις προσπάθειες των εργαζομένων εσωτερικά αλλά και να επεκτείνουν την δραστηριότητα της επιχείρησης σε πελάτες, συνεργάτες και προμηθευτές [82]. Στην περίπτωση αυτή, τα συστήματα προτάσεων υποστηρίζουν την χρήση κοινωνικών τεχνολογιών και την προσπάθεια των εταιριών να βελτιώσουν τις επιχειρησιακές δραστηριότητες τους και να εκμεταλλευτούν νέες ευκαιρίες στην αγορά. Ακόμη, οι κοινότητες ανάπτυξης λογισμικού αποτελούν χώρους όπου η διαχείριση της πληροφορίας αλλά και η αξιοποίηση των δεξιοτήτων των μελών έχει μεγάλη αξία για τους συμμετέχοντες.

Δεύτερον εξετάζουμε την πραγματοποίηση αγορών και πωλήσεων εντός και εκτός του παγκοσμίου ιστού. Οι συμμετέχοντες σε μια αγορά, αγοραστές και πωλητές, κινητοποιούνται από την επιθυμία τους να μεγιστοποιήσουν την ιδιωτική τους ωφέλεια ενώ συνολικά οδηγούν σε μια βέλτιστη κατανομή των μέσων παραγωγής [83]. Αν και οι ηλεκτρονικές αγορές προσφέρουν ένα περιβάλλον χαμηλής τριβής για την πραγματοποίηση αγορών, παρουσιάζουν και αυτές προκλήσεις στην λειτουργία τους. Κάποιες από τις προκλήσεις αντιμετωπίζονται από συστήματα προτάσεων τα οποία έχουν τη δυνατότητα να υποστηρίξουν τους αγοραστές στην επιλογή των προϊόντων που θα αγοράσουν, και τους πωλητές στην τιμολόγηση και την περιγραφή των αντικειμένων που πωλούν.

Στην εικόνα παρουσιάζεται μια γενική επισκόπηση των αξόνων στους οποίους τοποθετείται η παρούσα διατριβή (Εικόνα 4.1). Ακολούθως θα δούμε πως τοποθετούνται τα ερευνητικά ερωτήματα της διατριβής σε σχέση με τους άξονες. Επίσης θα παρουσιάσουμε την σχέση τους με τη συνεισφορά της διατριβής, και τις επιμέρους προσεγγίσεις που προτείνονται.



Εικόνα 4.1 Επισκόπηση της Διατριβής

4.2 Ερευνητικά Ερωτήματα

Με βάση τους άξονες που περιγράφηκαν στην προηγούμενη ενότητα ορίζουμε τα ερευνητικά ερωτήματά της διατριβής. Η γενική ερευνητική κατεύθυνση της διδακτορικής διατριβής αφορά την διερεύνηση της δυνατότητας για βελτιωμένα συστήματα προτάσεων στο εσωτερικό επιχειρήσεων, κοινοτήτων και στο ηλεκτρονικό εμπόριο με βάση λανθάνοντα θέματα.

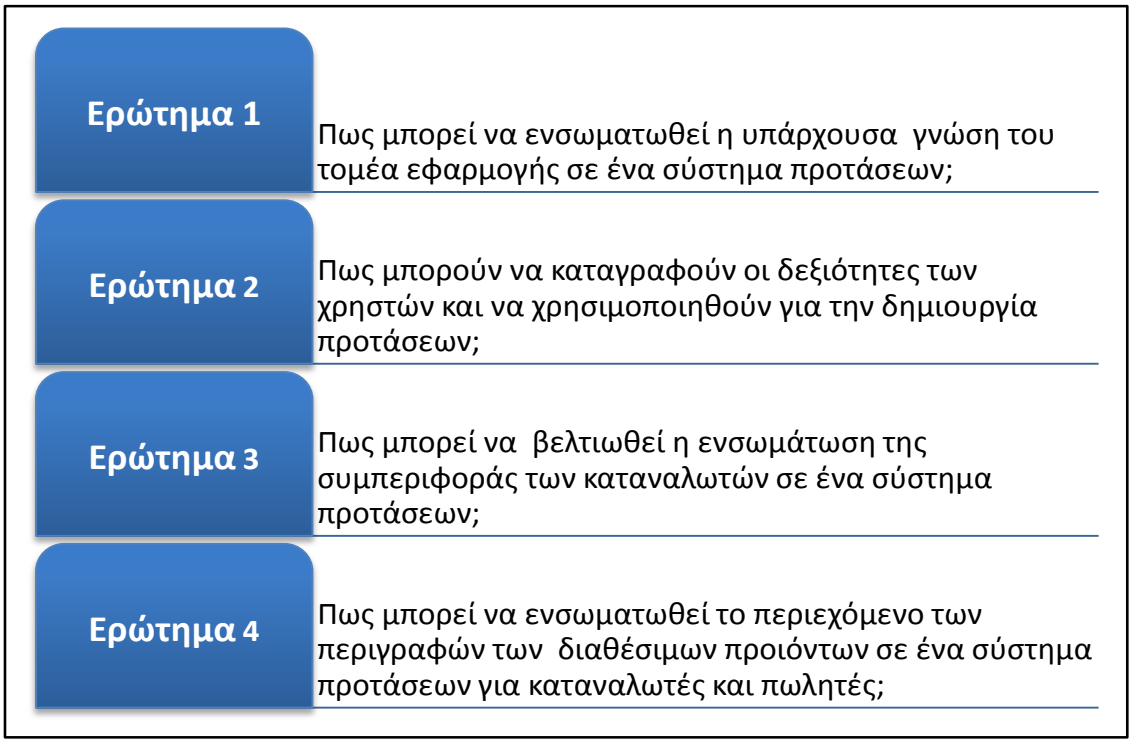
Πιο συγκεκριμένα, η κατεύθυνση αυτή αναλύεται σε τέσσερα ερευνητικά ερωτήματα.

Τα δύο πρώτα ερωτήματα αφορούν την διαχείριση γνώσης. Πρώτον, τίθεται το ερώτημα της ενσωμάτωσης υπάρχουσας γνώσης ενός συγκεκριμένου τομέα

εφαρμογής σε ένα σύστημα προτάσεων. Στο δεύτερο ερώτημα περιγράφεται η εξαγωγή και αξιοποίηση των δεξιοτήτων των χρηστών σε ένα σύστημα προτάσεων.

Τα επόμενα δύο ερωτήματα αφορούν την διεξαγωγή αγοραπωλησιών στο εμπόριο και στο ηλεκτρονικό εμπόριο. Στο τρίτο ερευνητικό ερώτημα αναλύεται η δυνατότητα ενσωμάτωσης της συμπεριφοράς των καταναλωτών σε ένα σύστημα προτάσεων. Τέλος, στο τέταρτο ερώτημα αναλύεται η δυνατότητα των συστημάτων προτάσεων να λάβουν υπόψη τις περιγραφές των αντικειμένων που διακινούνται.

Μια επισκόπηση των ερωτημάτων παρουσιάζεται στην Εικόνα 4.2.



Εικόνα 4.2 Ερευνητικά Ερωτήματα

1. Ενσωμάτωση υπάρχουσας γνώσης του τομέα εφαρμογής

Από τη σχετική βιβλιογραφία προκύπτει ότι οι προσπάθειες ενσωμάτωσης γνώσης στο παρελθόν αφορούν τεχνικές που απαιτούν τυπικές γνωσιακές δομές και διαρκείς παρεμβάσεις των χρηστών στα συστήματα διαχείρισης γνώσης, ενώ απουσιάζουν τεχνικές που περιλαμβάνουν μη δομημένες, μη τοπικές γνωσιακές δομές και κοινωνικό λογισμικό. Η ερευνητική προσπάθεια της παρούσας διδακτορικής διατριβής στοχεύει στην κάλυψη του κενού που παρουσιάζεται στη βιβλιογραφία με την ενσωμάτωση μοντέλων θεμάτων και αξιολόγηση των αντίστοιχων συστημάτων προτάσεων. Στην παρούσα διατριβή επιδιώκουμε μία

προσέγγιση για την ενσωμάτωση της υπάρχουσας γνώσης ενός πεδίου, όπως αυτή διατυπώνεται σε μη-δομημένο περιεχόμενο, σε ένα πλήρες σύστημα προτάσεων το οποίο χρησιμοποιεί τα πιθανοτικά μοντέλα θεμάτων. Επίσης θα παρουσιαστεί μια μεθοδολογία για τη σύνδεση των θεμάτων με ελαφρού τύπου γνωσιακές δομές.

2. Εξαγωγή, υπολογισμός και χρήση των δεξιοτήτων των χρηστών για την δημιουργία προτάσεων

Οι μέθοδοι που προτείνονται στη βιβλιογραφία κυρίως αντιμετωπίζουν το πρόβλημα καταγράφοντας χωριστά ποσοτικά και ποιοτικά χαρακτηριστικά των χρηστών, ενώ πολλές φορές αγνοούν την χρήση πολλών εργαλείων συνεργασίας. Στην παρούσα διατριβή επιχειρούμε να εκμεταλλευτούμε την εξαγωγή πιθανοτικών μοντέλων θεμάτων για την πλήρη και αποτελεσματική μοντελοποίηση της ικανότητας ενός εργαζομένου να αντιμετωπίσει ένα πρόβλημα. Η μεθοδολογία που προτείνουμε έχει δυο χαρακτηριστικά. Πρώτον, ενσωματώνει στοιχεία από πολλαπλές πηγές, δηλαδή από πολλά συνεργατικά εργαλεία στα οποία δραστηριοποιείται ο εργαζόμενος. Δεύτερον, συνδυάζει τόσο ποιοτικά όσο και ποσοτικά χαρακτηριστικά της δραστηριότητας του εργαζομένου, λαμβάνοντας υπόψη μετρικές δραστηριότητας αλλά και το κείμενο το οποίο εκείνος παράγει.

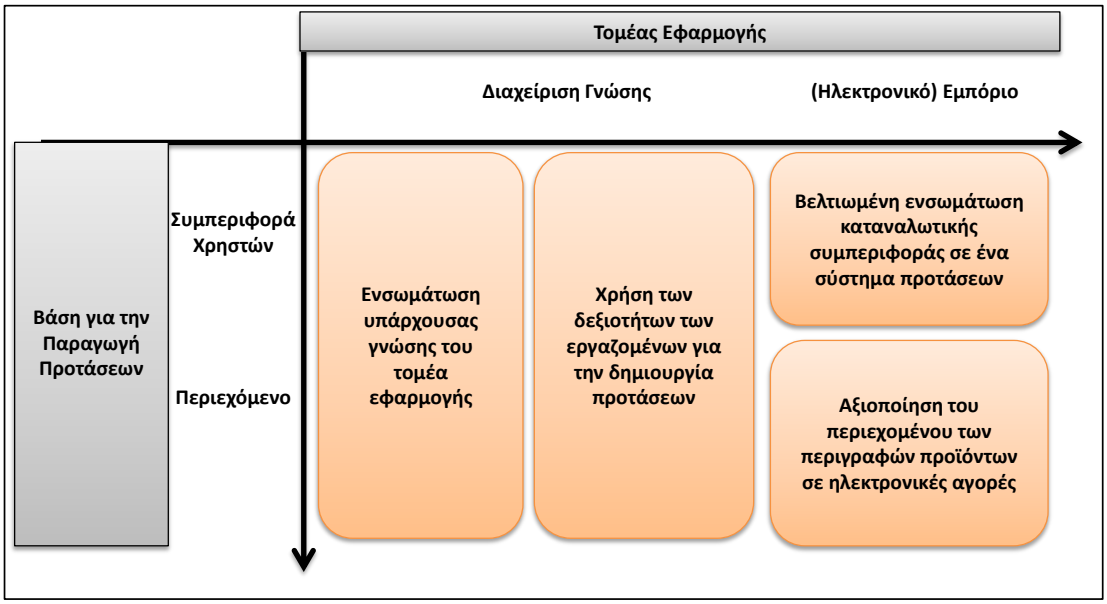
3. Ενσωμάτωση καταναλωτικής συμπεριφοράς σε ένα σύστημα προτάσεων

Ως τώρα στη βιβλιογραφία έχουν προταθεί διάφορες μέθοδοι για την εκμετάλλευση της καταγεγραμμένης συμπεριφοράς των καταναλωτών για την παραγωγή προτάσεων (εξαγωγή κανόνων συνάφειας, συνεργατική διήθηση, κ.α.). Απουσιάζει μια μέθοδος που να μπορεί να παρέχει μια εποπτική εικόνα των δραστηριοτήτων των καταναλωτών και να αποτελεί την βάση για την δημιουργία προτάσεων με ακρίβεια. Τα πιθανοτικά μοντέλα θεμάτων αποτελούν κομμάτι της προτεινόμενης μεθόδου, η οποία περιλαμβάνει δυο συναφείς τεχνικές: Πρώτον, την εξαγωγή θεμάτων από ομάδες προϊόντων που εμφανίζονται μαζί συχνά στο ιστορικό των αγοραστών (λανθάνοντες χρήστες). Δεύτερον, την εξαγωγή θεμάτων από προϊόντα που εμφανίζονται συχνά μαζί σε μεμονωμένες επισκέψεις των αγοραστών (λανθάνοντα καλάθια προϊόντων). Στη συνέχεια, τα εξαγόμενα μοντέλα χρησιμοποιούνται για παραγωγή προτάσεων.

4. Αξιοποίηση του περιεχομένου των περιγραφών των διαθέσιμων προϊόντων από ένα σύστημα προτάσεων σε ηλεκτρονικές αγορές

Η βιβλιογραφική έρευνα ως τώρα παρουσιάζει μια ποικιλία μεθόδων για την δημιουργία προτάσεων σε ηλεκτρονικές αγορές, κυριότερα με βάση τη συμπεριφορά των καταναλωτών. Απουσιάζει όμως μια μεθοδολογία για την εκμετάλλευση του μη δομημένου κειμένου που βρίσκεται σε ηλεκτρονικές αγορές δημοπρασιών για την παραγωγή προτάσεων που απευθύνονται σε καταναλωτές και πωλητές. Στην διατριβή ασχολούμαστε με προτάσεις που έχουν δυο σκέλη. Πρώτον, την πρόταση σχετικών προϊόντων που μπορεί να επιθυμεί ο επίδοξος αγοραστής. Δεύτερον, την πρόταση συναφών προϊόντων αλλά και σημαντικών όρων που αφορούν το προϊόν που θέλει να πουλήσει ένας πωλητής. Η δημιουργία αυτών των προτάσεων γίνεται με βάση τα πιθανοτικά μοντέλα θεμάτων που έχουν εξαχθεί από το περιεχόμενο της ηλεκτρονικής αγοράς.

Τα παραπάνω ερευνητικά ερωτήματα τοποθετούνται στους άξονες που περιγράφηκαν παραπάνω όπως φαίνεται στην Εικόνα 4.3.



Εικόνα 4.3 Τοποθέτηση των Ερευνητικών Ερωτημάτων

4.3 Συνεισφορά της Διατριβής

Η συνεισφορά της διατριβής συνίσταται στα αποτελέσματα της διερεύνησης για βελτιωμένα συστήματα προτάσεων στο εσωτερικό επιχειρήσεων, κοινοτήτων και στο ηλεκτρονικό εμπόριο με βάση πιθανοτικά μοντέλα θεμάτων.

Η ερευνητική εργασία που παρουσιάζεται διαφοροποιείται σημαντικά από την βιβλιογραφία τόσο στη στόχευση των ερωτημάτων που θέτει όσο και στις προτεινόμενες προσεγγίσεις και αξιολογήσεις. Στην παρούσα διατριβή για πρώτη φορά παρουσιάζεται μια ολοκληρωμένη προσέγγιση που συνδυάζει γνωσιακές δομές και πιθανοτικά μοντέλα θεμάτων για την παραγωγή βελτιωμένων προτάσεων. Επίσης για πρώτη φορά χρησιμοποιείται η τεχνολογία πιθανοτικών μοντέλων για την εξαγωγή πληροφοριών από διάφορα εργαλεία, για την ακόλουθη αποτύπωση των δεξιοτήτων των χρηστών και για τη χρήση τους για την δημιουργία προτάσεων. Επιπρόσθετα, προτείνεται η χρήση των μοντέλων θεμάτων για την εξαγωγή βελτιωμένων προφίλ καταναλωτών και παραγωγή προτάσεων. Τέλος σχεδιάζεται και υλοποιείται η πραγματοποίηση προτάσεων σε αγορές όπου μοντελοποιείται το μη-δομημένο περιεχόμενο των περιγραφών των αντικειμένων προς πώληση.

Οι προσεγγίσεις για τη δημιουργία προτάσεων που προτείνονται στα πλαίσια της διατριβής παρουσίασαν κάποια κοινά χαρακτηριστικά: (1) μειώνουν τις απαιτούμενες διαστάσεις του προβλήματος και παρέχουν γρήγορα προτάσεις αφού έχει προηγηθεί η εξαγωγή των μοντέλων θεμάτων, (2) ικανοποιούν τις απαιτήσεις των χρηστών για ακρίβεια και ανάκληση όλων των δεδομένων που τους ενδιαφέρουν, και (3) τα θέματα που εξάγονται μπορούν να αποτελέσουν σημαντική πληροφορία για τον ιδιοκτήτη ή τον διαχειριστή του συστήματος.

Η συνεισφορά της διατριβής αναλύεται με μεγαλύτερη ακρίβεια στις ακόλουθες ενότητες οι οποίες συνδέονται με τα αντίστοιχα ερευνητικά ερωτήματα.

4.3.1 Ενσωμάτωση Γνώσης

Διατύπωση Ερωτήματος

Το συγκεκριμένο ερευνητικό ερώτημα αφορά την δυνατότητα ενός συστήματος προτάσεων να ενσωματώσει την υπάρχουσα γνώση ενός τομέα εφαρμογής, όπως αυτή εκφράζεται από γνωσιακές δομές στο εσωτερικό ενός οργανισμού (π.χ. ταξονομίες ή φολκσονομίες). Για να είναι πλήρης αυτή η ενσωμάτωση πρέπει να περιλαμβάνει δυο τύπους αλληλεπιδράσεων: την δυνατότητα των γνωσιακών δομών να επηρεάζουν τα προτεινόμενα αντικείμενα και την δυνατότητα των προτάσεων να επηρεάζουν τον σχηματισμό και την εξέλιξη των γνωσιακών δομών.

Σχέση με τη Βιβλιογραφία

Από τη σχετική βιβλιογραφία προκύπτει η ανάγκη ενός ενημερωμένου συστήματος προτάσεων που συμβαδίζει με την εξελισσόμενη γνώση. Αυτή η ανάγκη γίνεται πιο εμφανής σε εταιρικά περιβάλλοντα έντασης γνώσης.

Πολύ συχνά οι επιχειρήσεις βασίζονται στην υποστήριξη κοινοτήτων εργαζομένων στο εσωτερικό και στο εξωτερικό τους και χρησιμοποιούν κοινωνικά και συμμετοχικά δικτυακά εργαλεία για τη συσχέτισή τους με ανάγκες της επιχείρησης [84]. Σε αντίστοιχη μελέτη [84] αναλύονται οι δραστηριότητες που λαμβάνουν χώρα στις επιχειρήσεις με βάση τον κοινωνικό ιστό και προτείνεται ένα πλαίσιο με το όνομα SLATES όπου περιγράφει τους βασικούς άξονες της χρήσης εταιρικού κοινωνικού λογισμικού. Η λέξη SLATES αποτελεί ένα ακρωνύμιο για τις λέξεις search (αναζήτηση), links (σύνδεσμοι), authoring (συγγραφή), tags (επισημειώσεις), extensions (επεκτάσεις) και signals (σήματα). Σημαντικό κομμάτι του εταιρικού κοινωνικού λογισμικού αποτελούν οι επεκτάσεις όπως τα συστήματα προτάσεων που μπορούν να προβλέψουν τι μπορεί να είναι χρήσιμο στους χρήστες και να το προτείνουν. Ο λόγος ύπαρξης τέτοιου τύπου επεκτάσεων είναι ακριβώς η ανάγκη για επεξεργασία μεγάλου όγκου περιεχομένου όπως αυτό παράγεται από τα κοινωνικά εργαλεία.

Στη βιβλιογραφία έχουν προταθεί διάφορες προσεγγίσεις που χρησιμοποιούν μοντέλα θεμάτων για να υποστηρίξουν την αναζήτηση και τα συστήματα προτάσεων. Τα λανθάνοντα θέματα έχουν χρησιμοποιηθεί για την εξόρυξη πληροφοριών από ιστολόγια και για την πρόταση σχετικών επισημειώσεων για συγκεκριμένες δημοσιεύσεις [67]. Ακόμη, στο εταιρικό περιβάλλον έχει προταθεί η χρήση των λανθανόντων θεμάτων ώστε να εντοπιστούν και να οπτικοποιηθούν λανθάνουσες κοινότητες με παρεμφερή ενδιαφέροντα [75], ενώ σε αντίστοιχη μελέτη προτείνεται η χρήση λανθανόντων θεμάτων για να αποτυπώνονται τα ενδιαφέροντα των εργαζομένων αλλά και για να προτείνονται σχετικά έγγραφα [65]. Το σύστημα που προτείνεται εντοπίζει τα θέματα για τα οποία ενδιαφέρονται οι χρήστες βραχυπρόθεσμα και μακροπρόθεσμα, και παράγει προτάσεις με έναν υψηλότερο βαθμό ποικιλότητας μεταξύ θεμάτων. Στη βιβλιογραφία εξετάζεται η πρόταση πόρων από κοινωνικά μέσα στο εσωτερικό της επιχείρησης με βάση τους ανθρώπους, τις επισημειώσεις και τις σχέσεις μεταξύ τους [85].

Επιπλέον, κάποιες μελέτες έχουν εστιάσει στην ανάδραση μεταξύ της συμπεριφοράς του χρήστη και των γνωσιακών δομών. Μια μορφή ανάδρασης που έχει προταθεί περιλαμβάνει την χρήση των επισημειώσεων από τον χρήστη για την βελτίωση της αναζήτησης στο εσωτερικό μιας εταιρίας [86]. Η προσέγγιση SemSLATES [87] χρησιμοποιεί τεχνολογίες σημασιολογικού ιστού για να βελτιώσει την αναζήτηση και την δημιουργία προτάσεων. Οι συγκεκριμένες τεχνικές που περιγράφονται αφορούν την σύνδεση πολύπλοκων σημασιολογικών δομών με απλό κείμενο σε εταιρικό περιβάλλον. Απαιτούν, όμως, την ύπαρξη τυπικών γνωσιακών δομών, τουλάχιστον στη μορφή των συνδεδεμένων δεδομένων, όπου θα είναι συνδεδεμένα με τις δραστηριότητες των χρηστών.

Παρατηρώντας τη σχετική βιβλιογραφία συμπεραίνουμε ότι οι προσπάθειες ενσωμάτωσης γνώσης στο παρελθόν αφορούν τεχνικές που είτε απαιτούν τυπικές γνωσιακές δομές και διαρκείς παρεμβάσεις των χρηστών στα συστήματα διαχείρισης γνώσης είτε δεν αλληλεπιδρούν εκτενώς με τις υπάρχουσες γνωσιακές δομές. Από την άλλη πλευρά, η χρήση των πιθανοτικών μοντέλων θεμάτων δεν συνδυάζεται επαρκώς με τις υπάρχουσες γνωσιακές δομές. Έτσι, με βάση την υπάρχουσα βιβλιογραφία εντοπίζουμε την έλλειψη μιας μεθοδολογίας που να μπορεί να συνδυάζει το μη δομημένο περιεχόμενο με τις υπάρχουσες γνωσιακές δομές αποδοτικά για την παραγωγή προτάσεων και την υποστήριξη της αναζήτησης.

Συνεισφορά της Διατριβής

Στην παρούσα διατριβή δίνεται μία προσέγγιση για την ενσωμάτωση της υπάρχουσας γνώσης ενός πεδίου σε ένα σύστημα προτάσεων.

Προτείνεται ένα σύστημα προτάσεων το οποίο χρησιμοποιεί τα πιθανοτικά μοντέλα θεμάτων για να αποτυπώσει την υπάρχουσα γνώση ενός τομέα όπως αυτή διατυπώνεται σε μη-δομημένο περιεχόμενο. Επίσης, παρουσιάζεται μια μεθοδολογία για τη σύνδεση του με ελαφρού τύπου γνωσιακές δομές με μικρή τυπικότητα. Πεδίο εφαρμογής αποτελεί η εταιρική γνώση στο εσωτερικό επιχειρήσεων.

Η μεθοδολογία που προτείνουμε βασίζεται στην ανάλυση εγγράφων με βάση την τεχνική εξαγωγής πιθανοτικών μοντέλων θεμάτων, και συγκεκριμένα της λανθάνουσας κατανομής Dirichlet για την επέκταση ερωτημάτων αναζήτησης αλλά και για την πραγματοποίηση προτάσεων εγγράφων σε χρήστες. Όταν ένα υπάρχον

αντικείμενο, π.χ. ένα έγγραφο, διαβάζεται από τον χρήστη το σύστημα μπορεί να συμπεράνει την κατανομή θεμάτων του. Όταν προστίθεται ένα καινούριο αντικείμενο, η κατανομή θεμάτων του μπορεί να εξαχθεί με βάση το υπάρχον μοντέλο.

Ο εντοπισμός των λανθανόντων θεμάτων για την πρόταση περιεχομένου και για την αναζήτηση είναι μια μέθοδος μη επιβλεπόμενη και προσφέρει μια σειρά από οφέλη, ιδιαίτερα σε σχέση με επιβλεπόμενες μεθόδους και με εκείνες που βασίζονται σε μοντέλα. Δεν εξαρτάται από ρητές γνωσιακές δομές όπως ταξονομίες ή οντολογίες και δεν απαιτεί προσπάθεια από τον χρήστη για την κατηγοριοποίηση των πόρων.

Ενώ τα λανθάνοντα θέματα δεν προϋποθέτουν άλλες γνωσιακές δομές, μπορούν να συνδεθούν με αυτές ώστε να καλύπτουν την οργανωσιακή γνώση η οποία εξελίσσεται χωρίς να εξαρτάται από συγκεκριμένες λέξεις. Η πρόταση επισημειώσεων με βάση τα λανθάνοντα θέματα επηρεάζει την δημιουργία των γνωσιακών δομών στο εταιρικό κοινωνικό λογισμικό και μπορεί να βοηθήσει την εξέλιξη της δομής ώστε να καλύψει νέα αναδυόμενα θέματα. Όσο καινούρια έγγραφα εισάγονται στο σύστημα και αναλύονται, τόσο καινούρια θέματα εντοπίζονται στο πιθανοτικό μοντέλο το οποίο επαναϋπολογίζεται για να αντιστοιχεί στο ενημερωμένο σύνολο δεδομένων. Οι κυρίαρχες λέξεις σε αυτά τα θέματα προτείνονται συνεχώς ως λέξεις κλειδιά για τα νέα έγγραφα που μένουν να επισημανθούν. Από την άλλη πλευρά, οι γνωσιακές δομές συνυπολογίζονται στην δημιουργία προτάσεων.

Τα πλεονεκτήματα της προσέγγισης που περιγράφεται γίνονται εμφανή σε ένα εταιρικό κοινωνικό λογισμικό, το οποίο διαθέτει γνωσιακές δομές διαφορετικής τυπικότητας. Τα πιθανοτικά μοντέλα θεμάτων μπορούν να βελτιώσουν την ευελιξία και την σταθερότητα του συστήματος προτάσεων, καθώς η επιχειρησιακή γνωσιακή δομή και οι αυθαίρετες δραστηριότητες των χρηστών συμπληρώνονται από την εξαγωγή λανθανόντων θεμάτων.

Το λογισμικό που αναπτύχθηκε με βάση τα παραπάνω έγινε διαθέσιμο με τη μορφή λογισμικού ανοιχτού κώδικα. Οι επαγγελματίες του χώρου μπορούν να τα χρησιμοποιήσουν και να τα ενσωματώσουν σε άλλα συστήματα διαχείρισης γνώσης, καθώς και να τα επεκτείνουν ώστε να προσθέσουν επιπλέον δυνατότητες. Επιπλέον, οι ερευνητές μπορούν να διαπιστώσουν κατά πόσο τα πιθανοτικά

μοντέλα θεμάτων μπορούν να είναι επωφελή για την διαχείριση γνώσης ιδίως όταν χρησιμοποιούνται γνωσιακές δομές.

Στο κεφάλαιο 5 της παρούσας διατριβής παρουσιάζουμε αναλυτικά τον τρόπο με τον οποίο τα πιθανοτικά μοντέλα θεμάτων μπορούν να λειτουργήσουν ως δομικά στοιχεία του κοινωνικού εταιρικού λογισμικού και να βελτιώσουν την δυνατότητα παροχής προτάσεων αλλά και να βελτιώσουν την απόδοση της αναζήτησης εσωτερικά. Η προσέγγισή μας επεκτείνει την αναζήτηση με χρήση ερωτημάτων και μπορεί να προτείνει σχετικά έγγραφα και επισημειώσεις, τα οποία με τη σειρά τους βοηθούν την δημιουργία και την συντήρηση γνωσιακών δομών με βάση την συμπεριφορά των χρηστών και το περιεχόμενο. Η προτεινόμενη προσέγγιση δεν απαιτεί επιπρόσθετη προσπάθεια από τους χρήστες.

Σχετικές Δημοσιεύσεις

Η συνεισφορά στο συγκεκριμένο τομέα έχει οδηγήσει σε τέσσερις δημοσιεύσεις:

- K. Christidis, G. Mentzas, and D. Apostolou, “Using latent topics to enhance search and recommendation in Enterprise Social Software,” Expert Systems with Applications, vol. 39, no. 10, pp. 9297–9307, Aug. 2012.
- K. Christidis, G. Mentzas, and D. Apostolou, “Supercharging Enterprise 2.0,” IEEE IT Professional, vol. 13, no. 4, pp. 29–35, 2011.
- K. Christidis and G. Mentzas, “Using Probabilistic Topic Models in Enterprise Social Software,” in Business Information Systems, 2010, pp. 23–34.
- K. Christidis, G. Mentzas, and D. Apostolou, “A Socially Intelligent Approach for Enterprise Information Search and Recommendation,” In Proceedings of 18th International ICE-Conference on Engineering, Technology and Innovation, 18 - 20 June 2012, Munich.

4.3.2 Δεξιότητες Συνεργατών

Διατύπωση Ερωτήματος

Το ερευνητικό ερώτημα αφορά την δυνατότητα ενός συστήματος προτάσεων να ενσωματώσει στους υπολογισμούς του τις δεξιότητες των χρηστών. Εδώ απαιτείται να εξαχθούν, να υπολογιστούν με βάση κάποιες μετρικές και να χρησιμοποιηθούν οι δεξιότητες των εργαζομένων ανά θεματικό τομέα ειδίκευσης σε μια ομάδα. Η ανάγκη για ένα τέτοιο σύστημα είναι εμφανής σε κοινότητες ανάπτυξης ελεύθερου λογισμικού όπου η αναζήτηση των κατάλληλων προγραμματιστών για να επιλύσουν συγκεκριμένα προβλήματα είναι καθοριστικής σημασίας.

Σχέση με τη Βιβλιογραφία

Από τη σχετική βιβλιογραφία προκύπτει η ανάγκη για μια προσέγγιση εξαγωγής δεξιοτήτων ανά θεματική περιοχή που θα μπορεί να πραγματοποιείται σε κοινότητες που χρησιμοποιούν πολλά εργαλεία για την συνεργασία μεταξύ των συμμετεχόντων. Ειδικότερα στον τομέα της ανάπτυξης λογισμικού έχει παρουσιαστεί ένας αριθμός συστημάτων προτάσεων τα οποία λαμβάνουν υπόψη είτε τις δραστηριότητες των χρηστών ή την σημασιολογία των συνεισφορών τους.

Τα προγενέστερα συστήματα προτάσεων βασίζονται κυρίως στις μετρικές δραστηριότητας των προγραμματιστών όπως αυτές καταγράφονται στις αλλαγές του πηγαίου κώδικα. Για παράδειγμα, έχει προταθεί η μέτρηση των δεξιοτήτων κάθε προγραμματιστή με βάση μια μετρική που ονομάζεται experience atoms (EA) [88]. Οι συγκεκριμένες μονάδες εμπειρίας αντιστοιχούν στις μικρότερες μετρήσιμες δραστηριότητες ενός ανθρώπου σε σχέση με το προϊόν μιας εργασίας και ο αριθμός τους μπορεί να απεικονίσουν την συσσωρευμένη εμπειρία του. Σε μια εναλλακτική προσέγγιση η δεξιότητα ενός προγραμματιστή για κάθε συγκεκριμένο αρχείο κωδικοποιείται σε έναν χάρτη όπου καταγράφεται η συχνότητα των συνεισφορών του προγραμματιστή στο συγκεκριμένο αρχείο [89]. Εδώ εισάγονται και κοινωνικά δεδομένα γιατί αυτή η συχνότητα κανονικοποιείται με βάση τον αριθμό των προγραμματιστών που έχουν συνεισφέρει στο συγκεκριμένο αρχείο στο παρελθόν. Στη βιβλιογραφία έχει διερευνηθεί η δυνατότητα χρήσης όχι μόνο της συχνότητας αλλά και της χρονικής απόστασης από την τελευταία δραστηριότητα του κάθε προγραμματιστή για τον υπολογισμό των δεξιοτήτων του [90]. Τέλος, έχει προταθεί

ο συνδυασμός διαφόρων μετρικών όπως ο αριθμός δραστηριοτήτων, η συχνότητα δραστηριοτήτων, η χρονική απόσταση από την τελευταία δραστηριότητα, ο αριθμός των αρχείων που αλλάζονται και ο αριθμός των γραμμών κώδικα που προστίθενται ή αφαιρούνται [91].

Από την άλλη πλευρά, υπάρχουν προσεγγίσεις που αφορούν περισσότερο το περιεχόμενο των συνεισφορών των προγραμματιστών. Οι όροι που υπάρχουν στον πηγαίο κώδικα μπορούν να χρησιμοποιηθούν με τη μορφή διανύσματος όρων για την πρόταση προγραμματιστών που έχουν την ανάλογη πείρα για να λύσουν ένα πρόβλημα [92]. Αυτό αποτελεί ένα παράδειγμα αμιγώς ποιοτικής διαφοροποίησης, όπου υπολογίζεται μόνον το πεδίο ειδίκευσης. Εναλλακτικά, ένα σύστημα προτάσεων μπορεί να βοηθά τους μηχανικούς λογισμικού να εναλλάσσουν αντικείμενο εργασίας με βάση τον τύπο της ανάπτυξης λογισμικού και το ιστορικό αλληλεπιδράσεων μεταξύ τους [93]. Ακόμη, έχει προταθεί μια παραλλαγή του LDA για την μοντελοποίηση της εξέλιξης των θεμάτων στις αποθήκες πηγαίου κώδικα λογισμικού [65].

Ένας σημαντικός αριθμός μελετών δείχνει ότι τα μοντέλα θεμάτων μπορούν να εφαρμοστούν με επιτυχία στο εσωτερικό κοινοτήτων ώστε να βοηθήσουν την εξαγωγή πληροφορίας από πολλαπλές πηγές σε ένα έργο, αλλά και από περισσότερο από ένα έργα. Έχει προταθεί η εφαρμογή μιας προσέγγισης μοντέλων θεμάτων ώστε να σχηματιστούν σχέσεις μεταξύ του πηγαίου κώδικα και αντικειμένων υψηλού επιπέδου, όπως απαιτήσεων λογισμικού [94]. Επίσης έχουν χρησιμοποιηθεί μοντέλα θεμάτων για να πραγματοποιηθεί η σύνδεση ανάμεσα στην δραστηριότητα των προγραμματιστών στα ιστολόγια και στις συνεισφορές τους [95], ενώ μια διαφορετική προσέγγιση εφαρμόζει την ανάλυση θεμάτων σε πολλά έργα και με τη χρήση μιας ταξονομίας μπορεί να παρέχει προτάσεις ονοματοδοσίας θεμάτων [96].

Στη κατεύθυνση των συστημάτων προτάσεων, μια μέθοδος με βάση την λανθάνουσα κατανομή Dirichlet (LDA) έχει χρησιμοποιηθεί για την ανάλυση του πηγαίου κώδικα του λογισμικού Eclipse και για να παρέχει εκτεταμένη επίγνωση της συνάφειας μεταξύ προγραμματιστών [97]. Επίσης, έχει παρουσιαστεί μια μεθοδολογία που περιορίζεται σε μια πηγή πληροφοριών για την εξαγωγή των δεξιοτήτων των προγραμματιστών με βάση τα μοντέλα θεμάτων του περιεχομένου και την δραστηριότητα των προγραμματιστών [66]. Παρ’ όλα αυτά, οι μέθοδοι λανθάνουσας σημασιολογικής ανάλυσης δεν χρησιμοποιούνται εκτενώς για την παραγωγή προτάσεων και δεν καλύπτουν τις διάφορες πηγές πληροφοριών.

Παρατηρώντας τη σχετική βιβλιογραφία συμπεραίνουμε ότι οι προσπάθειες για την αποτύπωση δεξιοτήτων και χρήση τους σε συστήματα προτάσεων στο παρελθόν βασίζονταν κυριότερα σε ποιοτικές ενδείξεις με βάση το κείμενο ή σε ποσοτικές με βάση την δραστηριότητα. Κάποιες προσπάθειες που έχουν γίνει στο παρελθόν για συνδυασμό των δυο κατηγοριών δεν καλύπτουν περισσότερα από ένα εργαλεία και δεν προσφέρουν μια εποπτική εικόνα των δεξιοτήτων του κάθε συνεργάτη.

Συνεισφορά της Διατριβής

Στην παρούσα διατριβή καλύπτεται εν μέρει αυτό το κενό καθώς προτείνεται μια μεθοδολογία που εκμεταλλεύεται την εξαγωγή πιθανοτικών μοντέλων θεμάτων για την πλήρη και αποτελεσματική μοντελοποίηση της ικανότητας ενός εργαζομένου να αντιμετωπίσει ένα πρόβλημα. Η μεθοδολογία αυτή έχει δυο χαρακτηριστικά. Πρώτον, ενσωματώνει στοιχεία από πολλαπλές πηγές, δηλαδή από πολλά συνεργατικά εργαλεία στα οποία δραστηριοποιείται ο εργαζόμενος. Δεύτερον, συνδυάζει τόσο ποιοτικά όσο και ποσοτικά χαρακτηριστικά της δραστηριότητας του εργαζομένου, λαμβάνοντας υπόψη μετρικές δραστηριότητας αλλά και το κείμενο το οποίο εκείνος παράγει.

Η μεθοδολογία αυτή υλοποιείται σε ένα υβριδικό σύστημα προτάσεων που προορίζεται για χρήση στο εσωτερικό μιας κοινότητας ανάπτυξης ελεύθερου λογισμικού ανοιχτού κώδικα. Το σύστημα αυτό συνυπολογίζει τόσο το πεδίο ειδίκευσης του κάθε προγραμματιστή όσο και τις μετρικές που δείχνουν τον βαθμό ενασχόλησης του με την κοινότητα. Επίσης, το συγκεκριμένο σύστημα έχει την δυνατότητα σύνδεσης με πολλαπλά εργαλεία συνεργασίας προγραμματιστών (λίστες ηλεκτρονικού ταχυδρομείου, αποθήκες πηγαίου κώδικα, συστήματα διαχείρισης ζητημάτων).

Η προτεινόμενη προσέγγιση συνδυάζει τα λανθάνοντα θέματα που εμφανίζονται στις συνεισφορές των προγραμματιστών με ποσοτικές μετρικές που αφορούν τις δραστηριότητες τους. Το αποτέλεσμα αυτού του συνδυασμού είναι ο σχηματισμός βαθμολογιών προγραμματιστών με βάση τα θέματα (topic-based competency score) που περιλαμβάνουν μια εποπτική εικόνα του τομέα εξειδίκευσης (σε τι είναι ικανός) αλλά και του βαθμού δεξιότητας (πόσο ικανός είναι). Ο σχηματισμός των προφίλ βασίζεται στην εξαγωγή πιθανοτικών μοντέλων θεμάτων.

Με βάση τη συγκεκριμένη βαθμολογία προκύπτει μια αντίστοιχη γραφική διεπαφή χρήστη η οποία μπορεί να παρέχει μια εποπτική εικόνα για τις δεξιότητες κάθε χρήστη. Τα λανθάνοντα θέματα έχουν την ικανότητα να περιγράψουν το πλήρες εύρος ενασχόλησης της κοινότητας και να ερμηνευτούν από τον χρήστη. Από την άλλη, η βαθμολογία που αφορά τον προγραμματιστή ανά θέμα, αντανακλά την ικανότητα που έχει επιδείξει στο παρελθόν και μπορεί να χρησιμοποιηθεί για να προβλέψουμε την ικανότητα του να λύσει ένα ζήτημα στο μέλλον.

Το σύστημα προτάσεων που προέκυψε με βάση αυτή τη βαθμολογία μπορεί να προβλέψει με σημαντική ακρίβεια τον πλέον κατάλληλο προγραμματιστή για να ασχοληθεί με ένα συγκεκριμένο ζήτημα. Για να γίνει αυτό, προβλέπεται η ανάλυση της περιγραφής του ζητήματος με βάση τα λανθάνοντα θέματα. Ακολούθως, αφού οριστεί το ζήτημα σε σχέση με τα μοντέλα θεμάτων, συγκρίνεται διανυσματικά με τις βαθμολογίες των προγραμματιστών. Η σύγκριση αυτή λαμβάνει υπόψη όλα τα στοιχεία των δεξιοτήτων των προγραμματιστών και οδηγεί σε πιο ισορροπημένες προτάσεις.

Η συγκεκριμένη μεθοδολογία αξιολογήθηκε με δεδομένα που προήλθαν από μια μεγάλη και από μια μικρότερη κοινότητα χρηστών με ενθαρρυντικά αποτελέσματα.

Σχετικές Δημοσιεύσεις

Η ερευνητική εργασία που αφορά το συγκεκριμένο ερευνητικό ερώτημα έχει παρουσιαστεί σε συνέδριο με τα παρακάτω στοιχεία:

- K. Christidis, F. Paraskevopoulos, D. Panagiotou and G. Mentzas. “Combining Activity Metrics and Contribution Topics for Software Recommendations”. In Proceedings of the 3rd International Workshop on Recommendation Systems for Software Engineering (RSSE '12). ACM, Zurich, Switzerland

Επίσης, μια ερευνητική εργασία έχει υποβληθεί σε ειδικό τεύχος του περιοδικού IEEE Software με τίτλο τεύχους «Software Analytics: So What?». Τα στοιχεία της εργασίας είναι τα παρακάτω:

- K. Christidis, F. Paraskevopoulos and G. Mentzas. “SOCRATES: A Topic-based Competency Score for Software Recommendations”. Submitted to IEEE Software, 15th of December 2012

4.3.3 Καταναλωτική Συμπεριφορά

Διατύπωση Ερωτήματος

Το ερευνητικό ερώτημα περιλαμβάνει την δυνατότητα ενός συστήματος προτάσεων να ενσωματώσει στους υπολογισμούς του τις δραστηριότητες των καταναλωτών. Εδώ απαιτείται να εξαχθούν και να καταγραφούν οι δραστηριότητες των καταναλωτών σε ένα μοντέλο περιορισμένων διαστάσεων το οποίο να περιγράφει τα χαρακτηριστικά τους. Το μοντέλο αυτό στη συνέχεια θα πρέπει να υποστηρίζει ένα σύστημα προτάσεων το οποίο να προβλέπει με ακρίβεια τις προτιμήσεις των καταναλωτών.

Σχέση με τη Βιβλιογραφία

Από τη σχετική βιβλιογραφία προκύπτει η συγκεκριμένη ανάγκη για μια προσέγγιση εξαγωγής της δραστηριότητας των καταναλωτών που θα μπορεί να χρησιμοποιείται για την παραγωγή προτάσεων στο χώρο του εμπορίου.

Στη περιοχή της έρευνας μάρκετινγκ έχουν εφαρμοστεί μεθοδολογίες εξόρυξης δεδομένων και μηχανικής μάθησης σε συναλλαγές λιανικής στις οποίες αναλύεται ένας μεγάλος όγκος δεδομένων αγορών [98]. Ο συγκεκριμένος κλάδος ονομάζεται ανάλυση καλαθιού αγοράς και αφορά στην ανακάλυψη μοτίβων συσχετίσεων σε συναλλαγές λιανικής. Η ανάλυση αυτή θέτει τις βάσεις για την ανάπτυξη εφαρμογών όπως η ομαδοποίηση προϊόντων, ο εντοπισμός εξαρτήσεων μεταξύ κατηγοριών αλλά και η δημιουργία προφίλ καταναλωτών. Μεταξύ αυτών, η εκμάθηση των προτιμήσεων των καταναλωτών έχει σκοπό να καθορίσει τις επιθυμίες των καταναλωτών ώστε, για παράδειγμα, να υποστηρίξει την πρόταση προϊόντων [99].

Η έλευση του ηλεκτρονικού εμπορίου άνοιξε το δρόμο για πολυάριθμες προόδους σε τεχνικές και μοντέλα τα οποία επιδιώκουν να βελτιώσουν την εμπειρία των καταναλωτών σε ηλεκτρονικά καταστήματα. Τα συστήματα προτάσεων χρησιμοποιούνται πλέον ευρέως σε εφαρμογές ηλεκτρονικού εμπορίου [100].

Στην ανάλυση των συναλλαγών των καταναλωτών, η αλλιώς ανάλυση καλαθιού αγοράς, η έρευνα έχει επικεντρωθεί αρχικά στην εφαρμογή της εξαγωγής κανόνων συσχέτισης. Η εξαγωγή κανόνων συσχέτισης περιλαμβάνει την ανάλυση συνολοστοιχείων (itemsets) και την εξαγωγή κανόνων που συσχετίζουν προϊόντα

μεταξύ τους. Σε αυτή την κατεύθυνση έχει επιτευχθεί σημαντική πρόοδος για παράδειγμα με την επιλογή μόνο των συνολοστοιχείων που δεν είναι παράγωγα άλλων [101]. Οι εξαγόμενοι κανόνες συσχέτισης είναι εύκολο να γίνουν κατανοητοί και πολλά νέα μοτίβα μπορούν να εντοπιστούν. Όμως, ο κατά κανόνα μεγάλος, αριθμός των κανόνων συσχέτισης δυσχεραίνει κατά πολύ την κατανόηση και την επεξήγηση τους από ανθρώπους.

Σε μια διαφορετική κατεύθυνση, έχουν διερευνηθεί διάφορες τεχνικές συνεργατικής διήθησης για χρήση σε δεδομένα καλαθιού αγοράς [102][103]. Η συνεργατική διήθηση όμως μπορεί να εμφανίσει χαμηλότερες επιδόσεις για την πραγματοποίηση υπολογισμών σε μεγάλα σύνολα δεδομένων, καθώς βασίζεται στην μνήμη. Επιπρόσθετα, η φύση του εμπορίου ως τομέας εφαρμογής των συστημάτων προτάσεων θέτει ορισμένους περιορισμούς στη συνεργατική διήθηση, όπως η έλλειψη αξιολογήσεων από τους καταναλωτές και η χαμηλή ποιότητα των προτάσεων [104], [105].

Η ανάπτυξη συστημάτων προτάσεων σε ανάλυση καλαθιού αγοράς έχει συναντήσει στο παρελθόν αρκετές προκλήσεις, δυο από τις οποίες είναι οι παρακάτω.

Πρώτον, οι τεχνικές που έχουν προταθεί στη βιβλιογραφία προσφέρουν μόνο μια περιορισμένη εποπτική εικόνα των προτιμήσεων των καταναλωτών. Αν και μπορούμε να χρησιμοποιήσουμε τους κανόνες συσχέτισης για να προβλέψουμε τα προϊόντα που θα απαρτίσουν το υπόλοιπο ενός καλαθιού αγοράς, αυτό που απουσιάζει είναι μια γενικότερη εικόνα των προτιμήσεων του χρήστη και των σχέσεων μεταξύ των προτιμήσεων αυτών.

Δεύτερον, η ποιότητα των προτάσεων που παράγονται αλλά και η ταχύτητα με την οποία παράγονται μπορεί να επηρεαστούν αρνητικά από τον τύπο του συνόλου δεδομένων. Οι κανόνες συσχέτισης τείνουν να αγνοούν μεγάλα στοιχειοσύνολα ενώ οι τεχνικές συνεργατικές διήθησης που βασίζονται στην μνήμη υστερούν στη δυνατότητα επέκτασης, καθώς όσο μεγαλώνει το σύνολο δεδομένων τόσο αυξάνονται οι απαιτήσεις του συστήματος σε μνήμη και επεξεργαστική ισχύ [106]. Ακόμη, τα συστήματα προτάσεων που βασίζονται στο περιεχόμενο δεν μπορούν να χρησιμοποιηθούν με ευκολία στις περισσότερες περιπτώσεις λιανικών συναλλαγών, καθώς η πληροφορία για τα αντίστοιχα προϊόντα είναι συνήθως ελλιπής ή μη διαθέσιμη.

Συνεισφορά της Διατριβής

Στην προσέγγιση που παρουσιάζεται στη διδακτορική διατριβή χρησιμοποιούμε την τεχνική των πιθανοτικών μοντέλων θεμάτων. Η συγκεκριμένη τεχνική οδηγεί σε ένα μοντέλο θεμάτων που να αντικατοπτρίζει τις προτιμήσεις των καταναλωτών και να αποτελεί τη βάση για να παραχθούν προτάσεις προϊόντων. Για την εξαγωγή αυτή χρησιμοποιούνται δυο συναφείς τεχνικές. Πρώτον, η ανάλυση ομάδων προϊόντων που εμφανίζονται μαζί συχνά στο ιστορικό των αγοραστών (λανθάνοντες χρήστες). Δεύτερον, η ανάλυση ομάδων προϊόντων που εμφανίζονται συχνά μαζί σε μεμονωμένες επισκέψεις των αγοραστών (λανθάνοντα καλάθια). Στη συνέχεια, τα μοντέλα θεμάτων που εξάχθηκαν χρησιμοποιούνται για την παραγωγή προτάσεων σε καταναλωτές.

Η μεθοδολογία που προτείνουμε διαθέτει δυο χαρακτηριστικά. Πρώτον προσφέρει μια εποπτική εικόνα της συμπεριφοράς των καταναλωτών με βάση των δεδομένων των καλαθιών αγοράς. Δεύτερον μπορεί να προβλέψει με ταχύτητα και με ακρίβεια τα προϊόντα που θα επιλέξει ένας καταναλωτής.

Η συγκεκριμένη μεθοδολογία υλοποιείται σε ένα σύστημα λογισμικού που παράγει προτάσεις. Παραλλαγές των τεχνικών δημιουργίας προτάσεων υλοποιήθηκαν και εφαρμόστηκαν στο σύνολο δεδομένων μιας ελληνικής υπεραγοράς όπου αξιολογήθηκαν θετικά ξεπερνώντας σε επίδοση τους κανόνες συσχέτισης.

Η προτεινόμενη προσέγγιση αποτελεί ένα μοντέλο καταναλωτικής συμπεριφοράς που μπορεί να χρησιμοποιηθεί για την πρόταση προϊόντων σε καταναλωτές. Η μελέτη που παρουσιάζεται οδηγεί στο συμπέρασμα ότι η λανθάνουσα ανάλυση θεμάτων αποτελεί έναν κατάλληλο και αποτελεσματικό τρόπο για την ανάλυση των δεδομένων αγορών, που ξεπερνά σε επιτυχία την εξαγωγή κανόνων συνάφειας που είναι η μέθοδος που συνήθως χρησιμοποιείται στην ανάλυση καλαθιού αγοράς. Τα πιθανοτικά μοντέλα θεμάτων όχι μόνο παρέχουν μια εικόνα των προτιμήσεων των καταναλωτών αλλά και μπορούν να υποστηρίξουν ένα σύστημα προτάσεων προϊόντων.

Σχετικές Δημοσιεύσεις

Η ερευνητική εργασία που αφορά στο συγκεκριμένο θέμα έχει παρουσιαστεί σε συνέδριο με τα παρακάτω στοιχεία:

- K. Christidis, D. Apostolou, and G. Mentzas, “Exploring Customer Preferences with Probabilistic Topics Models.” In Proceedings of Preference Learning Workshop, European Conference of Machine Learning September 2010, Barcelona, Spain

4.3.4 Περιγραφές των Διαθέσιμων Προϊόντων

Διατύπωση Ερωτήματος

Το ερευνητικό ερώτημα αφορά την δυνατότητα ενός συστήματος προτάσεων να υποστηρίξει του συμμετέχοντες σε μια ηλεκτρονική αγορά δημοπρασιών στις διαφορές δραστηριότητές τους με βάση το περιεχόμενο των περιγραφών των προϊόντων. Στην υποστήριξη αυτή περιλαμβάνεται τόσο η διαδικασία επιλογής προϊόντων για αγορά όσο και η διαδικασία περιγραφής, τιμολόγησης και πώλησης προϊόντων.

Σχέση με τη Βιβλιογραφία

Η συγκεκριμένη απουσία αντίστοιχων συστημάτων γίνεται εμφανής στους συμμετέχοντες μιας ηλεκτρονικής αγοράς δημοπρασιών ως δυο προκλήσεις, που αφορούν τους αγοραστές και τους πωλητές αντίστοιχα.

Η πρώτη πρόκληση αφορά την υπερφόρτωση πληροφορίας για τους αγοραστές η οποία οφείλεται στον μεγάλο αριθμό αντικειμένων που είναι διαθέσιμα για αγορά στις ηλεκτρονικές αγορές [2]. Ο αριθμός των διαθέσιμων αντικειμένων, ακόμη και στις πιο στενά ορισμένες κατηγορίες προϊόντων υπερβαίνει κατά πολύ τον αριθμό αντικειμένων που είναι διαθέσιμα σε μη ηλεκτρονικά καταστήματα αλλά και στα περισσότερα ηλεκτρονικά καταστήματα. Η πληροφορία που παρέχεται στην περιγραφή κάθε αντικειμένου είναι εκτενής αλλά και πυκνή. Απαιτεί από τον πιθανό αγοραστή να διαβάσει προσεκτικά και να πραγματοποιήσει έναν αριθμό συγκρίσεων με ανταγωνιστικά προϊόντα και αξιολογήσεων των χαρακτηριστικών του τρέχοντος προϊόντος. Συχνά ο αγοραστής δεν μπορεί να διακρίνει εύκολα αντικείμενα τα οποία του ταιριάζουν και ως αποτέλεσμα είτε εγκαταλείπει την αναζήτηση είτε συμβιβάζεται με κάτι που δεν καλύπτει τις προσδοκίες του. Επιπρόσθετα, καθώς ο αγοραστής δεν έχει μια εποπτική εικόνα των διαθέσιμων ενδιαφερόντων αντικειμένων, δυσκολεύεται να χαράξει μια στρατηγική προσφορών προς τον πωλητή.

Η δεύτερη πρόκληση αφορά την έλλειψη αντίληψης του ανταγωνισμού στο περιβάλλον των πωλητών. Η ευκολία της χρήσης ηλεκτρονικών αγορών έχει δημιουργήσει έναν καινούριο τύπο ερασιτεχνών πωλητών που πωλούν μεταχειρισμένα αντικείμενα. Σε αυτό το περιβάλλον πωλητές από διαφορετικές περιοχές και κράτη ανταγωνίζονται για την προσοχή των υποψηφίων αγοραστών. Αυτές οι συνθήκες δεν είναι ευνοϊκές για τους νέους επίδοξους πωλητές. Οι νεοεισερχόμενοι στις ηλεκτρονικές αγορές δεν έχουν ξεκάθαρη εικόνα των ανταγωνιστών τους και δεν έχουν κάποια υποστήριξη στις εργασίες που θα είναι οι πλέον κρίσιμες για την δραστηριότητά τους στην αγορά: την περιγραφή του αντικειμένου που θέλουν να πουλήσουν και την επιλογή της τιμής που θα θέσουν ως αρχική και ως τελική.

Για να αντιμετωπιστούν αυτές οι προκλήσεις στις ηλεκτρονικές αγορές έχει προταθεί η χρήση συστημάτων προτάσεων, τόσο συνεργατικής διήθησης όσο και με βάση το περιεχόμενο [37], [107].

Πιο συγκεκριμένα στην βιβλιογραφία έχουν αναπτυχθεί μέθοδοι οι οποίες οδηγούν σε μοντέλα και βάσεις γνώσης που χρησιμοποιούνται για την υποστήριξη συστημάτων προτάσεων. Έχει εφαρμοστεί η μέθοδος γενετικών αλγορίθμων k-μέσων ώστε να διαχωριστεί σε τμήματα η ηλεκτρονική αγορά, οδηγώντας στην ανάπτυξη ενός εργαλείου για προεπεξεργασία που απαιτείται στα συστήματα προτάσεων [108]. Σε μια διαφορετική μελέτη προτείνεται η χρήση μιας επιβλεπόμενης μεθόδου εκμάθησης η οποία οδηγεί στην εξαγωγή των σημασιολογικών κατηγοριών κάθε όρου που μπορεί να βρεθεί στους τίτλους των προϊόντων [109]. Οι κατηγορίες αυτές χρησιμοποιούνται στη συνέχεια για τη βελτίωση των αποτελεσμάτων της αναζήτησης και αυτή η βελτίωση παρατηρείται σε σύνολα δεδομένων από τον πραγματικό κόσμο. Τέλος έχει προταθεί ένα αποδοτικό πλαίσιο που επεκτείνει και χρησιμοποιεί μοντέλο θεμάτων. Το εκτεταμένο μοντέλο που προτείνεται διαθέτει τη δυνατότητα χρήσης ασταθών δυαδικών παρατηρήσεων και αξιολογείται θετικά σε σύνολα δεδομένων μεγάλων ιστοτόπων ηλεκτρονικού εμπορίου [110].

Στη γενική τους μορφή, τα συστήματα προτάσεων στο ηλεκτρονικό εμπόριο έχουν σχεδιαστεί ώστε να ομαδοποιούν τους καταναλωτές και τα προϊόντα αντίστοιχα, και να παράγουν κάποιους κανόνες συσχέτισης μεταξύ καταναλωτών και προϊόντων. Στη βιβλιογραφία έχει προταθεί ένα εννοιολογικό πλαίσιο υποστήριξης αποφάσεων για ένα online εμπορικό κέντρο, στο οποίο τοποθετείται μια μηχανή αναζήτησης εξωτερικά και ένα σύστημα προτάσεων εσωτερικά [111].

Επίσης, έχει προταθεί μια τεχνική άμεσης υποστήριξης αποφάσεων η οποία βασίζεται σε ένα μαθηματικό μοντέλο που περιγράφει τα χαρακτηριστικά των αγοραστών και τα κέρδη των προμηθευτών [112]. Αυτό το μοντέλο έχει αναπτυχθεί ώστε ένα προϊόν να μπορεί να προταθεί στον κατάλληλο άνθρωπο προσφέροντας το μέγιστο κέρδος για την επιχείρηση. Ακόμη, έχει περιγραφεί μια υπηρεσία επιλογής προϊόντος η οποία επιστρέφει μια ομάδα αποτελεσμάτων που ομαδοποιούνται σύμφωνα με την πιθανή τους συνάφεια με τον χρήστη [113]. Τέλος, έχει προταθεί μια λειτουργική μονάδα συστημάτων προτάσεων για το ηλεκτρονικό εμπόριο, όπου λαμβάνονται υπόψη οι στρατηγικές μάρκετινγκ και η πολυπλοκότητα των σχημάτων των διεπαφών χρήστη [114]. Στην ίδια μελέτη προτείνεται μια τεχνική τύπου συνεργατικής διήθησης με επιρροές κλίκας για την πρόβλεψη των προτιμήσεων των χρηστών, ενώ επιπρόσθετα πραγματοποιείται αξιολόγηση του συστήματος.

Οι σχετικές εργασίες που έχουν πραγματοποιηθεί στην περιοχή δεν αντιμετωπίζουν το πρόβλημα της εξαγωγής της λανθάνουσας σημασιολογίας από την τεράστια βάση δεδομένων ενός τόπου δημοπρασιών για την υποστήριξη ενός συστήματος προτάσεων. Οι υπάρχουσες υλοποιήσεις αποτυγχάνουν να εκμεταλλευτούν το μη δομημένο περιεχόμενο ώστε να υποστηρίξουν τις δραστηριότητες της πρότασης συναφών αντικειμένων και του εντοπισμού σημαντικών όρων για την συγγραφή της περιγραφής ενός αντικειμένου προς πώληση. Η πρόταση συναφών αντικειμένων έχει να κάνει με τη δυνατότητα του συστήματος να εντοπίζει αντικείμενα τα οποία μοιάζουν μεταξύ τους και μπορεί να ενδιαφέρουν τόσο τον πωλητή όσο και τον επίδοξο αγοραστή. Η πρόταση σημαντικών όρων αφορά την δυνατότητα για εντοπισμό και πρόταση λέξεων που αφορούν το αντικείμενο που περιγράφεται αλλά δεν βρίσκονται ακόμη στην περιγραφή του.

Συνεισφορά της Διατριβής

Στο πλαίσιο της διδακτορικής διατριβής προτείνεται μια μεθοδολογία για την εκμετάλλευση του μη δομημένου κειμένου που βρίσκεται σε ηλεκτρονικές αγορές δημοπρασιών για την παραγωγή προτάσεων που απευθύνονται σε καταναλωτές και πωλητές. Οι προτάσεις αυτές έχουν δυο σκέλη. Πρώτον, την πρόταση σχετικών προϊόντων με αυτό που βλέπει ο πιθανός πωλητής. Δεύτερον, την πρόταση συναφών αντικειμένων αλλά και σημαντικών όρων που αφορούν το προϊόν που θέλει να πουλήσει ένας πωλητής. Η δημιουργία αυτών των προτάσεων

γίνεται με βάση τα πιθανοτικά μοντέλα θεμάτων που έχουν εξαχθεί από το περιεχόμενο της ηλεκτρονικής αγοράς.

Η μεθοδολογία αυτή υλοποιήθηκε σε ένα λογισμικό σύστημα προτάσεων. Το σύστημα αυτό εφαρμόστηκε σε ένα σύνολο δεδομένων που προέρχεται από ένα διεθνή ιστότοπο ηλεκτρονικών δημοπρασιών υψηλής επισκεψιμότητας. Αξιολογήσαμε τόσο τα μοντέλα θεμάτων που δημιουργήθηκαν με βάση το σύνολο θεμάτων όσο και τη χρήση του συστήματος προτάσεων. Τα αποτελέσματα της αξιολόγησης οδηγούν στο συμπέρασμα ότι η μεθοδολογία μας μπορεί να οδηγήσει σε ένα σταθερό και χρήσιμο σύστημα προτάσεων σε μια ηλεκτρονική αγορά δημοπρασιών.

Σχετικές Δημοσιεύσεις

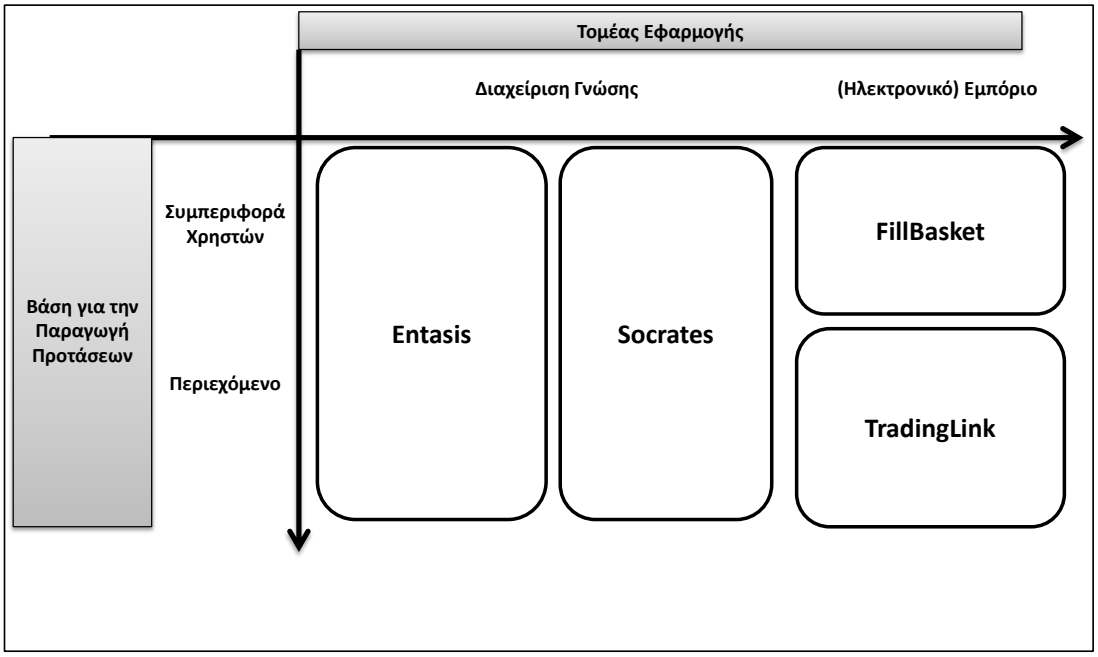
Η ερευνητική εργασία που αφορά το συγκεκριμένο θέμα έχει παρουσιαστεί στις παρακάτω εργασίες:

- K. Christidis and G.Mentzas, “A Topic-based Recommender System for Electronic Marketplace Platforms”, Expert Systems with Applications, in print, 2013
- K. Christidis and G. Mentzas. “A Topic-based Recommender System for Electronic Marketplace Platforms”, In Proceedings of the 24st IEEE International Conference on Tools with Artificial Intelligence (ICTAI ’12), IEEE, Athens, Greece

4.4 Συμπεράσματα

Τα συστήματα προτάσεων όπως περιγράφονται στη βιβλιογραφία αποτελούν ώριμες λύσεις που αντιμετωπίζουν εν μέρει το πρόβλημα της υπερφόρτωσης πληροφορίας. Εντούτοις, ένας αριθμός ελλείψεων και προβλημάτων παραμένουν να λυθούν. Οι ελλείψεις με τις οποίες ασχολούμαστε στην παρούσα διατριβή αφορούν κυρίως στις δυνατότητες των συστημάτων προτάσεων να συνεκτιμούν και να αλληλεπιδρούν με τις υπάρχουσες γνωσιακές δομές, με τους συνεργάτες και τις δεξιότητες τους, και τέλος με τους πωλητές και τους αγοραστές στο εμπόριο και στο ηλεκτρονικό εμπόριο.

Για να αντιμετωπιστούν αυτές οι ελλείψεις, προτείνονται τέσσερις διαφορετικές μεθοδολογίες που οδήγησαν στον σχεδιασμό, στην υλοποίηση και στην αξιολόγηση αντίστοιχων συστημάτων λογισμικού. Αυτά τα τέσσερα συστήματα έχουν τα διακριτικά ονόματα Entasis, Socrates, FillBasket και TradingLink και συνδέονται με τα αντίστοιχα ερευνητικά ερωτήματα (βλ. Εικόνα 4.4).



Εικόνα 4.4 Συνεισφορά της Διατριβής με Βάση τους Άξονες

Πρώτον, στη διατριβή παρουσιάζεται μια προσέγγιση που μπορεί να υποστηρίξει την ενσωμάτωση γνώσης από τον τομέα εφαρμογής στο σύστημα προτάσεων. Το Entasis αποτελεί μια μεθοδολογία και ένα αντίστοιχο λογισμικό σύστημα προτάσεων περιεχομένου και επισημειώσεων ενσωματωμένο σε εταιρικό κοινωνικό λογισμικό. Η πλήρης περιγραφή του Entasis καθώς και η αξιολόγησή του περιέχεται στο κεφάλαιο 5, σελίδα 141.

Δεύτερον, προτείνεται μια μεθοδολογία για την εξαγωγή και χρησιμοποίηση των δεξιοτήτων των χρηστών για την δημιουργία προτάσεων. Το Socrates αποτελεί ένα σύστημα προτάσεων για κοινότητες ανάπτυξης λογισμικού που βασίζεται σε πιθανοτικά μοντέλα θεμάτων και αποτυπώνει τις δεξιότητες των προγραμματιστών όπως αυτές λαμβάνονται από διαφορετικά εργαλεία συνεργασίας. Στο κεφάλαιο 6 στην σελίδα 165 παρουσιάζεται το σύστημα Socrates και η αξιολόγησή του.

Τρίτον προτείνουμε μια προσέγγιση για την βελτιωμένη ενσωμάτωση της καταναλωτικής συμπεριφοράς των επισκεπτών υπεραγορών σε ένα σύστημα

προτάσεων. Η προσέγγιση FillBasket αφορά σε ένα σύστημα προτάσεων που με βάση τα μοντέλα θεμάτων που παράγονται από την συμπεριφορά των καταναλωτών δημιουργεί προτάσεις σε ένα περιβάλλον υπεραγοράς. Η μεθοδολογία, το λογισμικό σύστημα FillBasket και η αξιολόγησή του παρουσιάζονται στο κεφάλαιο 7, στην σελίδα 183.

Τέταρτον, προτείνουμε μια προσέγγιση για την ενσωμάτωση στοιχείων των περιγραφών αντικειμένων προς πώληση σε ένα σύστημα προτάσεων μιας ηλεκτρονικής αγοράς. Το TradingLink είναι ένα σύστημα προτάσεων που με βάση τα μοντέλα θεμάτων που παράγονται από κείμενο δημιουργεί προτάσεις που απευθύνονται σε πωλητές και υποψήφιους αγοραστές σε ένα περιβάλλον ηλεκτρονικής αγοράς δημοπρασιών. Μία εκτενής περιγραφή της προτεινόμενης μεθοδολογίας καθώς και η αξιολόγησή του αντίστοιχου συστήματος περιλαμβάνεται στο κεφάλαιο 8, σελίδα 205.

Τα τέσσερα κεφάλαια που ακολουθούν περιγράφουν αναλυτικά την συνεισφορά της διατριβής σε κάθε ερευνητικό ερώτημα. Το περιεχόμενο κάθε κεφαλαίου αντιστοιχίζεται με ένα από τα ερευνητικά ερωτήματα που τέθηκαν παραπάνω. Η αντιστοίχιση παρουσιάζεται στην Εικόνα 4.5

Εικόνα 4.5 Αντιστοίχηση Κεφαλαίων με Προτεινόμενες Μεθοδολογίες

5 Entasis - Εταιρικό Κοινωνικό Λογισμικό

Στο κεφάλαιο αυτό παρουσιάζεται το Entasis, μια μεθοδολογία και το αντίστοιχο σύστημα για υποστήριξη αναζήτησης και συστημάτων προτάσεων σε εταιρικό κοινωνικό λογισμικό.

Στις παρακάτω σελίδες εισάγεται το αντικείμενο του Entasis, αναφέρεται η σχετική βιβλιογραφία και περιγράφεται η δική μας προσέγγιση. Ακολούθως, γίνεται μια εκτενής παρουσίαση μιας μελέτης εφαρμογής του συστήματος, και ακολουθεί η αξιολόγηση του συστήματος και τα συμπεράσματα.

Για να γίνει εφικτή η υποστήριξη των χρηστών στις καθημερινές τους δραστηριότητες που αφορούν στη διαχείριση γνώσης χρησιμοποιήθηκε η τεχνολογία των πιθανοτικών μοντέλων θεμάτων. Τη βάση των μοντέλων αποτέλεσε το περιεχόμενο που εισήγαγαν οι χρήστες με μορφή κειμένου, ενώ η γενικότερη περιοχή εφαρμογής του συστήματος είναι η διαχείριση γνώσης. Στη συνεισφορά του παρόντος κεφαλαίου συγκαταλέγεται η περιγραφή της προσέγγισης μας για την ενσωμάτωση της γνώσης του τομέα εφαρμογής στο σύστημα προτάσεων.

5.1 Εισαγωγή

Κατά τη δημιουργία του ο παγκόσμιος ιστός δεν αποτέλεσε πολλά παραπάνω από μια συλλογή ημι-στατικών σελίδων συνδεδεμένων μεταξύ τους, που δεχόντουσαν ανεξάρτητες επισκέψεις χρηστών. Σήμερα παρουσιάζει διαφορετικά χαρακτηριστικά από αυτά που παρουσίαζε κατά τη πρώτη περίοδο μετά τη δημιουργία του, έχοντας πάρει την ονομασία ιστός 2.0 (Web 2.0). Το γενικότερο πλαίσιο του ιστού 2.0 είναι η χρήση του ως ένα κοινωνικό και διαδραστικό εργαλείο [19], ενώ ένα ακόμα στοιχείο του είναι η ικανότητά του να εκμεταλλεύεται την συλλογική νοημοσύνη (collective intelligence).

Αυτή η τάση συναντάται και στο εσωτερικό των εταιρικών οργανισμών όπου ονομάστηκε επιχείρηση 2.0 (enterprise 2.0) [84]. Η επιχείρηση 2.0 βασίζεται στην υποστήριξη κοινοτήτων εργαζομένων στο εσωτερικό και στο εξωτερικό των εταιριών με χρήση κοινωνικών και συμμετοχικών δικτυακών εργαλείων και την ευθυγράμμιση τους με τις ανάγκες της επιχείρησης. Μια σειρά τεχνολογιών όπως σελίδες wiki, ιστολόγια και λογισμικό άμεσης επικοινωνίας μπορούν να

μετατρέψουν το παραδοσιακό εταιρικό δίκτυο σε μια πλατφόρμα συνεργασίας που αλλάζει μορφή για να συμβαδίζει με τον τρόπο που πραγματοποιούνται οι εργασίες.

Η επιχείρηση 2.0 είναι ήδη μια πραγματικότητα στο εσωτερικό πολλών εταιριών παγκοσμίως ενώ πολλές ακόμη εξετάζουν την χρήση τέτοιων εργαλείων [115]. Ένας νέος τύπος επιχειρήσεων έχει αρχίσει να αναδεικνύεται – επιχειρήσεις που χρησιμοποιούν τεχνολογίες του συνεργατικού ιστού 2.0 εντατικά ώστε να συντονίσουν τις προσπάθειες των εργαζομένων εσωτερικά αλλά και να επεκτείνουν την δραστηριότητα της επιχείρησης σε πελάτες, συνεργάτες και προμηθευτές [82]. Οι επιχειρήσεις συνολικά βελτιώνονται στη χρήση κοινωνικών τεχνολογιών, στην προσπάθεια τους να βελτιώσουν τις επιχειρησιακές δραστηριότητες τους και να εκμεταλλευτούν νέες ευκαιρίες στην αγορά [116].

Στη βιβλιογραφία με βάση τις λειτουργίες που προέκυψαν από τον Ιστό 2.0 προτείνεται ένα πλαίσιο με το όνομα SLATES όπου περιγράφει τους βασικούς άξονες της χρήσης εταιρικού κοινωνικού λογισμικού [84]. Η λέξη SLATES αποτελεί ένα ακρωνύμιο για τις λέξεις search (αναζήτηση), links (σύνδεσμοι), authoring (συγγραφή), tags (επισημειώσεις), extensions (επεκτάσεις) και signals (σήματα). Η αναζήτηση περιλαμβάνει στοιχεία πλοήγησης και αναζήτησης με λέξεις κλειδιά και παίζει ένα πρωταρχικό ρόλο στην ανάκτηση αποτελεσμάτων προηγούμενης εργασίας (εγγράφων, κ.α.). Οι σύνδεσμοι αναφέρονται στην ικανότητα μίας εκτεταμένης ομάδας ανθρώπων να σχηματίζουν δεσμούς στο εσωτερικό του συστήματος μεταξύ διαφορετικών πόρων. Τα εργαλεία συγγραφής επιτρέπουν την δημιουργία περιεχομένου τόσο από μεμονωμένα άτομα όσο και από ομάδες. Το περιεχόμενο αυτό μπορεί να συμπληρωθεί με επιπρόσθετη πληροφορία με τη μορφή των επισημειώσεων. Οι επισημειώσεις με τη σειρά τους δίνουν την δυνατότητα της πλοήγησης και της ανάκτησης περιεχομένου που έχει δημιουργηθεί στο παρελθόν (ιστολόγια, σελίδες wiki) ή που έχει εισαχθεί στο σύστημα (εξωτερικοί σύνδεσμοι, εικόνες, κ.α.). Οι επεκτάσεις αναφέρονται στην υποστήριξη εφαρμογών όπως τα συστήματα προτάσεων που μπορούν να προβλέψουν τι μπορεί να είναι χρήσιμο στους χρήστες και τους το προτείνουν. Τέλος, τα σήματα αφορούν σε τεχνολογίες όπως το RSS που επιτρέπουν στους χρήστες να μαθαίνουν άμεσα τις ειδήσεις από την επιχείρηση όπου εργάζονται.

Οι άξονες του πλαισίου SLATES, μερικές φορές σε συνδυασμό μεταξύ τους, παρέχουν ένα μέσο για τη βελτίωση της δημιουργίας, της επεξεργασίας και της διάχυσης της πληροφορίας στο εσωτερικό των οργανισμών. Τα νέα εργαλεία και

εφαρμογές, αλλά και μια σημαντική μεταστροφή στην κουλτούρα της συνεργασίας στο εσωτερικό των οργανισμών έχουν οδηγήσει στην δημιουργία μεγάλης ποσότητας πληροφορίας. Με το εταιρικό κοινωνικό λογισμικό, οι εργαζόμενοι και οι εξωτερικοί συνεργάτες δημιουργούν πληροφορία είτε άμεσα είτε σχηματίζοντας συνδέσμους στον ιστό. Το εταιρικό κοινωνικό λογισμικό δίνει πρόσβαση σε ένα μεγάλο όγκο πληροφορίας στο εσωτερικό της επιχείρησης και οι χρήστες μπορούν να επωφεληθούν αλλά και να αναλάβουν δράση. Όμως, καθώς ο όγκος των πληροφοριών μπορεί να είναι υπερβολικά μεγάλος για επεξεργασία και αξιολόγηση από ανθρώπους, αποτελεί πρόκληση η εκμετάλλευση αυτής της πληροφορίας με χρήση λογισμικού μέσω της μετατροπής της σε χρήσιμα κομμάτια γνώσης.

Η εργασία με πληροφορία η οποία είναι μη δομημένη ή αμφισβητούμενης ποιότητας αποτελεί μια ακόμη πρόκληση του εταιρικού κοινωνικού λογισμικού. Η διαχείριση της μη δομημένης πληροφορίας συνήθως απαιτεί μια ανθρώπινη προοπτική: την διατύπωση προτιμήσεων, τη διασύνδεση με εξωτερικές ή εσωτερικές αναφορές, την κατηγοριοποίηση, το φιλτράρισμα, την προσθήκη πρόσθετων πληροφοριών και την αξιολόγηση της ποιότητας. Αυτή η δραστηριότητα των ανθρώπων βοηθάει άλλους εργαζόμενους να εξετάσουν, να καταναλώσουν ή να χρησιμοποιήσουν την πληροφορία στην εργασία τους. Επιπρόσθετα, η πρόοδος στις τεχνολογίες διαχείρισης γνώσης έχει θέσει τις βάσεις για την μη επιβλεπόμενη επεξεργασία της μη-δομημένης πληροφορίας, η οποία βοηθάει τους χρήστες να αναζητήσουν πληροφορία αλλά και να δεχτούν χρήσιμες συστάσεις που αφορούν σχετικές πληροφορίες.

Στις επιχειρήσεις έντασης γνώσης πολύ συχνά υπάρχει μεγάλος όγκος πληροφορίας που δεν γίνεται διαθέσιμος στους εργαζομένους. Ταυτόχρονα, οι εργαζόμενοι εργάζονται με μικρές και ατελείς γνωσιακές δομές (π.χ. ταξονομίες και οντολογίες). Αυτό πολύ συχνά οδηγεί σε χαμένο χρόνο κατά τις αναζητήσεις, απώλεια πληροφοριών και επανάληψη εργασιών που έχουν ήδη πραγματοποιηθεί στο παρελθόν.

5.2 Σχετικές Εργασίες

Η αναζήτηση και τα συστήματα προτάσεων σε κοινωνικά περιβάλλοντα, τόσο στο εσωτερικό όσο και στο εξωτερικό της επιχείρησης, τελευταία έχουν αποτελέσει ένα αντικείμενο έντονης έρευνας. Έχει προταθεί η χρήση των επισημειώσεων ως ανάδραση από τον χρήστη για την βελτίωση της αναζήτησης στο

εσωτερικό μιας εταιρίας [86]. Προτείνεται τόσο η άμεση όσο και η έμμεση ανάδραση των χρηστών και αποδεικνύεται με πρώιμα πειράματα η θετική επίδραση της μεθόδου που προτείνουν στην ποιότητα της αναζήτησης στο εσωτερικό δίκτυο της IBM. Επίσης, έχουν παρουσιαστεί μέθοδοι για το συνδυασμό ετερογενούς πληροφορίας για την βελτίωση της αναζήτησης [117]. Τα δεδομένα που χρησιμοποιούνται για τον εμπλουτισμό της εσωτερικής εταιρικής αναζήτησης προέρχονται από εφαρμογές του ιστού 2.0. Ένα σύστημα προτάσεων για κοινωνικούς συνδέσμους μπορεί να πάρει την μορφή ενός παιχνιδιού και να χρησιμοποιηθεί για να επιτευχθούν ατομικοί, συνεργατικοί ή εταιρικοί στόχοι [118]. Σε μια διαφορετική μελέτη, εξετάζεται η πρόταση πόρων από κοινωνικά μέσα στο εσωτερικό της επιχείρησης με βάση τους ανθρώπους, τις επισημειώσεις και τις σχέσεις μεταξύ τους [85]. Οι πληροφορίες αυτές συλλέγονται από διαφορετικές πηγές στο εσωτερικό της επιχείρησης και κάθε αντικείμενο που προτείνεται συνοδεύεται από μια σύντομη εξήγηση των ανθρώπων και των επισημειώσεων που οδήγησαν στην πρόταση. Έχει δειχθεί ότι η χρήση των τεχνολογιών του Ιστού 2.0 υπό όρους επιτρέπει σε εταιρίες την αύξηση της παραγωγικότητας αλλά και του συγκριτικού τους πλεονεκτήματος [119]. Η συγκεκριμένη μελέτη περιλαμβάνει αποτελέσματα συνεντεύξεων και παρατηρήσεις από επιχειρήσεις στο Ηνωμένο Βασίλειο. Επίσης στη βιβλιογραφία έχει σχεδιαστεί και αξιολογηθεί ένα σύστημα προτάσεων για την πρόταση άγνωστων ως εξωτερικούς συνεργάτες σε εργαζομένους [120]. Η προσέγγισή τους αξιολογείται θετικά από 516 συμμετέχοντες στο εσωτερικό μιας μεγάλης εταιρίας με ποιοτικά και ποσοτικά αποτελέσματα. Τέλος, υπάρχουν μέθοδοι που χρησιμοποιούν σημασιολογικές τεχνολογίες για να βελτιώσουν την αναζήτηση και την δημιουργία προτάσεων στα πλαίσια της προσέγγισης SemSLATES [87]. Αυτές οι προσπάθειες κυρίως χρησιμοποιούν ετερογενές περιεχόμενο φτιαγμένο από χρήστες, πληροφορίες από κοινωνικούς γράφους και σημασιολογικές τεχνολογίες.

Στα πλαίσια της προτεινόμενης προσέγγισης χρησιμοποιούμε τα αποτελέσματα της λανθάνουσας σημασιολογικής ανάλυσης για τον εντοπισμό λανθάνουσας σημασιολογίας στο περιεχόμενο. Τα πιθανοτικά μοντέλα θεμάτων έχουν ήδη χρησιμοποιηθεί για την υποστήριξη συστημάτων προτάσεων αλλά και αναζήτησης. Η πιθανοτική λανθάνουσα σημασιολογική ανάλυση (probabilistic latent semantic analysis) έχει χρησιμοποιηθεί για την εξόρυξη μοτίβων στην δραστηριότητα των χρηστών [121]. Προτείνεται ένα ενιαίο πλαίσιο για την ανακάλυψη και την ανάλυση των μοτίβων στην πλοήγηση των χρηστών στον ιστό,

και αποδεικνύεται η ευελιξία του πλαισίου όσον αφορά στον χαρακτηρισμό σχέσεων μεταξύ χρηστών και σελίδων. Τα μοντέλα θεμάτων έχουν εφαρμοστεί και για την πρόταση κοινοτήτων στις οποίες μπορεί να συμμετέχουν οι χρήστες [68]. Οι εμπειρικές συγκρίσεις που πραγματοποιούν με βάση τις k κορυφαίες προτάσεις δείχνουν ότι η λανθάνουσα κατανομή Dirichlet υπερτερεί σε απόδοση της εξαγωγής κανόνων συσχέτισης όταν προτείνεται μια λίστα 4 ή παραπάνω κοινοτήτων. Επίσης, έχει παρουσιαστεί μια προσέγγιση για την πρόταση άρθρων της Wikipedia, όπου τα αρχικά πειραματικά αποτελέσματα δείχνουν πως μπορεί να παράγει πολλά σχετικά αποτελέσματα τα οποία δεν καλύπτονται από τους συνδέσμους σε ένα δεδομένο άρθρο [72].

Σε διαφορετικές μελέτες έχουν προταθεί μεθοδολογίες για την πρόταση επισημειώσεων στους χρήστες. Έχει παρατηρηθεί ότι η χρήση πιθανοτικών μοντέλων θεμάτων παρουσιάζει καλύτερη ακρίβεια σε σχέση με την εξαγωγή κανόνων συσχέτισης [73]. Ακόμη, στη βιβλιογραφία έχει γίνει μια αξιολόγηση εκτός σύνδεσης στο σύνολο δεδομένων του Bibsonomy καθώς και μια online αξιολόγηση με χρήστες που αποδεικνύει την αποτελεσματικότητα της μεθόδου [74].

Τα λανθάνοντα θέματα έχουν χρησιμοποιηθεί για την εξόρυξη πληροφοριών από ιστολόγια και για την πρόταση σχετικών επισημειώσεων για συγκεκριμένες δημοσιεύσεις [67]. Στο εταιρικό περιβάλλον, έχει προταθεί η χρήση των λανθανόντων θεμάτων ώστε να εντοπιστούν και να οπτικοποιηθούν λανθάνουσες κοινότητες με παρεμφερή ενδιαφέροντα [75]. Επίσης, τα λανθάνοντα θέματα μπορούν να αποτυπώσουν τα ενδιαφέροντα των εργαζομένων αλλά και να χρησιμοποιηθούν για να προτείνονται σχετικά έγγραφα [65]. Το αντίστοιχο σύστημα που προτείνεται εντοπίζει τα βραχυπρόθεσμα και τα μακροπρόθεσμα θέματα για τα οποία ενδιαφέρονται οι χρήστες και παράγει προτάσεις με έναν υψηλότερο βαθμό ποικιλότητας μεταξύ θεμάτων.

Στην σχετική βιβλιογραφία μπορούμε να εντοπίσουμε δύο βασικές προκλήσεις που δεν έχουν αντιμετωπιστεί και αφορούν στην ανάπτυξη συστημάτων υποστήριξης χρηστών του εταιρικού κοινωνικού λογισμικού. Πρώτον η απουσία ενός ολοκληρωμένου πλαισίου υποστήριξης των χρηστών σε επίπεδο εταιρικού λογισμικού για την χρήση μη-δομημένου περιεχομένου. Δεύτερον, η έλλειψη μιας μεθοδολογίας που να μπορεί να συνδυάζει το μη δομημένο περιεχόμενο με τις υπάρχουσες γνωσιακές δομές. Σε αυτές τις προκλήσεις επιχειρούμε να ανταποκριθούμε στο παρόν κεφάλαιο.

5.3 Προτεινόμενη Προσέγγιση

5.3.1 Πλαίσιο

Για τα συστήματα διαχείρισης γνώσης πρωταρχικός στόχος είναι να εξασφαλιστεί ότι «οι σωστές πληροφορίες είναι διαθέσιμες ή παραδίδονται στους σωστούς ανθρώπους τη σωστή στιγμή». Δυο στρατηγικές παροχής πληροφοριών μπορούν να το πετύχουν αυτό, η προώθηση (push) και η αίτηση (pull) [122]. Με την εμφάνιση του εταιρικού κοινωνικού λογισμικού, οι προκλήσεις που αφορούν στο στόχο αυτό γίνονται ακόμη πιο έντονες λόγω της δραματικής αύξησης του διαθέσιμου περιεχομένου στον οργανισμό. Δύο τεχνολογίες που υποστηρίζουν τις παραπάνω στρατηγικές είναι η αναζήτηση και τα συστήματα συστάσεων.

Η αναζήτηση αφορά την εύρεση υλικού (συνήθως εγγράφων) μη δομημένης μορφής (συνήθως κειμένου) όπου ικανοποιεί μια ανάγκη για πληροφορία που αφορά μεγάλες συλλογές δεδομένων που συνήθως έχουν αποθηκευτεί σε υπολογιστές [16]. Η αναζήτηση αφορά οποιαδήποτε μορφή διεπαφής που επιτρέπει την πραγματοποίηση μιας σειράς από εργασίες ανάκτησης πληροφορίας από τον χρήστη. Τα συστήματα προτάσεων από την άλλη πλευρά, αντιμετωπίζουν το πρόβλημα της εκτίμησης της χρησιμότητας ή της αξιολόγησης αντικειμένων τα οποία ο χρήστης δεν έχει δει ακόμη. Για την επίλυση αυτού του προβλήματος, μια σειρά από διαφορετικούς τύπους συστημάτων προτάσεων έχουν προταθεί: συστήματα βασισμένα στο περιεχόμενο και στις προτιμήσεις των χρηστών, αλλά και συστήματα βασισμένα στην μνήμη ή σε μοντέλα [14].

Η αναζήτηση και τα συστήματα προτάσεων αποτελούν, κατά μια έννοια, δυο όψεις του ίδιου νομίσματος που χρησιμοποιούν παρόμοιες τεχνολογίες. Οι Belkin και Croft [15], μετά από εξέταση των δυο τομέων, εντοπίζουν μικρή διαφορά μεταξύ της ανάκτησης πληροφορίας και του φιλτραρίσματος της σε ένα αφηρημένο επίπεδο, καθώς ο στόχος αλλά και το πλαίσιο λειτουργίας τους είναι παρόμοιο.

Με τον όρο αναζήτηση συνήθως αναφερόμαστε σε μη προσωποποιημένες τεχνικές ανάκτησης περιεχομένου με χρήση ενός αριθμού λέξεων-κλειδιά. Με τον όρο συστήματα προτάσεων, υπονοούμε ότι ο χρήστης δεν ζήτησε άμεσα για υποστήριξη με κάποια δραστηριότητα μέσα στο σύστημα. Το σύστημα προτάσεων δρα σαν μηχανή αναζήτησης με την έννοια ότι ψάχνει για αντικείμενα που θα μπορούσαν να ενδιαφέρουν τον χρήστη. Τα συστήματα προτάσεων συνήθως

λαμβάνουν υπόψη το προφίλ του χρήστη, το πλαίσιο στο οποίο ενεργεί και άλλες πληροφορίες οι οποίες επιτρέπουν την ευφυή επεξεργασία και παρουσίαση της πληροφορίας.

Η σχεδίαση και ανάπτυξη συστημάτων διαχείρισης γνώσης εξαρτάται σε μεγάλο βαθμό από την αποτύπωση της πληροφορίας σε γνωσιακές δομές.

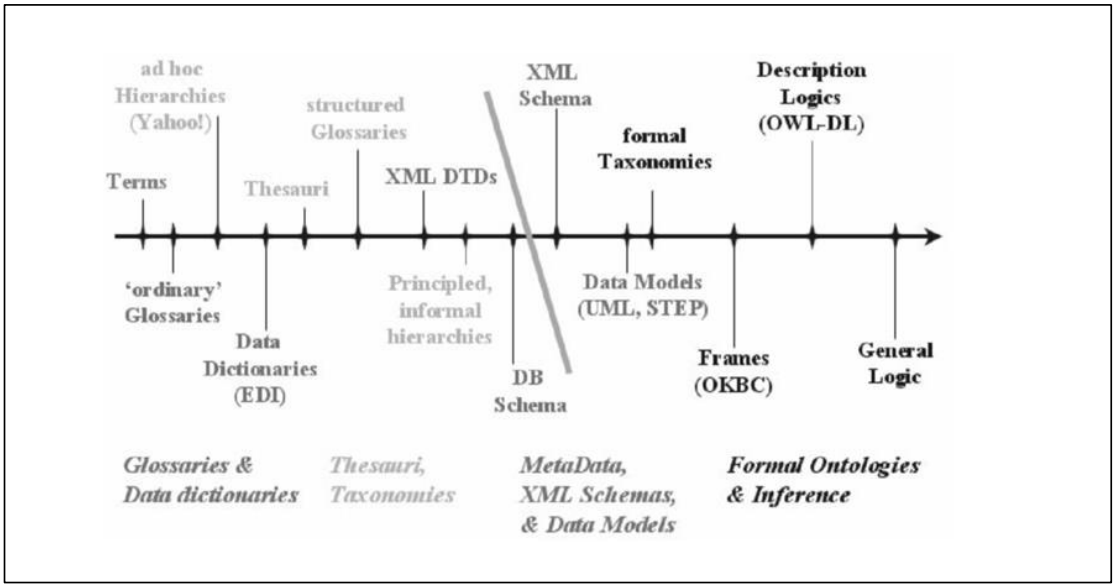
5.3.2 Γνωσιακές Δομές

Με τον όρο γνωσιακές δομές αναφερόμαστε στην ευρύτερη περιοχή των οντολογιών όπου αφορά [43] ένα λεξικό όρων που αναφέρονται σε αντικείμενα ενδιαφέροντος σε έναν τομέα αλλά και κάποιον ορισμό εννοιών για τους όρους αυτούς που ιδανικά βασίζεται στη λογική.

Οι οντολογίες πολύ συχνά αναπαριστούν διαφορετικούς τύπους αντικειμένων σε έναν δεδομένο θεματικό τομέα. Οι κλάσεις (ή έννοιες) αποτελούν ομάδες αντικειμένων και πολύ συχνά οργανώνονται με διαφορετικούς βαθμούς τυπικότητας.

Αυτός ο βαθμός τυπικότητας καθώς και η μέθοδος ορισμού των εννοιών των όρων διαφοροποιεί τα είδη γνωσιακών δομών. Αυτό επιτρέπει την απεικόνιση ενός συνεχούς φάσματος στο οποίο μπορούν να τοποθετηθούν τα διάφορα είδη γνωσιακών δομών (βλ. Εικόνα 5.1 [43]). Στο ένα άκρο εμφανίζονται οι δομές ελαφρού τύπου, όπου περιλαμβάνουν μόνο όρους. Στο άλλο έχουμε εκτενώς ορισμένες τυπικές οντολογίες. Ανάμεσα μπορούν να τοποθετηθούν διάφορες παραλλαγές γνωσιακών δομών όπως ταξονομίες, φολκσονομίες (κατάλογοι λέξεων), τυπικές οντολογίες, κ.α..

Η αναζήτηση και τα συστήματα προτάσεων δεν είναι υποχρεωτικό να συνεργάζονται με μη δομημένες πληροφορίες. Οι γνωσιακές δομές μπορούν να προσδώσουν ένα βαθμό οργάνωσης καθώς χρησιμοποιούνται για την κατηγοριοποίηση των διάφορων πόρων. Τόσο η επιλογή της πλατφόρμας λογισμικού όσο και η γενικότερη προσέγγιση που ακολουθείται από τον οργανισμό καθορίζουν το βαθμό της τυπικότητας στις γνωσιακές δομές.



Εικόνα 5.1 Συνεχές Φάσμα Γνωσιακών Δομών

Η αναζήτηση και τα συστήματα προτάσεων παρουσιάζουν συνάφεια στις τεχνολογίες που τα υποστηρίζουν. Και τα δυο περιλαμβάνουν την χρήση ενός ερωτήματος που απευθύνεται σε μία βάση περιεχόμενου και χρησιμοποιούν, στη γενική τους μορφή, αλγορίθμους ανάκτησης πληροφορίας για να εκτιμήσουν τη συνάφεια των αποτελεσμάτων με το ερώτημα. Για παράδειγμα, τόσο μια αναζήτηση με βάση ερωτήματα όσο και μια πρόταση βασισμένη στο περιεχόμενο μπορεί να μοντελοποιηθεί ως ένα πρόβλημα ταξινόμησης κειμένου όπου το περιεχόμενο που βασίζεται σε κείμενο αντιστοιχίζεται σε κλάσεις. Τα συστήματα προτάσεων που βασίζονται στην συμπεριφορά των χρηστών απαιτούν εξειδικευμένες καταγραφές των σχετικών συμβάντων (προτίμηση αντικειμένου, αξιολόγηση αντικειμένου, κ.α.) αλλά και επεξεργασία τους. Οι γνωσιακές δομές μπορούν να υποστηρίξουν τις διαδικασίες που ακολουθούνται τόσο από την αναζήτηση όσο και από τα συστήματα προτάσεων. Στα συστήματα προτάσεων, στοιχεία όπως οι έννοιες ή οι επισημειώσεις μπορούν να χρησιμοποιηθούν για τον υπολογισμό της σχετικότητας μεταξύ των αντικειμένων ενώ στην αναζήτηση η ύπαρξη μιας γνωσιακής δομής μπορεί να επιτρέχει την αναζήτηση με χρήση ευρύτερων ή στενότερων εννοιών (επέκταση ερωτημάτων).

Οι κλάσεις των πληροφοριακών πόρων μοντελοποιούνται με χρήση διάφορων τύπων γνωσιακών δομών. Παραδοσιακά, για παράδειγμα στα σύγχρονα συστήματα διαχείρισης γνώσης, το πρόβλημα της δημιουργίας γνωσιακών δομών για τον κάθε τομέα ενδιαφέροντος μιας επιχείρησης αντιμετωπίζεται με την χρήση

επιχειρηματικών ταξονομιών [123], δηλαδή σχημάτων ταξινόμησης που οργανώνουν έννοιες του τομέα της εταιρίας σε ιεραρχικές δενδρικές δομές. Οι ταξονομίες συνήθως χτίζονται με μια προσέγγιση από πάνω προς τα κάτω (top-down) όπου η περιοχή των θεμάτων διαχωρίζεται σε ολοένα και στενότερες κατηγορίες όπου αναλύονται με μεγαλύτερη λεπτομέρεια προς τα κάτω.

Όμως, το εταιρικό κοινωνικό λογισμικό, βασίζεται συχνά στη χρήση φολκσονομιών (folksonomies), μιας κατηγορίας δομών γνώσης που δημιουργούνται μη-τυπικά και συνεργατικά για την κατηγοριοποίηση των πληροφοριών. Οι φολκσονομίες δημιουργούνται από κάτω προς τα πάνω, με έναν τρόπο όπου η θεματική περιοχή χωρίζεται σε επιμέρους έννοιες που μπορούν να συντεθούν για να σχηματιστεί ένα πολύπλοκο θέμα, μέσω των κατάλληλων συνόλων κανόνων [124]. Διάφορες επεκτάσεις των παραπάνω βασικών τύπων γνωσιακών δομών έχουν προταθεί στη βιβλιογραφία, όπως η οντολογία με βάση χάρτες γνώσης [125], οι υβριδικές ταξονομίες-φολκσονομίες για ταξινομήσεις [126] και το κοινωνικό σημασιολογικό νέφος επισημειώσεων [20].

Μια πρόκληση για οποιαδήποτε εταιρική ταξονομία είναι η συντήρηση της και η εξέλιξη της ώστε να είναι επίκαιρη. Οι εταιρικές ταξονομίες μπορούν να γίνουν εύκολα παρωχημένες λόγω των ταχέως μεταβαλλόμενων συνθηκών στο επιχειρηματικό περιβάλλον. Από την άλλη πλευρά οι φολκσονομίες, οι οποίες είναι εξ ορισμού αυθαίρετες και εξελίσσονται, μπορεί να μην είναι σε θέση να αναπαριστούν με ακρίβεια τα θέματα που διέπουν το τομέα ενδιαφέροντος.

Οι περιορισμοί και οι προκλήσεις στην εφαρμογή των γνωσιακών δομών, καθώς και η αύξηση του όγκου των πληροφοριών που παράγονται από χρήστες εταιρικού κοινωνικού λογισμικού, μπορούν να οδηγήσουν σε μια υποδεέστερη ποιότητα αναζήτησης και αντικειμένων που προτείνονται [127]. Χρειαζόμαστε, λοιπόν, νέους τρόπους για να αποκαλύψουμε τη συνεχώς μεταβαλλόμενη γνώση του τομέα ενδιαφέροντος και να χτίσουμε γνωσιακές δομές.

Κίνητρό μας στο συγκεκριμένο σύστημα είναι η βελτίωση της αναζήτησης και της λειτουργίας των συστημάτων προτάσεων σε εταιρικό κοινωνικό λογισμικό. Για το σκοπό αυτό, έχουμε πραγματοποιήσει εφαρμογή σύγχρονων στατιστικών μεθόδων για την υποστήριξη της αναζήτησης και της δημιουργίας προτάσεων που λειτουργούν σε συνδυασμό και με τις δύο δομές, φολκσονομιών και ταξονομιών, με βάση τη γνώση. Χτίζουμε ένα μοντέλο θεμάτων με τις παρούσες γνώσεις στην πλατφόρμα και το αξιοποιούμε, προκειμένου να βελτιωθεί η αναζήτηση αλλά και

για να παρέχουμε συστάσεις για έγγραφα και επισημειώσεις, και με αυτό τον τρόπο στηρίζουμε τη δημιουργία περισσότερο κατάλληλων γνωσιακών δομών.

Προκειμένου να βελτιωθούν οι λειτουργίες αναζήτησης και δημιουργίας προτάσεων στο εταιρικό κοινωνικό λογισμικό, έχουμε επικεντρωθεί στην επέκταση των προσεγγίσεων που αφορούν τις γνωσιακές δομές, όπως φολκσονομίες και ταξονομίες, με προσθήκη λανθανόντων θεμάτων. Χρησιμοποιούμε πιθανοτικά μοντέλα θεμάτων ως τεχνικό υπόβαθρο για την αποκάλυψη αυτών των λανθανόντων θεμάτων.

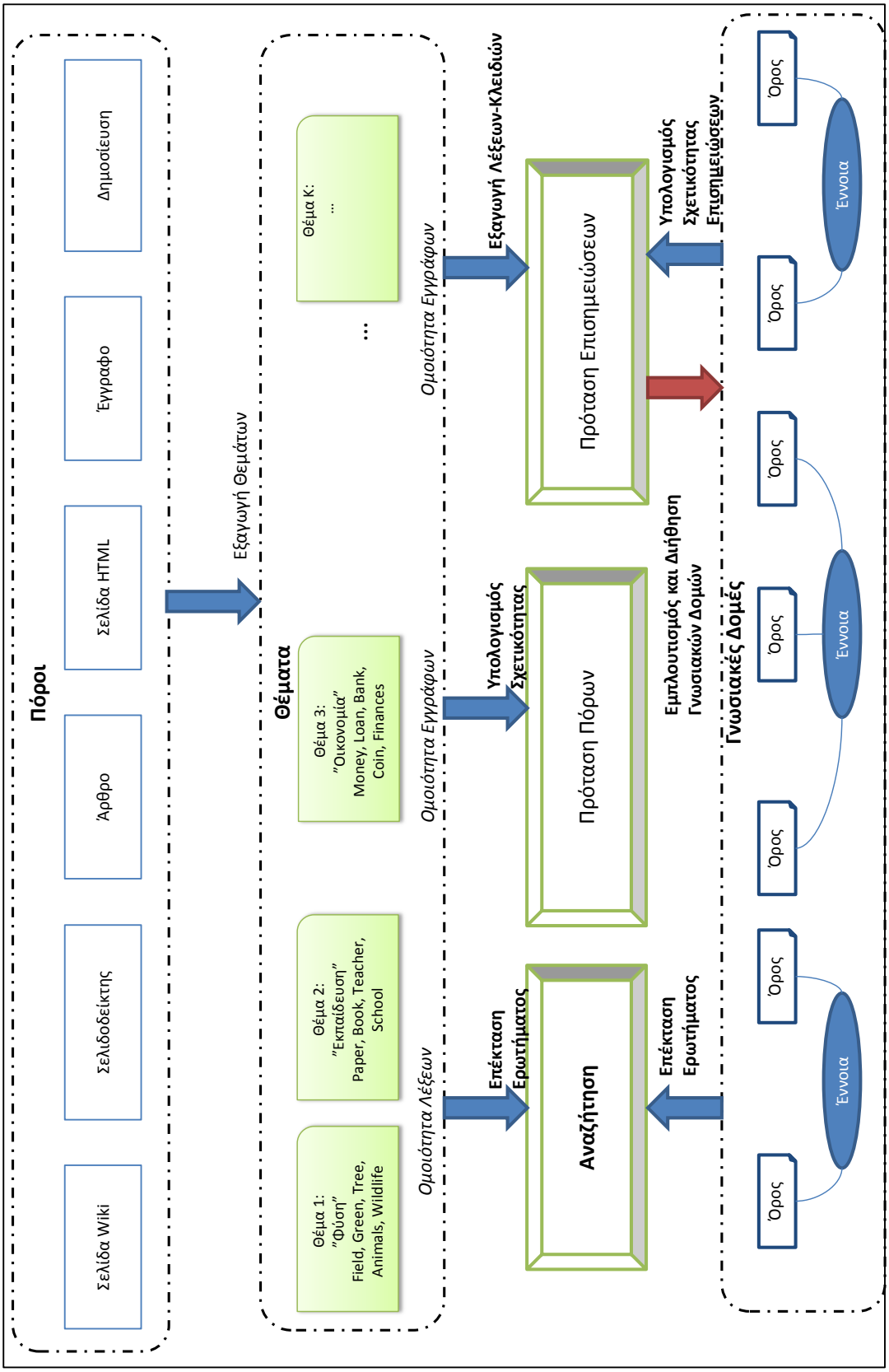
Τα πιθανοτικά μοντέλα θεμάτων βασίζονται στην ιδέα ότι τα έγγραφα είναι μείγματα θεμάτων, ενώ ένα θέμα ορίζεται ως μια κατανομή πιθανότητας ανάμεσα σε πιθανές λέξεις. Στατιστικές μέθοδοι μπορούν να χρησιμοποιηθούν για να ανακαλύψουν ένα μοντέλο που περιγράφει τον τρόπο με τον οποίο τα έγγραφα μπορούν να δημιουργηθούν [53]. Συγκεκριμένα χρησιμοποιούμε λανθάνουσα κατανομή Dirichlet (Latent Dirichlet Allocation) [54] για να εξάγουμε λανθάνοντα θέματα, τα οποία χρησιμοποιούμε για τον υπολογισμό των ομοιοτήτων. Οι ομοιότητες στη συνέχεια χρησιμοποιούνται για την σύσταση εγγράφων και επισημειώσεων και για την επέκταση των αποτελεσμάτων αναζήτησης.

Περισσότερες πληροφορίες πάνω στη μέθοδο των πιθανοτικών μοντέλων θεμάτων μπορούν να βρεθούν στο κεφάλαιο 3.

5.3.3 Αναζήτηση και Συστήματα Προτάσεων

Η προσέγγισή μας επικεντρώνεται σε συνδυασμό δομημένων ταξονομιών και φολκσονομιών με λανθάνοντα θέματα (Εικόνα 7).

Στο κάτω μέρος της εικόνας εμφανίζονται οι γνωσιακές δομές οι οποίες μπορεί να περιλαμβάνουν ταξονομίες ή φολκσονομίες. Οι δομές αυτές χρησιμοποιούνται για την παραγωγή προτάσεων πόρων και για την επέκταση ερωτημάτων στην αναζήτηση. Με την περιοδική επεξεργασία των πόρων και μη επιβλεπόμενη εξαγωγή θεμάτων, τα θέματα που ανακαλύπτονται χρησιμοποιούνται ως βάση για την ενίσχυση της αναζήτησης και των λειτουργιών προτάσεων. Επίσης τα θέματα που εξάγονται με τη έγκριση των χρηστών συμμετέχουν στην επέκταση των γνωσιακών δομών.



Εικόνα 5.2 Γνωσιακές Δομές και Μοντέλα Θεμάτων

5.3.3.1 Αναζήτηση

Η αναζήτηση μέσα σε μια βάση πληροφοριών ενός οργανισμού είναι μια απαιτητική εφαρμογή. Μια σημαντική πρόκληση είναι ότι οι λέξεις που χρησιμοποιούνται στα ερωτήματα μερικές φορές δεν εμφανίζονται στο περιεχόμενο που κρίνεται σχετικό από τον χρήστη. Οι προτάσεις για την αντιμετώπιση αυτού του προβλήματος περιλαμβάνουν σημασιολογική αναζήτηση αλλά και τεχνικές επέκτασης ερωτημάτων που κάνουν χρήση των συνωνύμων, ιδιοσυγκρασιακών όρων ή σημασιολογικά παρόμοιων όρων, προκειμένου να είναι σε θέση να ανακτήσουν τα αποτελέσματα που είναι σχετικά, αλλά δεν περιέχουν τις ακριβείς λέξεις που χρησιμοποιούνται στο ερώτημα [128]. Τέτοιες προσεγγίσεις απαιτούν την ύπαρξη μιας ρητής γνωσιακής δομής η οποία συμβάλλει στον υπολογισμό ομοιοτήτων μεταξύ όρων.

Τα πιθανοτικά μοντέλα θεμάτων μπορούν να χρησιμοποιηθούν ως τεχνική για την μη-επιβλεπόμενη κατηγοριοποίηση εγγράφων, και στη συνέχεια ως βάση για την επέκταση ερωτημάτων αναζήτησης. Σε αυτήν την προσέγγιση, ζητήματα όπως οι σχέσεις μεταξύ λέξεων, η ύπαρξη συνωνύμων και οι λέξεις με πολλαπλά νοήματα μπορούν να αντιμετωπιστούν με τη χρήση στατιστικής για την ανακάλυψη σχέσεων μεταξύ λέξεων. Δεν υπάρχει ανάγκη για τη διατήρηση μιας γνωσιακής δομής καθώς αυτή προέρχεται (αναδύεται) από το πιθανοτικό μοντέλο. Μια πιθανή πρόκληση σε αυτή τη τεχνική είναι ότι ο επαναλαμβανόμενος υπολογισμός πιθανοτήτων σχέσεων σε κάθε ερώτημα αναζήτησης μπορεί να καταναλώνει μεγάλη επεξεργαστική ισχύ.

Στο παρόν κεφάλαιο, οι ομοιότητες που προέρχονται από τα μοντέλα θεμάτων αποθηκεύονται σε ευρετήρια λέξεων, προκειμένου να βελτιωθεί η ταχύτητα της επέκτασης ερωτήματος. Η προσέγγισή μας υλοποιεί μια λύση βασισμένη σε έναν ευρετήριο συναφών λέξεων (thesaurus) στην οποία οι λέξεις που βρέθηκαν μαζί σε ένα λανθάνον θέμα θεωρούνται παρόμοιες και αποθηκεύονται μαζί με το μέτρο ομοιότητας τους. Ακόμη, γίνεται χρήση ενός κατωφλιού και μιας διαδικασίας κλαδέματος (pruning) για τον περιορισμό του μεγέθους του ευρετηρίου λέξεων. Με αυτό τον τρόπο, όταν ο χρήστης εκτελεί ένα ερώτημα που περιλαμβάνει έναν αριθμό όρων, το ερώτημα επεκτείνεται με χρήση μερικών ιδιαίτερα συναφών όρων.

Για να υπολογίσουμε την ομοιότητα κάθε εγγράφου με το ερώτημα που διατυπώνεται χρησιμοποιούμε την εξίσωση (5.1), όπως περιγράφεται στο [129].

$$S_d(Q) = \mu S_d(E) + (1-\mu) S_d(Q) \qquad (5.1)$$

Στη σχέση (5.1), Q είναι οι S όροι του ερωτήματος και E είναι η επέκταση με βάση το ευρετήριο συναφών λέξεων που βασίστηκε στο μοντέλο θεμάτων. S_D είναι τα αποτελέσματα της αξιολόγησης της ομοιότητας των εγγράφων που βρίσκονται στο εταιρικό κοινωνικό λογισμικό και μ (0 <μ <1) είναι η παράμετρος ανάμειξης.

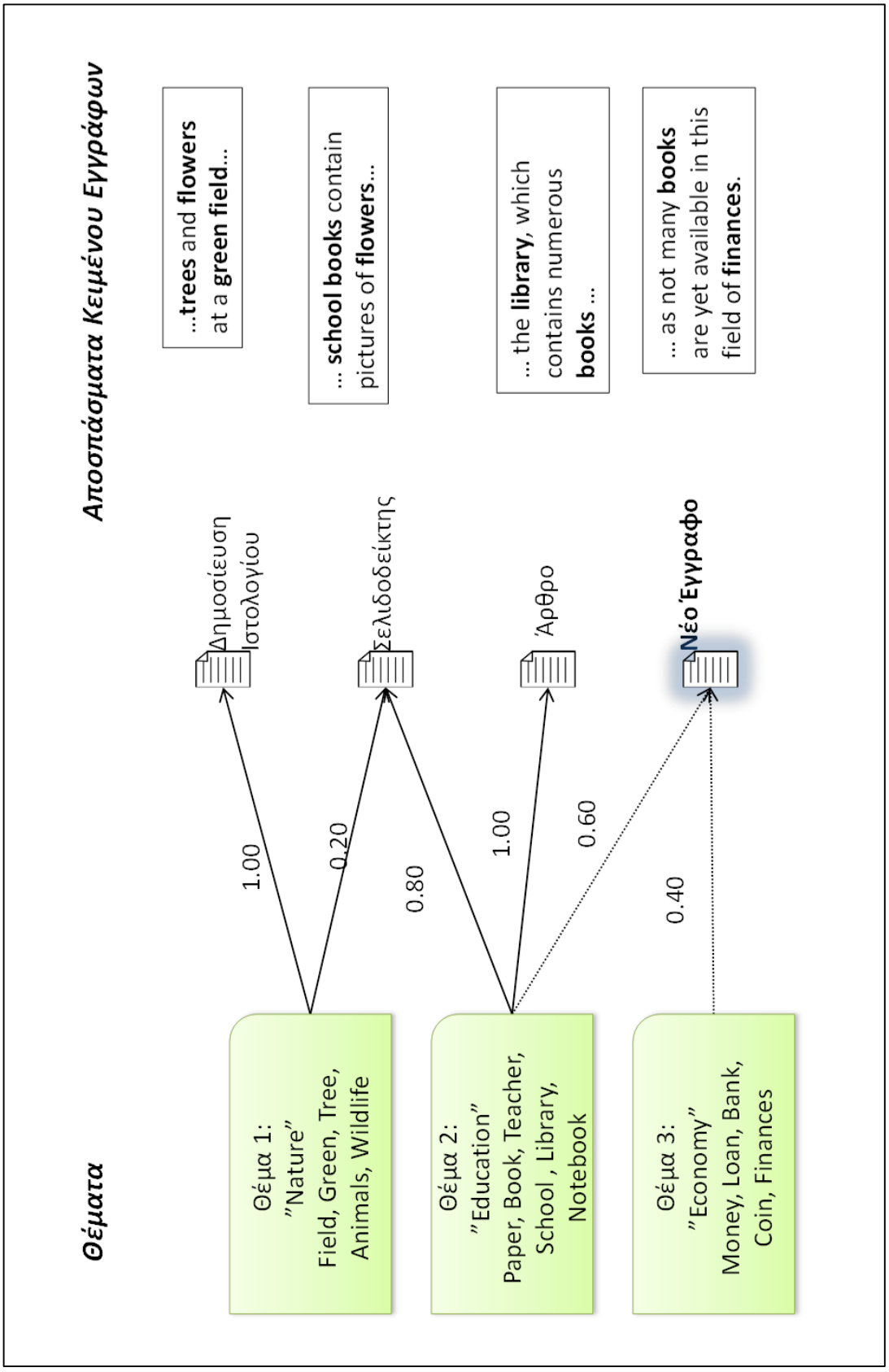
5.3.3.2 Προτάσεις Εγγράφων[20]

Η μέθοδος λανθάνουσας κατανομής Dirichlet (LDA), όταν εκπαιδεύεται, παράγει δύο πιθανοτικές κατανομές που απεικονίζουν την πιθανότητα εμφάνισης λέξεων με δεδομένο ένα θέμα και την πιθανότητα εμφάνισης θεμάτων σε ένα έγγραφο. Οι πιθανότητες που υπολογίζονται στις κατανομές χρησιμοποιούνται ως μέτρα για τον υπολογισμό των ομοιοτήτων για την δημιουργία προτάσεων πόρων. Τα έγγραφα που ήταν στο αρχικό σύνολο δεδομένων με το οποίο εκπαιδεύτηκε η μέθοδος έχουν ήδη περιγραφεί σε θέματα. Στην περίπτωση ενός νέου, προηγουμένως άγνωστου, εγγράφου το μοντέλο συνάγει τα λανθάνοντα θέματα που σχετίζονται με αυτόν.

Για να περιγράψει τα έγγραφα, η κατανομή θεμάτων χρησιμοποιείται με τη μορφή ενός διανύσματος. Αυτό το διάνυσμα έχει τόσες διαστάσεις όσα και τα λανθάνοντα θέματα που βρέθηκαν στο σύνολο δεδομένων. Όταν ένα νέο έγγραφο προστίθεται στο σύστημα, η κατανομή θεμάτων που τον αφορούν συνάγεται με βάση τις λέξεις που παρατηρήθηκαν. Για τον υπολογισμό της σημασιολογικής απόστασης μεταξύ των πόρων υπολογίζουμε την ομοιότητα συνημίτονου μεταξύ των αντίστοιχων διανυσμάτων που περιγράφουν τα θέματά τους. Στο (5.2) θ_A και θ_B είναι τα διανύσματα θεμάτων που περιγράφουν τα έγγραφα A και B και χρησιμοποιούνται για τον υπολογισμό της ομοιότητας μεταξύ τους.

$$Similarity(A,B) = \cos(\theta_A, \theta_B) = \frac{\theta_A \cdot \theta_B}{\|\theta_A\| \cdot \|\theta_B\|} \qquad (5.2)$$

[20] Στην ενότητα αυτή χρησιμοποιούμε τον όρο «έγγραφο» για να αναφερθούμε σε κάθε τύπου πληροφοριακό πόρο που μπορεί να βρεθεί στο εσωτερικό και στο εξωτερικό ενός εταιρικού συστήματος.



Εικόνα 5.3 Προτάσεις Αντικειμένων - Προσθήκη Νέου Εγγράφου

Στο παράδειγμα που φαίνεται στην Εικόνα 5.3, τρία έγγραφα υπάρχουν ήδη στο σύστημα: μια δημοσίευση σε ιστολόγιο (blog post), ένας σύνδεσμος και ένα άρθρο. Οι κατανομές θεμάτων τους έχουν ήδη υπολογιστεί και φαίνονται τα βέλη που τα συνδέουν με τα θέματα. Όταν ένα νέο έγγραφο προστίθεται, τα θέματα με τα οποία συσχετίζεται συνάγονται με βάση τις λέξεις που περιέχει το έγγραφο.

Σε αυτό το παράδειγμα, το έγγραφο περιέχει λέξεις όπως *book* και *finances*, και, συνεπώς, συνδέεται με το θέμα 2 που αφορά την εκπαίδευση και το θέμα 3 που σχετίζεται με την οικονομία. Ο βαθμός της ομοιότητας ενός εγγράφου με ένα συγκεκριμένο θέμα εξαρτάται από το πόσες λέξεις που βρέθηκαν σε αυτό ανήκουν στο θέμα. Για τον υπολογισμό ομοιότητας μεταξύ των εγγράφων, έχουμε υπολογίσει την ομοιότητα συνημίτονου τους, χρησιμοποιώντας ως διάνυσμα την κατανομή θεμάτων (Πίνακας 5.1).

Έγγραφο	**Έγγραφο**	**Ομοιότητα Συνημίτονου**
Δημοσίευση Ιστολογίου	Σελιδοδείκτης	0,243
Σελιδοδείκτης	Άρθρο	0,865
Άρθρο	***Νέο Έγγραφο***	***0,832***
…	…	…

Πίνακας 5.1 Υπολογισμός Ομοιότητας Αντικειμένων

5.3.3.3 Προτάσεις Επισημειώσεων

Η επισημείωση ορίζεται ως η σύνδεση λέξεων-κλειδιών με ένα έγγραφο. Οι χρήστες που θέλουν να προσθέσουν επισημειώσεις (ή αλλιώς, μετα-δεδομένα) σε ένα έγγραφο μπορούν να εισάγουν είτε δικές τους λέξεις-κλειδιά, στην περίπτωση των φολκσονομιών, είτε να επιλέξουν από μια προ-υπάρχουσα γνωσιακή δομή. Το εταιρικό κοινωνικό λογισμικό επιτρέπει συνήθως την ελεύθερη επισημείωση η οποία οδηγεί στο σχηματισμό των φολκσονομιών. Ακόμα, πολλά συστήματα διαχείρισης γνώσης χρησιμοποιούν κάποιο είδος ταξονομίας ή καταλόγου όρων. Η διαδικασία επεξεργασίας και ανάλυσης των φολκσονομιών είναι ένας τομέας που έχει ερευνηθεί διεξοδικά [130]. Επιπλέον, μια σειρά από τεχνικές έχουν εξετασθεί οι οποίες συνδυάζουν μη επιβλεπόμενες μεθόδους λανθάνουσας σημασιολογικής ανάλυσης με φολκσονομίες. Συγκεκριμένα, η λανθάνουσα κατανομή Dirichlet LDA έχει αξιολογηθεί για τη πρόταση επισημειώσεων σε φολκσονομίες όπου φαίνεται να ξεπερνά σε απόδοση τεχνολογίες εξόρυξης κανόνων συσχέτισης [73].

Η πρώτη τεχνική προτάσεων επισημειώσεων που χρησιμοποιούμε στην προσέγγισή μας βασίζεται στην ομοιότητα μεταξύ εγγράφων όπως εκείνη υπολογίζεται στη πρόταση πόρων. Το σύστημα συνάγει τα σχετικά έγγραφα από τη κατανομή θεμάτων, και στη συνέχεια προτείνει τις επισημειώσεις που έχουν ήδη προστεθεί στα πιο συναφή σημασιολογικά έγγραφα. Η προσέγγιση αυτή επαναχρησιμοποιεί ένα μεγάλο ποσοστό επισημειώσεων (ή όρων σε περίπτωση τυπικών γνωσιακών δομών, όπως ταξονομίες), ως εκ τούτου βοηθάει την ενίσχυση των όρων που χρησιμοποιούνται από πριν στο σύστημα. Στην εξίσωση (5.3), το διάνυσμα που περιέχει τη ομοιότητα της κάθε επισημείωσης με το έγγραφο A υπολογίζεται πολλαπλασιάζοντας τον πίνακα με τις επισημειώσεις κάθε εγγράφου (TD), με την πλήρη κατανομή θεμάτων εγγράφου στο σύνολο δεδομένων (Θ), και στη συνέχεια με το διάνυσμα που περιγράφει την κατανομή θεμάτων του συγκεκριμένου εγγράφου (A).

$$tagSim_{\mathrm{A}} = TD \cdot \Theta \cdot \theta_{\mathrm{A}} \qquad (5.3)$$

Η δεύτερη τεχνική βασίζεται στο γεγονός ότι τα θέματα που προκύπτουν από την ανάλυση αναπαριστώνται από κατανομές λέξεων. Στο πλαίσιο των προτάσεων επισημειώσεων αυτές οι κατανομές μπορούν να γίνουν οι χώροι που περιέχουν πιθανές λέξεις-κλειδιά. Οι κυρίαρχες λέξεις ενός θέματος μπορούν να χρησιμοποιηθούν για να κατηγοριοποιηθούν έγγραφα που σε κάποιο βαθμό παράγονται από αυτό το θέμα. Εάν το έγγραφο είναι στενά συνδεδεμένο με ένα θέμα τότε οι κυρίαρχες λέξεις του θέματος προτείνονται ως πιθανές επισημειώσεις. Αυτή η τεχνική μπορεί να λειτουργήσει τόσο με φολκσονομίες όσο και ταξονομίες. Ο υπολογισμός αυτός διατυπώνεται στην εξίσωση (5.4), όπου το διάνυσμα που περιγράφει την κατανομή θεμάτων του συγκεκριμένου εγγράφου (A), πολλαπλασιάζεται με το πίνακα κατανομής λέξεων ανά θέμα στο σύνολο δεδομένων.

$$tagSim_{\mathrm{A}} = \theta_{\mathrm{A}}{}^{T} \cdot \Phi \qquad (5.4)$$

Επιπλέον, οι δύο αυτές τεχνικές μπορούν να συνδυαστούν για να προταθεί μια λίστα πιθανών επισημειώσεων χρησιμοποιώντας μια προσέγγιση σταθμισμένου μέσου. Αυτή η μέθοδος υβριδικών προτάσεων μπορεί να υποστηρίξει τόσο την διατήρηση των υφιστάμενων όρων αλλά και να προτείνει νέες λέξεις-κλειδιά με βάση τις κυρίαρχες λέξεις σε μια εξελισσόμενη βάση γνώσης.

5.4 Μελέτη Εφαρμογής

5.4.1 Αρχιτεκτονική

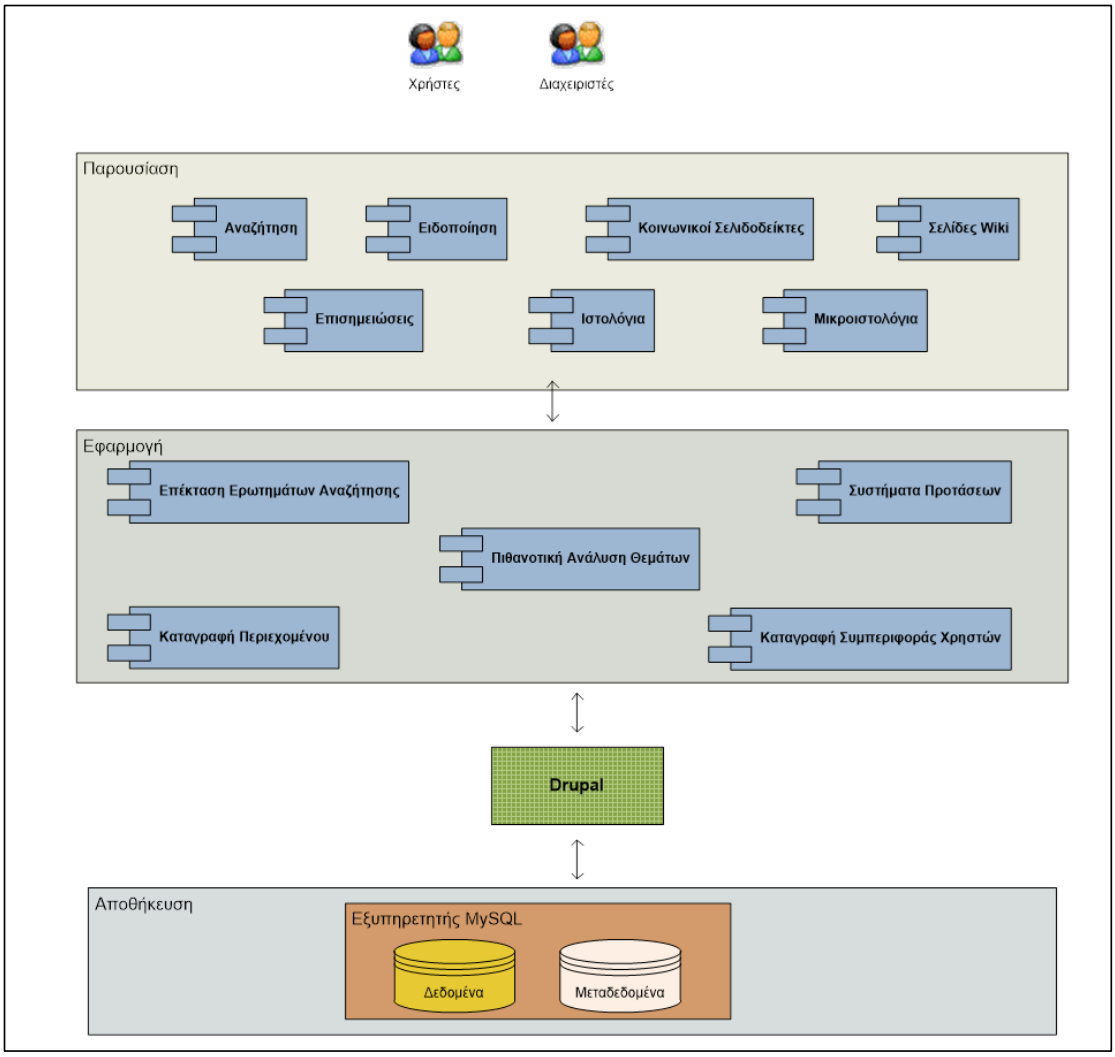
Έχουμε δοκιμάσει την προσέγγιση του Entasis, υλοποιώντας την στο εσωτερικό ενός συστήματος εταιρικού κοινωνικού λογισμικού ανοιχτού κώδικα. Ως βάση ανάπτυξης της εφαρμογής χρησιμοποιείται το δημοφιλές σύστημα διαχείρισης περιεχομένου Drupal το οποίο παρέχει μια συλλογή κοινωνικών εφαρμογών για το εταιρικό περιβάλλον και υποστηρίζει τόσο τις ταξονομίες όσο και τις φολκσονομίες για την επισημείωση πόρων [131].

Το Entasis αποτελεί ένα ολοκληρωμένο πακέτο λογισμικού διαχείρισης γνώσης που προορίζεται για χρήση στο εσωτερικό επιχειρήσεων και επιτρέπει την διάχυση και την ανταλλαγή πληροφορίας μεταξύ των εργαζομένων. Το συνολικό σύστημα περιλαμβάνει κοινωνικά και ευφυή χαρακτηριστικά ως βάση της διαχείρισης γνώσης στην επιχείρηση. Το Entasis είναι βασισμένο σε τεχνολογίες ιστού και είναι προσπελάσιμο από τον περιηγητή που χρησιμοποιεί ο κάθε χρήστης. Αποτελεί ένα κεντρικό σημείο στο οποίο μπορεί να μοιράζεται η πληροφορία στο εσωτερικό της επιχείρησης και επιτρέπει εύκολη εγκατάσταση, αναβάθμιση και συντήρηση. Η αρχιτεκτονική του Entasis φαίνεται αναλυτικά στην Εικόνα 5.4.[21]

[21] Το Entasis σχεδιάστηκε και αναπτύχθηκε στα πλαίσια του ερευνητικού προγράμματος OrganiK που χρηματοδοτήθηκε από την Ευρωπαική Επιτροπή. Το OrganiK έχει σαν στόχο την έρευνα και ανάπτυξη ενός καινοτόμου συστήματος διαχείρισης γνώσης που να επιτρέπει την σημασιολογική σύνδεση μεταξύ διαφόρων εφαρμογών εταιρικού κοινωνικού λογισμικού. Το σύστημα αυτό συγκεντρώνει πληροφορία που ανταλλάσσεται μεταξύ των εργαζομένων στο εσωτερικό αλλά και στο εξωτερικό εταιριών και επιτρέπει μια αποδοτική διαχείριση της οργανωσιακής γνώσης, ενώ μπορεί να εξυπηρετεί τις ανάγκες των μικρομεσαίεων επιχειρήσεων στην Ευρώπη.

Λόγω της συμμετοχής του αυτής το Entasis εγκαταστάθηκε και χρησιμοποιήθηκε σε 5 εταιρίες έντασης γνώσης:

- CAS Software AG, εταιρία ανάπτυξης λογισμικού στην Γερμανία
- LeserAuskunft Gmbh, εταιρία πώλησης περιεχομένου στην Γερμανία
- LTC - Language Technology Centre Ltd, εταιρία παροχής υπηρεσιών μεταφράσης στο Ηνωμένο Βασίλειο
- Syria Informatica S.r.l., εταιρία ανάπτυξης λογισμικού στην Ιταλία
- Quality Maritime Services Ltd, εταιρία υπηρεσιών ναυτιλίας στην Ελλάδα



Εικόνα 5.4 Αρχιτεκτονική Entasis

Το Entasis έχει εγκατασταθεί και χρησιμοποιηθεί σε πέντε μικρές και μεσαίες επιχειρήσεις στην Ευρώπη, συμπεριλαμβανομένων μιας επιχείρησης μεταφράσεων και τοπικοποιήσεων, δυο εταιριών υπηρεσιών πληροφορικής, ενός παρόχου περιεχομένου και μιας συμβουλευτικής επιχείρησης ναυτιλίας. Οι εταιρίες αυτές είχαν έναν ενεργό ρόλο στην ανάπτυξη, εφαρμογή και χρήση του συστήματος. Οι εργαζόμενοι ήταν παρόντες στις πρώιμες συζητήσεις για το πώς θα δουλεύει το σύστημα. Επίσης, κατά τη διάρκεια της ανάπτυξης του, το σύστημα ήταν διαθέσιμο στους εργαζομένους σύμφωνα με το υπόδειγμα της συνεχούς βήτα έκδοσης στο λογισμικό (Perpetual Beta Paradigm). Αφού το σύστημα έγινε σταθερό, οι χρήστες άρχισαν να το εντάσσουν στην καθημερινότητά τους.

5.4.2 Περιήγηση του Συστήματος

Όλα τα έγγραφα που εισάγονται στο Entasis αναλύονται συνεχώς και λανθάνοντα θέματα έρχονται στην επιφάνεια. Οι κατανομές των θεμάτων αποθηκεύονται στη βάση του συστήματος αλλά και στο μοντέλο το οποίο μπορεί να χρησιμοποιηθεί για να συναχθούν τα θέματα από καινούρια έγγραφα.

Εικόνα 5.5 Προτάσεις Επισημειώσεων και Αντικειμένων στο Entasis

Ένα στιγμιότυπο από την εγκατάσταση του συστήματος στην εταιρία παροχής συμβουλευτικών υπηρεσιών σε ναυτιλιακές φαίνεται στην Εικόνα 5.5. Έστω ότι ένας εργαζόμενος βρίσκει μια ενδιαφέρουσα πληροφορία στον παγκόσμιο ιστό η οποία παρουσιάζει ενδιαφέρον για την εταιρία. Την εισάγει, μαζί με τις προσωπικές του εκτιμήσεις, στο εταιρικό κοινωνικό λογισμικό στο αντίστοιχο κομμάτι του ιστολογίου (τμήμα 1 της Εικόνα 5.5). Το κείμενο που εισάγεται αναλύεται και οι συσχετίσεις με τα υπάρχοντα θέματα εξάγονται. Ο χρήστης μπορεί τότε να εντοπίσει σχετικά έγγραφα με βάση τα λανθάνοντα θέματα. Οι προτάσεις εμφανίζονται στο τμήμα 2 της Εικόνα 5.5 και εξελίσσονται όσο ο χρήστης εισάγει καινούριες πληροφορίες. Ο χρήστης μπορεί να βελτιώσει το περιεχόμενο του εγγράφου προσθέτοντας περιεχόμενο από συναφή αντικείμενα ή προσθέτοντας συνδέσμους προς αυτά.

Όσο ο χρήστης εισάγει κείμενο στη σελίδα, διάφορες επισημειώσεις προτείνονται (τμήμα 3 της Εικόνα 5.5). Το σύστημα προτάσεων βοηθά τον

εργαζόμενο να κατηγοριοποιήσει τον συγκεκριμένο έγγραφο. Οι επισημειώσεις εξάγονται τόσο από την υπάρχουσα γνωσιακή δομή όσο και από τις λέξεις των υπαρχόντων μοντέλων θεμάτων. Μερικά λεπτά αργότερα, όταν ένας χρήστης πραγματοποιήσει μια αναζήτηση μπορεί να πλοηγηθεί στην σελίδα που δημιουργήθηκε, έστω κι αν εκείνη δεν περιέχει τις ακριβείς λέξεις του ερωτήματος. Το ερώτημα επεκτείνεται ώστε να καλύψει σχετικά έγγραφα με τη χρήση των μοντέλων θεμάτων.

5.5 Πειραματική Αξιολόγηση

Πριν την αξιολόγηση του συστήματος, πραγματοποιήθηκε εγκατάσταση του συστήματος σε πέντε μικρομεσαίες επιχειρήσεις και αρχικοποίηση με δεδομένα. Κατά την αρχικοποίηση έγινε εξαγωγή των θεμάτων από δεδομένα που δόθηκαν από τις εταιρίες. Η αρχική εκτέλεση της εξαγωγής θεμάτων από τα δεδομένα δημιουργεί κατανομές λέξεων όπως εκείνες που φαίνονται στον σχετικό πίνακα (Πίνακας 5.2). Συγκεκριμένα σε αυτό το πίνακα φαίνονται θέματα που εξάχθηκαν από τα δεδομένα μιας ναυτιλιακής εταιρίας.

Πίνακας 5.2 Κατανομή Λέξεων σε Θέματα σε Ναυτιλιακή Εταιρία

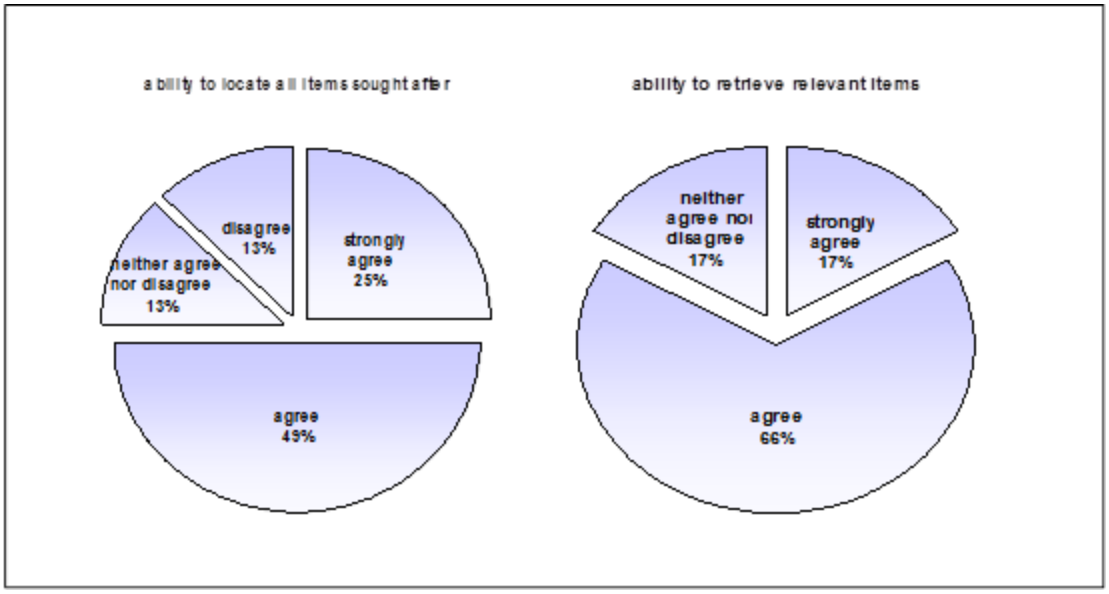
Αριθμός Θέματος	Λέξεις
5	canal panama panagiotopoulou sailor
23	securing devices csm fixed loading portable
127	certificate certificates reply solid cargoes bulk
409	water external contact scheduled check marked
357	fire model life equipments saving trader

Οι λέξεις που ανήκουν στα θέματα που εξάχθηκαν από τα δεδομένα της εταιρίας παρουσιάζουν μια σημασιολογική συνάφεια που είναι εμφανής. Επίσης η λογαριθμική πιθανότητα και λογαριθμική πιθανότητα ανά token παρουσιάζεται ίση με -6,0929. Η τιμή αυτή είναι αρκετά χαμηλή και δείχνει την ευστάθεια του μοντέλου θεμάτων. Η εκτίμηση της συγκεκριμένης πιθανότητας έγινε με τη χρήση της μεθόδου left-to-right [132].

Επίσης, παρακολουθήσαμε τη χρήση του συστήματος σε πέντε μικρομεσαίες επιχειρήσεις. Το σύστημα χρησιμοποιήθηκε από 32 χρήστες κατά τη διάρκεια μιας περιόδου έξι μηνών. Εδώ παρουσιάζονται τα αποτελέσματα σε μία από τις πέντε επιχειρήσεις. Το ερωτηματολόγιο που χρησιμοποιήθηκε βρίσκεται στο τέλος της διατριβής (σελ. 251) κάτω από τον τίτλο «Παράρτημα 1: Ερωτηματολόγιο Αξιολόγησης Συστήματος Entasis».

Οι χρήστες κλήθηκαν να αξιολογήσουν την απόδοση της αναζήτησης και του συστήματος προτάσεων, αλλά και συγκεκριμένων λειτουργιών του εταιρικού κοινωνικού λογισμικού. Έτσι, δυο τύποι ερωτηματολογίων χρησιμοποιήθηκαν : ένα που περιλάμβανε συγκεκριμένες ερωτήσεις για τα στοιχεία του συστήματος για τα οποία οι χρήστες διατύπωσαν τη γνώμη τους με βάση μια κλίμακα Likert, και ένα που περιλάμβανε ερωτήσεις ανοιχτού τύπου που αφορούν την χρησιμότητα και την ευστοχία των λειτουργιών του συστήματος.

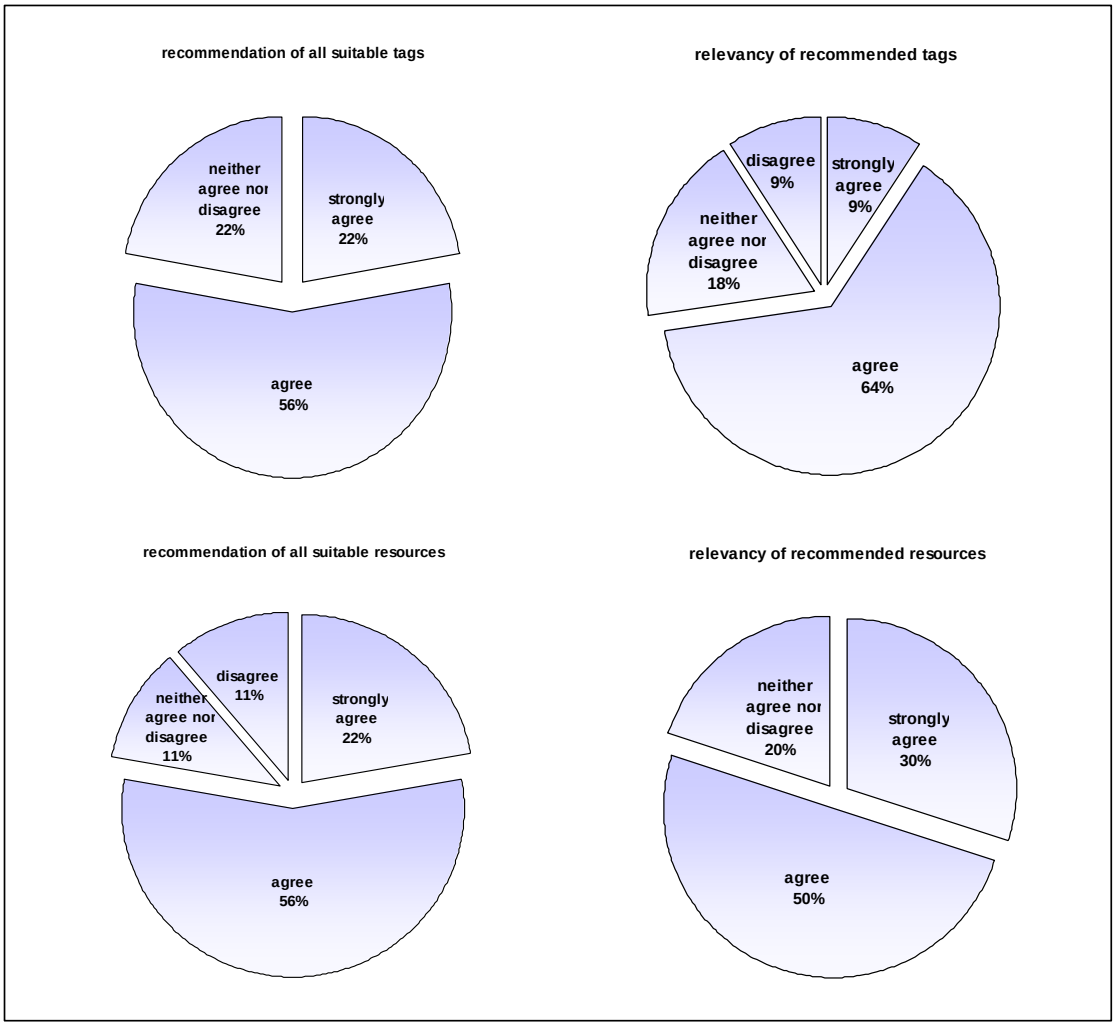
Η αναζήτηση αξιολογήθηκε θετικά καθώς οι χρήστες αξιολόγησαν θετικά την ικανότητα του συστήματος να εντοπίσει με ακρίβεια τα έγγραφα που έψαχναν (74% θετικές αποκρίσεις) και να ανακαλέσει όλα αυτά που έψαχναν (83% θετικές απαντήσεις), βλέπε Εικόνα 5.6.



Εικόνα 5.6 Αξιολόγηση της Αναζήτησης

Οι προτάσεις γενικά έγιναν δεκτές με θετικά σχόλια (Εικόνα 5.7). 78% των εργαζομένων απάντησαν ότι το σύστημα τους βοήθησε να βρουν όλες τις πιθανές επισημειώσεις για να προσθέσουν στα έγγραφά τους ενώ το 73% βρήκε ότι οι προτεινόμενες επισημειώσεις ήταν σχετικές και μπορούσαν να χρησιμοποιηθούν.

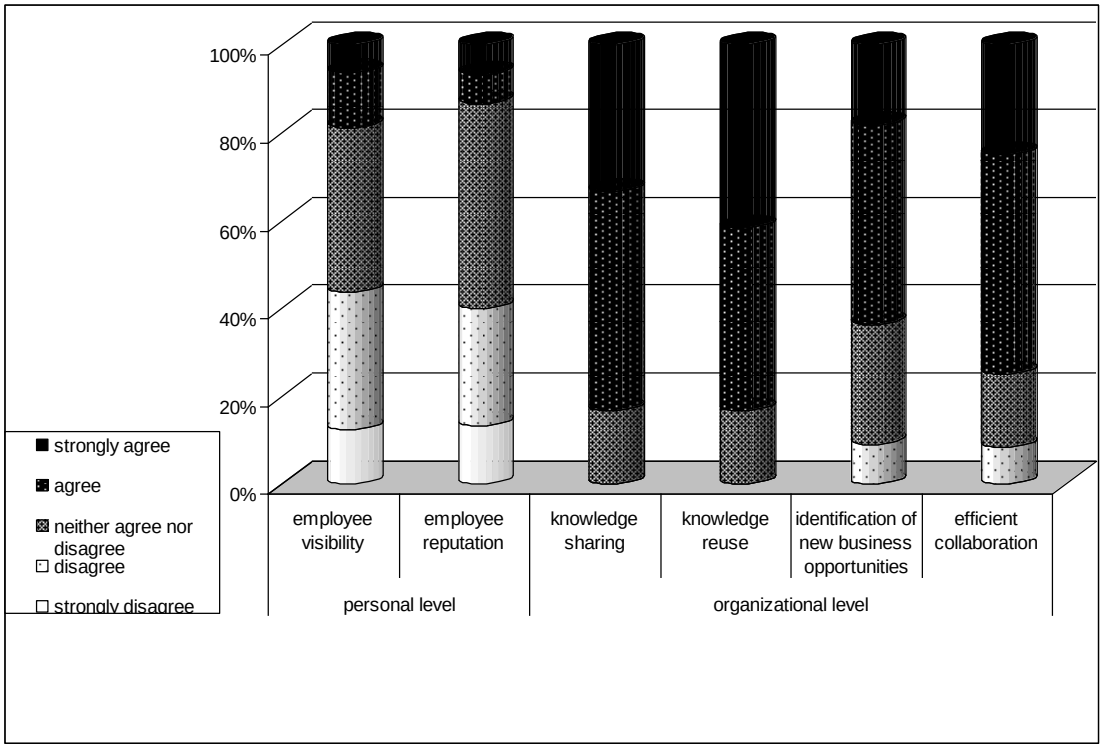
Για την πρόταση πόρων, το 78% των εργαζομένων θεωρεί ότι το σύστημα τους βοήθησε να εντοπίσουν όλους τα έγγραφα που θα περίμεναν ενώ το 80% θεώρησε ότι τα έγγραφα που προτείνονται είναι σχετικά με τις ανάγκες τους.



Εικόνα 5.7 Αξιολόγηση των Προτάσεων

Στο δεύτερο μέρος της έρευνας, οι χρήστες αξιολόγησαν πως το σύστημα θα επιδρούσε στην προσωπική αλλά και στην εταιρική απόδοση. (Εικόνα 5.8). Σε προσωπικό επίπεδο, οι χρήστες φαίνεται να αμφιβάλλουν κατά πόσο το σύστημα θα βελτίωνε την ορατότητά τους (40% διαφώνησαν) ή τη φήμη τους (37.5% διαφώνησαν). Αντίθετα, τα αποτελέσματα που αφορούν την απόδοση σε εταιρικό επίπεδο δείχνουν ότι οι χρήστες αναμένουν η χρήση του συστήματος ενδοεταιρικά να έχει ένα σημαντικό θετικό αντίκτυπο. Αυτή η θετική άποψη φαίνεται να σχετίζεται με την αύξηση της γνώσης που μοιράζεται και ξαναχρησιμοποιείται

(85%), τον εντοπισμό νέων επιχειρηματικών ευκαιριών (69%) και την καλύτερη συνεργασία(78%).



Εικόνα 5.8 Αξιολόγηση σε Προσωπικό και Οργανωσιακό Επίπεδο

Οι απαντήσεις των εργαζομένων στις ερωτήσεις ανοιχτού τύπου συμφωνούν με τα αποτελέσματα που αναλύθηκαν παραπάνω. Οι χρήστες θεωρούν ότι το σύστημα θα μπορεί να τους βοηθάει να είναι πιο καλά ενημερωμένοι για δραστηριότητες που λαμβάνουν χώρα στον οργανισμό τους. Παρ' όλα αυτά ανέφεραν ότι οι άνθρωποι πρέπει να το χρησιμοποιούν τακτικά. Ρωτώντας τους εργαζόμενους για το βαθμό που η χρήση του συστήματος βοηθάει την δημιουργία κοινωνικών δεσμών, εκείνοι το θεώρησαν χρήσιμο αλλά όχι πολύ σημαντικό με δεδομένο το μικρό μέγεθος των οργανισμών και την δυνατότητα άμεσης επικοινωνίας. Ένας χρήστης αναφέρει : «Το σύστημα σίγουρα με βοηθάει να είμαι συνδεδεμένος με τους συνεργάτες μου, αλλά εφόσον είμαστε ένας μικρός οργανισμός, θεωρώ ότι τα οφέλη του είναι κυρίως αυτά που σχετίζονται με το μοίρασμα της γνώσης και την συνεργασία». Οι εργαζόμενοι θεωρούν ότι το σύστημα μπορεί να βοηθήσει την σύλληψη και την οργάνωση της γνώσης αλλά και της εξειδίκευσης στο εσωτερικό του οργανισμού. «Οι άνθρωποι πρέπει να ενθαρρύνονται συστηματικά ώστε να το χρησιμοποιούν», αναφέρει κάποιος. Τέλος,

για τις καθημερινές εργασίες, το μήνυμα που λαμβάνεται είναι ότι το σύστημα μπορεί να αποτελέσει μια πλατφόρμα συνεργασίας. Οι εργαζόμενοι θεωρούν ότι η χρήση του συστήματος από ομάδες συνεργατών θα τους επιτρέψει να «μοιράζονται πολύ περισσότερες πληροφορίες και να συζητούν για την δουλειά».

5.6 Συμπεράσματα

Η μεθοδολογία που προτείνουμε βασίζεται στην ανάλυση εγγράφων με βάση την τεχνική εξαγωγής πιθανοτικών μοντέλων θεμάτων, και συγκεκριμένα της λανθάνουσας κατανομής Dirichlet για την επέκταση ερωτημάτων αναζήτησης αλλά και για την πραγματοποίηση προτάσεων πόρων σε χρήστες. Όταν ένα υπάρχον αντικείμενο, π.χ. ένα έγγραφο, διαβάζεται από τον χρήστη το σύστημα μπορεί να χρησιμοποιήσει την κατανομή θεμάτων του. Όταν προστίθεται ένα καινούριο αντικείμενο, η κατανομή θεμάτων του μπορεί να εξαχθεί από τα υπάρχοντα.

Ο εντοπισμός των λανθανόντων θεμάτων για την πρόταση περιεχομένου και για την αναζήτηση είναι μια μέθοδος μη επιβλεπόμενη και προσφέρει μια σειρά από οφέλη, ιδιαίτερα σε σχέση με επιβλεπόμενες μεθόδους και με εκείνες που βασίζονται σε μοντέλα. Δεν εξαρτάται από ρητές γνωσιακές δομές όπως ταξονομίες ή οντολογίες και δεν απαιτεί προσπάθεια από τον χρήστη για την κατηγοριοποίηση των πόρων. Επιπρόσθετα μπορεί να επεκταθεί ώστε να καλύπτει την οργανωσιακή γνώση η οποία εξελίσσεται χωρίς να εξαρτάται από συγκεκριμένες λέξεις

Η πρόταση επισημειώσεων μπορεί να επηρεάσει την δημιουργία των γνωσιακών δομών στο εταιρικό κοινωνικό λογισμικό. Η πρόταση επισημειώσεων που ανήκαν στο μοντέλο θεμάτων μπορεί να βοηθήσει την εξέλιξη της δομής ώστε να καλύψει αναδυόμενα θέματα. Όσο καινούρια έγγραφα εισάγονται στο σύστημα και αναλύονται, τόσο καινούρια θέματα εντοπίζονται στο πιθανοτικό μοντέλο που επαναϋπολογίζεται για να αντιστοιχεί στο σύνολο δεδομένων. Οι κυρίαρχες λέξεις σε αυτά τα θέματα προτείνονται συνεχώς ως λέξεις κλειδιά για τα νέα έγγραφα που μένουν να επισημανθούν.

Τα πλεονεκτήματα της στατιστικής προσέγγισης που περιγράφεται γίνονται εμφανή σε ένα εταιρικό κοινωνικό λογισμικό, το οποίο διαθέτει γνωσιακές δομές διαφορετικής τυπικότητας. Τα πιθανοτικά μοντέλα θεμάτων μπορούν να βελτιώσουν την ευελιξία και την σταθερότητα, καθώς η επιχειρησιακή γνωσιακή

δομή και οι αυθαίρετες δραστηριότητες των χρηστών συμπληρώνονται από την εξαγωγή λανθανόντων θεμάτων.

Οι τεχνικές που ακολουθήσαμε για την αναζήτηση και την δημιουργία προτάσεων βασίζονται στα πιθανοτικά μοντέλα θεμάτων και είναι διαθέσιμες με τη μορφή λογισμικού ανοιχτού κώδικα και μαζί με το Entasis, που είναι επίσης διαθέσιμο[22]. Οι επαγγελματίες του χώρου μπορούν να τα χρησιμοποιήσουν και να τα ενσωματώσουν σε άλλα συστήματα διαχείρισης γνώσης, καθώς και να τα επεκτείνουν ώστε να προσθέσουν επιπλέον δυνατότητες. Επιπλέον, οι ερευνητές μπορούν να δουν κατά πόσο τα πιθανοτικά μοντέλα θεμάτων μπορούν να είναι επωφελή για την διαχείριση γνώσης ιδίως όταν χρησιμοποιούνται γνωσιακές δομές που δεν περιγράφηκαν εδώ (π.χ. οντολογίες).

Σε αυτή την εργασία έχουμε δείξει πως πιθανοτικά μοντέλα θεμάτων μπορούν να λειτουργήσουν ως δομικά στοιχεία του κοινωνικού εταιρικού λογισμικού και να βελτιώσουν την δυνατότητα παροχής προτάσεων αλλά και να βελτιώσουν την απόδοση της αναζήτησης εσωτερικά. Η προσέγγισή μας αντιμετωπίζει προβλήματα στην επέκταση ερωτημάτων και μπορεί να προτείνει σχετικά έγγραφα και επισημειώσεις, τα οποία με τη σειρά τους βοηθούν την δημιουργία και την συντήρηση γνωσιακών δομών (ταξονομιών, φολκσονομιών, κ.α.). Επίσης παρέχουν μια βάση για την δημιουργία προτάσεων με βάση την συμπεριφορά των χρηστών και το περιεχόμενο. Η προσέγγιση μας δεν απαιτεί επιπρόσθετη προσπάθεια από τους χρήστες καθώς το περιεχόμενο που εξελίσσεται καλύπτει διάφορα λανθάνοντα θέματα.

Στα πλαίσια της μεθοδολογίας Entasis, παρουσιάστηκε μια προσέγγιση που οδηγεί στην αλληλεπίδραση ενός συστήματος προτάσεων με ένα σύστημα διαχείρισης γνωσιακών δομών, καθώς τα λανθάνοντα θέματα συνδέθηκαν με τις υπάρχουσες γνωσιακές δομές ελαφρού τύπου. Παραμένουν όμως αρκετές προκλήσεις που αφορούν τον συνδυασμό των γνωσιακών δομών και της ρητής δραστηριότητας των χρηστών με την εξαγόμενη γνώση από τα λανθάνοντα θέματα. Μια πρόκληση αφορά την σταθερή και συνεχή σύνδεση της μη-επιβλεπόμενης εξαγωγής γνώσης χωρίς τη χρήση ενός συστήματος προτάσεων.

[22] http://imu.ntua.gr/software/organik

6 Socrates - Κοινότητες Ανάπτυξης ΕΛΛΑΚ

Οι προγραμματιστές που συνεργάζονται στα πλαίσια κοινοτήτων ελεύθερου λογισμικού και λογισμικού ανοιχτού κώδικα λαμβάνουν μέρος σε ποικίλες δραστηριότητες. Στέλνουν μηνύματα ηλεκτρονικού ταχυδρομείου, σχολιάζουν τα προβλήματα που εμφανίζονται στο λογισμικό, τα επιδιορθώνουν και συνεισφέρουν πηγαίο κώδικα. Τα μέλη της κοινότητας είναι σημαντικό να γνωρίζουν τις δραστηριότητες της κοινότητας αλλά και την εξέλιξη του κώδικα.

Στο συγκεκριμένο κεφάλαιο παρουσιάζουμε μια προσέγγιση για τον συνδυασμό των λανθανόντων θεμάτων των συνεισφορών των προγραμματιστών με τις μετρικές της δραστηριότητας τους για να υποστηρίξουμε την ανάθεση εργασιών στο εσωτερικό της κοινότητας. Το σύστημα προτάσεων που περιγράφουμε παρέχει μια βαθμολογία προγραμματιστών με βάση τα θέματα (topic-based competency score) η οποία χρησιμοποιείται για την ανάθεση προβλημάτων στους πιο κατάλληλους προγραμματιστές. Στα πλαίσια της συγκεκριμένης εργασίας παρουσιάζουμε την αντίστοιχη διεπαφή χρήστη και αξιολογούμε την προτεινόμενη προσέγγιση σε δυο κοινότητες ανοιχτού λογισμικού.

6.1 Εισαγωγή

Οι προγραμματιστές στις κοινότητες ανάπτυξης λογισμικού ανοιχτού κώδικα συνήθως χρησιμοποιούν μια σειρά εφαρμογών για τη συνεργασία τους: λίστες ηλεκτρονικού ταχυδρομείου, συστήματα καταγραφής προβλημάτων (issue tracking systems), συστήματα διαχείρισης πηγαίου κώδικα (source code management system) και ιστοτόπους συζητήσεων (discussion forum). Η ανάλυση της δραστηριότητας στο εσωτερικό μιας κοινότητας μπορεί να προσφέρει σημαντική βοήθεια στους συμμετέχοντες ώστε να μπορούν να λάβουν αποφάσεις. Για την υποστήριξη των προγραμματιστών στην αντιμετώπιση τέτοιων ερωτημάτων έχει αναπτυχθεί ο τομέας των συστημάτων προτάσεων για την ανάπτυξη λογισμικού [133]. Η συγκεκριμένη απόφαση με την οποία ασχολούμαστε στο τρέχον κεφάλαιο είναι η ανάθεση της επίλυσης ενός προβλήματος στον καταλληλότερο προγραμματιστή.

Κάποιες από τις υπάρχουσες προσεγγίσεις αφορούν στην ποσοτική καταμέτρηση της δραστηριότητας των προγραμματιστών για την εξαγωγή βαθμολογίας που να αφορά τις ικανότητες τους [88]. Η μέτρηση της δραστηριότητας των προγραμματιστών μπορεί να παρέχει μια εκτίμηση της γενικότερης εξειδίκευσης τους αλλά και του βαθμού εμπειρίας και των ικανοτήτων τους. Παρ' όλα αυτά δε μπορεί να αποτυπώσει την ικανότητα ενός προγραμματιστή να αντιμετωπίσει ένα συγκεκριμένο θέμα εφ' όσον δεν αντικατοπτρίζει την περιοχή που ειδικεύεται.

Εναλλακτικά, έχουν προταθεί προσεγγίσεις που προβλέπουν την παροχή προτάσεων με βάση το περιεχόμενο. Οι συγκεκριμένες μεθοδολογίες βασίζονται στην ανάλυση της πληροφορίας που οι προγραμματιστές έχουν δημοσιεύσει με μορφή κειμένου ως τώρα στα πλαίσια της εργασίας τους [134]. Οι μετρικές που αφορούν στο περιεχόμενο ενώ αποτυπώνουν τον τομέα της ειδίκευσης του κάθε προγραμματιστή, δε λαμβάνουν υπόψη τις ικανότητες του να λύσει έναν μικρό ή μεγαλύτερο αριθμό προβλημάτων.

Ακόμη, έχουν διατυπωθεί υβριδικές προσεγγίσεις που προβλέπουν τον συνδυασμό χαρακτηριστικών που εξάγονται από το κείμενο με μετρικές δραστηριότητας [97].

Η προσέγγιση που προτείνουμε στοχεύει στην υποστήριξη των δραστηριοτήτων των προγραμματιστών λαμβάνοντας υπόψη τόσο την περιοχή της εξειδίκευσης του κάθε προγραμματιστή αλλά και τις μετρικές που αποτυπώνουν το βαθμό ενασχόλησης του και τις ικανότητες του. Για το σκοπό αυτό λαμβάνονται πληροφορίες από πολλαπλές πηγές και σχηματίζεται μια βαθμολογία με βάση τα λανθάνοντα θέματα.

Σε αυτό το κεφάλαιο ορίζουμε ένα υβριδικό σύστημα προτάσεων για προγραμματιστές που εξετάζει τόσο τον τομέα εξειδίκευσης όσο και τις μετρικές δραστηριότητας κάθε προγραμματιστή και μπορεί να λάβει υπόψη πληροφορίες από διάφορες πηγές. Το σύστημα προτάσεων που προτείνεται συνδυάζει τη λανθάνουσα σημασιολογία των δεδομένων που παρέχουν οι προγραμματιστές με ποσοτικές μετρικές που αφορούν τις δραστηριότητες των χρηστών. Η βαθμολογία που προκύπτει ονομάζεται **βαθμολογία προγραμματιστή με βάση τα θέματα** (topic-based competency score) και περιγράφει πόσο ικανός είναι ο προγραμματιστής και σε ποιους τομείς. Παρουσιάζουμε την βαθμολογία στα πλαίσια μιας διεπαφής χρήστη που μπορεί να ενσωματωθεί άμεσα σε εξωτερικά

εργαλεία. Επιπλέον αξιολογούμε πως το προτεινόμενο σύστημα μπορεί να χρησιμοποιηθεί στα πλαίσια μεγάλων ή μικρών κοινοτήτων για την δημιουργία προτάσεων

Παρακάτω εξηγούμε την εφαρμογή της σημασιολογικής ανάλυσης μέσω των μοντέλων θεμάτων καθώς και την συγκέντρωση των μετρικών δραστηριοτήτων. Στη συνέχεια περιγράφουμε τον συνδυασμό τους σε μια ενοποιημένη βαθμολογία, παρουσιάζουμε την προτεινόμενη διεπαφή χρήστη και την αξιολόγηση σε δυο κοινότητες ανάπτυξης λογισμικού ανοιχτού κώδικα.

6.2 Σχετικές Εργασίες

Τα προγενέστερα συστήματα προτάσεων για την υποστήριξη της ανάπτυξης λογισμικού βασίζονται κυρίως στις μετρικές δραστηριότητας των προγραμματιστών όπως αυτές καταγράφονται στις αλλαγές του πηγαίου κώδικα.

Για παράδειγμα, έχει προταθεί η μέτρηση των δεξιοτήτων κάθε προγραμματιστή με βάση μια μετρική που ονομάζεται experience atoms (EA) [88]. Οι συγκεκριμένες μονάδες εμπειρίας ορίζονται από τους συγγραφείς ως οι ελάχιστες μετρήσιμες δραστηριότητες ενός ανθρώπου σε σχέση με το προϊόν μιας εργασίας και ο αριθμός τους μπορεί να απεικονίσει την συσσωρευμένη εμπειρία του. Αντίστοιχα, η δραστηριότητα ενός προγραμματιστή που αφορά κάθε συγκεκριμένο αρχείο μπορεί να κωδικοποιηθεί σε έναν χάρτη όπου καταγράφει τη συχνότητα των συνεισφορών του στο συγκεκριμένο αρχείο [89]. Εδώ εισάγονται κοινωνικά δεδομένα καθώς η συγκεκριμένη συχνότητα κανονικοποιείται και με βάση τον αριθμό των προγραμματιστών που έχουν συνεισφέρει στο συγκεκριμένο αρχείο στο παρελθόν. Επίσης έχει διερευνηθεί η δυνατότητα χρήσης όχι μόνο της συχνότητας αλλά και της χρονικής απόστασης από την τελευταία δραστηριότητα του κάθε προγραμματιστή για τον υπολογισμό των δεξιοτήτων του [90]. Τέλος, έχει προταθεί ο συνδυασμός διαφορετικών μετρικών όπως ο αριθμός δραστηριοτήτων, η συχνότητα δραστηριοτήτων, η χρονική απόσταση από την τελευταία δραστηριότητα, ο αριθμός των αρχείων που αλλάζονται και ο αριθμός των γραμμών κώδικα που προστίθενται ή αφαιρούνται [91].

Από την άλλη πλευρά, υπάρχουν προσεγγίσεις που εστιάζουν περισσότερο το περιεχόμενο των συνεισφορών των προγραμματιστών. Στη βιβλιογραφία έχει προταθεί η χρήση όρων που μπορούν να βρεθούν στον πηγαίο κώδικα με τη μορφή

διανύσματος όρων για την πρόταση προγραμματιστών που έχουν την ανάλογη πείρα για να λύσουν ένα πρόβλημα [92]. Αυτό αποτελεί ένα παράδειγμα αμιγώς ποιοτικής διαφοροποίησης, όπου μόνον το πεδίο ειδίκευσης υπολογίζεται. Αντίστοιχα, έχει περιγραφεί ένα σύστημα προτάσεων που βοηθά τους μηχανικούς λογισμικού να εναλλάσσουν αντικείμενο εργασίας με βάση τον τύπο της ανάπτυξης λογισμικού και το ιστορικό αλληλεπιδράσεων μεταξύ τους [93]. Τέλος μια παραλλαγή της λανθάνουσας κατανομής Dirichlet μπορεί να υποστηρίξει τη μοντελοποίηση της εξέλιξης των θεμάτων στις αποθήκες πηγαίου κώδικα λογισμικού [65].

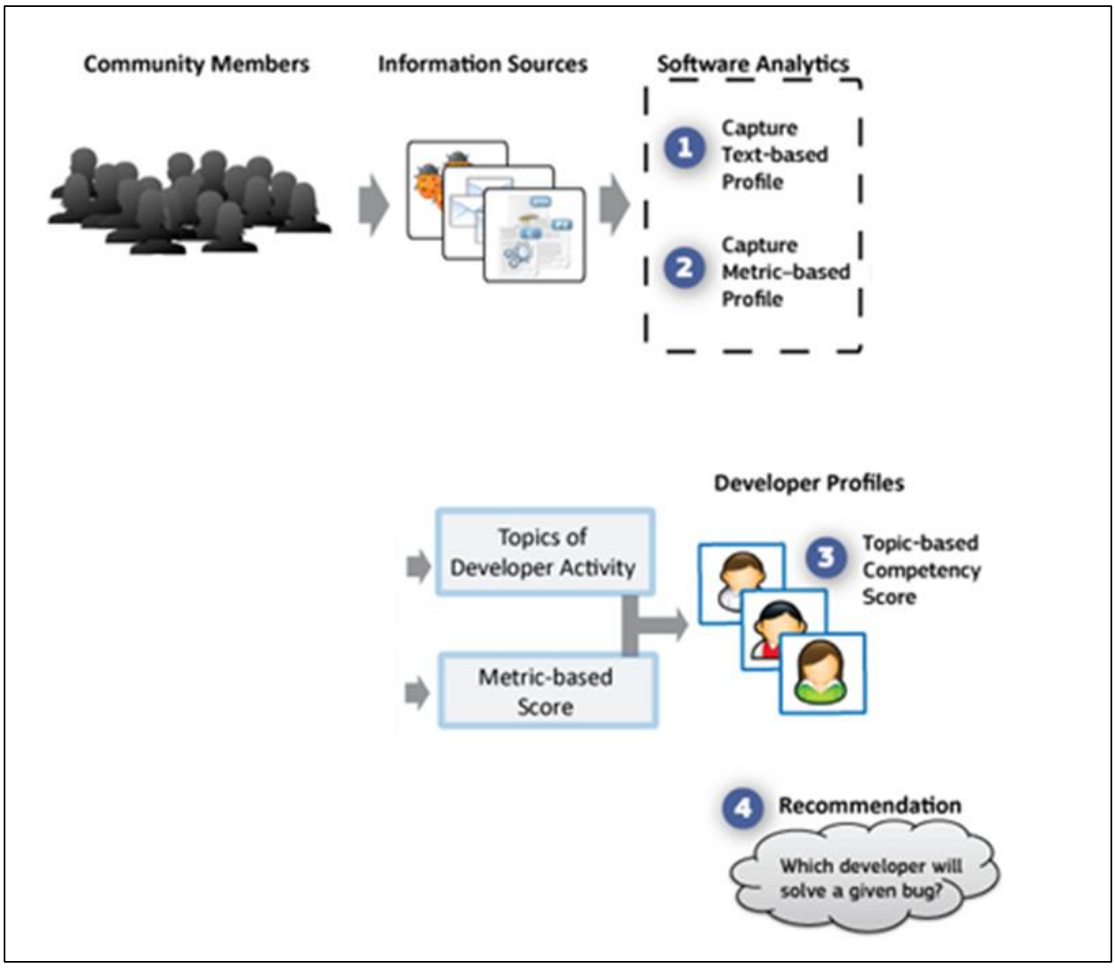
Ένας σημαντικός αριθμός μελετών δείχνει ότι τα μοντέλα θεμάτων μπορούν να εφαρμοστούν με επιτυχία στο εσωτερικό κοινοτήτων ώστε να βοηθήσουν την εξαγωγή πληροφορίας από πολλαπλές πηγές σε ένα έργο, αλλά και από περισσότερο από ένα έργα. Μια προσέγγιση μοντέλων θεμάτων έχει εφαρμοστεί ώστε να εντοπιστούν οι σχέσεις μεταξύ του πηγαίου κώδικα και αντικειμένων υψηλού επιπέδου, όπως απαιτήσεις λογισμικού [94]. Τα μοντέλα θεμάτων έχουν χρησιμοποιηθεί για να δείξουν την σχέση ανάμεσα στην δραστηριότητα των προγραμματιστών στα ιστολόγια και στις συνεισφορές τους [95], ενώ μια εναλλακτική προσέγγιση εφαρμόζει την ανάλυση θεμάτων σε πολλά έργα και με τη χρήση μιας ταξονομίας μπορεί να παρέχει προτάσεις ονοματοδοσίας θεμάτων [96].

Μια μέθοδος με βάση την λανθάνουσα κατανομή Dirichlet (LDA) έχει χρησιμοποιηθεί για την ανάλυση του πηγαίου κώδικα του λογισμικού Eclipse και για να παρέχει εκτεταμένη επίγνωση της συνάφειας μεταξύ προγραμματιστών [97]. Επίσης, έχει παρουσιάζεται μια μεθοδολογία που βασίζεται σε μια πηγή πληροφοριών για την εξαγωγή των δεξιοτήτων των προγραμματιστών με βάση τα μοντέλα θεμάτων του περιεχομένου και την δραστηριότητα των προγραμματιστών [66]. Παρ’ όλα αυτά, οι μέθοδοι λανθάνουσας σημασιολογικής ανάλυσης δεν χρησιμοποιούνται εκτενώς με αυτόν τον τρόπο και δεν καλύπτουν τις διάφορες πηγές πληροφοριών.

Παρατηρώντας τη σχετική βιβλιογραφία συμπεραίνουμε ότι οι προσπάθειες για την αποτύπωση δεξιοτήτων και χρήση τους σε συστήματα προτάσεων στο παρελθόν βασίζονταν κυριότερα σε ποιοτικές ενδείξεις με βάση το κείμενο ή σε ποσοτικές με βάση την δραστηριότητα. Κάποιες προσπάθειες που έχουν γίνει στο παρελθόν για συνδυασμό των δυο κατηγοριών δεν καλύπτουν περισσότερα από ένα εργαλεία και δεν προσφέρουν μια εποπτική εικόνα των δεξιοτήτων του κάθε συνεργάτη.

6.3 Προσέγγιση

Στην τρέχουσα ενότητα περιγράφουμε την προτεινόμενη προσέγγιση για την εξαγωγή χρήσιμων συμπερασμάτων από την δραστηριότητα στο εσωτερικό μιας κοινότητας με τη μορφή βαθμολογιών προγραμματιστών. Τα συμπεράσματα αυτά χρησιμοποιούνται για την πραγματοποίηση προτάσεων προγραμματιστών για να επιλύσουν συγκεκριμένα προβλήματα. Η Εικόνα 6.1 παρουσιάζει μια εποπτική εικόνα της προτεινόμενης διαδικασίας.



Εικόνα 6.1 Διαδικασία Δημιουργίας Προτάσεων Προγραμματιστών

Η διαδικασία μπορεί να χωριστεί σε στάδια. Το πρώτο στάδιο της διαδικασίας περιλαμβάνει την ανάλυση της δραστηριότητας του προγραμματιστή στην κοινότητα. Το στάδιο αυτό χωρίζεται σε δυο επιμέρους βήματα: (1) την εξαγωγή των ποιοτικών στοιχείων της δραστηριότητας του προγραμματιστή από το κείμενο το οποίο έχει συνεισφέρει στα διάφορα εργαλεία και (2) την μέτρηση της

ικανότητας του προγραμματιστή με χρήση κατάλληλων μετρικών δραστηριότητας στα εργαλεία που χρησιμοποιούνται στην κοινότητα. Η πληροφορία που παρέχεται από το βήμα αυτό συνδυάζεται και ενοποιείται σε μια βαθμολογία θεμάτων για προγραμματιστές που μπορεί να ερμηνευθεί εύκολα από ανθρώπους. Τέλος, οι βαθμολογίες αυτές χρησιμοποιούνται για την δημιουργία προτάσεων προγραμματιστών για ένα δεδομένο πρόβλημα.

Βήμα 1: Εντοπισμός των Θεμάτων στα οποία συνεισφέρει κάθε προγραμματιστής

Για την πραγματοποίηση της σημασιολογικής ανάλυσης, ένας αριθμός μη επιβλεπόμενων τεχνικών έχουν προταθεί που προχωρούν πέρα από την ανάλυση των μοντέλων διανυσματικού χώρου λέξεων (όπως το TF*IDF). Η λανθάνουσα σημασιολογική ανάλυση έχει αναπτυχθεί ακριβώς για να εντοπίζει τις νοηματικές συσχετίσεις μεταξύ λέξεων σε μια συλλογή εγγράφων κειμένου, ενώ η πιθανοτική λανθάνουσα σημασιολογική ανάλυση (pLSA) ενσωματώνει ένα πιθανοτικό υπόβαθρο [51]. Η λανθάνουσα κατανομή Dirichlet (LDA) [54] έχει προταθεί ως μια εξέλιξη της pLSA η οποία περιλαμβάνει εκ των προτέρων κατανομές στην δημιουργία θεμάτων και λέξεων.

Σε αυτό το στάδιο εφαρμόζουμε την λανθάνουσα κατανομή Dirichlet η οποία βασίζεται στην υπόθεση ότι ένα γενετικό μοντέλο μπορεί να περιγράψει επαρκώς το σώμα των κειμένων που διαθέτουμε καθώς και ότι το μοντέλο αυτό μπορεί να ανακαλυφθεί: τα έγγραφα παράγονται από δειγματοληψία θεμάτων από μια κατανομή θεμάτων σε έγγραφα, και οι λέξεις λαμβάνονται μέσω δειγματοληψίας λέξεων από μια κατανομή λέξεων με δεδομένο το θέμα. Οι μεταβλητές που αναπαριστούν την σχέση μεταξύ λέξεων και θεμάτων αλλά και θεμάτων και εγγράφων έχουν μια εκ των προτέρων κατανομή Dirichlet και πρέπει να προσδιοριστούν για να περιγραφεί πλήρως το μοντέλο. Έτσι, η LDA εκπαιδεύεται από την ανάλυση ενός αριθμού εγγράφων κειμένου και οι κατανομές πιθανοτήτων συγκλίνουν ώστε το μοντέλο να μπορεί να χρησιμοποιηθεί σε εφαρμογές.

Εδώ λαμβάνουμε πληροφορίες από τα εργαλεία που έχουν χρησιμοποιηθεί για την ανταλλαγή πληροφοριών μεταξύ προγραμματιστών. Σε αυτά περιλαμβάνονται οι λίστες ηλεκτρονικού ταχυδρομείου, συστήματα διαχείρισης πηγαίου κώδικα, συστήματα καταγραφής προβλημάτων και ιστότοποι συζητήσεων. Έτσι, τα έγγραφα μπορούν να πάρουν διάφορες μορφές: μηνύματα ηλεκτρονικού ταχυδρομείου, συνεισφορά κώδικα, περιγραφές προβλημάτων και δημοσιεύσεις σε

συζητήσεις. Εφαρμόζουμε σημασιολογική ανάλυση σε όλα τα διαφορετικά αντικείμενα που περιέχουν κείμενο. Η εφαρμογή αυτή οδηγεί στην εξαγωγή ενός κοινού μοντέλου θεμάτων καθώς και στον υπολογισμό των κατανομών θεμάτων για κάθε έγγραφο. Τα λανθάνοντα θέματα αναπαριστούν τις περιοχές ειδίκευσης των προγραμματιστών και αντιστοιχούν σε ομάδες πιθανών λέξεων όπως «*methods*, *parameters*, *api»* και «*mount*, *location*, *directory»*.

Βήμα 2: Μέτρηση της Δραστηριότητας των Προγραμματιστών

Το επόμενο βήμα για την εκτίμηση της πείρας και των ικανοτήτων των προγραμματισμών είναι η δημιουργία προφίλ με βάση τις ποσοτικές πληροφορίες που έχουν εξαχθεί κατά τη διάρκεια της ανάπτυξης λογισμικού. Οι συγκεκριμένες πληροφορίες μπορούν να καταγραφούν και να επεξεργαστούν από τα εργαλεία που χρησιμοποιούνται για την συνεργασία μεταξύ των χρηστών. Κάθε εργαλείο αντιμετωπίζεται ως μια διαφορετική πηγή πληροφορίας από την οποία μπορούν να εξαχθούν συμπεράσματα με τη μορφή μετρικών δραστηριότητας. Μια εκτενής λίστα μετρικών αξιολόγησης προγραμματιστών παρουσιάζεται στο [136].

Εδώ προτείνουμε την μέτρηση των δραστηριοτήτων των προγραμματιστών στα διαθέσιμα εργαλεία συνεργασίας μιας κοινότητας. Οι μετρικές που εξάγονται περιλαμβάνουν τον αριθμό των γραμμών κώδικα που παράγεται, τον αριθμό των διεπαφών που δημιουργήθηκαν (API), τον αριθμό των σχολίων στο σύστημα καταγραφής προβλημάτων, τον αριθμό των προβλημάτων που επιλύθηκαν, τον αριθμό απαντήσεων σε συζητήσεις και στη λίστα ηλεκτρονικού ταχυδρομείου και τον χρόνο που πέρασε από την τελευταία συνεισφορά. Κάθε μετρική πολλαπλασιάζεται με ένα βάρος συνεισφοράς για να δοθεί μεγαλύτερη επιρροή στις πιο σημαντικές μετρικές. Στην περίπτωση μας το μεγαλύτερο βάρος δίνεται στον αριθμό των προβλημάτων που έχουν λυθεί από τον συγκεκριμένο προγραμματιστή. Το αποτέλεσμα που προκύπτει είναι μια κανονικοποιημένη βαθμολογία δραστηριότητας.

Βήμα 3: Υπολογισμός Βαθμολογίας Προγραμματιστή με βάση Θέματα

Στο συγκεκριμένο βήμα συνδυάζουμε τα χαρακτηριστικά που βασίζονται στο κείμενο με τις διάφορες μετρικές δραστηριότητας ώστε να σχηματίσουμε μια βαθμολογία που θα αναπαριστά με ακρίβεια την εμπειρία ενός προγραμματιστή. Η βαθμολογία αυτή θα μπορεί να απαντήσει στις ερωτήσεις (α) ποιος είναι ικανός σε

τι; και (β) πόσο ικανός είναι; Στη συνέχεια η βαθμολογία αυτή χρησιμοποιείται για την πραγματοποίηση προτάσεων.

Αυτές οι δυο μορφές πληροφορίας ενοποιούνται σε ένα διάνυσμα ικανοτήτων. Χρησιμοποιούμε τα θέματα της δραστηριότητας των καταναλωτών ως διαστάσεις για το συγκεκριμένο διάνυσμα. Οι αρχικές τιμές του διανύσματος αναπαριστούν την ομοιότητα της δραστηριότητας του προγραμματιστή με κάθε θέμα και κυμαίνονται από 0 εως 1. Σε αυτό το στάδιο το άθροισμα όλων των διαστάσεων είναι 1. Πολλαπλασιάζουμε το διάνυσμα με την κανονικοποιημένη βαθμολογία που βασίζεται στη δραστηριότητα του προγραμματιστή ώστε να έχουμε ένα πιο ακριβές αποτέλεσμα.

Η εξίσωση (6.1) περιλαμβάνει έναν τυπικό ορισμό του συγκεκριμένου υπολογισμού: $\vec{s}$ είναι το διάνυσμα της βαθμολογίας προγραμματιστή με βάση τα θέματα, K είναι ο αριθμός των διάφορων μετρικών, m_i είναι η κάθε μετρική δραστηριότητας και w_i το αντίστοιχο βάρος, ενώ $\overrightarrow{tp}$ είναι το διάνυσμα που περιγράφει την κατανομή θεμάτων του κάθε προφίλ χρήστη.

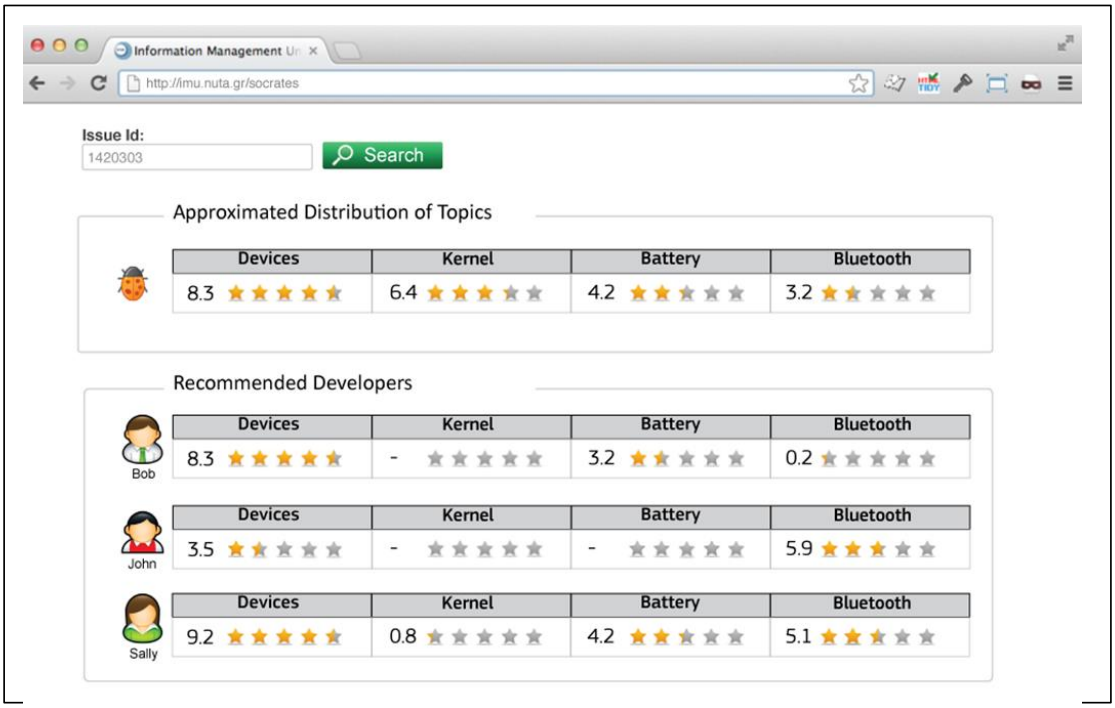
$$\vec{s} = normalize(\sum_{i=1}^{K} w_i m_i)\ \overrightarrow{tp} \qquad (6.1)$$

Ένα παράδειγμα που δείχνει τόσο τις βαθμολογίες αλλά και την αντίστοιχη διεπαφή χρήστη εμφανίζεται στην Εικόνα 6.2. Χρησιμοποιώντας την αντίστοιχη διεπαφή ο χρήστης μπορεί να αναζητήσει οποιοδήποτε πρόβλημα που είναι ενεργό στην κοινότητα. Στη συνέχεια, το προτεινόμενο σύστημα προτάσεων παρέχει μια απεικόνιση των προτεινόμενων προγραμματιστών και των αντίστοιχων βαθμολογιών τους.

Βήμα 4: Πρόταση Προγραμματιστών

Το τελευταίο βήμα στην προτεινόμενη προσέγγιση περιλαμβάνει την πρόταση ενός προγραμματιστή που σύμφωνα με το σύστημα είναι ο πλέον κατάλληλος να αναλάβει την επίλυση ενός συγκεκριμένου προβλήματος. Για να παράγουμε την λίστα με τους προτεινόμενους προγραμματιστές, υπολογίζουμε την ομοιότητα μεταξύ του διανύσματος που περιγράφει το θέμα προς επίλυση και του διανύσματος που περιγράφει τον προγραμματιστή. Υπολογίζουμε το εσωτερικό γινόμενο για να αναπαραστήσουμε ακριβώς την ομοιότητα μεταξύ των δυο διανυσμάτων, τόσο την ικανότητα του προγραμματιστή (το μήκος του διανύσματος) αλλά και τον τομέα που ασχολείται (την διεύθυνση του). Πολλαπλασιάζουμε το διάνυσμα περιγραφής του προβλήματος με μια σταθερά αρκετά μεγάλη (ίση με την

μέγιστη κανονικοποιημένη βαθμολογία δραστηριότητας ενός προγραμματιστή) για να βεβαιωθούμε ότι πάντοτε ανταμείβεται ο ικανότερος προγραμματιστής.

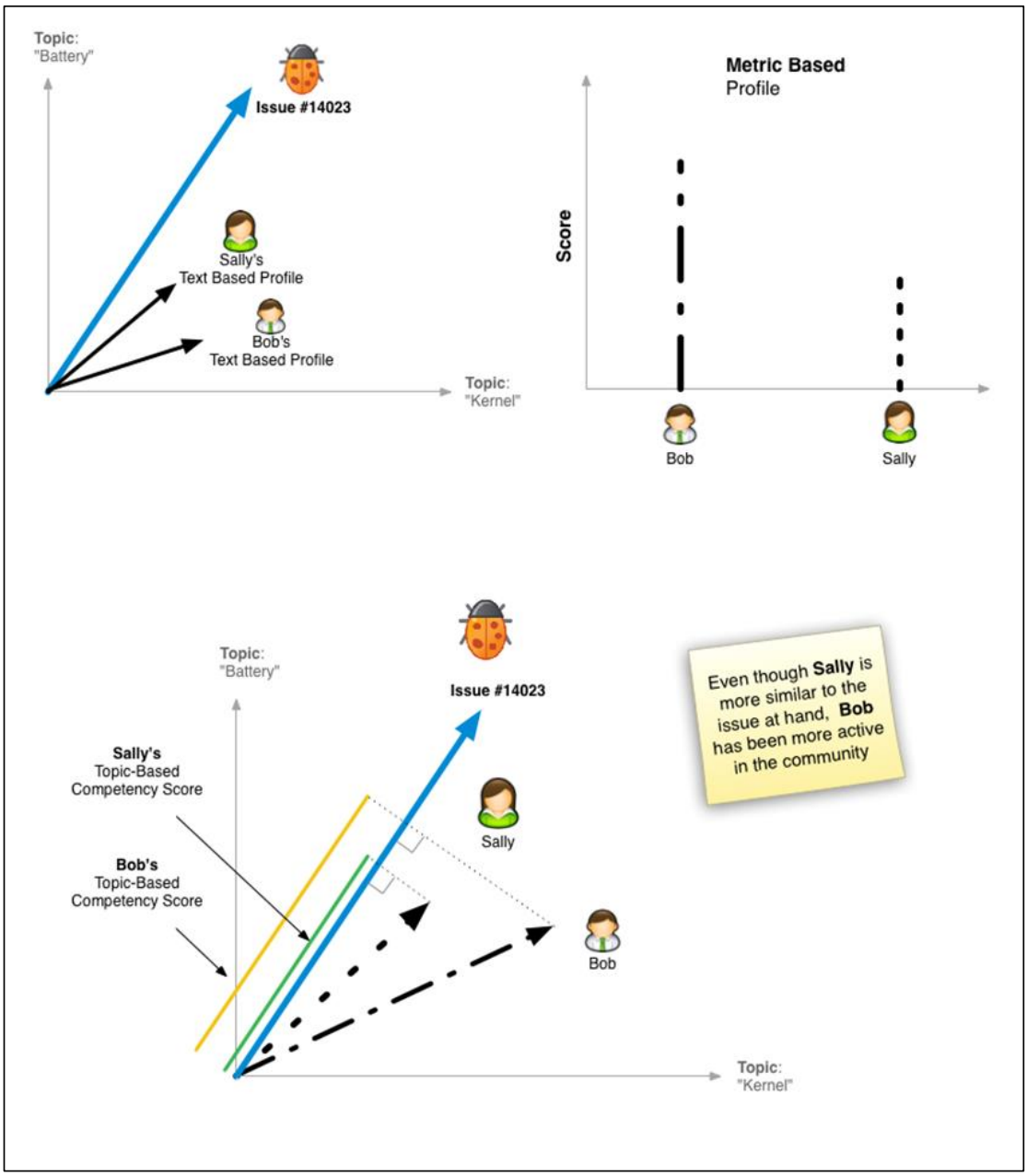


Εικόνα 6.2 Βαθμολογία Προγραμματιστή με Βάση Θέματα

Η εξίσωση (6.2) περιγράφει με τυπικό τρόπο τον υπολογισμό: το $\vec{s}$ είναι το διάνυσμα που αναπαριστά την βαθμολογία με θέματα του προγραμματιστή, το $\vec{\imath}$ είναι το διάνυσμα που αναπαριστά την περιγραφή του προβλήματος και φ είναι η μεταξύ τους γωνία.

$$similarity = cos(\varphi)\,|\vec{s}\,||\vec{\imath}\,| \qquad (6.2)$$

Ένα παράδειγμα απεικονίζεται στην Εικόνα 6.3. Τα ενδιαφέροντα της Sally είναι πιο κοντινά με το συγκεκριμένο πρόβλημα. Όμως ο Bob επειδή ήταν πιο ενεργός στην κοινότητα προτείνεται για να αναλάβει την επίλυση του προβλήματος.



Εικόνα 6.3 Υπολογισμός Ομοιότητας Προβλήματος και Προγραμματιστών

6.4 Αξιολόγηση

Στη συνέχεια παρουσιάζουμε την αξιολόγηση της προτεινόμενης διαδικασίας σε δυο κοινότητες ανάπτυξης λογισμικού ανοιχτού κώδικα. Η αξιολόγηση αυτή στοχεύει στον έλεγχο της εφαρμογής και της χρησιμότητας της προτεινόμενης προσέγγισης. Περιγράφουμε τα χαρακτηριστικά των κοινοτήτων και

στην συνέχεια δείχνουμε τις διαφορές στην αποτελεσματικότητα μεταξύ της προτεινόμενης προσέγγισης και άλλων αλγορίθμων που χρησιμοποιούμε ως βάση αναφοράς.

6.4.1 Κοινότητες

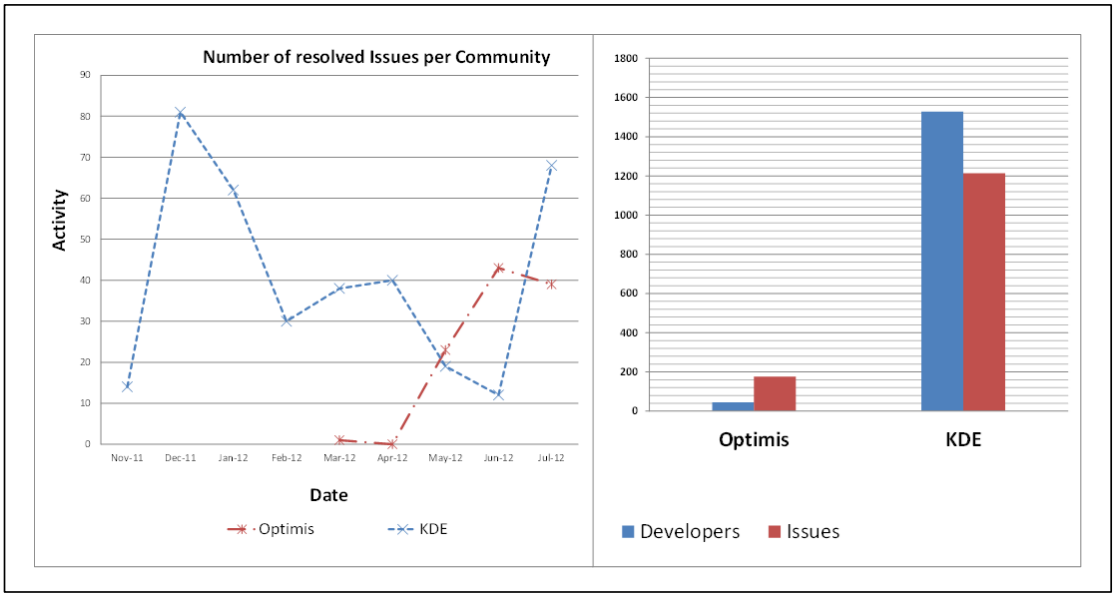
Εξετάζουμε την εφαρμογή του συστήματος προτάσεων Socrates σε δυο κοινότητες ανοιχτού λογισμικού διαφορετικού μεγέθους: στην κοινότητα του KDE [23] και στην κοινότητα OPTIMIS[24].

Η κοινότητα KDE είναι μια διεθνής ομάδα που συνεργάζεται στην ανάπτυξη και διανομή ελεύθερου λογισμικού ανοιχτού κώδικα για εφαρμογές προσωπικών σταθερών και φορητών υπολογιστών. Πιο συγκεκριμένα εξετάζουμε την δραστηριότητα 1.529 προγραμματιστών της κοινότητας και συγκεκριμένα το πώς επιλύουν μια σειρά από 1.213 προβλήματα τα οποία αφορούν ένα μόνο προιόν, το KDE Solid[25]. Το προιόν αυτό εξειδικεύεται στην παροχή υποστήριξης ενσωμάτωσης υλικού και λογισμικού στην επιφάνεια εργασίας του KDE.

Η κοινότητα OPTIMIS είναι μικρή ομάδα ανάπτυξης λογισμικού που εστιάζει την εργασία της σε υπηρεσίες νέφους που σχετίζονται με την εμπιστοσύνη, την διαχείριση ρίσκου, την οικολογικότητα και την επίγνωση κόστους. Συνολικά εξετάζουμε την δραστηριότητα 46 προγραμματιστών στα πλαίσια της κοινότητας και την επίλυση 176 προβλημάτων.

Μια εποπτική εικόνα των διαφορών στα μεγέθη και στη δραστηριότητα των κοινοτήτων μπορεί να βρεθεί στην Εικόνα 6.4. Στα αριστερά της εικόνας παρατηρούμε τον αριθμό των λυμένων προβλημάτων σε διάστημα 10 μηνών και στις δυο κοινότητες. Στα δεξιά, παρουσιάζουμε την διαφορά στον αριθμό των προγραμματιστών και στον αριθμό των προβλημάτων σε κάθε κοινότητα.

[23] http://www.kde.org/
[24] http://www.optimis-project.eu/
[25] http://solid.kde.org/



Εικόνα 6.4 Χαρακτηριστικά Κοινοτήτων

6.4.2 Πειραματική Αξιολόγηση

Έχουμε αξιολογήσει την προτεινόμενη μεθοδολογία στις κοινότητες ανοιχτού λογισμικού KDE και Optimis. Επίσης, ως βάση αναφοράς έχουμε εφαρμόσει τόσο μετρικές που βασίζονται στη δραστηριότητα όσο και μετρικές που βασίζονται στο κείμενο.

Οι μετρικές δραστηριότητας που χρησιμοποιήσαμε αφορούν έναν προγραμματιστή και περιλαμβάνουν τα παρακάτω: τον αριθμό των προβλημάτων που έχει λύσει, τον χρόνο που έχει περάσει από το τελευταίο πρόβλημα που έλυσε, τον αριθμό των μηνυμάτων ηλεκτρονικού ταχυδρομείου που έχει στείλει, τον χρόνο που πέρασε από το τελευταίο μήνυμα που έχει στείλει και τον αριθμό των σχολίων που έχει δημοσιεύσει. Οι μετρικές αυτές έχουν συντεθεί σε μια μοναδική βαθμολογία που είναι ένας σταθμισμένος μέσος των παραπάνω και είναι συγκρίσιμος με τις απλές μετρικές σε αποτελεσματικότητα ενώ είναι πιο αντιπροσωπευτικός της δραστηριότητας του προγραμματιστή.

Επίσης πειραματιστήκαμε με μεθόδους που βασίζονται στο κείμενο και αναλύουν το περιεχόμενο που έχει συνεισφέρει ο προγραμματιστής στην κοινότητα. Πριν την επεξεργασία του κειμένου έχουμε απομονώσει τις λέξεις –

κλειδιά χρησιμοποιώντας το λογισμικό Alert-KEUI[26] στις διάφορες συνεισφορές που έχουν δημοσιευτεί. Στη συνέχεια χρησιμοποιούμε την συχνότητα των όρων και την ομοιότητα μοντέλων θεμάτων για να παρέχουμε προτάσεις.

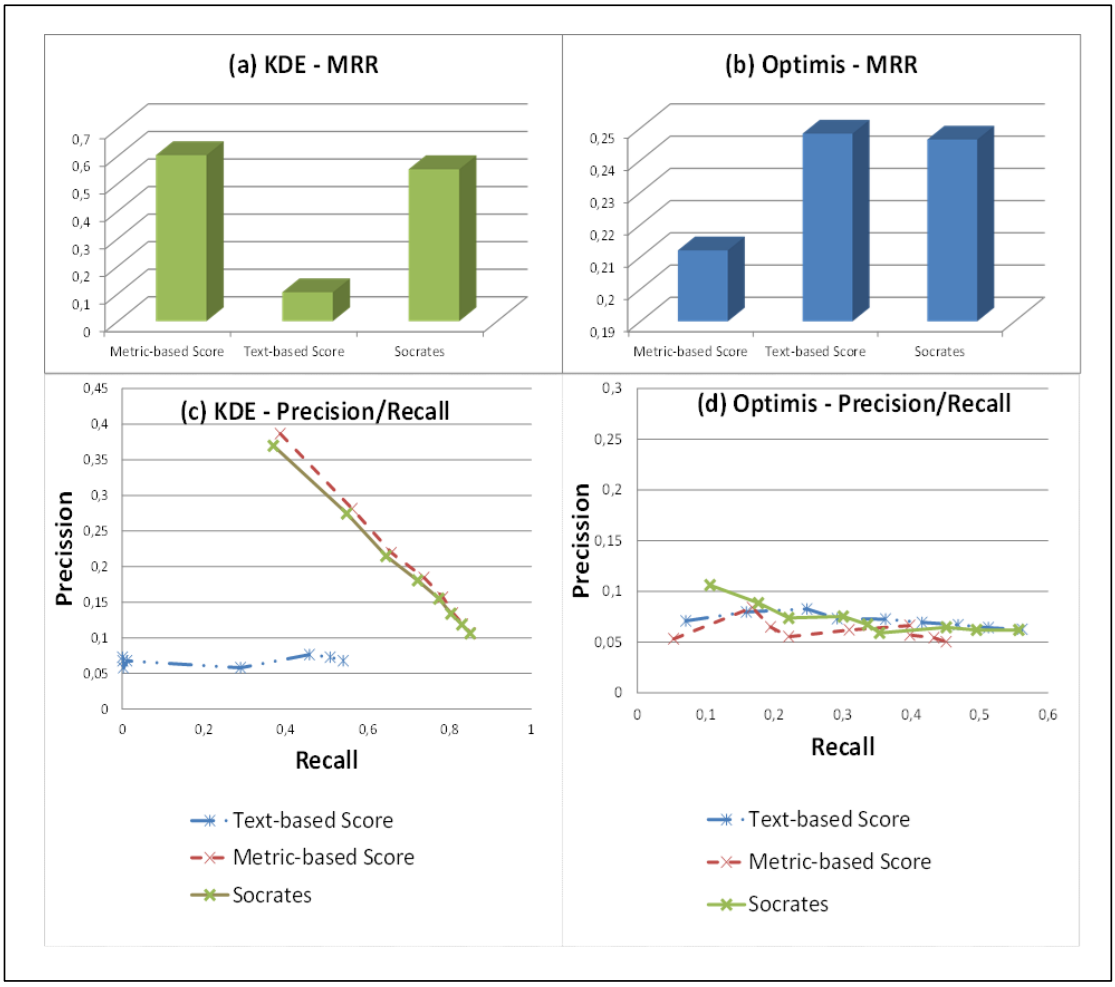
Συνοπτικά παρουσιάζουμε την εφαρμογή μιας μεθόδου βασισμένη σε κείμενο, μιας μεθόδου βασισμένης σε μετρικές δραστηριότητας και τέλος της υβριδικής μεθόδου που αποτελεί και τη συνεισφορά του συγκεκριμένου κεφαλαίου. Τα αποτελέσματα της σύγκρισης μπορούν να βρεθούν στην Εικόνα 6.5.

Δυο μετρικές έχουν χρησιμοποιηθεί για την συγκεκριμένη σύγκριση, η μέση αμοιβαία κατάταξη (mean reciprocal rank, MRR) και τα γραφήματα ακρίβειας – ανάκλησης (precision-recall graphs). Και οι δυο μετρικές αξιολογούν τις απαντήσεις των συστημάτων προτάσεων στην ερώτηση: «Ποιος θα επιλύσει το συγκεκριμένο πρόβλημα λογισμικού;». Σε αυτή την κατεύθυνση, η MRR αποκαλύπτει πόσο γρήγορα ένα σύστημα προτάσεων θα ανακτήσει την σωστή απάντηση. Τα γραφήματα ακρίβειας - ανάκλησης δείχνουν πόσο γρήγορα θα βρεθεί η σωστή απάντηση αλλά και πόση ακρίβεια παρουσιάζουν οι προτάσεις. Στην Εικόνα 6.5 συγκρίνουμε την μετρική MRR στις κοινότητες KDE (a) και Optimis (b), ενώ παρουσιάζουμε τα γραφήματα ακρίβειας - ανάκλησης στο KDE (c) και στο Optimis (d).

Ένα αρχικό συμπέρασμα που παρέχεται από τα αποτελέσματα που παρουσιάζονται στην Εικόνα 6.5 είναι η διαφορά της απόδοσης μεταξύ των μετρικών που βασίζονται στο κείμενο και στη δραστηριότητα ανάλογα με την κοινότητα που αναφερόμαστε. Οι μέθοδοι που βασίζονται στο κείμενο ξεπερνούν σημαντικά τις μεθόδους που βασίζονται σε μετρικές δραστηριότητας όταν αναφερόμαστε στην μικρή κοινότητα Optimis. Ωστόσο, τα ευρήματα αυτά αντιστρέφονται όταν αξιολογούμε την εφαρμογή στην μεγαλύτερη κοινότητα KDE. Αυτό μπορεί να εξηγηθεί από το γεγονός ότι σε μια μικρή κοινότητα, οι περισσότεροι προγραμματιστές έχουν παρόμοια ποσότητα προσφοράς και οι ποιοτικές διαφορές τους, όπως αυτές αντανακλώνται στο κείμενο των συνεισφορών τους, μπορούν να διαχωρίσουν τον κατάλληλο για την επίλυση ενός προβλήματος. Αντιθέτως, σε μια μεγαλύτερη κοινότητα οι μετρικές δραστηριότητας μπορούν να ξεχωρίσουν τα ενεργά μέλη της κοινότητας από τους περιστασιακούς προγραμματιστές.

[26] http://www.mediafire.com/download.php?g6eao8tc72hy88b

Η απόδοση της προτεινόμενης προσέγγισης είναι παρόμοια με εκείνη της καλύτερης προσέγγισης ανά περίπτωση. Στην περίπτωση του Optimis, και κυρίως στην κορυφή της λίστας των προτάσεων η προσέγγιση μας ξεπερνά σε απόδοση τις άλλες δυο, όπως φαίνεται στα γραφήματα ακρίβειας - ανάκλησης.



Εικόνα 6.5 Σύγκριση Μεθόδων Δημιουργίας Προτάσεων

6.5 Συμπεράσματα

Στο κεφάλαιο αυτό προτείναμε ένα πλαίσιο δημιουργίας προτάσεων για κοινότητες ανάπτυξης ελεύθερου λογισμικού και λογισμικού ανοιχτού κώδικα. Η προτεινόμενη μεθοδολογία μπορεί να εφαρμοστεί σε κοινότητες με διαφορετικά μεγέθη και με διάφορα εργαλεία συνεργασίας. Περιγράψαμε την προσέγγιση μας και δείξαμε μια αντίστοιχη διεπαφή χρήστη. Επίσης αξιολογήσαμε το σύστημα και δείξαμε την αποτελεσματικότητά του σε μικρές και μεγαλύτερες κοινότητες ανάπτυξης λογισμικού.

Για την επέκταση της μεθοδολογίας η προτεινόμενη προσέγγιση μπορεί να αξιολογηθεί σε επιπλέον κοινότητες με χρήση πειραμάτων εκτός σύνδεσης αλλά και με χρήστες. Στις αξιολογήσεις εκτός σύνδεσης σκοπεύουμε να αξιολογήσουμε το σύστημα σε επιπλέον διαφορετικά σύνολα δεδομένων ενώ να αυξήσουμε και τον αριθμό των μετρικών τις οποίες συνδυάζουμε. Στις αξιολογήσεις με χρήστες θα επιδιωχθεί η εξαγωγή ποιοτικών συμπερασμάτων που να αφορούν την λειτουργία του συστήματος.

Επίσης ένας τομέας που μπορεί να αναλυθεί, είναι ο τομέας των κοινωνικών δεσμών μεταξύ των προγραμματιστών. Το Socrates μπορεί να επεκταθεί ώστε να ανακαλύψει υπάρχοντες δεσμούς μεταξύ των προγραμματιστών και να δημιουργήσει προτάσεις με βάση το κοινωνικό δίκτυο του χρήστη. Τέτοιες προτάσεις μπορεί να λαμβάνουν υπόψη όχι μόνο τις δεξιότητες τους αλλά και φιλικές σχέσεις και σχέσεις εμπιστοσύνης [137].

7 FillBasket - Πελάτες Υπεραγορών

Η εκμάθηση των προτιμήσεων των καταναλωτών και η δημιουργία προτάσεων τόσο στο εμπόριο όσο και στο ηλεκτρονικό εμπόριο είναι ένα πρόβλημα το οποίο έχει ερευνηθεί διεξοδικά και για την επίλυση του οποίου έχουν προταθεί διάφορες προσεγγίσεις.

Σε αυτό το κεφάλαιο προτείνουμε μια προσέγγιση στο πρόβλημα της εξαγωγής και μοντελοποίησης των προτιμήσεων των χρηστών στο εμπόριο με βάση τα λανθάνοντα θέματα. Διερευνούμε την χρήση πιθανοτικών μοντέλων σε στοιχειοσύνολα συναλλαγών στα οποία θεωρούμε τόσο το τρέχον καλάθι αγορών των καταναλωτών όσο και το προηγούμενο ιστορικό των πελατών. Συμπεραίνουμε ότι τα μοντέλα που εξάγονται όχι μόνο μπορούν να παρέχουν μια εποπτική εικόνα της συμπεριφοράς των καταναλωτών αλλά μπορούν επίσης να υποστηρίξουν ένα σύστημα προτάσεων προϊόντων.

7.1 Εισαγωγή

Στη περιοχή της έρευνας μάρκετινγκ έχουν εφαρμοστεί μεθοδολογίες εξόρυξης δεδομένων και μηχανικής μάθησης σε συναλλαγές λιανικής στις οποίες αναλύεται ένας μεγάλος όγκος δεδομένων αγορών [138]. Η ανάλυση καλαθιού αγοράς είναι ο κλάδος που αφορά την ανακάλυψη μοτίβων συσχετίσεων σε συναλλαγές λιανικής. Ο συγκεκριμένος κλάδος έχει θέσει τις βάσεις για την ανάπτυξη εφαρμογών όπως η ομαδοποίηση προϊόντων, ο εντοπισμός εξαρτήσεων μεταξύ κατηγοριών αλλά και η δημιουργία προφίλ καταναλωτών [139], [140].

Η έλευση του ηλεκτρονικού εμπορίου άνοιξε το δρόμο για πολυάριθμες προόδους σε τεχνικές και μοντέλα τα οποία επιδιώκουν να βελτιώσουν την εμπειρία των καταναλωτών σε ηλεκτρονικά καταστήματα. Μεταξύ αυτών, η εκμάθηση προτιμήσεων έχει σκοπό να προσδιορίσει τις επιθυμίες των καταναλωτών με ένα ρητό τρόπο [99]. Η εκμάθηση των προτιμήσεων των καταναλωτών μπορεί να αποτελέσει την βάση για διαφορές εφαρμογές, όπως για παράδειγμα, την πρόταση προϊόντων.

Δυο βασικές κατευθύνσεις έχουν ακολουθηθεί στην πραγματοποίηση προτάσεων στο πεδίο της ανάλυσης των συναλλαγών των καταναλωτών (ή αλλιώς ανάλυση καλαθιού αγοράς).

Αρχικά η έρευνα έχει επικεντρωθεί στην εφαρμογή της εξαγωγής κανόνων συσχέτισης. Η εξαγωγή κανόνων συσχέτισης περιλαμβάνει την ανάλυση συνολοστοιχείων (ομάδων προϊόντων που αγοράζονται μαζί) και την εξαγωγή κανόνων που συσχετίζουν προϊόντα μεταξύ τους. Η αρχική προσέγγιση των κανόνων συσχέτισης μπορεί να παράγει κανόνες που συσχετίζουν την εμφάνιση προϊόντων, όπου, για παράδειγμα, αν ένας αγοραστής αγοράσει *γάλα* και *βούτυρο* πιθανότατα θα αγοράσει και *ψωμί.* Σε αυτή την κατεύθυνση έχει επιτευχθεί σημαντική πρόοδος για παράδειγμα με την επιλογή μόνον των συνολοστοιχείων που δεν είναι παράγωγα άλλων [101]. Οι εξαγόμενοι κανόνες συσχέτισης είναι εύκολο να γίνουν κατανοητοί και πολλά νέα μοτίβα μπορούν να εντοπιστούν. Όμως, ο κατά κανόνα μεγάλος, αριθμός των κανόνων συσχέτισης δυσχεραίνει κατά πολύ την κατανόηση και την επεξήγηση τους από ανθρώπους.

Σε μια διαφορετική κατεύθυνση, έχουν διερευνηθεί διάφορες τεχνικές συνεργατικής διήθησης για χρήση σε δεδομένα καλαθιού αγοράς [102][103]. Η συνεργατική διήθηση όμως μπορεί να εμφανίσει χαμηλότερες επιδόσεις για την πραγματοποίηση υπολογισμών σε μεγάλα σύνολα δεδομένων, καθώς βασίζεται στην μνήμη [104]. Οι μέθοδοι αυτές έχουν εξ’ ορισμού σημαντικούς περιορισμούς στην επεκτασιμότητά τους. Επιπρόσθετα, η φύση του εμπορίου ως τομέας εφαρμογής των συστημάτων προτάσεων θέτει ορισμένους περιορισμούς στη συνεργατική διήθηση, όπως η έλλειψη αξιολογήσεων από τους καταναλωτές και η απαιτούμενη ποιότητα των προτάσεων [105].

Η ανάπτυξη συστημάτων προτάσεων για να καλύψουν τις ανάγκες ενός καλαθιού αγοράς έχει συναντήσει στο παρελθόν αρκετές προκλήσεις, δυο από τις οποίες είναι οι παρακάτω.

Πρώτον, οι τεχνικές που έχουν προταθεί στη βιβλιογραφία προσφέρουν μόνο μια περιορισμένη εποπτική εικόνα των προτιμήσεων των καταναλωτών. Αν και μπορούμε να χρησιμοποιήσουμε τους κανόνες συσχέτισης για να προβλέψουμε τα προϊόντα που θα απαρτίσουν το υπόλοιπο ενός καλαθιού αγοράς, αυτό που απουσιάζει είναι μια γενικότερη εικόνα των προτιμήσεων του χρήστη και των σχέσεων μεταξύ των προτιμήσεων αυτών.

Δεύτερον, η ποιότητα των προτάσεων που παράγονται αλλά και η ταχύτητα με την οποία παράγονται μπορεί να επηρεαστούν αρνητικά από τον τύπο του συνόλου δεδομένων. Οι κανόνες συσχέτισης τείνουν να αγνοούν μεγάλα στοιχειοσύνολα ενώ οι τεχνικές συνεργατικής διήθησης που βασίζονται στην μνήμη υστερούν στη δυνατότητα επέκτασης, καθώς όσο μεγαλώνει το σύνολο δεδομένων τόσο αυξάνονται οι απαιτήσεις του συστήματος σε μνήμη και επεξεργαστική ισχύ [106]. Ακόμη, τα συστήματα προτάσεων που βασίζονται στο περιεχόμενο δεν μπορούν να χρησιμοποιηθούν με ευκολία στις περισσότερες περιπτώσεις λιανικών συναλλαγών καθώς η πληροφορία που περιγράφει τα προϊόντα είναι συνήθως ελλιπής ή μη διαθέσιμη.

Στην προσέγγιση που προτείνεται στα πλαίσια της διατριβής χρησιμοποιούμε την τεχνική των πιθανοτικών μοντέλων θεμάτων. Πρόκειται για στατιστικά μοντέλα που χρησιμοποιούνται για την ανάλυση των σχέσεων μεταξύ ενός συνόλου εγγράφων και των όρων που περιέχουν, μέσω της παραγωγής μιας ομάδας εννοιών που αφορούν τόσο τα έγγραφα όσο και τις λέξεις. Σε αυτό το κεφάλαιο εφαρμόζουμε την συγκεκριμένη τεχνική ώστε να παράγουμε ένα μοντέλο θεμάτων που να αντικατοπτρίζει τις προτιμήσεις των καταναλωτών αλλά και που να μπορεί να χρησιμοποιηθεί ως βάση για να παραχθούν προτάσεις προϊόντων.

Για την εξαγωγή των λανθανόντων θεμάτων χρησιμοποιούνται δυο συναφείς προσεγγίσεις. Πρώτον, γίνεται εξαγωγή θεμάτων από ομάδες προϊόντων που εμφανίζονται μαζί συχνά στο ιστορικό των αγοραστών (*λανθάνοντες χρήστες*). Δεύτερον, πραγματοποιείται εξαγωγή θεμάτων από ομάδες προϊόντων που εμφανίζονται συχνά μαζί σε μεμονωμένες επισκέψεις των αγοραστών (*λανθάνοντα καλάθια*). Στη συνέχεια τα μοντέλα θεμάτων που εξάχθηκαν χρησιμοποιούνται για την παροχή προτάσεων σε καταναλωτές.

Η μεθοδολογία που προτείνουμε διαθέτει δυο συγκριτικά πλεονεκτήματα. Πρώτον προσφέρει μια εποπτική εικόνα της συμπεριφοράς των καταναλωτών με βάση τα δεδομένα των καλαθιών αγοράς. Δεύτερον μπορεί να προβλέψει με ταχύτητα και σημαντική ακρίβεια τα προϊόντα που θα επιλέξει ένας καταναλωτής.

Η συγκεκριμένη μεθοδολογία υλοποιείται σε ένα σύστημα λογισμικού που παράγει προτάσεις. Παραλλαγές των τεχνικών δημιουργίας προτάσεων υλοποιήθηκαν και εφαρμόστηκαν στο σύνολο δεδομένων μιας ελληνικής υπεραγοράς όπου αξιολογήθηκαν θετικά ξεπερνώντας σε επίδοση τους κανόνες συσχέτισης.

Το κεφάλαιο αυτό δομείται ως εξής. Η επόμενη ενότητα παρέχει μια εικόνα των σχετικών εργασιών, ενώ ακολούθως αναλύεται η μεθοδολογία που ακολουθήσαμε. Στη συνέχεια περιγράφεται το προτεινόμενο σύστημα και ακολούθως παρατίθενται τα στοιχεία που αφορούν την πειραματική αξιολόγηση και τα αποτελέσματά της. Τέλος κλείνουμε το κεφάλαιο με συμπεράσματα και προτάσεις για μελλοντικές ερευνητικές εργασίες.

7.2 Σχετικές Εργασίες

Τα συστήματα προτάσεων στη γενική τους μορφή επιδιώκουν να προβλέψουν την αξιολόγηση ενός αντικειμένου που ο χρήστης δεν έχει δει ακόμη [14]. Στο εμπόριο που πραγματοποιείται μέσω του παγκόσμιου ιστού η υπερφόρτωση πληροφορίας έχει δημιουργήσει προβλήματα στους καταναλωτές καθώς καλούνται να επιλέξουν τα προϊόντα που θα αγοράσουν από μια τεράστια ποικιλία. Από την άλλη πλευρά οι πωλητές προσπαθούν να εντοπίσουν και να αποτυπώσουν αποτελεσματικά τις προτιμήσεις των καταναλωτών [141].

Τα συστήματα προτάσεων χρησιμοποιούνται ευρέως σε εφαρμογές ηλεκτρονικού εμπορίου [100]. Πολλοί από τους μεγαλύτερους ιστοτόπους ηλεκτρονικού εμπορίου χρησιμοποιούν ήδη συστήματα προτάσεων για να βοηθήσουν τους χρήστες τους να εντοπίσουν τα προϊόντα που επιθυμούν να αγοράσουν.

Για να βελτιωθούν τα συστήματα προτάσεων καταναλωτών έχει προταθεί η μοντελοποίηση ενός προφίλ χρήστη το οποίο εκτός από τα προσωπικά χαρακτηριστικά του καταναλωτή, λαμβάνει υπόψη και ένα ομαδικό προφίλ που απεικονίζει καλύτερα τα ενδιαφέροντα ομάδων χρηστών με παρόμοια χαρακτηριστικά [142]. Οι προτάσεις παράγονται μετά από συσχέτιση του προφίλ χρήστη με τα προϊόντα.

Στην ανάλυση των συναλλαγών των καταναλωτών, η αλλιώς ανάλυση καλαθιού αγοράς, η πρώιμη έρευνα έχει επικεντρωθεί στην εφαρμογή της *εξαγωγής κανόνων συσχέτισης* (association rule mining). Η εξαγωγή κανόνων συσχέτισης περιλαμβάνει την ανάλυση συνολοστοιχείων και την εξαγωγή κανόνων που συσχετίζουν προϊόντα μεταξύ τους. Στη βιβλιογραφία έχει επιδιωχθεί ο εντοπισμός των κανόνων που έχουν δυο στοιχεία [143]: Πρώτον, περιέχουν προϊόντα τα οποία είναι αρκετά διαδεδομένα σε όλες τις συναλλαγές που καταγράφονται στο σύνολο

δεδομένων. Δεύτερον, περιέχουν στο πρώτο σκέλος τους προϊόντα που είναι αρκετά διαδεδομένα στο σύνολο δεδομένων. Επίσης έχει προταθεί μια προσέγγιση εξαγωγής κανόνων που διευκολύνει την είσοδο και έξοδο δεδομένων ενώ ταυτόχρονα έχει περιορισμένες απαιτήσεις σε υπολογιστική ισχύ [144]. Επίσης, μια διαφορετική μελέτη στοχεύει στην επιλογή μόνον των συνολοστοιχείων που δεν είναι παράγωγα άλλων [101].

Οι εξαχθέντες κανόνες συσχέτισης είναι εύκολο να γίνουν κατανοητοί και πολλά νέα μοτίβα μπορούν να εντοπιστούν. Όμως, ο αριθμός και μόνο των κανόνων συσχέτισης δυσχεραίνει κατά πολύ την κατανόηση και την επεξήγηση τους από ανθρώπους [145].

Σε μια διαφορετική κατεύθυνση, έχουν διερευνηθεί τεχνικές συνεργατικής διήθησης για χρήση σε δεδομένα καλαθιών αγοράς. Στην βιβλιογραφία έχει ελεγχθεί η δυνατότητα των μεθόδων συνεργατικής διήθησης να διαχειριστούν σύνολα δεδομένων που προέρχονται από αγοραστικές συνήθειες καταναλωτών και έχει παρουσιαστεί μια επέκταση του γενικού αλγορίθμου k-πλησιέστερων γειτόνων [103]. Επίσης έχει προταθεί ένα σχήμα συνεργατικής διήθησης που χρησιμοποιεί δυαδική λογιστική παλινδρόμηση (binary logistic regression) και αντιμετωπίζει την επιλογή του καταναλωτή ως ένα πρόβλημα ταξινόμησης [102].

Παρά τις επιτυχημένες εφαρμογές, η συνεργατική διήθηση εμφανίζει χαμηλότερες επιδόσεις για την πραγματοποίηση υπολογισμών σε μεγάλα σύνολα δεδομένων, καθώς βασίζεται στην μνήμη [104]. Οι μέθοδοι αυτές έχουν εξ’ ορισμού σημαντικούς περιορισμούς στην επεκτασιμότητά τους πέρα από ένα σημείο. Επιπρόσθετα, η φύση του εμπορίου ως τομέας εφαρμογής των συστημάτων προτάσεων θέτει ορισμένους περιορισμούς στη συνεργατική διήθηση, όπως η έλλειψη αξιολογήσεων από τους καταναλωτές και η απαιτούμενη ποιότητα των προτάσεων [105].

Στη βιβλιογραφία έχουν προταθεί συστήματα που βασίζονται σε πράκτορες λογισμικού για την επίλυση του συγκεκριμένου προβλήματος. Συγκεκριμένα έχουν χρησιμοποιηθεί πράκτορες λογισμικού (software agents) για να υποστηριχθούν δυο δραστηριότητες που αφορούν τις ηλεκτρονικές αγορές [146]: (1) την παραγωγή προτάσεων για την αγορά αντικειμένων και (2) την αυτόματη διαπραγμάτευση τιμών σε σύστημα ηλεκτρονικών δημοπρασιών. Αντίστοιχα έχει προταθεί ένα σύστημα βασισμένο σε πράκτορες για την ενημέρωση των συμμετεχόντων για το ιστορικό δημοπρασιών και τιμών [147]. Με βάση τα συγκεκριμένα δεδομένα

βελτιώνεται η διαδικασία λήψης αποφάσεων από τους συμμετέχοντες σε δημοπρασίες. Επίσης, τεχνικές όπως η μείωση διαστάσεων, τα γενετικά μοντέλα και η ανάλυση δεσμών [37] έχουν προταθεί για την πραγματοποίηση προτάσεων προϊόντων και υπηρεσιών σε περιβάλλοντα ηλεκτρονικού εμπορίου. Τέλος, έχει ερευνηθεί ένας αριθμός πιθανοτικών μοντέλων που επικεντρώνονται στους χρήστες και αξιολογηθεί η χρήση τους σε έναν ιστότοπο για κατέβασμα περιεχομένου [148].

Παρά τις προσπάθειες των ερευνητών, το εμπόριο, ηλεκτρονικό ή όχι, ως τομέας εφαρμογής συνεχίζει να δημιουργεί μια σειρά προκλήσεων για τα συστήματα προτάσεων όπως η έλλειψη αξιολογήσεων, η επεκτασιμότητα και η ποιότητα των προτάσεων.

7.3 Προσέγγιση

7.3.1 Πλαίσιο

Το ερευνητικό πεδίο που ασχολείται με την καταγραφή και την μοντελοποίηση των καταναλωτικών συνηθειών ονομάζεται ανάλυση καλαθιού αγοράς (market basket analysis). Όπως αναφέρθηκε στην προηγούμενη ενότητα, μια από τις πλέον διαδομένες προσεγγίσεις για την ανάλυση των αγορών είναι οι κανόνες συσχέτισης (association rules).

7.3.1.1 Ανάλυση Καλαθιού Αγοράς

Η ανάλυση καλαθιού αγοράς αποτελεί έναν ερευνητικό πεδίο που έχει στόχο να κατανοήσει την σύνθεση των καλαθιών αγορών των διαφόρων καταναλωτών και να προσδιορίσει τη λογική με την οποία οι καταναλωτές τοποθετούν προϊόντα σε ένα καλάθι αγορών και τα αγοράζουν. Ως καλάθι αγορών ορίζεται ένα σύνολο προϊόντων – «συνολοστοιχείο» (itemset) – που ένας καταναλωτής τοποθετεί στο ίδιο καλάθι και αγοράζει μαζί κατά τη διάρκεια μίας επίσκεψης του σε ένα κατάστημα [145].

Η ανάλυση καλαθιού αγοράς αποτελεί μια μεθοδολογία μοντελοποίησης δεδομένων που χρησιμοποιείται για την ανακάλυψη σχέσεων ανάμεσα σε αγοραζόμενα προϊόντα ή ομάδες προϊόντων. Η μεθοδολογία αυτή χρησιμοποιείται για να παρατηρηθούν οι αγοραστικές συνήθειες των καταναλωτών μέσω της εξαγωγής σχέσεων προτιμήσεων και προσδιορισμού προϊόντων που συχνά

αγοράζονται μαζί. Σε κάποιες περιπτώσεις οι προτιμήσεις αυτές είναι προφανείς λόγω της φύσης των αντικειμένων ενώ σε άλλες οι συσχετισμοί είναι δύσκολο να γίνουν αντιληπτοί.

Με βάση την ανάλυση προτιμήσεων, προτείνονται στους καταναλωτές προϊόντα ή λίστες προϊόντων και υπηρεσιών τα οποία μπορεί να τους ενδιαφέρουν, με ποικίλες εφαρμογές [142] σε τομείς προώθησης προϊόντων και υπηρεσιών αλλά και επιχειρησιακής έρευνας [140].

Η ανάλυση του καλαθιού αγοράς είναι ένα ισχυρό εργαλείο που μπορεί να χρησιμοποιηθεί για την λήψη αποφάσεων στο marketing των προϊόντων αλλά και για να ενισχύσει την εφαρμογή των cross-selling στρατηγικών.

7.3.1.2 Κανόνες Συσχέτισης

Η μεθοδολογία των κανόνων συσχέτισης, όπως έχει προταθεί στη βιβλιογραφία [143], ορίζει ότι με δεδομένα δύο μη επικαλυπτόμενα συνόλων προϊόντων, T_1 και T_2, μία σχέση της μορφής $T_1 \rightarrow T_2$ υποδεικνύει πως αν ένας καταναλωτής αγοράζει τα προϊόντα που περιέχονται στο T_1 αγοράζει και τα προϊόντα του T_2. Οι μετρικές που χρησιμοποιούνται συνήθως για το χαρακτηρισμό και στη συνεχεία την επιλογή της συσχέτισης είναι η υποστήριξη (support) και η εμπιστοσύνη (confidence) [149].

Η υποστήριξη αποτελεί μέτρο του μεγέθους της συχνότητας που μία συναλλαγή περιέχει και τα δύο σύνολα T_1, T_2 ενώ η εμπιστοσύνη μετράει την ακρίβεια της σχέσης διασύνδεσης ως λόγου του αριθμού των συναλλαγών που περιέχουν τα T_1 και T_2 προς το πλήθος των εγγραφών συναλλαγών που περιέχουν μόνο το T_1.

Οι κανόνες συσχέτισης λόγω της εύκολης εφαρμογής αλλά και της εύκολης κατανόησής τους έχουν χρησιμοποιηθεί ευρέως σε διάφορες πρακτικές εφαρμογές στο εμπόριο.

7.3.2 Η Μεθοδολογία του FillBasket

Παρουσιάζουμε την προσέγγιση FillBasket που προβλέπει την χρήση της λανθάνουσας κατανομής Dirichlet σε δεδομένα συναλλαγών λιανικής. Στη συνέχεια περιγράφουμε την μεθοδολογία της εξαγωγής των μοντέλων θεμάτων καθώς και

της χρήσης τους για την συγκέντρωση και ομαδοποίηση των προτιμήσεων των καταναλωτών στο χρόνο.

Ένα καλάθι αγοράς αποτελείται από αντικείμενα που αγοράζονται κατά τη διάρκεια μίας επίσκεψης ενός πελάτη σε κάποιο κατάστημα. Αγνοώντας την ποσότητα που αγοράζει και την τιμή που πληρώνει ο καταναλωτής, κάθε συναλλαγή αντιστοιχεί σε μία μαζική αγορά προϊόντων που συνέβη σε συγκεκριμένη ώρα και μέρος.

Ένα σύνολο δεδομένων που περιέχει συναλλαγές μπορεί να ενταχθεί σε έναν πίνακα. Ο πίνακας είναι ένα σύνολο όλων των συναλλαγών που έχουν πραγματοποιηθεί (7.1) και κάθε συναλλαγή μπορεί να μοντελοποιηθεί σαν ένα διάνυσμα ***T*** που περιέχει μόνο δυαδικές τιμές.

$$T = \{T_1, T_2, T_3, ..., T_n\} \qquad (7.1)$$

Το ***T*** είναι ένα διάνυσμα με μήκος ***n***, όπου ***n*** είναι τα διαθέσιμα προϊόντα στην κατάστημα. Τα διαθέσιμα προϊόντα στο κατάστημα περιέχονται στο σύνολο ***I*** όπως εμφανίζεται στην εξίσωση (7.2). Για το διάνυσμα ***T*** ορίζουμε ότι για το προϊόν ***k*** ισχύει *t[k] = 1* αν το αντικείμενο I_k περιέχεται στη συναλλαγή και *t[k] = 0* διαφορετικά.

$$I = \{I_1, I_2, I_3, ..., I_n\} \qquad (7.2)$$

Βασισμένο σε αυτά τα χαρακτηριστικά (συναλλαγή, αντικείμενο), το καλάθι αγοράς μπορεί να οριστεί ως ***N*** αντικείμενα που αγοράζονται μαζί από έναν καταναλωτή. Ο πίνακας των συναλλαγών περιέχει το άθροισμα πολλών τέτοιων συναλλαγών από διάφορους καταναλωτές.

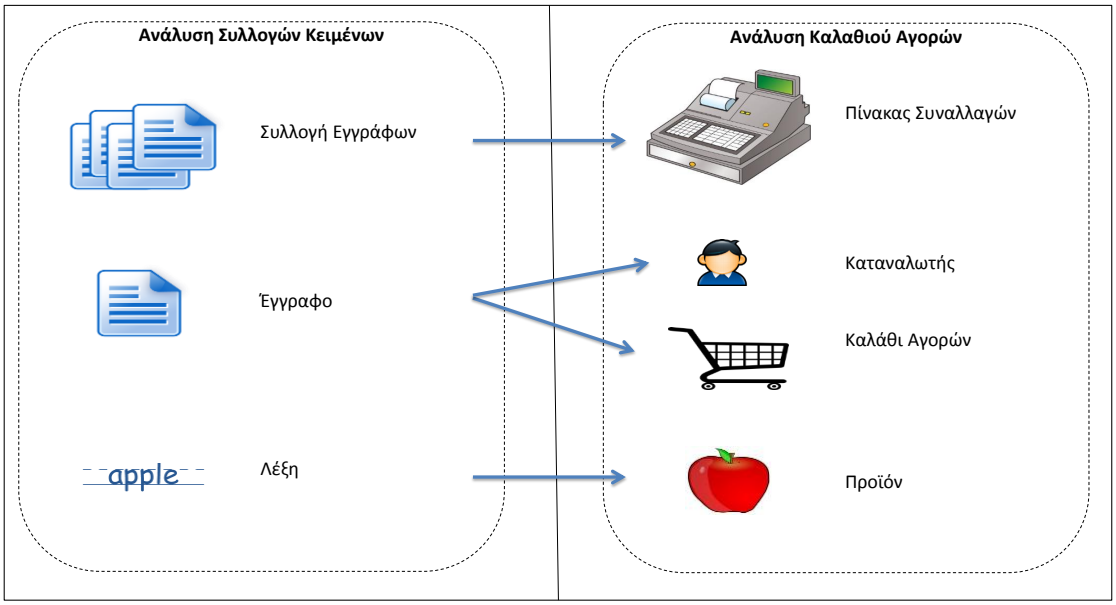
Το πρόβλημα που προσπαθούν να επιλύσουν τα συστήματα προτάσεων και πιο συγκεκριμένα οι τεχνικές ανάλυσης καλαθιού αγορών είναι υποστήριξη των καταναλωτών στην επιλογή των προϊόντων που τους ενδιαφέρουν. Το πρόβλημα που αντιμετωπίζουμε ορίζεται ως εξής: με δεδομένα τα *N-m* αντικείμενα που έχουν τοποθετήσει στο καλάθι τους οι καταναλωτές επιχειρούμε να προτείνουμε την αγορά ***m*** αντικειμένων που λείπουν.

7.3.2.1 Εξαγωγή μοντέλου

Στις κλασικές εφαρμογές των πιθανοτικών μοντέλων θεμάτων συνήθως θεωρούμε μια συλλογή εγγράφων που περιέχουν κείμενο. Στην προσέγγιση που

προτείνουμε δεν περιλαμβάνονται κείμενα αλλά οι προτιμήσεις των καταναλωτών στο παρελθόν. Έτσι για να εφαρμόσουμε την συγκεκριμένη μεθοδολογία περιγράφουμε ένα διαφορετικό μοντέλο εφαρμογής.

Αντιστοιχίζουμε την δραστηριότητα των καταναλωτών με τα στοιχεία που περιέχει ένα σύνολο εγγράφων [76]. Οι όροι στην περίπτωσή μας είναι τα προϊόντα που είναι διαθέσιμα προς πώληση. Θεωρούμε ότι κάθε κείμενο σχηματίζεται είτε από τα προϊόντα που αγοράζονται μαζί σε μία συναλλαγή είτε από τα προϊόντα που αγοράζονται από έναν πελάτη. Η αντιστοίχιση φαίνεται στην Εικόνα 7.1.



Εικόνα 7.1 Αντιστοιχία Ανάλυσης Κειμένων - Ανάλυσης Καλαθιού Αγορών

Για την εξαγωγή των πιθανοτικών μοντέλων θεμάτων χρησιμοποιούμε την λανθάνουσα κατανομή Dirichlet. Πιο συγκεκριμένα, χρησιμοποιούμε μια παραλλαγή της μεθόδου η οποία τοποθετεί μια εκ των προτέρων συνάρτηση Dirichlet τόσο στις κατανομές θεμάτων σε έγγραφα όσο και σε κατανομές λέξεων σε θέματα [54]. Στη συνέχεια εφαρμόζεται η δειγματοληψία Gibbs για την εξαγωγή μοντέλων θεμάτων.

Στην διαδικασία της εξαγωγής θεμάτων ακολουθούμε δυο πιθανές μεθόδους, την ανάλυση των λανθανόντων καλαθιών και των λανθανόντων χρηστών.

Λανθάνοντα Καλάθια

Στην πρώτη περίπτωση, αντιμετωπίζουμε κάθε συναλλαγή ενός πελάτη ως ένα κείμενο. Τα προϊόντα που αγοράζονται μαζί από έναν πελάτη κατά την διάρκεια

μιας επίσκεψης αντιμετωπίζονται ως ένα ξεχωριστό κείμενο που δημιουργήθηκε χρησιμοποιώντας λέξεις από το λεξιλόγιο των προϊόντων.

Θεωρούμε ότι κάθε επίσκεψη στο κατάστημα ως μια αυτοτελή οντότητα. Το σύνολο των προϊόντων σε ένα καλάθι αγορών θεωρείται ότι είναι ένα αποτέλεσμα ενός γενετικού πιθανοτικού μοντέλου θεμάτων το οποίο προσπαθούμε να υπολογίσουμε. Στην υπόθεση αυτή θεωρούμε ότι κάθε καλάθι που φεύγει από το κατάστημα περιέχει προϊόντα που παράγονται πιθανοτικά από μια συλλογή θεμάτων. Αυτά τα θέματα μπορούν να θεωρηθούν ότι αντικατοπτρίζουν λανθάνοντα καλάθια.

Λανθάνοντες Χρήστες

Στη δεύτερη περίπτωση αντιμετωπίζουμε κάθε σειρά από συναλλαγές που έχουν γίνει στην πορεία του χρόνου από τον ίδιο πελάτη ως ένα κείμενο. Τα προϊόντα που έχουν αγοραστεί από έναν πελάτη σε ολόκληρο το ιστορικό του θεωρούνται ως ένα κείμενο που δημιουργήθηκε χρησιμοποιώντας λέξεις από το λεξιλόγιο προϊόντων.

Στη συγκεκριμένη περίπτωση αντιμετωπίζουμε το άθροισμα των επισκέψεων ενός καταναλωτή ως μια αυτοτελή οντότητα. Αυτό το σύνολο των προϊόντων θεωρείται ως ένα αποτέλεσμα του γενετικού μοντέλου θεμάτων όπου προσπαθούμε να περιγράψουμε χρησιμοποιώντας θεματικά μοντέλα. Στην υπόθεση αυτή θεωρούμε ότι κάθε καταναλωτής αγοράζει προϊόντα που παράγονται πιθανοτικά από μια συλλογή θεμάτων. Τα θέματα αυτά συντίθενται για να δημιουργήσουν τις προτιμήσεις των καταναλωτών σε μία χρονική περίοδο και θεωρούμε ότι αντικατοπτρίζουν λανθάνοντα προφίλ καταναλωτών ή λανθάνοντες χρήστες.

7.3.3 Δημιουργία Προτάσεων

Στην ενότητα αυτή περιγράφουμε πως αξιοποιούμε τα εξαγόμενα πιθανοτικά μοντέλα θεμάτων για να παρέχουμε προτάσεις προϊόντων στους χρήστες.

Θεωρούμε γνωστή μια ομάδα *N-m* προϊόντων από ένα καλάθι ***N*** προϊόντων και επιχειρούμε να προβλέψουμε τα υπόλοιπα ***m*** προϊόντα που θα αγοράσει ο καταναλωτής. Για το σκοπό αυτό χρησιμοποιούμε δυο βασικές μεθόδους.

Πρώτον, χρησιμοποιούμε την στατιστική επαγωγή για να βρούμε τα λανθάνοντα θέματα που ταιριάζουν περισσότερο με το τρέχον καλάθι. Αντιμετωπίζουμε τα *N-m* προϊόντα που γνωρίζουμε ως ένα νέο έγγραφο και υπολογίζουμε την αναλογία των θεμάτων που θα μπορούσαν να το έχουν παράγει. Με βάση αυτή την αναλογία εντοπίζουμε πιθανά προϊόντα που δεν έχουν τοποθετηθεί στο καλάθι του καταναλωτή.

Δεύτερον, υπολογίζουμε την ομοιότητα κάθε προϊόντος με τα υπόλοιπα προϊόντα που είναι διαθέσιμα για αγορά. Ο αριθμός αυτός συσχετίζει τα προϊόντα μεταξύ τους και αποθηκεύεται σε μορφή ευρετηρίου. Με βάση την υπολογιζόμενη ομοιότητα υπολογίζουμε τα πιο συναφή αντικείμενα σε σχέση με τα προϊόντα που έχουν ήδη τοποθετηθεί στο καλάθι.

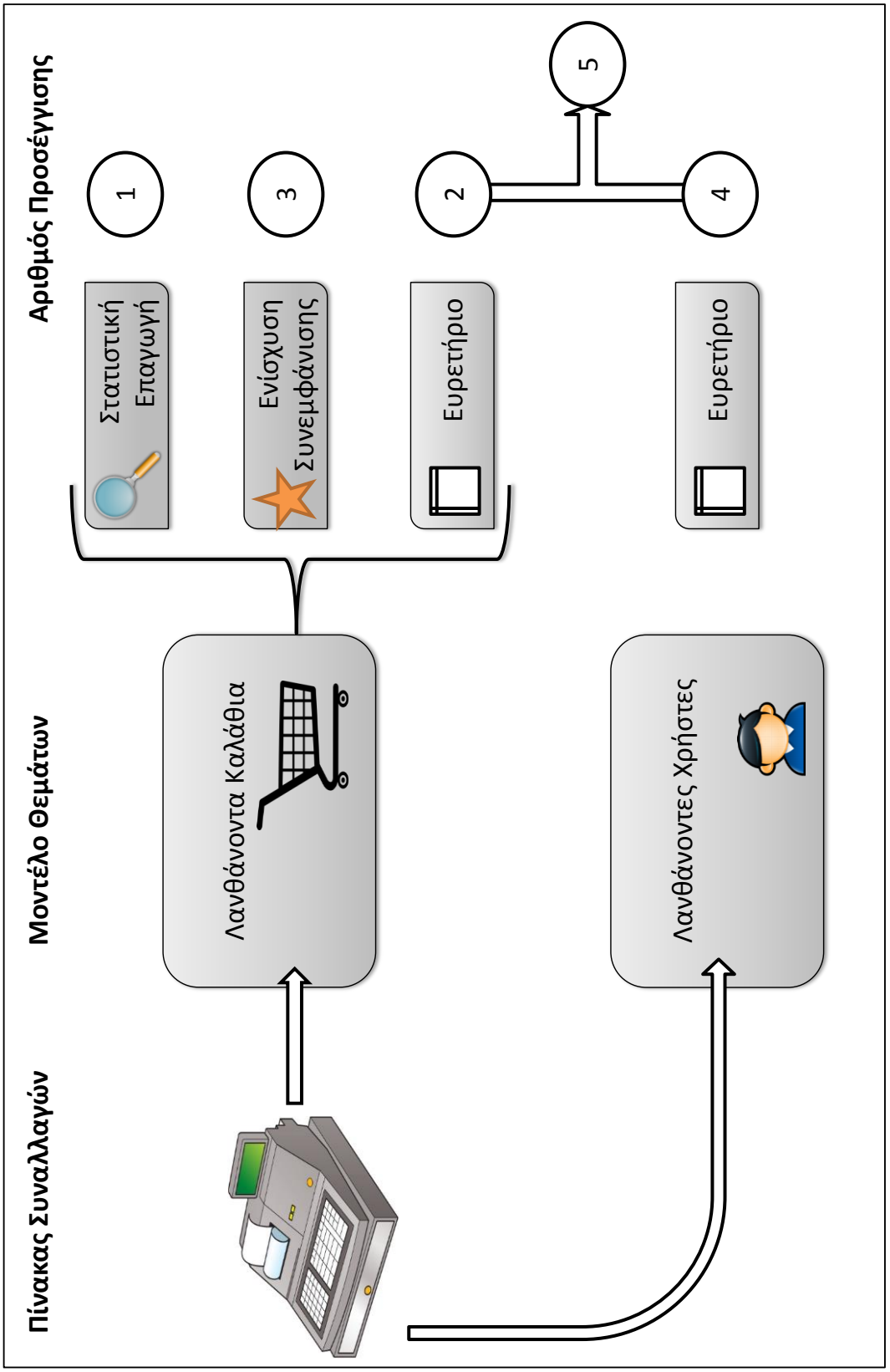
Οι μέθοδοι αυτές μπορούν να τροποποιηθούν με δυο τρόπους. Πρώτον, μπορούμε να χρησιμοποιήσουμε ως βάση τόσο τα λανθάνοντα καλάθια όσο και τους λανθάνοντες χρήστες. Δεύτερον, όταν χρησιμοποιούμε ένα ευρετήριο με συσχετίσεις μπορούμε να επιβραβεύσουμε την σύνδεση ενός προϊόντος με πολλά προϊόντα στο καλάθι του καταναλωτή με έναν αντίστοιχο όρο. Στη συνέχεια ονομάζουμε αυτή την αλλαγή «ενίσχυση συνεμφάνισης»

Η προσέγγισή μας περιλαμβάνει πέντε διαφορετικές τεχνικές για την δημιουργία προτάσεων που αναλύονται παρακάτω. Οι συγκεκριμένες μέθοδοι παρουσιάζονται εποπτικά στην Εικόνα 7.2.

1) <u>Λανθάνοντα καλάθια με στατιστική επαγωγή</u>

Σε αυτή την περίπτωση τα λανθάνοντα καλάθια χρησιμοποιούνται για να προβλέψουν την συμπεριφορά του χρήστη με δεδομένα τα αντικείμενα που έχει το καλάθι του αυτή τη στιγμή.

Η δειγματοληψία Gibbs χρησιμοποιείται για να εξάγει την πιθανοτική κατανομή των γνωστών αντικειμένων που περιέχονται στο καλάθι. Προτείνονται τα αντικείμενα με την μεγαλύτερη συσχέτιση με αυτή την κατανομή που δεν βρίσκονται ακόμη στο καλάθι του καταναλωτή. Η μετρική συσχέτισης με την κατανομή χρησιμοποιείται για να ταξινομηθεί η λίστα των προτάσεων.



Εικόνα 7.2 Προσεγγίσεις Δημιουργίας Προτάσεων FillBasket

2) Λανθάνοντα καλάθια με χρήση ευρετηρίου

Όπως και στην προηγούμενη περίπτωση, τα λανθάνοντα καλάθια χρησιμοποιούνται με σκοπό να προβλέψουν την συμπεριφορά του χρήστη. Όμως αντί να χρησιμοποιήσουν στατιστική επαγωγή, ένα ευρετήριο δημιουργείται από την δημιουργία θεματικών μοντέλων.

Εδώ θεωρούμε ότι το $S_{LBi,j}$ είναι η υπολογισμένη ομοιότητα μεταξύ δύο αντικειμένων ***i*** και ***j*** κάτι που έχει προκύψει από το θεματικό μοντέλο λανθανόντων καλαθιών. Αυτή η ομοιότητα βρίσκεται αποθηκευμένη σε ένα ευρετήριο στο σύστημα προτάσεων. Χρησιμοποιούμε την εξίσωση (7.3) για να υπολογίσουμε την ομοιότητα των γνωστών αντικειμένων του καλαθιού του χρήστη με όλα τα ***n*** διαφορετικά πιθανά αντικείμενα που είναι διαθέσιμα στο κατάστημα. Στην συνέχεια οι επιλογές που παρουσιάζουν την μεγαλύτερη ομοιότητα προστίθενται σε μια λίστα προτάσεων προς τον καταναλωτή.

$$\{w_1, w_2, w_3, ..., w_n\} = \{\sum_{j \in KI} S_{LB1,j}, \sum_{j \in KI} S_{LB2,j}, \sum_{j \in KI} S_{LB3,j}, ..., \sum_{j \in KI} S_{LBn,j}\} \qquad (7.3)$$

3) Λανθάνοντα καλάθια με ενίσχυση συνεμφάνισης

Αυτή η περίπτωση είναι παρόμοια με την μέθοδο του ευρετηρίου λανθανόντων καλαθιών με τη διαφορά ότι σε αυτή την περίπτωση τα αντικείμενα που συσχετίζονται με περισσότερα του ενός των γνωστών αντικειμένων από το καλάθι του χρήστη παίρνουν μία μικρή ενίσχυση.

Μετράμε τον αριθμό των παρόμοιων αντικειμένων μέσα στο καλάθι του χρήστη και χρησιμοποιούμε αυτόν των αριθμό ως μία δύναμη για τον παράγοντα ενίσχυσης. Στην εξίσωση (7.4), το ***M*** είναι ο αριθμός των αντικειμένων που έχουν βρεθεί παρόμοια με αντικείμενο ***l***. Με βάση τους υπολογισμούς μας ταξινομούμε τα προϊόντα στην λίστα προτάσεων προς τον καταναλωτή.

$$w_l = b^{M-1} \sum_{j \in KI} S_{LBl,j} \qquad (7.4)$$

4) Λανθάνοντες χρήστες με χρήση ευρετηρίου

Σε αυτή την περίπτωση χρησιμοποιούμε μοντέλα θεμάτων που έχουν εξαχθεί από το ιστορικό των προτιμήσεων των καταναλωτών και όχι από μεμονωμένες επισκέψεις.

Αντίστοιχα με παραπάνω σχηματίζουμε ένα ευρετήριο όπου το $S_{LUi,j}$ είναι η υπολογισμένη ομοιότητα μεταξύ των αντικειμένων ***i*** και ***j*** προερχόμενη από το θεματικό μοντέλο λανθανόντων χρηστών. Χρησιμοποιούμε τις ομοιότητες για να προβλέψουμε προτεινόμενα αντικείμενα, δεδομένων των γνωστών αντικειμένων στο καλάθι του χρήστη (7.5). Τα αντικείμενα κατατάσσονται με βάση τη συγκεκριμένη ομοιότητα.

$$\{w_1, w_2, w_3, ..., w_n\} = \{\sum_{j \in KI} S_{LU1,j}, \sum_{j \in KI} S_{LU2,j}, \sum_{j \in KI} S_{LU3,j}, ..., \sum_{j \in KI} S_{LUn,j}\} \quad (7.5)$$

5) Λανθάνοντα καλάθια συνδυασμένα με λανθάνοντες χρήστες

Εδώ συνδυάζουμε τα αποτελέσματα των δυο διαφορετικών προσεγγίσεων, της αντιστοίχισης των εγγράφων με τους καταναλωτές και της αντιστοίχισης με μεμονωμένες συναλλαγές.

Σε αυτή την περίπτωση η πρόταση του μοντέλου θεμάτων λανθανόντων καλαθιών συμπληρώνεται από ένα μοντέλο λανθανόντων χρηστών, χρησιμοποιώντας μία παράμετρο ανάμειξης μ, 0<μ<1. Για να υπολογίσουμε την ομοιότητα του αντικειμένου l με τα γνωστά αντικείμενα στο καλάθι του χρηστή χρησιμοποιούμε την (7.6). Τα αντικείμενα στην λίστα προτάσεων προς τον καταναλωτή κατατάσσονται με βάση τη συγκεκριμένη ομοιότητα.

$$w_l = (1-\mu)\sum_{j \in KI} S_{LBl,j} + \mu \sum_{j \in KI} S_{LUl,j} \quad (7.6)$$

7.4 Πειραματική Αξιολόγηση

Ακολουθεί η περιγραφή της πειραματικής αξιολόγησης της προτεινόμενης μεθόδου. Περιγράφεται το σύνολο δεδομένων που χρησιμοποιήθηκε και αναλύονται τα αποτελέσματα της αξιολόγησης της μεθόδου σε αυτό.

7.4.1 Σύνολο Δεδομένων

Τα δεδομένα συναλλαγών συλλέχθηκαν από μια μεγάλη ελληνική υπεραγορά στην διάρκεια ενός έτους. Ο αριθμός των συναλλαγών ήταν 1.057.076 ενώ εκείνος των διαφορετικών καταναλωτών που είχαν εγγραφεί στο σύστημα ήταν 17.672. Τα διαφορετικά προϊόντα που ήταν διαθέσιμα για αγορά ήταν 102.142.

Κατά τη διάρκεια της πειραματικής αξιολόγησης έχουμε χρησιμοποιήσει τις κατηγορίες των προϊόντων ώστε να εξάγουμε χρήσιμα συμπεράσματα τα οποία να αφορούν τύπους προϊόντων χωρίς να διαχωρίζουν συγκεκριμένες μάρκες. Για παράδειγμα αντί να αντιμετωπίζονται οι διαφορετικές μάρκες και τα μεγέθη γάλακτος ως διαφορετικά προϊόντα, θεωρούμε το γάλα χαμηλών λιπαρών ως μια κατηγορία και συνεπώς ως ένα τύπο αντικειμένου. Αυτός ο βαθμός ανάλυσης οδήγησε στην εξαγωγή 473 διαφορετικών προϊόντων.

Αριθμός Λανθάνοντος Καλαθιού	**Προϊόντα**
23	White paper napkins, Body shampoo, Snack, Soda
25	Toilet paper, Kitchen paper, White paper napkins, Oily hair shampoo
29	Toilet paper, Kitchen paper, White paper napkins, Bleach
34	Spoons/forks/knives, kitchen utensils, daily use, Plastic utensils
40	Cleaning sponges, Cleaning towels, Scourers
42	Olives, Pickles, Precooked food can, Canned fish

Πίνακας 7.1 Παραδείγματα Λανθανόντων Καλαθιών

Εφαρμόσαμε της τεχνικές θεματικής ανάλυσης στο συγκεκριμένο σύνολο δεδομένων και πιο συγκεκριμένα στους καταναλωτές και στα καλάθια αγορών τους. Τα αποτελέσματα της ανάλυσης δεν είναι μόνο η βάση για την παραγωγή προτάσεων αλλά προσφέρουν και μια εποπτική εικόνα στις προτιμήσεις των καταναλωτών. Η εξαγωγή θεμάτων με χρήση της λανθάνουσας κατανομής Dirichlet και το πακέτο λογισμικού Mallet [150] διήρκησε 4 ώρες και 30 λεπτά για τον αριθμό

των 2.000 επαναλήψεων σε έναν υπολογιστή με επεξεργαστή Intel Core 2 Duo T9300 CPU με 2.0 GB μνήμης RAM.

Στον αντίστοιχο πίνακα παρουσιάζονται ενδεικτικά παραδείγματα λανθανόντων καλαθιών χρηστών τα οποία αντιστοιχούν σε ομάδες προϊόντων που αγοράζονται συχνά μαζί σε μια επίσκεψη στην υπεραγορά (Πίνακας 7.1). Ακολούθως παρουσιάζονται ενδεικτικά παραδείγματα λανθανόντων χρηστών, ομάδες προϊόντων που αγοράζονται συχνά από τον ίδιο χρήστη κατά τη διάρκεια του έτους (Πίνακας 7.2).

Αριθμός Λανθάνοντος Χρήστη	**Προϊόντα**
3	Toothpaste, Cleaning Sponges, Bleach, Water descaler
6	Gouda cheese, Kaser cheese, edam cheese, feta cheese
7	Brooms, Broom sticks, mops, mop towels
9	School accessories, Ravioli pasta, Fresh milk
24	Hygiene Cotton, baby napkins, Colored napkins
39	Cake bases, truffle, Turkish delights, semolina

Πίνακας 7.2 Παραδείγματα Λανθανόντων Χρηστών

7.4.2 Αποτελέσματα

Σε αυτή την ενότητα περιγράφουμε τα αποτελέσματα της αξιολόγησης του συστήματος προτάσεων που βασίζεται στα πιθανοτικά μοντέλα θεμάτων.

Χρησιμοποιούμε την μέθοδο της εξαγωγής κανόνων συσχέτισης FP-growth με διαφορετικές ρυθμίσεις ως βάση αναφοράς [151] όπως αυτή είναι διαθέσιμη από την πλατφόρμα rapidminer [152]. Από αυτή τη διαδικασία προέκυψαν 12.522 κανόνες, με χρήση 1.000 συχνών συνολοστοιχείων και ελάχιστη βεβαιότητα 0,1. Οι κανόνες αυτοί κατατάσσονται με βάση την βεβαιότητα τους και εφαρμόζονται στα προϊόντα που είναι γνωστά από κάθε καλάθι αγοράς.

Στη συνέχεια επιχειρούμε να εντοπίσουμε εμπειρικά τον αριθμό των επαναλήψεων της μεθόδου δειγματοληψίας Gibbs που απαιτούνται για να

επιτευχθεί η σύγκλιση σε μοντέλο θεμάτων που να είναι αντιπροσωπευτικό της πραγματικότητας (Πίνακας 7.3). Αξιολογούμε την καλύτερη απόδοση για τον αλγόριθμο προτάσεων με χρήση ευρετηρίου λέξεων για την δημιουργία προτάσεων. Παρατηρούμε ότι τα βέλτιστα αποτελέσματα εμφανίζονται στις 4.000 επαναλήψεις, ενώ συγκρίσιμα αποτελέσματα μπορούμε να πάρουμε και με μόλις 2.000 επαναλήψεις.

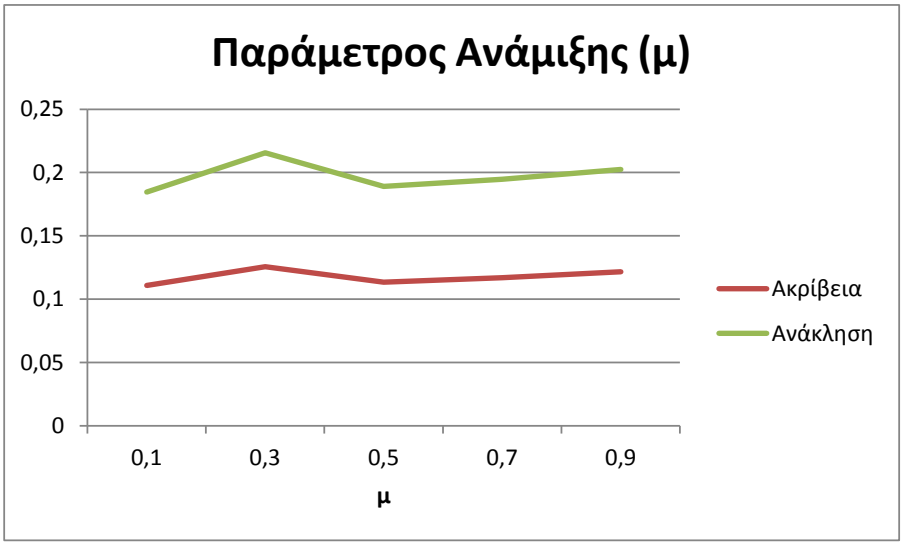
Επαναλήψεις Δειγματοληψίας GIbbs	**500**	**1000**	**2000**	**3000**	**4000**
Ακρίβεια	0,0976	0,1151	0,1322	0,1358	0,1385
Ανάκληση	0,3228	0,3835	0,4406	0,4526	0,4618
F- measure	0,1499	0,1771	0,2034	0,2089	0,2131

Πίνακας 7.3 Αξιολόγηση με Διαφορετικό Αριθμό Επαναλήψεων

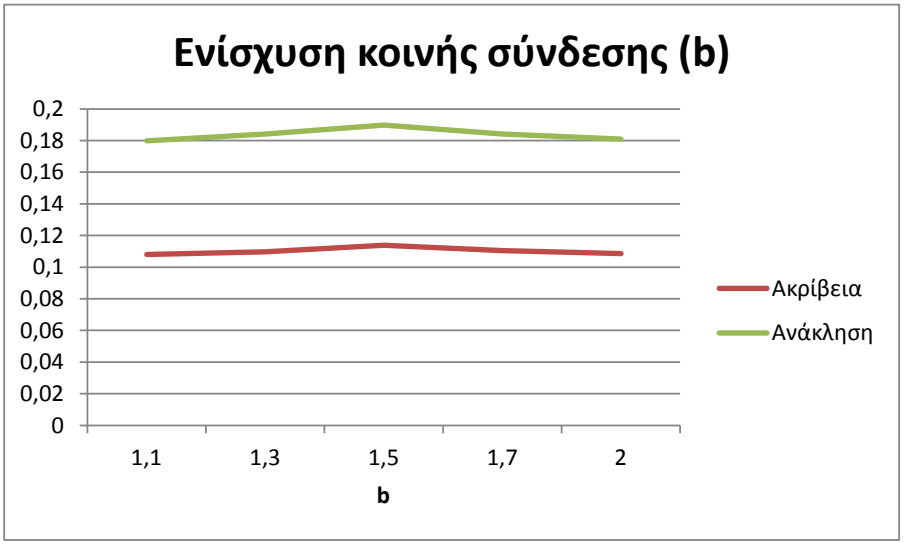
Ακολούθως αξιολογούμε το αποτέλεσμα της πραγματοποίησης προτάσεων για διαφορετικές τιμές παραμέτρων. Οι παράμετροι αυτές είναι η παράμετρος ενίσχυσης συνεμφάνισης στα λανθάνοντα καλάθια και η παράμετρος ανάμειξης για συνδυασμό λανθανόντων καλαθιών με λανθάνοντες χρήστες. Τα αποτελέσματα φαίνονται στους αντίστοιχους πίνακες (Πίνακας 7.4 και Πίνακας 7.5) και στις αντίστοιχες εικόνες (Εικόνα 7.3 και Εικόνα 7.4).

Ενίσχυση συνεμφάνισης (b)	**1,1**	**1,3**	**1,5**	**1,7**	**2**
Ακρίβεια	0,1080	0,1097	0,1139	0,1105	0,1086
Ανάκληση	0,1799	0,1842	0,1898	0,1842	0,1810

Πίνακας 7.4 Ενίσχυση Συνεμφάνισης στα Λανθάνοντα Καλάθια



Εικόνα 7.3 Υπολογισμός Παραμέτρου Ανάμειξης



Εικόνα 7.4 Υπολογισμός Παραμέτρου Ενίσχυσης Συνεμφάνισης

Παράμετρος Ανάμειξης (μ)	0,1	0,3	0,5	0,7	0,9
Ακρίβεια	0,1108	0,1255	0,1135	0,1169	0,1215
Ανάκληση	0,1847	0,2155	0,1891	0,1948	0,2025

Πίνακας 7.5 Παράμετρος Ανάμειξης για Λανθάνοντα Καλάθια και Χρήστες

Οι βέλτιστες τιμές των παραμέτρων, όπως προέκυψαν από την διαδικασία που περιγράφηκε, είναι ***b****=1,5* και ***μ****=0,3*. Χρησιμοποιώντας τις συγκεκριμένες τιμές πραγματοποιήσαμε την αξιολόγηση των διαφορετικών μεθόδων. Στα συγκεκριμένα πειράματα, με δεδομένο ένα καλάθι αγορών αφήνουμε έναν αριθμό από προϊόντα εκτός του καλαθιού και ζητούμε από το σύστημα να παρέχει έναν αριθμό προτάσεων.

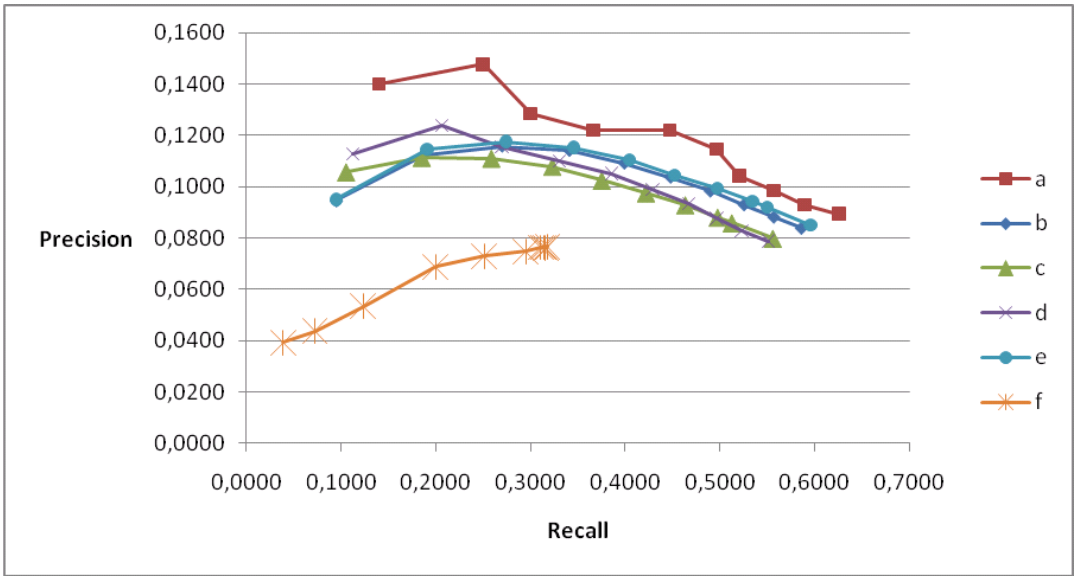
	Προτεινόμενα προϊόντα	**5**	**9**	**13**	**19**
a. Λανθάνοντα καλάθια με στατιστική επαγωγή	Ακρίβεια	0,1476	0,1221	0,1146	0,0930
	Ανάκληση	0,2499	0,3662	0,4967	0,5893
b. Λανθάνοντα καλάθια με χρήση ευρετηρίου	Ακρίβεια	0,1124	0,1140	0,1037	0,0883
	Ανάκληση	0,1875	0,3408	0,4475	0,5563
c. Λανθάνοντα καλάθια με ενίσχυση κοινής σύνδεσης	Ακρίβεια	0,1114	0,1079	0,0976	0,0858
	Ανάκληση	0,1856	0,3235	0,4225	0,5130
d. Λανθάνοντες χρήστες με χρήση ευρετηρίου	Ακρίβεια	0,1239	0,1101	0,0990	0,0826
	Ανάκληση	0,2065	0,3303	0,4289	0,5232
e. Λανθάνοντα καλάθια συνδυασμένα με λανθάνοντες χρήστες	Ακρίβεια	0,1145	0,1152	0,1044	0,0919
	Ανάκληση	0,1908	0,3457	0,4524	0,5502
f. Κανόνες Συσχέτισης FP-Growth	Ακρίβεια	0,0436	0,0687	0,0747	0,0763
	Ανάκληση	0,0723	0,2000	0,2952	0,3116

Πίνακας 7.6 Αποτελέσματα ανά Μέθοδο Προτάσεων

Το μέγεθος της λίστας προτάσεων κυμάνθηκε από 3 εως 21 αντικείμενα με βήμα 3. Όλες οι τεχνικές εκμάθησης που αναφέρθηκαν προηγουμένως

εφαρμόστηκαν στο 80% του συνόλου δεδομένων, ενώ το υπόλοιπο 20% χρησιμοποιήθηκε για αξιολόγηση (διασταυρωμένη επικύρωση). Ενδεικτικά αποτελέσματα για k=5, 9, 13 και 19 παρουσιάζονται στον αντίστοιχο πίνακα (Πίνακας 7.6). Οι καμπύλες ακρίβειας - ανάκλησης των διαφορετικών τεχνικών μπορούν να δώσουν μια εικόνα των αποτελεσμάτων των πειραμάτων και βρίσκονται στην Εικόνα 7.5

Τα αποτελέσματα της αξιολόγησης συνοψίζονται στην Εικόνα 7.5 όπου παρουσιάζονται οι καμπύλες ακρίβειας - ανάκλησης για τις διαφορετικές τεχνικές προτάσεων που εφαρμόστηκαν. Για τις πρώτες ***k*** προτάσεις που παράγονται από κάθε τεχνική αξιολογούμε και συγκρίνουμε με τα πραγματικά αντικείμενα τα οποία αγοράστηκαν και υπολογίζουμε την ακρίβεια και την ανάκληση. Για διαφορετικές τιμές του ***k***, χαράζουμε την καμπύλη ακρίβειας - ανάκλησης για κάθε μέθοδο που αποκαλύπτει πόσο ακριβείς αλλά και πόσο πλήρεις ήταν οι προβλέψεις της.



Εικόνα 7.5 Καμπύλες Ακρίβειας- Ανάκλησης για Εναλλακτικές Προσεγγίσεις

Στην Εικόνα 7.5 παρουσιάζονται οι καμπύλες ακρίβειας - ανάκλησης για κάθε προσέγγιση ανάλυσης καλαθιού αγοράς και για τιμές του k από 1 εως και 10. Οι μέθοδοι που απεικονίζονται είναι οι εξής: a. Λανθάνοντα καλάθια με στατιστική επαγωγή b. Λανθάνοντα καλάθια με χρήση ευρετηρίου c. Λανθάνοντα καλάθια με ενίσχυση συνεμφάνισης d. Λανθάνοντες χρήστες με χρήση ευρετηρίου e. Λανθάνοντα καλάθια συνδυασμένα με λανθάνοντες χρήστες f. Κανόνες Συσχέτισης FP-Growth.

Οι καμπύλες που εμφανίζονται στην Εικόνα 7.5 δείχνουν ότι όλες οι μέθοδοι που βασίζονται σε πιθανοτικά μοντέλα θεμάτων και εφαρμόστηκαν στην παρούσα εφαρμογή υπερτερούν σημαντικά σε σχέση την εξαγωγή κανόνων συσχέτισης. Πρέπει να παρατηρηθεί ότι η μέθοδος λανθανόντων καλαθιών με χρήση στατιστικής επαγωγής (καμπύλη (a)) κατορθώνει να προβλέψει σωστά πάνω από το 60% των αντικειμένων που αγοράζονται τελικά. Από την άλλη πλευρά, η δειγματοληψία Gibbs δεν είναι αρκετά αποδοτική σε ταχύτητα και ανάγκες για αποθήκευση καθώς βασίζεται στην στατιστική επαγωγή. Οι εναλλακτικές προσεγγίσεις με βάση ευρετήρια λέξεων που φαίνονται στις καμπύλες (b), (c), (d) και (e) παρέχουν ικανοποιητικά αποτελέσματα σε σχέση με την καμπύλη (a) και μπορούν να χρησιμοποιηθούν σε πρακτικές εφαρμογές.

Η Εικόνα 7.5 επίσης παρουσιάζει τις διαφορές στα αποτελέσματα των διάφορων τεχνικών που χρησιμοποιούν ευρετήρια συσχετίσεων. Και οι δυο προσεγγίσεις για την ανάλυση συναλλαγών, η μια που θεωρεί μόνο μια ομάδα συναλλαγών που συμβαίνουν ταυτόχρονα (λανθάνοντα καλάθια) και η άλλη που θεωρεί ολόκληρο το ιστορικό του χρήστη (λανθάνοντες χρήστες), παράγουν μοντέλα θεμάτων που μπορούν να προβλέψουν με σημαντική ακρίβεια την συμπεριφορά των χρηστών. Πιο συγκεκριμένα, τα θέματα που παράγονται από τους λανθάνοντες χρήστες (d) παρέχουν πιο ακριβείς προτάσεις από τα λανθάνοντα καλάθια (b) όταν ο αριθμός προτάσεων παραμένει μικρός. Το φαινόμενο αυτό αντιστρέφεται καθώς το μέγεθος της λίστας των προτάσεων γίνεται μεγαλύτερο (βλ. Πίνακας 7.6). Η απόδοση των μοντέλων των λανθανόντων καλαθιών βελτιώνεται ελαφρά αν πραγματοποιήσουμε μια ώθηση του σκορ των αντικειμένων που συνυπάρχουν σε πολλά θέματα (c). Ωστόσο, καθώς ο αριθμός των προτάσεων που παρέχονται αυξάνει, αυτή η βελτίωση εξανεμίζεται.

Η μέθοδος που αναπτύξαμε και πραγματοποιεί έναν γραμμικό συνδυασμό λανθανόντων χρηστών και λανθανόντων καλαθιών (e) παρέχει ένα αποτελεσματικό σύστημα προτάσεων τόσο για μικρές όσο και για μεγαλύτερες λίστες προτεινόμενων προϊόντων. Τέλος, αναφορικά με τους κανόνες συσχέτισης που εξήχθησαν με την μεθοδολογία FP-Growth (f), παρατηρούμε ότι η ακρίβεια και η ανάκληση τους παραμένουν σταθερή όσο μεγαλώνει το μέγεθος της λίστας των προτεινόμενων αντικειμένων, καθώς ο αριθμός των κανόνων που μπορούν να εφαρμοστούν δεν μπορούν να προβλέψουν επιπλέον προτάσεις.

7.5 Συμπεράσματα

Σε αυτό το κεφάλαιο εφαρμόζουμε πιθανοτικά μοντέλα θεμάτων σε δεδομένα αγορών ώστε να παράγουμε ένα μοντέλο καταναλωτικής συμπεριφοράς με βάση το οποίο προτείνουμε προϊόντα σε καταναλωτές.

Η ερευνά μας οδηγεί στο συμπέρασμα ότι η λανθάνουσα ανάλυση θεμάτων αποτελεί έναν κατάλληλο και αποτελεσματικό τρόπο για την ανάλυση των δεδομένων αγορών, που ξεπερνά σε αποτελεσματικότητα την εξαγωγή κανόνων συνάφειας που είναι η μέθοδος που συνήθως χρησιμοποιείται στην ανάλυση καλαθιού αγοράς. Η λανθάνουσα ανάλυση θεμάτων δεν παρέχει μόνο μια εποπτική εικόνα των προτιμήσεων των καταναλωτών αλλά μπορεί και να υποστηρίξει ένα σύστημα προτάσεων προϊόντων. Συγκεκριμένα η δυνατότητα της ανακάλυψης λανθανόντων καλαθιών και λανθανόντων χρηστών θέτει τις βάσεις για την κατανόηση των προτιμήσεων των χρηστών αλλά και των σχέσεων που έχουν με τα αντικείμενα. Επιπροσθέτως, τα πιθανοτικά μοντέλα θεμάτων μπορούν να είναι αποτελεσματικά στην πρόταση αντικειμένων στους χρήστες ακόμα και όταν εφαρμόζονται σε μεγάλα σύνολα δεδομένων και σε μεγάλα συνολοστοιχεία.

Ως μελλοντικά ζητήματα διερεύνησης προτείνουμε δυο διαφορετικές προσεγγίσεις. Πρώτον, θεωρούμε την δυνατότητα για υποστήριξη αλλαγών στον τρόπο ζωής των ανθρώπων. Η σημερινή κοινωνία συχνά οδηγεί σε αποφάσεις που δεν λαμβάνουν υπόψη τα αρνητικά μακροπρόθεσμα αποτελέσματα στο περιβάλλον και στην ανθρώπινη υγεία [153]. Το σύστημα FillBasket θα μπορούσε να επεκταθεί ώστε να στοχεύσει και στην αλλαγή του τρόπου ζωής. Το σύστημα θα μπορούσε μέσω της αγοράς προϊόντων να ωθήσει τους χρήστες προς τις αποφάσεις που ωφελούν τα προσωπικά τους μακροπρόθεσμα συμφέροντα. Αυτό μπορεί να επιτευχθεί μέσω μιας κατάλληλης δόμησης των δυνατών επιλογών («Αρχιτεκτονική Επιλογών») [154].

Δεύτερον, είναι ενδιαφέρον να δούμε πως τα μοντέλα θεμάτων μπορούν να τροποποιηθούν ώστε να υποστηρίξουν ένα σύστημα προτάσεων σε μια ηλεκτρονική αγορά και να λάβουν υπόψη τις αξιολογήσεις των χρηστών.

8 TradingLink - Ηλεκτρονικές Αγορές Δημοπρασιών

Ένας μεγάλος αριθμός αντικειμένων τοποθετούνται, αγοράζονται και πωλούνται καθημερινά σε αγορές δημοπρασιών στον ιστό. Η ποσότητα των πληροφοριών και ο αριθμός των διαθέσιμων αντικειμένων προς αγορά κάνει την εύρεση των επιθυμητών αντικειμένων δύσκολη. Επίσης, λόγω του μεγάλου πλήθους διαφορετικών ανταγωνιστικών αντικειμένων που πωλούνται, η πράξη της περιγραφής ενός αντικειμένου προς πώληση αποτελεί μια πρόκληση για τους συμμετέχοντες στην αγορά. Στο κεφάλαιο αυτό προτείνουμε ένα σύστημα προτάσεων βασισμένο σε πιθανοτικά μοντέλα θεμάτων που εκμεταλλεύεται τη λανθάνουσα σημασιολογία στις περιγραφές των αντικειμένων για να υποστηρίξει τις δραστηριότητες των αγοραστών και των πωλητών σε μια ηλεκτρονική αγορά δημοπρασιών. Στη συνέχεια παρουσιάζεται ο σχεδιασμός του συστήματος, η χρήση του σε πραγματικά σενάρια και η αξιολόγηση του.

Για να γίνει εφικτή η υποστήριξη των συμμετεχόντων σε ηλεκτρονικές αγορές χρησιμοποιήθηκε η τεχνολογία των πιθανοτικών μοντέλων θεμάτων. Τη βάση των μοντέλων αποτέλεσε το περιεχόμενο που εισήγαγαν οι χρήστες στην ηλεκτρονική αγορά με μορφή κειμένου, ενώ η γενικότερη περιοχή εφαρμογής του συστήματος είναι οι αγοραπωλησίες. Στη συνεισφορά του παρόντος κεφαλαίου συγκαταλέγεται η περιγραφή της προσέγγισης μας για την ενσωμάτωση των πληροφοριών που εισάγονται σε μια ηλεκτρονική αγορά στο σύστημα προτάσεων.

8.1 Εισαγωγή

Σε μια ηλεκτρονική αγορά οι αγοραστές και οι πωλητές συναντούνται και συναλλάσσονται όπως και σε μια παραδοσιακή αγορά [155]. Σε αντίθεση με τις παραδοσιακές αγορές, οι ηλεκτρονικές αγορές εν γένει εκμεταλλεύονται τις τεχνολογίες που σχετίζονται με την πληροφορία για να συνδέσουν αγοραστές με πωλητές αποτελεσματικά με χαμηλότερο κόστος συναλλαγών, οδηγώντας σε πιο αποδοτικές μορφές αγορών.

Οι συμμετέχοντες σε μια αγορά, αγοραστές, πωλητές και ενδιάμεσοι, κινητοποιούνται από την επιθυμία τους να μεγιστοποιήσουν την ιδιωτική τους ωφέλεια ενώ συνολικά οδηγούν σε μια βέλτιστη κατανομή των μέσων παραγωγής [83]. Αν και οι ηλεκτρονικές αγορές προσφέρουν ένα περιβάλλον χαμηλής τριβής για την πραγματοποίηση αγορών, παρουσιάζουν δυο προκλήσεις στην λειτουργία τους.

Η πρώτη πρόκληση αφορά στην υπερφόρτωση πληροφορίας για τους αγοραστές η οποία οφείλεται στον μεγάλο αριθμό αντικειμένων που είναι διαθέσιμα για αγορά στις ηλεκτρονικές αγορές [2]. Ο αριθμός των διαθέσιμων αντικειμένων, ακόμη και στις πιο στενά ορισμένες κατηγορίες προϊόντων υπερβαίνει κατά πολύ τον αριθμό αντικειμένων που είναι διαθέσιμα σε μη ηλεκτρονικά καταστήματα αλλά και στα περισσότερα ηλεκτρονικά καταστήματα. Η πληροφορία που παρέχεται στην περιγραφή κάθε αντικειμένου είναι εκτενής αλλά και πυκνή, και απαιτεί από τον πιθανό αγοραστή να διαβάσει προσεκτικά και να πραγματοποιήσει έναν αριθμό συγκρίσεων και αξιολογήσεων. Ο συνολικός όγκος πληροφορίας που περιέχεται στις περιγραφές των αντικειμένων μπορεί να εκτιμηθεί με ασφάλεια ότι απαιτεί έναν μεγάλο αριθμό ωρών προσεκτικής εξέτασης από τον αγοραστή ώστε να βρει τα αντικείμενα για τα οποία πραγματικά ενδιαφέρεται. Συχνά ο αγοραστής δεν μπορεί να διακρίνει εύκολα αντικείμενα τα οποία του ταιριάζουν και ως αποτέλεσμα είτε εγκαταλείπει την αναζήτηση είτε συμβιβάζεται με κάτι που δεν καλύπτει τις προσδοκίες του. Επιπρόσθετα, καθώς ο αγοραστής δεν έχει μια γενική εικόνα των διαθέσιμων ενδιαφερόντων αντικειμένων, δυσκολεύεται να χαράξει μια στρατηγική προσφορών προς τον πωλητή.

Η δεύτερη πρόκληση αφορά την έλλειψη αντίληψης του ανταγωνισμού από τους πωλητές. Η ευκολία της χρήσης ηλεκτρονικών αγορών έχει δημιουργήσει έναν καινούριο τύπο μη-επαγγελματιών πωλητών που πωλούν μεταχειρισμένα αντικείμενα. Σε αυτό το περιβάλλον πωλητές από διαφορετικές περιοχές ή ακόμη και κράτη ανταγωνίζονται για την προσοχή των υποψηφίων αγοραστών. Αυτές οι συνθήκες δημιουργούν ένα ανεξερεύνητο τοπίο, όχι ιδιαίτερα φιλόξενο για τους νέους επίδοξους πωλητές. Οι νεοεισερχόμενοι στις ηλεκτρονικές αγορές δεν έχουν μια ξεκάθαρη εικόνα για τους ανταγωνιστές τους και δεν έχουν κάποια υποστήριξη στις εργασίες που είναι οι πλέον κρίσιμες για την δραστηριότητά τους στην αγορά: την περιγραφή του αντικειμένου που θέλουν να πουλήσουν και την επιλογή της τιμής που θα θέσουν ως αρχική και ως τελική.

Για να αντιμετωπιστούν αυτές οι προκλήσεις στις ηλεκτρονικές αγορές έχει προταθεί η χρήση συστημάτων προτάσεων, τόσο συνεργατικής διήθησης όσο και με βάση το περιεχόμενο [37], [107]. Εντούτοις, οι υπάρχουσες υλοποιήσεις αποτυγχάνουν να εκμεταλλευτούν το μη δομημένο περιεχόμενο ώστε να υποστηρίξουν τις δραστηριότητες της πρότασης συναφών αντικειμένων και του εντοπισμού σημαντικών όρων για την συγγραφή της περιγραφής ενός αντικειμένου προς πώληση. Η πρόταση συναφών αντικειμένων έχει να κάνει με τη δυνατότητα του συστήματος να εντοπίζει αντικείμενα τα οποία μοιάζουν μεταξύ τους και μπορεί να ενδιαφέρουν τόσο τον πωλητή όσο και τον επίδοξο αγοραστή. Η πρόταση σημαντικών όρων αφορά την δυνατότητα για εντοπισμό και πρόταση λέξεων που αφορούν το αντικείμενο που περιγράφεται αλλά δεν βρίσκονται ακόμη στην περιγραφή του.

Σε αυτό το κεφάλαιο θεωρούμε δυο δραστηριότητες που μπορούν να υποστηριχθούν από ένα σύστημα προτάσεων στο εσωτερικό μιας ηλεκτρονικής αγοράς. Πρώτον, θεωρούμε την πρόταση σχετικών αντικειμένων στον πιθανό αγοραστή που πλοηγείται στις διάφορες προσφορές. Δεύτερον, προτείνουμε την παροχή σχετικών αντικειμένων αλλά και σχετικών όρων στον πωλητή που περιγράφει ένα προϊόν που θέλει να πουλήσει. Η συνεισφορά μας αποτελείται από την πρόταση του συστήματος TradingLink που εκμεταλλεύεται τα λανθάνοντα θέματα στη μη δομημένη πληροφορία ώστε να υποστηρίξει αυτές τις δραστηριότητες. Για το σκοπό αυτό προτείνουμε μια διαδικασία με τρία βήματα για την εξαγωγή πιθανοτικών μοντέλων θεμάτων ώστε να αποκαλυφθεί η λανθάνουσα σημασιολογία, τη χρήση της ως βάση για τον υπολογισμό της ομοιότητας μεταξύ όρων και αντικειμένων και την παραγωγή των αντίστοιχων προτάσεων.

8.1.1 Ηλεκτρονικές Αγορές

Το ηλεκτρονικό εμπόριο, ως το εμπόριο μέσω του ιστού, αποτελείται από διάφορες συνιστώσες: την παρουσίαση των αγαθών, την προσέλκυση πελατών και την αλληλεπίδραση μαζί τους, τις ηλεκτρονικές συναλλαγές, την υποστήριξη μετά την συναλλαγή και την online επικοινωνία με τους προμηθευτές. Μια ηλεκτρονική αγορά μπορεί να ιδωθεί ως μέσο που αναθέτει διαφορετικούς ρόλους στο εσωτερικό μιας κοινότητας, πρωτίστως στους αγοραστές και στους πωλητές αλλά και κατά δεύτερο λόγο στους πάροχους υπηρεσιών logistics, τράπεζες και άλλους τύπους ενδιάμεσων. Οι αγορές ως μέσα διευκολύνουν την ανταλλαγή

πληροφοριών, αγαθών, υπηρεσιών και πληρωμών ενώ προσφέρουν και τη σχετική υποδομή [156].

Οι αγορές έχουν ως στόχο την αντιστοίχιση προσφοράς και ζήτησης. Η διαδικασία της αντιστοίχισης της ζήτησης των αγοραστών με αντίστοιχη προσφορά αντικειμένων από τους πωλητές αποτελείται από τρία βασικά τμήματα: τον εντοπισμό των προϊόντων που προσφέρονται, την αναζήτηση και την ανακάλυψη των τιμών. Οι δημοπρασίες είναι ένας τρόπος πραγματοποίησης αντιστοιχίσεων τέτοιου τύπου. Στη γενικότερη μορφή τους οι ηλεκτρονικές δημοπρασίες απαιτούν τη συμμετοχή τριών μερών, του δημοπράτη (ενδιάμεσου), του αγοραστή (πελάτη) και του πωλητή (προμηθευτή). Η συμπεριφορά των αγοραστών, των πωλητών και των ενδιάμεσων έχει ως κίνητρο τη μεγιστοποίηση της προσωπικής τους ωφέλειας [155]. Σε αντίθεση με ηλεκτρονικά καταστήματα σταθερής τιμής, οι τιμές αλλάζουν ανάλογα με τη σχέση μεταξύ προσφοράς και ζήτησης σε οποιαδήποτε δεδομένη χρονική στιγμή.

Στις αγορές η δυναμική τιμολόγηση των προϊόντων διαφοροποιείται και μπορεί να πάρει μία από τέσσερις δυνατές μορφές [157]. Οι πιθανοί συνδυασμοί δυναμικής τιμολόγησης με βάση τον αριθμό των συμμετεχόντων είναι: ένας αγοραστής και ένας πωλητής, ένας αγοραστής και πολλοί πιθανοί πωλητές, ένας πωλητής και πολλοί πιθανοί αγοραστές και πολλοί πωλητές με πολλούς αγοραστές. Πιο αναλυτικά οι δημοπρασίες μπορούν να πάρουν μια από τις παρακάτω μορφές:

• **Δημοπρασίες Αγγλικού τύπου** (ή πλειστηριασμοί ή forward δημοπρασίες) όπου ένας πωλητής χειρίζεται διάφορες προσφορές από τους πιθανούς αγοραστές. Τα προϊόντα καταχωρούνται με μια αρχική τιμή, από την οποία ξεκινά η κατάθεση προσφορών αλλά και μια τιμή άμεσης αγοράς στην οποία ο αγοραστής μπορεί να αγοράσει άμεσα το προϊόν.

• **Δημοπρασία τύπου haggle** στην οποία ένας πωλητής και ένας αγοραστής διαπραγματεύονται μέχρι να καταλήξουν σε μια κοινώς αποδεκτή τιμή.

• **Αντίστροφη δημοπρασία** (διαδικασία προσφορών), όπου ένας δυνητικός αγοραστής περιγράφει με ποιους όρους θα αγοράσει από κάθε πωλητή και οι πωλητές πραγματοποιούν προσφορές.

• **Διπλή δημοπρασία** είναι ένα είδος πλειστηριασμού όπου πολλοί αγοραστές και οι τιμές προσφοράς τους αντιστοιχούνται σε πολλούς πωλητές και τις ζητούμενες τιμές τους, εξετάζοντας τις ποσότητες και στις δύο πλευρές.

Στην συνέχεια του παρόντος κεφαλαίου υποθέτουμε τη χρήση των δημοπρασιών αγγλικού τύπου, στις οποίες οι πωλητές επιλέγουν τον αγοραστή με την υψηλότερη προσφορά και για απλότητα αναφερόμαστε σε αυτές απλώς ως «δημοπρασίες». Ωστόσο, οι τεχνικές που παρουσιάζονται εδώ μπορούν να είναι χρήσιμες και σε άλλους τύπους δημοπρασιών καθώς καλύπτουν την κοινή ανάγκη για αντιστοίχιση μεταξύ πωλητών και αγοραστών.

Μια σειρά από οφέλη προκύπτουν από τη χρήση μιας ηλεκτρονικής αγοράς για τους πωλητές, τους αγοραστές και τους δημοπράτες [158].

Οι πωλητές μπορούν να παρατηρήσουν αύξηση των εσόδων τους, εφόσον μπορούν να επικοινωνήσουν με ένα μεγαλύτερο αριθμό πελατών και η αγορά μπορεί να συντομεύσει το χρόνο που απαιτείται για την ολοκλήρωση των συναλλαγών. Επιπλέον οι πωλητές έχουν την δυνατότητα για αύξηση εσόδων και επέκταση του μεριδίου αγοράς βρίσκοντας νέους συνεργάτες, προμηθευτές και από την άμεση πρόσβασή τους στην αγορά [159]. Ακόμη, οι πωλητές μπορούν να κερδίσουν περισσότερα χρήματα καθώς η προμήθεια σε ηλεκτρονική αγορά είναι πολύ χαμηλότερη από τις προμήθειες σε μεσάζοντες ή σε φυσική δημοπρασία. Όταν έχουν επείγουσα ανάγκη για μετρητά, οι πωλητές μπορούν να ρευστοποιήσουν γρήγορα τα αποθέματά τους. Τέλος μπορούν να χτίσουν μια σχέση εμπιστοσύνης με τους πελάτες με την παροχή υπηρεσιών υψηλής ποιότητας και την εκμάθηση των προτιμήσεών τους.

Από την άλλη πλευρά, οι αγοραστές μπορούν να βρουν μοναδικά ή συλλεκτικά αντικείμενα που ταιριάζουν στο γούστο τους ενώ η συμμετοχή σε ηλεκτρονικές δημοπρασίες μπορεί να είναι συναρπαστική και διασκεδαστική. Μπορούν, επίσης, εύκολα και ανώνυμα να συμμετέχουν σε αγορές. Τελικά, το μικρότερο κόστος που έχει μια αναζήτηση σε μια αγορά για τους αγοραστές σημαίνει ότι υπάρχει ανταγωνισμός στις τιμές μεταξύ των πωλητών [155].

Οι δημοπράτες – οι υπεύθυνοι των ιστοτόπων που φιλοξενούν τις ηλεκτρονικές αγορές - αποκτούν έναν καινούριο ρόλο καθώς ως ενδιάμεσοι πρέπει να πραγματοποιούν λειτουργίες όπως αντιστοίχιση μεταξύ πωλητών και αγοραστών, παροχή πληροφοριών για τα αντικείμενα προς πώληση σε επίδοξους αγοραστές, ολοκλήρωση των διαδικασιών των συναλλαγών, διαχείριση παραδόσεων και πληρωμών, δημιουργία και διατήρηση σχέσεων εμπιστοσύνης καθώς και συντήρηση της ακεραιότητας της αγοράς. Οι δημοπράτες, εκτός από τις προμήθειες, μπορούν να εισπράττουν χρήματα λόγω της πραγματοποίησης

πολλαπλών συναλλαγών ενώ επίσης μπορούν να διευρύνουν το κοινωνικό τους δίκτυο γνωριμιών.

Αυτά τα οφέλη μπορούν να εξηγήσουν την θεαματική αύξηση που παρουσιάστηκε στις αγορές δημοπρασίας τα τελευταία χρόνια. Το EBay[27],ο μεγαλύτερος ιστότοπος δημοπρασιών, φιλοξενεί πάνω από 14 εκατομμύρια δημοπρασίες σε κάθε δεδομένη στιγμή. Ένας μεγάλος αριθμός από άλλους δικτυακούς τόπους έχει δημιουργηθεί για να φιλοξενεί δημοπρασίες, όπως το eBid[28], το Online Auction[29] και το Overstock[30].

Προκειμένου να αντιμετωπίσουν τα προβλήματα που αφορούν στον μεγάλο αριθμό αντικειμένων και κατηγοριών, οι ηλεκτρονικές αγορές συνήθως χρησιμοποιούν ταξονομίες. Εκτεταμένες πρωτοβουλίες έχουν ξεκινήσει για να καθοριστούν κοινές γνωσιακές δομές (ταξονομίες ή οντολογίες) ως μέσο για τις κατηγοριοποιήσεις του ηλεκτρονικού εμπορίου [160]. Ωστόσο, αυτές οι ταξονομίες παραμένουν ως επί το πλείστον διάσπαρτες και προσαρμοσμένες σε κάθε ξεχωριστή περίπτωση. Στις περισσότερες περιπτώσεις δεν έχει δημιουργηθεί συναίνεση σε ζητήματα που είναι κοινά σε όλες τις ηλεκτρονικές αγορές, όπως ποιά προϊόντα συνθέτουν ένα τομέα ενδιαφέροντος, πώς πρέπει να περιγραφούν και ποιες είναι οι κατάλληλες δομές του καταλόγου των προϊόντων τους.

Η συμμετοχή μεγάλου αριθμού ανθρώπων και η επακόλουθη υπερφόρτωση πληροφοριών έχει δημιουργήσει προβλήματα στους αγοραστές που επιλέγουν προϊόντα για να αγοράσουν online και στους πωλητές που προσπαθούν να εντοπίσουν τις προτιμήσεις των πελατών [141].

8.2 Σχετικές Εργασίες

Τα συστήματα προτάσεων δεν αποτελούν πλέον σπάνιες εξαιρέσεις αλλά ευρέως διαδεδομένα εργαλεία στον τομέα του ηλεκτρονικού εμπορίου [100]. Ένας μεγάλος αριθμός τεχνικών έχουν σχεδιαστεί, αναπτυχθεί και αξιολογηθεί για την παραγωγή προτάσεων στις περιοχές των δημοπρασιών και του ηλεκτρονικού εμπορίου.

[27] http://www.ebay.com
[28] http://www.ebid.net
[29] http://www.onlineauctioon.com
[30] http://www.overstock.com

Κάποιες από τις δραστηριότητες των χρηστών στο ηλεκτρονικό εμπόριο υποστηρίζονται στη βιβλιογραφία από συστήματα βασισμένα σε πράκτορες (agent–based systems). Τα συγκεκριμένα συστήματα δημιουργούν υπολογιστικά μοντέλα που προσομοιώνουν τις αλληλεπιδράσεις των αυτόνομων πρακτόρων. Επιπρόσθετα έχουν προταθεί συστήματα που βασίζονται σε μαθηματικά μοντέλα και μηχανική μάθηση για να αναλύσουν και να βοηθήσουν την παραγωγή προτάσεων.

Στη βιβλιογραφία έχει προταθεί μια μεθοδολογία η οποία καταδεικνύει την απόδοση ενός ευφυούς πράκτορα και παρουσιάζει τη χρήση του σε μια διαδικασία διαπραγματεύσεων μεταξύ εταιριών και πελατών στο ηλεκτρονικό εμπόριο [161]. Η πειραματική αξιολόγηση που ακολουθεί επιβεβαιώνει τα πλεονεκτήματα της χρήσης τέτοιων συστημάτων. Επίσης έχει προταθεί η χρήση πρακτόρων λογισμικού για να υποστηριχθούν δυο πτυχές των δραστηριοτήτων που αφορούν τις ηλεκτρονικές αγορές [146]: (1) η διαδραστική παραγωγή προτάσεων για την αγορά αντικειμένων και (2) η αυτόματη διαπραγμάτευση τιμών. Τα αποτελέσματα που παρουσιάζονται στην αντίστοιχη μελέτη δείχνουν την αποδοτικότητα του συστήματος αλλά και τις αυξημένες του δυνατότητες. Ακόμη, ένα σύστημα βασισμένο σε πράκτορες μπορεί να ενημερώνει τους συμμετέχοντες για το ιστορικό δημοπρασιών και τιμών [147]. Τα συγκεκριμένα δεδομένα χρησιμοποιούνται για να βελτιωθεί η διαδικασία λήψης αποφάσεων από τους ανθρώπους που συμμετέχουν σε δημοπρασίες.

Στη βιβλιογραφία έχει προταθεί ένα μοντέλο ωφέλειας που βασίζεται σε ικανοποίηση ασαφών περιορισμών για να επιβεβαιώσει την εύρεση μιας λύσης που είναι δίκαια και για τα δυο μέρη [162]. Το μοντέλο χρησιμοποιεί προτεραιότητες μεταξύ ασαφών περιορισμών ώστε να εντοπίσει πως διάφορες υποχωρήσεις πρέπει να γίνουν όταν αυτό είναι αναγκαίο, ενώ επιπροσθέτως ενσωματώνει την ιδέα ενός επιχειρήματος στο μοντέλο αξιολόγησης ώστε οι πράκτορες να μπορούν να οδηγηθούν σε συμφωνίες ενώ αλλιώς αυτό δε θα ήταν δυνατόν. Εναλλακτικά σε μια διαφορετική μελέτη προτείνεται ένα σύστημα με πολλαπλούς πράκτορες λογισμικού με το όνομα ARSEC στο οποίο κάθε συσκευή η οποία χρησιμοποιείται από έναν αγοραστή συνδέεται με έναν πράκτορα συσκευής ο οποίος αυτόνομα καταγράφει την συμπεριφορά του [163]. Κάθε καταναλωτής συνδέεται με έναν πράκτορα καταναλωτή ο οποίος συλλέγει σε ένα καθολικό προφίλ τις πληροφορίες από όλους τους πράκτορες συσκευής του συγκεκριμένου καταναλωτή. Τέλος κάθε ιστότοπος πωλητή, συνδέεται με έναν πράκτορα πωλητή. Με βάση την ομοιότητα μεταξύ των καθολικών προφίλ, οι καταναλωτές ομαδοποιούνται και

εκπροσωπούνται από ένα σύμβουλο πράκτορα. Οι προτάσεις που παράγονται από το σύστημα είναι το αποτέλεσμα της συνεργασίας μεταξύ του πράκτορα πωλητή και των συμβούλων πρακτόρων που συσχετίζονται με τον πελάτη. Με χρήση του πράκτορα συσκευής, λαμβάνεται υπόψη η χρήση της συσκευής – ενώ το συνολικό σύστημα είναι ολοκληρωτικά αποκεντρωμένο.

Επιπλέον στην βιβλιογραφία έχουν αναπτυχθεί μέθοδοι οι οποίες οδηγούν σε μοντέλα και βάσεις γνώσης που μπορούν να χρησιμοποιηθούν για την υποστήριξη συστημάτων προτάσεων. Η μέθοδος γενετικών αλγορίθμων K-μέσων μπορεί να χρησιμοποιηθεί για τον διαχωρισμό της ηλεκτρονικής αγοράς σε τμήματα [108]. Η συγκεκριμένη μελέτη οδήγησε στην ανάπτυξη ενός εργαλείου για την προεπεξεργασία που απαιτείται στα συστήματα προτάσεων. Ακόμη, έχει προταθεί ένα αποδοτικό πλαίσιο που επεκτείνει και χρησιμοποιεί μοντέλο θεμάτων [110] που διαθέτει τη δυνατότητα χρήσης ασταθών δυαδικών παρατηρήσεων και αξιολογείται θετικά σε σύνολα δεδομένων μεγάλων ιστοτόπων ηλεκτρονικού εμπορίου. Στη βιβλιογραφία περιγράφεται και η χρήση μιας επιβλεπόμενης μεθόδου εκμάθησης η οποία οδηγεί στην εξαγωγή των σημασιολογικών κατηγοριών κάθε όρου που περιέχεται στους τίτλους των προϊόντων [109]. Οι κατηγορίες αυτές χρησιμοποιούνται στη συνέχεια για τη βελτίωση των αποτελεσμάτων της αναζήτησης και αυτή η βελτίωση παρατηρείται σε σύνολα δεδομένων από τον πραγματικό κόσμο.

Στη γενική τους μορφή, τα συστήματα προτάσεων στο ηλεκτρονικό εμπόριο έχουν σχεδιαστεί ώστε να ομαδοποιούν τους καταναλωτές και τα προϊόντα αντίστοιχα, και να παράγουν κάποιους κανόνες συσχέτισης μεταξύ καταναλωτών και προϊόντων. Έχει προταθεί ένα εννοιολογικό πλαίσιο υποστήριξης αποφάσεων για ένα online εμπορικό κέντρο, στο οποίο τοποθετείται μια μηχανή αναζήτησης εξωτερικά και ένα σύστημα προτάσεων εσωτερικά [111]. Επίσης έχει αναπτυχθεί μια τεχνική άμεσης υποστήριξης αποφάσεων που βασίζεται σε ένα μαθηματικό μοντέλο που περιγράφει τα χαρακτηριστικά των αγοραστών και τα κέρδη των προμηθευτών [112]. Αυτό το μοντέλο έχει αναπτυχθεί ώστε το σωστό προϊόν να μπορεί να προταθεί στον σωστό άνθρωπο προσφέροντας το μέγιστο κέρδος για την επιχείρηση. Ακόμη, έχει αναπτυχθεί μια υπηρεσία επιλογής προϊόντος η οποία επιστρέφει μια ομάδα αποτελεσμάτων που ομαδοποιούνται σύμφωνα με την πιθανή τους συνάφεια με τον χρήστη [113]. Στα πλαίσια της ίδιας μελέτης προτείνεται μια υπηρεσία που παρέχει στους χρήστες πληροφορίες που αφορούν τις σχέσεις μεταξύ των χαρακτηριστικών των προϊόντων μιας κατηγορίας.

Κάποιες μέθοδοι εξαγωγής ωφέλειας έχουν αναπτυχθεί με βάση την θεωρία ωφέλειας πολλαπλών χαρακτηριστικών (MAUT) για την αναπαράσταση της προτίμησης του ανθρώπου που αποφασίζει. Σε μια από τις σχετικές μελέτες διερευνάται αν αυτές οι τεχνικές που βασίζονται στην ωφέλεια υπερτερούν μεθόδων βασισμένων στο περιεχόμενο για προτάσεις σε πραγματικό χρόνο [164]. Για την επιβεβαίωση των αποτελεσμάτων πραγματοποιήθηκαν πειραματικές αξιολογήσεις σε δυο διαφορετικά περιβάλλοντα ηλεκτρονικού εμπορίου. Τα αποτελέσματα των αξιολογήσεων δείχνουν ότι η απόδοση των τεχνικών εξαρτώνται από τα συμφραζόμενα και το πλαίσιο στο οποίο παρέχεται η πρόταση. Σε μια διαφορετική μελέτη περιγράφεται μια λειτουργική μονάδα συστημάτων προτάσεων για το ηλεκτρονικό εμπόριο, όπου λαμβάνονται υπόψη οι στρατηγικές μάρκετινγκ και η πολυπλοκότητα των σχημάτων των διεπαφών χρήστη [114]. Επίσης προτείνεται μια τεχνική τύπου συνεργατικής διήθησης με επιρροές κλίκας για την πρόβλεψη των προτιμήσεων των χρηστών, ενώ επιπρόσθετα πραγματοποιείται αξιολόγηση του συστήματος.

Οι σχετικές εργασίες που έχουν πραγματοποιηθεί στην περιοχή δεν αντιμετωπίζουν το πρόβλημα της εξαγωγής της λανθάνουσας σημασιολογίας από την τεράστια βάση δεδομένων ενός τόπου δημοπρασιών. Έτσι δεν χρησιμοποιείται το κείμενο που βρίσκεται εκεί για την πρόταση αντικειμένων. Επίσης, είναι σημαντικό ότι στην βιβλιογραφία δεν παρουσιάζονται μέθοδοι για τον εντοπισμό και την πρόταση σημαντικών όρων για να χρησιμοποιηθούν στην περιγραφή αντικειμένων.

8.3 Προσέγγιση

Ως βάση για την προσέγγιση μας θεωρούμε έναν ιστότοπο δημοπρασιών όπου πραγματοποιούνται δημοπρασίες αγγλικού τύπου. Ο πωλητής περιγράφει το αντικείμενο που επιθυμεί να πουλήσει και ορίζει μια αρχική τιμή καθώς και μια τιμή άμεσης αγοράς. Από την άλλη πλευρά, ο αγοραστής πλοηγείται μεταξύ αντικειμένων ψάχνοντας για αυτό το οποίο θα καλύψει τις απαιτήσεις του.

Για την υποστήριξη τέτοιου τύπου δραστηριοτήτων σε έναν ιστότοπο τα αντικείμενα και οι περιγραφές τους αντιμετωπίζονται ως έγγραφα. Τα δεδομένα αυτά εισάγονται σε ένα μοντέλο θεμάτων και προ-υπολογίζονται οι ομοιότητες μεταξύ των αντικειμένων και μεταξύ αντικειμένων και λέξεων. Οι ομοιότητες αυτές χρησιμοποιούνται στη συνέχεια στο σύστημα προτάσεων ώστε: (1) να προτείνει

ενδιαφέροντα αντικείμενα στους επίδοξους αγοραστές και (2) να προτείνει σχετικά αντικείμενα αλλά και λέξεις στους πωλητές. Η χρήση των μοντέλων θεμάτων επιτρέπει την ανάκτηση ομοιοτήτων που δεν ήταν εμφανείς προηγουμένως.

8.3.1 Μοντέλο Θεμάτων και Ομοιότητα

Το πρώτο μέρος της επεξεργασίας αφορά την εξαγωγή του πιθανοτικού μοντέλου θεμάτων και τον αριθμητικό υπολογισμό των ομοιοτήτων που θα χρησιμοποιηθούν για τη δημιουργία προτάσεων.

Για την εκπαίδευση του μοντέλου θεμάτων χρησιμοποιείται η τεχνική της λανθάνουσας κατανομής Dirichlet [54]. Η εκπαίδευση του μοντέλου οδηγεί στον καθορισμό των κατανομών πιθανοτήτων των λέξεων σε συγκεκριμένα θέματα αλλά και των θεμάτων που αποτελούν τα έγγραφα. Για να πραγματοποιηθούν προτάσεις όμως απαιτείται ο υπολογισμός των ομοιοτήτων μεταξύ αντικειμένων και λέξεων.

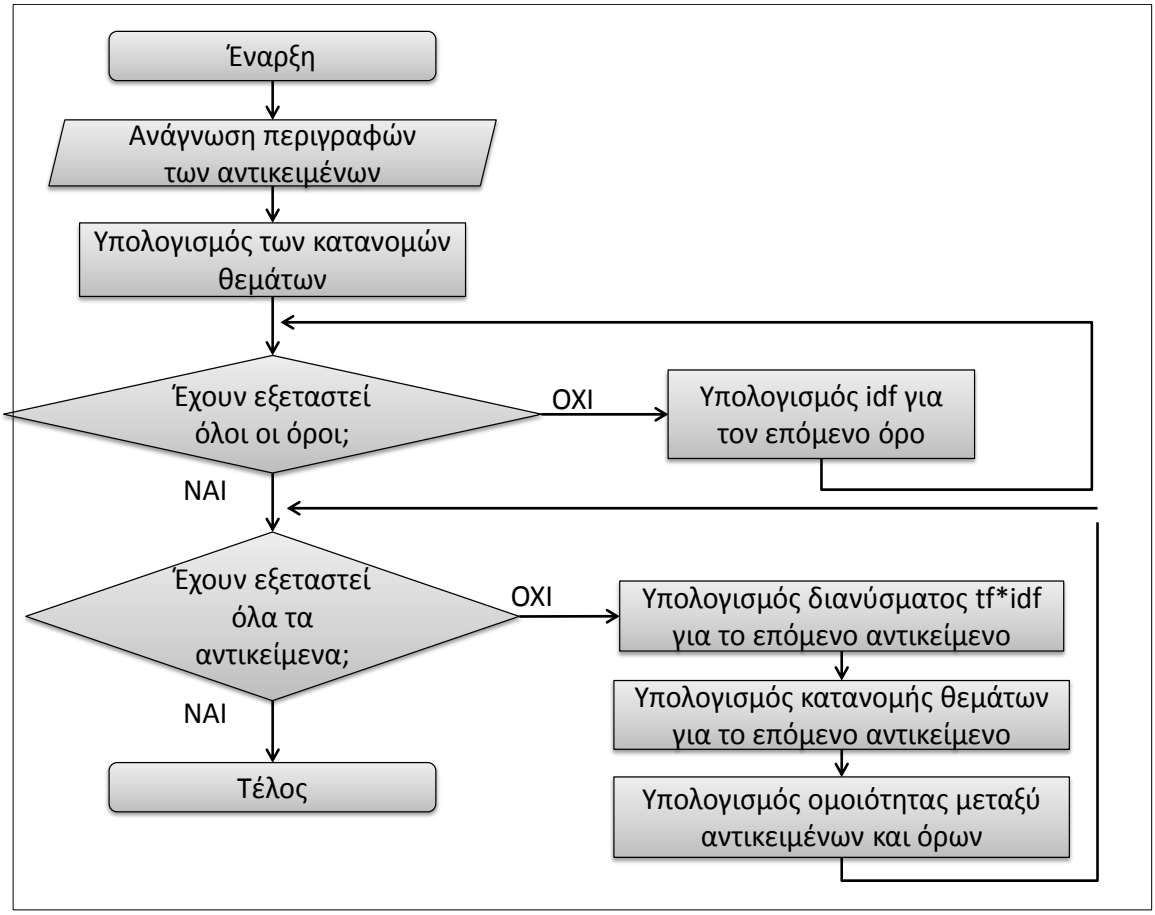
Πρώτον, για να υπολογιστεί η ομοιότητα μεταξύ των αντικειμένων χρησιμοποιείται η κατανομή θεμάτων κάθε αντικειμένου. Το κείμενο της περιγραφής του κάθε αντικειμένου εκτιμάται ότι είναι συναφές με κάποια από τα λανθάνοντα θέματα, με διαφορετικό βαθμό συνάφειας με το καθένα. Η ομοιότητα συνημίτονου μεταξύ των κατανομών θεμάτων υπολογίζεται και χρησιμοποιείται ως μετρική ομοιότητας, βλέπε (8.1).

$$item\_sim^{topic}_{ij} = \frac{\sum_{\text{topics that relate with both items}} S_{topic\_item\_i} S_{topic\_item\_j}}{\sum_{\text{topics of item } i} \left(S_{topic\_item\_i}\right)^2 \sum_{\text{topics of item } j} \left(S_{topic\_item\_j}\right)^2} \quad (8.1)$$

Δεύτερον, χρειάζεται να υπολογίσουμε την ομοιότητα μεταξύ του κειμένου της περιγραφής ενός αντικειμένου και των συγκεκριμένων όρων που υπάρχουν στο σύστημα TradingLink. Για να περιγράψουμε το κείμενο χρησιμοποιούμε τα λανθάνοντα θέματα όπως παραπάνω. Με τον ίδιο τρόπο, η κατανομή των λέξεων σε κάθε θέμα χρησιμοποιείται για να σχηματιστεί ένα διάνυσμα θεμάτων στα οποία μπορεί να βρεθεί η κάθε λέξη. Έτσι μπορούμε να υπολογίσουμε την ομοιότητα συνημίτονου μεταξύ των λέξεων και των θεμάτων (8.2).

$$tag\_sim^{topic}_{il} = \frac{\sum_{\text{topics with word } l} S_{topic\_item\_i} S_{topic\_word\_l}}{\sum_{\text{topics of item } i} \left(S_{topic\_item\_i}\right)^2 \sum_{\text{topics with word } l} \left(S_{topic\_word\_l}\right)^2} \quad (8.2)$$

Μια επισκόπηση της διαδικασίας παρουσιάζεται στην Εικόνα 8.1.



Εικόνα 8.1 Διαδικασία Επεξεργασίας Δεδομένων της Ηλεκτρονικής Αγοράς

8.3.2 Πρόταση Αντικειμένου σε Αγοραστές

Ένας χρήστης, ο οποίος είναι πιθανός αγοραστής, σε μια ηλεκτρονική αγορά μπορεί να έχει πρόσβαση σε διάφορα αντικείμενα μέσα από διάφορους τρόπους: κοιτώντας σε συγκεκριμένες κατηγορίες ή κοιτώντας τα πιο πρόσφατα αντικείμενα που έγιναν διαθέσιμα. Όταν φτάνει σε ένα αντικείμενο που τον ενδιαφέρει θεωρούμε ότι μπορεί να μελετήσει τα διάφορα χαρακτηριστικά του, και πιο συγκεκριμένα την περιγραφή του και την τιμή που χρειάζεται να πληρώσει για να συμμετέχει στη δημοπρασία ή για να το αγοράσει άμεσα.

Απομονώνουμε το κομμάτι αυτό της διαδικασίας όπου ο χρήστης πλοηγείται σε ένα αντικείμενο και το επεκτείνουμε με τη χρήση ενός συστήματος προτάσεων που βασίζεται σε μοντέλα θεμάτων. Η περιγραφή ενός αντικειμένου αναλύεται και η ομοιότητα της κατανομής θεμάτων του σε σχέση με τα υπάρχοντα αντικείμενα

υπολογίζεται όπως στο (8.1). Η ομοιότητα συνδυάζεται με την ομοιότητα tf*idf όπως στο (8.3) με τη χρήση μιας παραμέτρου ανάμειξης μ. Ο υπολογισμός της tf*idf ομοιότητας πραγματοποιήθηκε χρησιμοποιώντας διανύσματα κάθε αντικειμένου που παράγονται από τις λέξεις που περιλαμβάνονται στην περιγραφή του αντικειμένου. Η ομοιότητα αυτή συνδυάζεται με τα μοντέλα θεμάτων ώστε να δοθεί έμφαση στο γεγονός ότι ακριβώς οι ίδιες λέξεις μπορούν να βρεθούν σε δυο αντικείμενα.

Μια λίστα με τα πιο συναφή αντικείμενα δημιουργείται και προωθείται στον χρήστη ως λίστα προτάσεων.

$$item\_sim_{ij} = (1-\mu)\cdot item\_sim^{tfidf}{}_{ij} + \mu \cdot item\_sim^{topic}{}_{ij} \qquad (8.3)$$

8.3.3 Πρόταση Όρων και Αντικειμένων σε Πωλητές

Ένας πωλητής σε μια αγορά μπορεί να είναι ένας άνθρωπος ή μια εταιρία που επιχειρεί να πουλήσει ένα αντικείμενο που έχει στην κατοχή του. Επιδιώκει, συνήθως, να εισπράξει την μεγαλύτερη δυνατή αξία σε χρήματα για το δεδομένο αντικείμενο βρίσκοντας κάποιον που θα εκτιμήσει τα μοναδικά του χαρακτηριστικά. Έτσι προκύπτουν δυο ανάγκες: η ανάγκη για ξεκάθαρη και πλήρη περιγραφή του αντικειμένου και η ανάγκη για σωστή τιμολόγηση.

Σε μια ηλεκτρονική αγορά δημοπρασιών τόσο οι αγοραστές όσο και οι πωλητές επιδιώκουν να ελαχιστοποιήσουν τα κόστη συναλλαγών, τα οποία γενικά ορίζονται ως τα κόστη της αναζήτησης της σωστής εναλλακτικής, της διαπραγμάτευσης και της εφαρμογής μιας συμφωνίας αγοράς [165]. Από την περιγραφή του αντικειμένου μπορεί να απουσιάζουν σημαντικά στοιχεία για το αντικείμενο. Οι πιθανοί αγοραστές μπορεί να χάσουν το ενδιαφέρον τους και να προχωρήσουν στον επόμενο πωλητή αν δεν βρουν την πληροφορία που θα ήθελαν.

Για να αντιμετωπιστεί αυτό το πρόβλημα το σύστημα παρέχει τους πιο σχετικούς όρους οι οποίοι δεν περιέχονται στην τρέχουσα περιγραφή του αντικειμένου. Κατά τη διάρκεια της πληκτρολόγησης της περιγραφής από τον χρήστη, το κείμενο αναλύεται και εξάγεται η κατανομή θεμάτων που αντιστοιχεί σε αυτό. Αυτά τα θέματα χρησιμοποιούνται για να εντοπιστούν όροι οι οποίοι χρησιμοποιούνται κατ' επανάληψη σε παρόμοια αντικείμενα, με χρήση της εξίσωσης (8.2). Στην προσέγγισή μας εκτιμούμε ότι αυτοί οι όροι αναπαριστούν σημαντικά κομμάτια πληροφορίας τα οποία μπορεί να βρεθούν σε παρόμοια και

ενδεχομένως ανταγωνιστικά προϊόντα, άρα πρέπει να λαμβάνονται υπόψη κατά τη συγγραφή της περιγραφής του αντικειμένου από τον πωλητή.

Οι δημοπρασίες που διαδραματίζονται στο εσωτερικό της ηλεκτρονικής αγοράς περιλαμβάνουν δυο στοιχεία τα οποία πρέπει να ορίζονται από τον πωλητή, την τιμή εκκίνησης και την τιμή άμεσης αγοράς. Αυτές οι τιμές βασίζονται στην εικόνα που έχει ο πωλητής για την αξία του προϊόντος το οποίο θέλει να πουλήσει στην αγορά. Η σωστή τοποθέτηση του προϊόντος στην αγορά μπορεί να διασφαλίσει ότι ο πωλητής θα εισπράξει την μέγιστη δυνατή τιμή για το προϊόν του στην αγορά. Για την υποστήριξη αυτής της λειτουργίας, όσο ο πωλητής γράφει την περιγραφή του προϊόντος στο αντίστοιχο κομμάτι της διεπαφής, το σύστημα βρίσκει παρόμοια αντικείμενα, πιθανώς ανταγωνιστικά, και τα εμφανίζει στον χρήστη μαζί με τις τιμές διαπραγμάτευσης (8.1). Αυτή η πληροφορία μπορεί να βοηθήσει τον πωλητή να θέσει τις τιμές για το προϊόν που θέλει να πουλήσει.

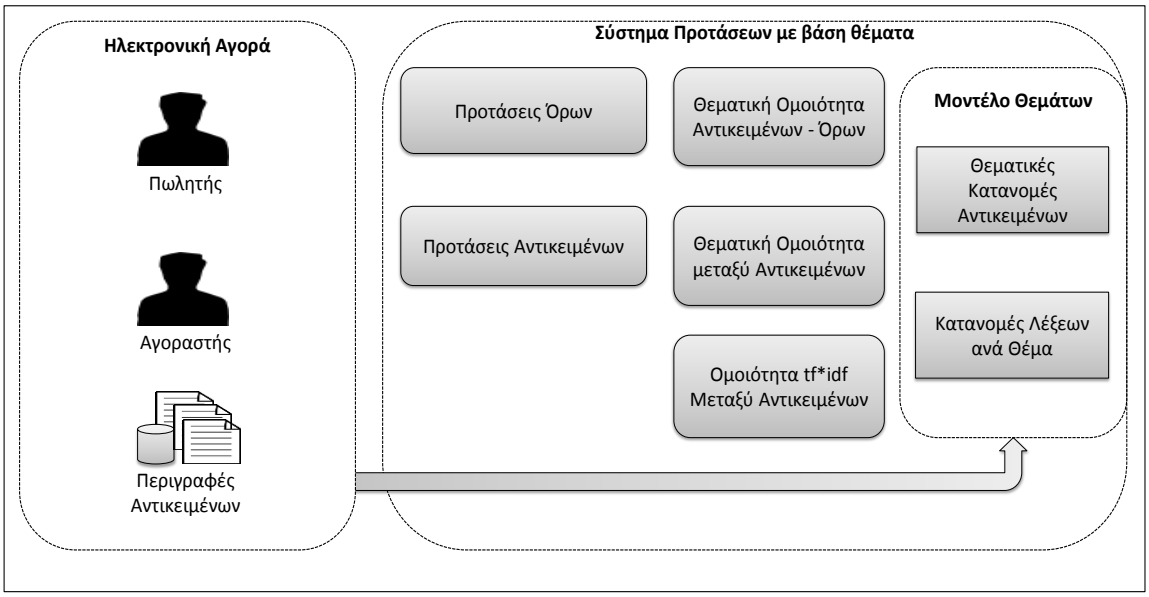
8.3.4 Περιγραφή Συστήματος TradingLink

Παρακάτω περιγράφεται το σύστημα TradingLink που προτείνεται ως μέρος μιας ηλεκτρονικής αγοράς για την υποστήριξη των συμμετεχόντων με βάση την προσέγγιση που αναπτύχθηκε προηγουμένως.

Το σύστημα αυτό βασίζεται σε μια σχεσιακή βάση δεδομένων και αναπτύσσεται στο πλαίσιο ενός εξυπηρετητή εφαρμογών Java (java application server). Το σύστημα TradingLink παρέχει δυο βασικές λειτουργίες: (1) την υποστήριξη των αγοραστών στην εύρεση αντικειμένων που μπορεί να τον ενδιαφέρουν και (2) υποστήριξη των πωλητών στην περιγραφή και στην τιμολόγηση των προϊόντων που θέλουν να πουλήσουν. Για να παράγονται αυτές οι προτάσεις έχουν τοποθετηθεί αντίστοιχες υπηρεσίες. Πλέον των υπηρεσιών που αφορούν τις προτάσεις, έχει αναπτυχθεί και μια υπηρεσία που αφορά την εξαγωγή των θεμάτων και τον υπολογισμό της ομοιότητας.

Στην Εικόνα 8.2 παρουσιάζεται μια εποπτική εικόνα του συστήματος. Οι περιγραφές των αντικειμένων που βρίσκονται στην ηλεκτρονική αγορά δημοπρασιών αναλύονται ώστε να παραχθούν ομοιότητες tf*idf και να εξαχθεί ένα πιθανοτικό μοντέλο θεμάτων. Το μοντέλο θεμάτων αποθηκεύεται με τη μορφή των κατανομών αντικειμένων σε θέματα και με τη μορφή θεμάτων σε λέξεις. Αυτές οι κατανομές με τη σειρά τους χρησιμοποιούνται για την παραγωγή προτάσεων με βάση τα θέματα μεταξύ αντικειμένων και λέξεων. Οι προτάσεις εμφανίζονται στους

πωλητές και στους αγοραστές όταν αυτοί πλοηγούνται στην αγορά δημοπρασιών. Η μεθοδολογία των υπολογισμών έχει παρουσιαστεί στη προηγούμενη ενότητα.



Εικόνα 8.2 Αρχιτεκτονική Συστήματος

8.4 Μελέτη Εφαρμογής

Σε αυτή την ενότητα περιγράφουμε πως ένα σύστημα προτάσεων που βασίζεται σε μοντέλα θεμάτων μπορεί να ενσωματωθεί σε μια πραγματική ηλεκτρονική αγορά δημοπρασιών. Περιγράφουμε την μελέτη εφαρμογής, το πιθανοτικό μοντέλο που εξήχθηκε, ένα σενάριο χρήσης και την αξιολόγηση που πραγματοποιήθηκε.

8.4.1 EBid: Ηλεκτρονική Αγορά Δημοπρασιών

Το EBid είναι μια μεγάλη και αρκετά γνωστή ηλεκτρονική αγορά δημοπρασιών. Μέχρι και τον Ιούνιο του 2012 ο ιστότοπος αυτός περιλάμβανε πάνω από 1.950.000 δημοπρασίες ενώ περίπου 50.000 νέα αντικείμενα προσθέτονταν κάθε 24 ώρες και η συνολική αξία των αντικειμένων που υπήρχαν στον ιστότοπο ξεπερνούσε τα 500 εκατομμύρια ευρώ.

Το σύνολο δεδομένων που χρησιμοποιήθηκε για την αξιολόγηση της προτεινόμενης μεθόδου έχει ληφθεί από το γνωστό ιστότοπο ηλεκτρονικής αγοράς

δημοπρασιών, Ebid. Στο χρονικό διάστημα 20 έως 25 Μαΐου 2012 καταγράφηκαν 65.205 αντικείμενα που δόθηκαν σε δημοπρασίες στον συγκεκριμένο ιστότοπο. Οι περιγραφές των αντικειμένων αλλά και οι τιμές έναρξης διαπραγμάτευσης και άμεσης πώλησης καταγράφηκαν στο σύνολο δεδομένων. Οι δημοπρασίες που εντάχθηκαν στο σύνολο δεδομένων είναι οι ανοιχτές δημοπρασίες τη στιγμή της λήψης ενώ δεν καταγράφονται ούτε χρησιμοποιούνται οι κατηγορίες στις οποίες ανήκουν τα αντικείμενα.

8.4.2 Ανάλυση Θεμάτων

Η λανθάνουσα κατανομή Dirichlet εφαρμόστηκε στο σύνολο των περιγραφών των αντικειμένων. Για τον σκοπό αυτό χρησιμοποιήθηκε το πακέτο λογισμικού Mallet [150].

Πριν την εξαγωγή του θεματικού μοντέλου, οι περιγραφές των αντικειμένων υφίστανται προ-επεξεργασία. Οι συνήθεις λέξεις όπως οι λέξεις «το», «όταν» και «ο» αφαιρούνται καθώς θεωρείται ότι δεν έχουν αξιοπρόσεκτη σημασιολογική αξία. Τα λήμματα από τα οποία προέρχονται οι λέξεις αντικαθιστούν τις λέξεις καθαυτές όπου αυτό είναι εφικτό, με τη χρήση της τεχνικής Porter [32] για την αγγλική γλώσσα. Δεν λαμβάνεται υπόψη η σύνταξη των προτάσεων και η σειρά των λέξεων, καθώς χρησιμοποιείται η υπόθεση «συνόλου λέξεων» (bag-of-words).

Αριθμός Θέματος	**Πιθανότερες Λέξεις Θεμάτων**
6	light, night, day, sides, led, secret
40	isbn,book,dog,publisher,978,product,pages,english,weight
28	items,postage,standard,delivery,vintage,postcards,unposted,posted
52	battery, charger, batteries, ion
68	comic, bagged, major, promotional, free, copy
95	vinyl, producer, written, genre, art, condition, garfunkel

Πίνακας 8.1 Λανθάνοντα Θέματα στις Περιγραφές των Αντικειμένων

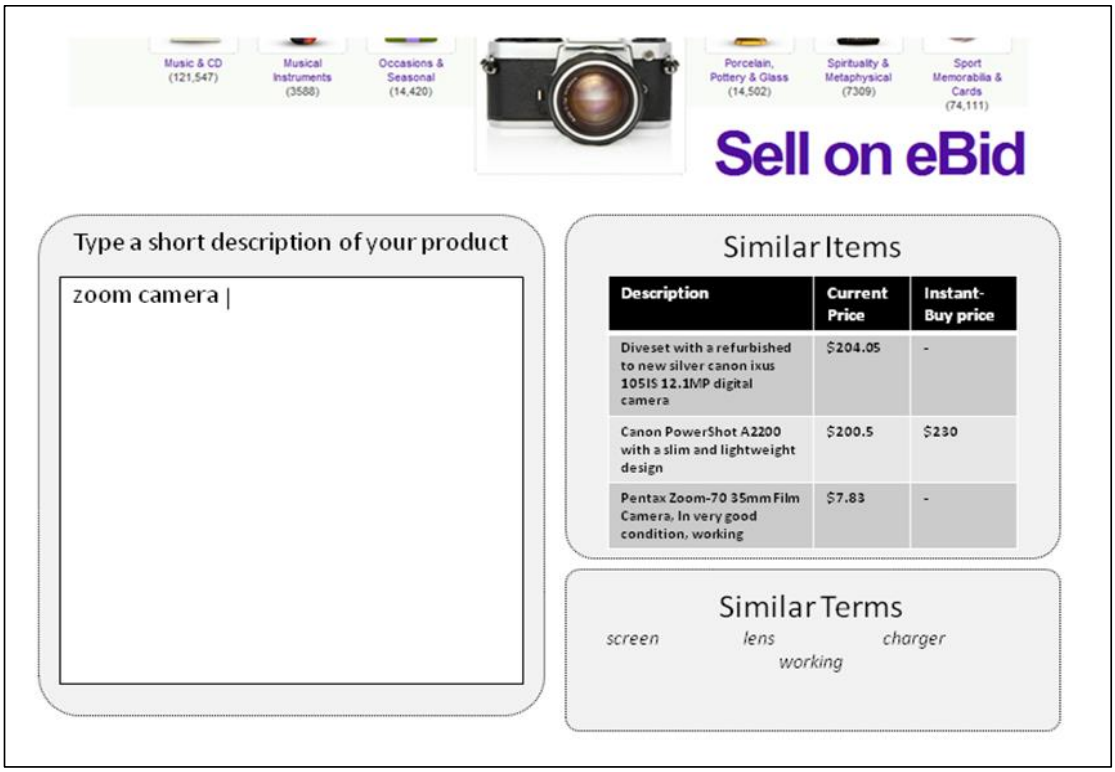
Για την δημιουργία του μοντέλου επιλέχθηκε να εξαχθούν 500 θέματα, άνω για την εκτίμηση της σύνθεσης των θεμάτων έγιναν 2.000 επαναλήψεις της αντίστοιχης δειγματοληψίας. Παρατηρήθηκε ότι με αυτές τις παραμέτρους το

σύστημα μπορούσε να παρέχει σταθερά θέματα που παρουσιάζουν σύγκλιση. Το μοντέλο θεμάτων που εξάχθηκε περιέχει τις πιθανοτικές σχέσεις μεταξύ των όρων και των θεμάτων αλλά και μεταξύ θεμάτων και αντικειμένων.

Ένα παράδειγμα αντιπροσωπευτικών θεμάτων που εξάγονται από τις περιγραφές των αντικειμένων φαίνονται στον σχετικό πίνακα (Πίνακας 8.1).

8.4.3 Σενάριο Χρήσης

Εδώ περιγράφουμε την λειτουργικότητα που παρέχεται από το σύστημα σε ένα υποθετικό σενάριο χρήσης. Το σενάριο που ακολουθεί αφορά την δημιουργία και την πώληση ενός προϊόντος σε μια ηλεκτρονική αγορά. Για να μπορέσουμε να παρουσιάσουμε με μεγαλύτερη ακρίβεια τη χρήση του συστήματος προτάσεων στην ηλεκτρονική αγορά δημοπρασιών έχουμε ανακατασκευάσει δυο ιστοσελίδες από τον πραγματικό ιστότοπο eBid. Δείχνουμε το σύστημα όπως θα το χρησιμοποιούσαν δυο χρήστες: ένας πωλητής και ένας αγοραστής, η Sally και ο Bob.



Εικόνα 8.3 Διεπαφή Πωλητή

Η Sally έχει μια παλιά ψηφιακή φωτογραφική μηχανή. Αφού αγόρασε ένα καινούριο μοντέλο την προηγούμενη εβδομάδα, αποφασίζει να πουλήσει το παλιό

μοντέλο που είχε. Οι φίλοι της και το οικογενειακό της περιβάλλον δεν ενδιαφέρονται να αγοράσουν το παλιό αυτό μοντέλο. Αποφασίζει να χρησιμοποιήσει τον ιστότοπο ηλεκτρονικών αγορών EBid για να βρει πιθανούς αγοραστές για την φωτογραφική του μηχανή.

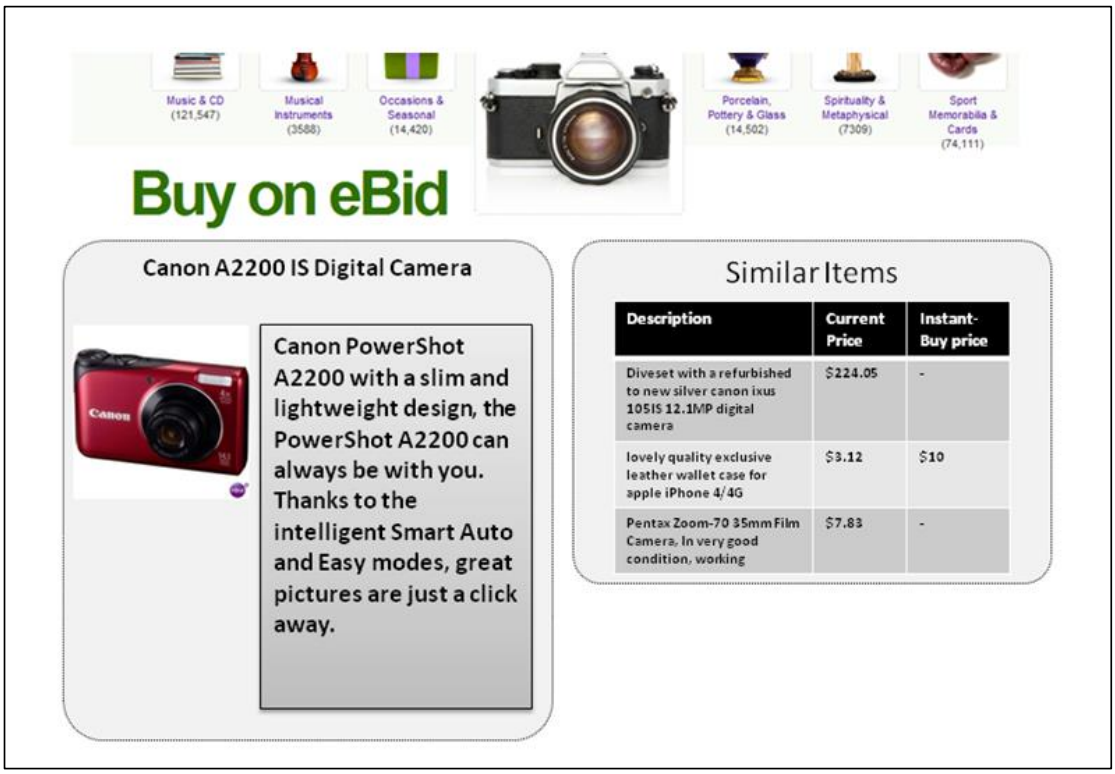
Καθώς η Sally γράφει τις λέξεις *camera και zoom* που αναφέρεται στην μάρκα κατασκευής του προϊόντος ένα κομμάτι στα δεξιά της σελίδας ανανεώνεται (Εικόνα 8.3). Συναφείς σημασιολογικά όροι παράγονται από το σύστημα οι οποίοι μπορούν να βοηθήσουν τον χρήστη στη περιγραφή της φωτογραφικής μηχανής. Αυτοί οι όροι είναι στην περίπτωση αυτή *screen*, *lens* και *charger*. Αυτοί οι όροι την βοηθούν να εισάγει αυτές τις πληροφορίες στην προσφορά της: γράφει μια πλήρη περιγραφή του φακού και της οθόνης της φωτογραφικής μηχανής, και επισημαίνει το γεγονός ότι στο πακέτο προς πώληση περιλαμβάνεται κι ένας φορτιστής.

Ενώ η Sally ολοκληρώνει την περιγραφή της φωτογραφικής μηχανής, τα παρόμοια αντικείμενα στα δεξιά της περιγραφής ενημερώνονται. Τα παρόμοια αντικείμενα ανακτώνται από την βάση των αντικειμένων που είναι διαθέσιμα στο eBid με βάση τα μοντέλα θεμάτων. Κοιτώντας τα παρόμοια αντικείμενα ο χρήστης μπορεί να συγκρίνει την τιμή έναρξης της δημοπρασίας και την τιμή έναρξης αγοράς με εκείνες που θέλει να θέσει στο προϊόν του αντίστοιχα. Η Sally που θέλει να πουλήσει την φωτογραφική μηχανή της, παρατηρεί ότι οι δημοπρασίες για παρόμοια αντικείμενα που δεν περιλαμβάνουν φορτιστή κινούνται στα $200. Έτσι αποφασίζει να επιτρέψει την έναρξη από τα $190 και την άμεση αγορά με $210 και καταχωρεί το προϊόν.

Ο Bob, εντωμεταξύ, πλοηγείται στον ιστότοπο του eBid ψάχνοντας για μια μεταχειρισμένη φωτογραφική μηχανή για το δεκαπεντάχρονο παιδί του. Κοιτώντας ένα άλλο μοντέλο φωτογραφικής μηχανής, στο σύστημα αυτόματα ανακτά παρόμοιες προσφορές τις οποίες εμφανίζει στο δεξί τμήμα της οθόνης (Εικόνα 8.4). Η φωτογραφική μηχανή της Sally προτείνεται. Ο Bob βρίσκει την προσφορά συμφέρουσα καθώς περιλαμβάνεται και ο φορτιστής και αγοράζει το πακέτο άμεσα.

Σε αυτό το σενάριο βλέπουμε ότι το σύστημα μπορεί να βοηθήσει στην αντιστοίχιση μεταξύ των αγοραστών και των πωλητών με δυο τρόπους. Πρώτον, οι συμμετέχοντες μπορούν να χρησιμοποιήσουν την δυνατότητα του συστήματος και να βρουν συναφή αντικείμενα με αυτά που ψάχνουν ή που θέλουν να πουλήσουν. Δεύτερον, οι πωλητές μπορούν να χρησιμοποιήσουν τις προτάσεις όρων για να

παρέχουν πληρέστερες περιγραφές των προϊόντων που θέλουν να πουλήσουν, αλλά και ώστε να ορίσουν σωστά τις τιμές τους.



Εικόνα 8.4 Διεπαφή Αγοραστή

8.4.4 Αξιολόγηση

Δυο μέθοδοι εφαρμόστηκαν για την αξιολόγηση της προτεινόμενης προσέγγισης. Στην πρώτη μέθοδο αξιολογούμε την ποιότητα των μοντέλων θεμάτων που εξήχθησαν με βάση μια μετρική που έχουν παρουσιαστεί στη βιβλιογραφία [132]. Στη δεύτερη μέθοδο, αξιολογούμε τη χρήση του συστήματος προτάσεων το οποίο βασίστηκε στο μοντέλο θεμάτων.

Μια προσέγγιση στην αξιολόγηση της ποιότητας του μοντέλου θεμάτων που παράχθηκε είναι εκείνη της εκτίμησης της πιθανότητας εμφάνισης των κρατημένων εγγράφων (held-out probability). Ο αριθμός αυτός εκφράζει την πιθανότητα το μοντέλο που εξάγεται να μπορούσε να είχε παράγει τα αντικείμενα τα οποία ανήκουν στο σύνολο δεδομένων αλλά δεν είχαν τροφοδοτηθεί στο σύστημα για την εκπαίδευση του μοντέλου. Στη βιβλιογραφία έχει προταθεί η προσέγγιση «αριστερά προς τα δεξιά» για την εκτίμηση της πιθανότητας αυτής [132]. Στον αντίστοιχο πίνακα (Πίνακας 8.2) παρουσιάζουμε την διαφορετική πιθανότητα που εμφανίζεται

όταν χρησιμοποιούμε διαφορετικό αριθμό θεμάτων κατά την εκπαίδευση του μοντέλου θεμάτων που χρησιμοποιούμε. Η υψηλότερη λογαριθμική πιθανοφάνεια ανά λήμμα, που υπολογίζεται διαιρώντας την πιθανότητα με τον συνολικό αριθμό των λημμάτων που βρίσκονται στο σύνολο δεδομένων, σηματοδοτεί ένα καλύτερο μοντέλο. Οι υπολογισμοί που αφορούν στην αξιολόγηση του μοντέλου θεμάτων έχουν γίνει με το πακέτο λογισμικού Mallet [150].

Πίνακας 8.2 Αξιολόγηση Αριστερά προς Δεξιά Μοντέλου Θεμάτων

Αριθμός Θεμάτων	Λογαριθμική Πιθανοφάνεια	Λογαριθμική Πιθανοφάνεια / Λήμμα
300	-6.924.695	-1,3087
500	-6.882.160	-1,3006
1000	-7.162.187	-1,3532
2000	-6.949.959	-1,3132

Τα αποτελέσματα που παρουσιάζονται (Πίνακας 8.2) δείχνουν ότι η επιλογή του αριθμού των 500 θεμάτων στο μοντέλο συντελεί στην εμφάνιση υψηλής πιθανότητας κρατημένων εγγράφων. Επίσης, σε σχέση με προηγούμενα αποτελέσματα [54], [132] συμπεραίνουμε ότι το μοντέλο που εξήχθηκε είναι σταθερό και μπορεί να χρησιμοποιηθεί ως βάση για παραγωγή προτάσεων.

Στην δεύτερη μέθοδο αξιολόγησης επιλέξαμε τυχαία 15 σενάρια χρήσης του συστήματος προτάσεων. Αυτά τα σενάρια περιλαμβάνουν την δραστηριότητα των πωλητών και των αγοραστών στην ηλεκτρονική αγορά, κατά τη διάρκεια της οποίας τους προσφέρονται προτάσεις αντικειμένων και όρων. Αυτά τα σενάρια περιγράφονται με ακρίβεια σε ένα ερωτηματολόγιο που έχει παραδοθεί σε 32 συμμετέχοντες. Το ερωτηματολόγιο μπορεί να βρεθεί στο τέλος της παρούσας διατριβής, στο «Παράρτημα 2: Σενάρια Αξιολόγησης Συστήματος TradingLink», σελ. 253. Οι συμμετέχοντες ήταν 32 άτομα ελληνικής καταγωγής κάτοικοι της Αθήνας, 11 κάτοχοι πτυχίων πανεπιστημίων, 15 απόφοιτοι μεταπτυχιακών σπουδών και 5 κάτοχοι διδακτορικών διπλωμάτων. Οι ηλικίες κυμαίνονται από 22 εως 51 έτη ενώ ο μέσος όρος είναι 31,5 έτη. Η έρευνα έλαβε χώρα κατά τον Ιούνιο του 2012 και διήρκησε 2 εβδομάδες. Η συμπλήρωση έγινε μέσω του διαδικτύου και συγκεκριμένα μέσω τη εφαρμογής google forms όπου οι συμμετέχοντες διάβαζαν την περιγραφή του σεναρίου χρήσης και την αντίστοιχη λειτουργία του προτεινόμενου συστήματος.

Οι συμμετέχοντες αξιολόγησαν την χρησιμότητα των προτάσεων με χρήση μιας κλίμακας Likert, όπου το 1 σημαίνει καθόλου χρήσιμο ενώ το 5 πολύ χρήσιμο. Οι αξιολογήσεις αυτές αυθαίρετα ανατοποθετήθηκαν σε ένα εύρος από [0,1], ώστε να διευκολύνουν τον αναγνώστη. Στον αντίστοιχο πίνακα (Πίνακας 8.3) βλέπουμε τον αριθμό του κάθε σεναρίου, τον ρόλο που καλείται να παίξει ο συμμετέχων, τον τύπο της πρότασης και τον μέσο όρο και την τυπική απόκλιση των απαντήσεων.

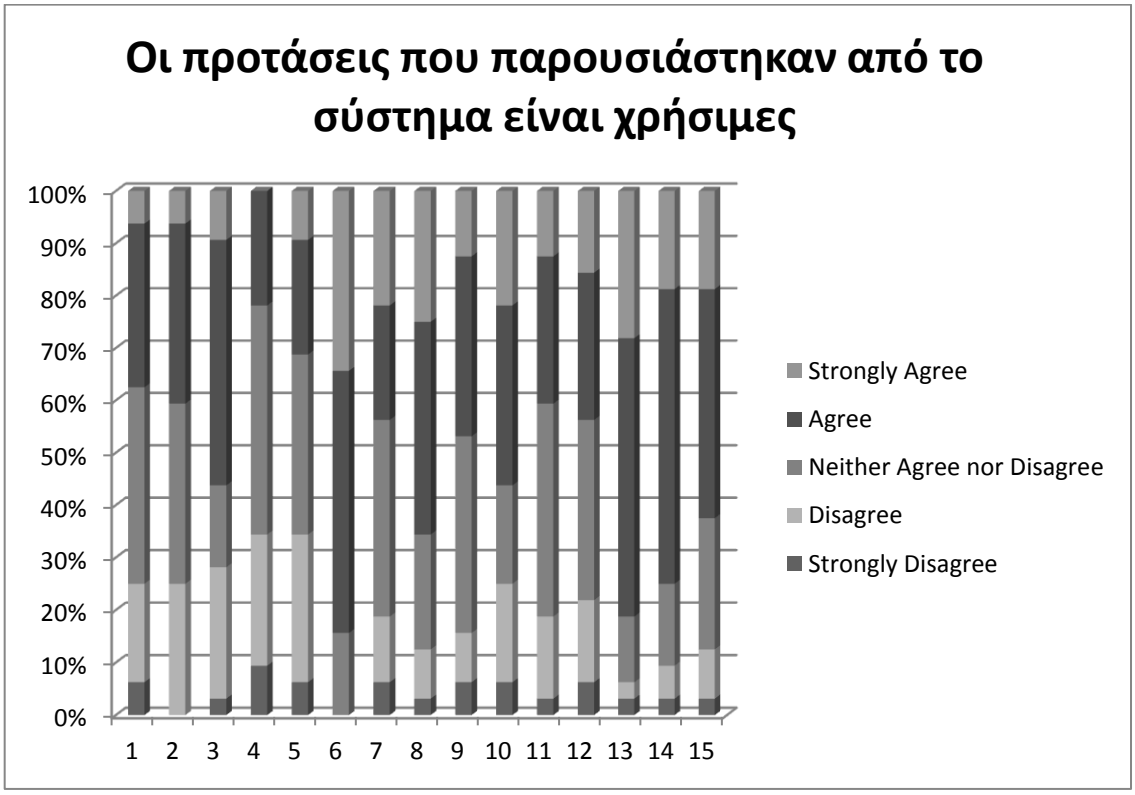
Αριθμός Σεναρίου	Ρόλος Χρήστη	Είδος Προτάσεων	Μέσος Όρος	Τυπική Απόκλιση
1	Πωλητής	Όροι	0,531	0,252
2	Πωλητής	Όροι	0,555	0,227
3	Πωλητής	Όροι	0,586	0,266
4	Πωλητής	Όροι	0,445	0,227
5	Πωλητής	Όροι	0,500	0,269
6	Πωλητής	Αντικείμενα	0,797	0,173
7	Πωλητής	Αντικείμενα	0,602	0,290
8	Πωλητής	Αντικείμενα	0,688	0,262
9	Πωλητής	Αντικείμενα	0,594	0,260
10	Πωλητής	Αντικείμενα	0,617	0,304
11	Αγοραστής	Αντικείμενα	0,578	0,249
12	Αγοραστής	Αντικείμενα	0,578	0,280
13	Αγοραστής	Αντικείμενα	0,750	0,229
14	Αγοραστής	Αντικείμενα	0,703	0,233
15	Αγοραστής	Αντικείμενα	0,664	0,251

Πίνακας 8.3 Αξιολόγηση Σεναρίων από Χρήστες

Στην Εικόνα 8.5 παρουσιάζονται με εποπτικό τρόπο τα αποτελέσματα της αξιολόγησης των χρηστών στα σενάρια χρήσης που τους παρουσιάστηκαν και οι αντιδράσεις στην δήλωση «Οι προτάσεις που παρουσιάστηκαν από το σύστημα είναι χρήσιμες».

Τα αποτελέσματα της αξιολόγησης των χρηστών δείχνουν ότι οι προτάσεις αντικειμένων και όρων που παράχθηκαν από το σύστημα προτάσεων που βασίζεται

σε μοντέλα θεμάτων γενικά αντιμετωπίζονται ως χρήσιμες και πολύ χρήσιμες με χαμηλές τιμές τυπικής απόκλισης.



Εικόνα 8.5 Γράφημα Απεικόνισης Αξιολόγησης Χρηστών με Κλίμακα Likert

8.5 Συμπεράσματα

Σε αυτή την εργασία αντιμετωπίζουμε το πρόβλημα της υποστήριξης δυο διαδικασιών σε μια ηλεκτρονική αγορά και συγκεκριμένα την αναζήτηση αντικειμένων από αγοραστές και την περιγραφή και τιμολόγηση προϊόντων από πωλητές. Προτείνουμε την δημιουργία προτάσεων που αφορούν σχετικά αντικείμενα στον χρήστη που ψάχνει να αγοράσει και την πρόταση σχετικών αντικειμένων και όρων στον πιθανό πωλητή.

Για το σκοπό αυτό έχουμε σχεδιάσει και αναπτύξει ένα σύστημα προτάσεων που βασίζεται στα μοντέλα θεμάτων και υποστηρίζει αυτές τις εργασίες. Έχουμε υλοποιήσει το σύστημα και εφαρμόσει τους αλγορίθμους στα δεδομένα που ανακτήθηκαν από ένα γνωστό ιστότοπο δημοπρασιών. Αξιολογήσαμε τόσο τα

μοντέλα θεμάτων που δημιουργήθηκαν με βάση το σύνολο θεμάτων όσο και τη χρήση του συστήματος προτάσεων. Τα αποτελέσματα της αξιολόγησης οδηγούν στο συμπέρασμα ότι η μεθοδολογία μας μπορεί να οδηγήσει σε ένα σταθερό και χρήσιμο σύστημα προτάσεων σε μια ηλεκτρονική αγορά δημοπρασιών.

Η μελλοντική έρευνα σε αυτή την περιοχή θα μπορούσε να περιλαμβάνει την εφαρμογή επιπρόσθετης αξιολόγησης σε διαφορετικές ηλεκτρονικές αγορές δημοπρασιών. Επίσης θα μπορούσαν να ληφθούν υπόψη δημογραφικές και άλλες διαφορές μεταξύ των χρηστών. Ακόμη, στην πραγματοποίηση των προτάσεων θα μπορούσαν να ληφθούν υπόψη επιπλέον παράγοντες, εκτός του περιεχομένου των προτάσεων. Μια τέτοια προσθήκη θα ήταν το πλαίσιο (context) στο οποίο λαμβάνονται οι αποφάσεις. Για παράδειγμα, λαμβάνοντας υπόψη τον χώρο που βρίσκεται ο χρήστης στον οποίο δίνεται μια πρόταση, το σύστημα προτάσεων θα πρότεινε περισσότερα τοπικά προϊόντα.

9 Συμπεράσματα και Μελλοντική Εργασία

Στο συγκεκριμένο κεφάλαιο παρουσιάζουμε τα συμπεράσματα της παρούσας διατριβής. Επίσης, περιγράφουμε τους περιορισμούς της έρευνας που πραγματοποιήθηκε και πιθανές επεκτάσεις. Κλείνουμε με μια αναφορά σε κατευθύνσεις για μελλοντική έρευνα.

9.1 Συμπεράσματα

Στην παρούσα διατριβή παρουσιάστηκε η ανάπτυξη συστημάτων προτάσεων με χρήση πιθανοτικών μοντέλων θεμάτων.

Με βάση βιβλιογραφική μελέτη στα γνωστικά πεδία της χρήσης τεχνικών μηχανικής μάθησης και της πιθανοτικής λανθάνουσας σημασιολογικής ανάλυσης για την πραγματοποίηση προτάσεων, σχεδιάστηκε και υλοποιήθηκε ένας αριθμός συστημάτων προτάσεων με βάση πιθανοτικά μοντέλα θεμάτων. Τα συστήματα αυτά αφορούν διαφορετικές θεματικές περιοχές: το εταιρικό περιβάλλον με χρήση κοινωνικών τεχνολογιών, το ηλεκτρονικό εμπόριο στον παγκόσμιο ιστό, το εμπόριο και την συνεργατική ανάπτυξη λογισμικού. Για την ανάπτυξη των ανωτέρω συστημάτων έγινε χρήση τεχνικών που λαμβάνουν υπόψη την δραστηριότητα των χρηστών αλλά και το περιεχόμενο των προτιμήσεων.

Η συνεισφορά της διατριβής παρουσιάζει μια αντιστοιχία με τα ερευνητικά ερωτήματα. Παρουσιάστηκε μία προσέγγιση για την ενσωμάτωση της υπάρχουσας γνώσης ενός πεδίου σε ένα σύστημα προτάσεων. Ακόμη, προτάθηκε μια μεθοδολογία που εκμεταλλεύεται την εξαγωγή πιθανοτικών μοντέλων θεμάτων για την πλήρη και αποτελεσματική μοντελοποίηση της ικανότητας ενός εργαζομένου να αντιμετωπίσει ένα πρόβλημα. Περιγράφεται μια μεθοδολογία εξαγωγής προτιμήσεων για καταναλωτές σε υπεραγορές από ένα σύνολο δεδομένων με χρήση λανθανόντων θεμάτων. Τέλος προτείνεται και αξιολογείται μια μεθοδολογία για την εκμετάλλευση του μη δομημένου κειμένου που βρίσκεται σε ηλεκτρονικές αγορές δημοπρασιών για την παραγωγή προτάσεων που απευθύνονται σε αγοραστές και πωλητές.

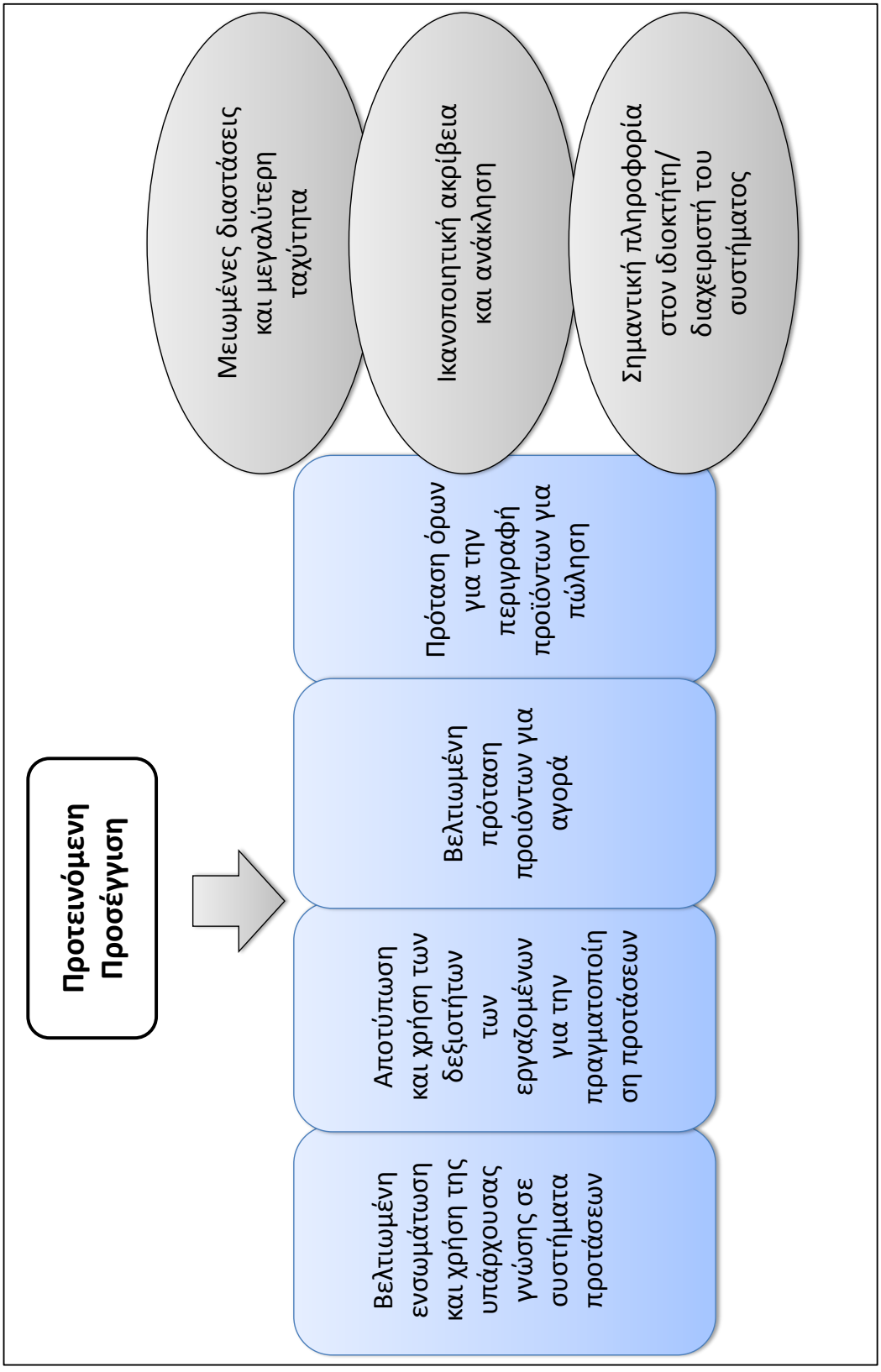
Η πραγματοποίηση της έρευνας που εντάσσεται στην παρούσα διατριβή συνεισφέρει στα παρακάτω:

- Στη βελτιωμένη ενσωμάτωση και χρήση της υπάρχουσας γνώσης σε συστήματα προτάσεων.
- Στην αποτύπωση και χρήση των δεξιοτήτων των χρηστών για την πραγματοποίηση προτάσεων.
- Στην βελτιωμένη πρόταση προϊόντων για αγορά.
- Στην πρόταση όρων για την περιγραφή προϊόντων για πώληση.

Τα συστήματα προτάσεων που προτείνονται παρουσίασαν τα παρακάτω χαρακτηριστικά:

- Μειώνουν τις απαιτούμενες διαστάσεις του προβλήματος και παρέχουν γρήγορα προτάσεις αφού έχει προηγηθεί η εξαγωγή των μοντέλων θεμάτων.
- Ικανοποιούν τις απαιτήσεις των χρηστών για ακρίβεια και ανάκληση όλων των δεδομένων που τους ενδιαφέρουν.
- Μπορούν να προσφέρουν σημαντική πληροφορία για τον ιδιοκτήτη ή τον διαχειριστή του συστήματος.

Μια εποπτική εικόνα των αποτελεσμάτων της διατριβής δίνεται στην Εικόνα 9.1.



Εικόνα 9.1 Αποτελέσματα της Διατριβής

9.2 Περιορισμοί και Πιθανές Επεκτάσεις

Κατά την αντιμετώπιση των ερευνητικών ερωτημάτων που τέθηκαν στην διατριβή ήταν αναγκαία η χρήση μιας σειράς παραδοχών που οδήγησαν σε αντίστοιχους περιορισμούς. Οι περιορισμοί αφορούν στην πληρέστερη αξιολόγηση των προτεινόμενων προσεγγίσεων και στην καλύτερη σύνδεση μεταξύ γνωσιακών δομών και μοντέλων θεμάτων.

Αναφορικά με την αξιολόγηση των προτεινόμενων συστημάτων, η εμπειρική αξιολόγηση που παρουσιάζεται καλύπτει επαρκώς την ανάγκη για επαλήθευση της συνεισφοράς της διατριβής. Παρ' όλα αυτά, υπάρχει η δυνατότητα για πιθανές επεκτάσεις. Πρώτον η αξιολόγηση μπορεί να επεκταθεί σε μεγαλύτερη κλίμακα και να συμπεριλάβει διαφορετικές συνθήκες αποδοχής των δημιουργούμενων προτάσεων. Μεταβλητές που αφορούν στα δημογραφικά των χρηστών (ηλικία, χώρα καταγωγής, πολιτισμικά χαρακτηριστικά, κ.α.) αλλά και στην διαδικασία παραγωγής προτάσεων (χρόνος, τόπος, διάθεση, κ.α.) μπορούν να ληφθούν υπόψη [166]. Δεύτερον, για την αξιολόγηση των συστημάτων μπορεί να χρησιμοποιηθεί έλεγχος A/B (A/B test). Ο συγκεκριμένος τύπος αξιολόγησης μπορεί να επιτρέψει την εξαγωγή περισσότερων συμπερασμάτων από την εκτενή χρήση των προτεινόμενων συστημάτων προτάσεων στο πλαίσιο ενός ολοκληρωμένου συστήματος. Αυτό συμβαίνει καθώς στον συγκεκριμένο τύπο αξιολόγησης οι χρήστες δεν γνωρίζουν ότι αξιολογούν κάποιο σύστημα αλλά η προτίμηση τους αξιολογείται με βάση μετρικές δραστηριότητας [167].

Η δυνατότητα των πιθανοτικών μοντέλων θεμάτων για ερμηνεία από ανθρώπους μπορεί να επιτρέψει την αλληλεπίδραση ενός συστήματος προτάσεων με ένα σύστημα διαχείρισης γνωσιακών δομών. Στα πλαίσια της μεθοδολογίας Entasis, όπως παρουσιάστηκε στο κεφάλαιο 5, προτάθηκαν τα πρώτα βήματα προς τη συγκεκριμένη κατεύθυνση καθώς τα λανθάνοντα θέματα συνδέθηκαν με τις υπάρχουσες γνωσιακές δομές ελαφρού τύπου. Παραμένουν όμως αρκετές προκλήσεις που αφορούν τον συνδυασμό των γνωσιακών δομών και της ρητής γνώσης των χρηστών με την εξαγόμενη γνώση από τα λανθάνοντα θέματα. Μια δυνατότητα αφορά την σταθερή και συνεχή σύνδεση της μη-επιβλεπόμενης εξαγωγής γνώσης με τις αντίστοιχες δομές χωρίς τη χρήση ενός συστήματος προτάσεων (π.χ. με βάση τεχνολογίες εκμάθησης οντολογιών [168]). Ακόμη, υπάρχει η δυνατότητα για βελτιωμένη εκπαίδευση των μοντέλων θεμάτων με βάση τις αλληλεπιδράσεις με τους χρήστες. Τέλος, η εξέλιξη των συνδεδεμένων

δεδομένων (linked open data) και των δημόσια διαθέσιμων οντολογιών μπορεί να υποστηρίξει την διαχείριση γνώσης στο εσωτερικό ομάδων [169] και να συνδυαστεί με λανθάνοντα θέματα [170].

9.3 Μελλοντική Έρευνα

Με βάση τη συνεισφορά της παρούσας διατριβής θεωρούμε ότι δίνεται σημαντική δυνατότητα για μελλοντική έρευνα σε τρεις κατευθύνσεις.

Η πρώτη κατεύθυνση είναι η ενσωμάτωση γνώσεων διαφορετικού τύπου, όπως πλαισίου (context) και κοινωνικών δικτύων για την βελτίωση των προτάσεων. Στην προτεινόμενη διατριβή παρουσιάστηκαν συστήματα που εντάσσουν την δραστηριότητα των χρηστών και το περιεχόμενο που παράγουν σε ένα πιθανοτικό πλαίσιο. Στην πράξη όμως, δεν είναι μόνο αυτοί οι δυο παράγοντες που επηρεάζουν τις αποφάσεις των χρηστών. Διαφορετικοί παράγοντες μπορούν να ενταχθούν σε αντίστοιχες μεθοδολογίες, όπως το πλαίσιο (τα συμφραζόμενα) στο οποίο λαμβάνονται οι αποφάσεις αλλά και η ύπαρξη κοινωνικών δεσμών. Για παράδειγμα, λαμβάνοντας υπόψη το χρονικό πλαίσιο στο οποίο δίνεται μια πρόταση, ένα σύστημα προτάσεων ταξιδιών μπορεί να παρέχει διαφορετικές προτάσεις το χειμώνα από ότι το καλοκαίρι [25]. Ακόμη, τα συστήματα προτάσεων μπορούν να ανακαλύψουν υπάρχοντες δεσμούς μεταξύ των ανθρώπων και να δημιουργήσουν προτάσεις με βάση το κοινωνικό δίκτυο του χρήστη. Τέτοιες προτάσεις μπορεί να βασίζονται σε φιλικές σχέσεις ή σχέσεις εμπιστοσύνης [137].

Η δεύτερη κατεύθυνση αφορά την ενσωμάτωση ποικιλίας (diversity) και εύνοιας τυχαίων ανακαλύψεων (serendipity) στην προτεινόμενη προσέγγιση. Τα στοιχεία αυτά αποτελούν στόχους της ερευνητικής κοινότητας των συστημάτων προτάσεων [171]. Η ποικιλία αφορά το βαθμό που οι προτάσεις καλύπτουν με σχετική πληρότητα τις διαφοροποιήσεις μεταξύ των πιθανών αντικειμένων, ενώ η εύνοια τυχαίων ανακαλύψεων αφορά το βαθμό που τα προτεινόμενα αντικείμενα θα είναι απρόβλεπτα από τον χρήστη και θα τον εκπλήξουν θετικά. Σημαντικές μελέτες επιχειρούν να μετρήσουν τη χρησιμότητα του απροσδόκητου περιεχομένου για την εύνοια τυχαίων ανακαλύψεων [172]. Η ποικιλία και το απροσδόκητο περιεχόμενο που εμφανίζεται στα συστήματα προτάσεων μπορούν να ευνοήσουν τυχαίες ανακαλύψεις ενδιαφερόντων στοιχείων που ο χρήστης δεν γνωρίζει ή δεν περιμένει να εμφανιστούν. Οι τυχαίες ανακαλύψεις που μπορεί να προέλθουν από ένα σύστημα προτάσεων μπορούν με τη σειρά τους να ευνοήσουν την

πραγματοποίηση δημιουργικών σκέψεων [173]. Κάποιες μελέτες χρησιμοποιούν τα πιθανοτικά μοντέλα θεμάτων για την υποστήριξη ποικιλίας στις προτάσεις που παρέχονται από τα συστήματα προτάσεων [174]

Τέλος, η τρίτη κατεύθυνση αφορά την εκμετάλλευση των συστημάτων προτάσεων για την υποστήριξη αλλαγών στον τρόπο ζωής των ανθρώπων. Στη σημερινή κοινωνία, και ιδιαίτερα σε ανεπτυγμένες χώρες, ο τρόπος ζωής επηρεάζεται από την ύπαρξη της τεχνολογίας και την αφθονία των υλικών πόρων. Η συγκεκριμένη κατάσταση οδηγεί σε αποφάσεις που δεν λαμβάνουν υπόψη τα αρνητικά μακροπρόθεσμα αποτελέσματα στο περιβάλλον και στην ανθρώπινη υγεία [153]. Κατά συνέπεια, σε ατομικό επίπεδο παρατηρείται επιδείνωση των ασθενειών που σχετίζονται με τον τρόπο ζωής (διαβήτης, παχυσαρκία, κ.α.) ενώ σε επίπεδο κοινωνίας οι ατομικές επιλογές βλάπτουν το περιβάλλον και οδηγούν σε κλιματική αλλαγή και εξάντληση των φυσικών πόρων. Τα συστήματα που φιλοδοξούν να αλλάξουν τον τρόπο ζωής των ανθρώπων έχουν σαν στόχο να συνθέτουν και να προτείνουν ακολουθίες δραστηριοτήτων. Επίσης οφείλουν να διατηρούν την προσοχή του χρήστη, να εξηγούν τις προτάσεις που παρέχουν και να παρέχουν ενημέρωση για τις θετικές αλλαγές που εκείνος καταφέρνει [175]. Οι τεχνολογίες που στοχεύουν στην αλλαγή του τρόπου ζωής επιχειρούν να ωθήσουν τους χρήστες προς αποφάσεις που ωφελούν τα προσωπικά τους μακροπρόθεσμα συμφέροντα.

Ευρετήριο Όρων

Α/Β Αξιολόγηση: A/B Test
Ακρίβεια: Precision
Άλγεβρα Boole: Boolean Algebra
Αλυσίδες Markov Monte Carlo: Monte Carlo Markov Chains
Ανάκληση: Recall
Ανάκτηση Πληροφορίας: Information Retrieval
Ανάλυση Δεσμών: Link Analysis
Ανάλυση Ιδιοτιμών: Singular Value Decomposition
Ανάλυση Καλαθιού Αγοράς: Basket Market Analysis
Ανάλυση Καλαθιού Αγορών: Market Basket Analysis
Ανάλυση Κοινωνικών Δικτύων: Social Network Analysis
Ανάλυση Κύριων Συνιστωσών: Principal Component Analysis
Ανοικτά Συνδεδεμένα Δεδομένα: Linked Open Data
Αριστερά προς τα Δεξιά Αξιολόγηση: Left to Right Evaluation
Αρχική Φάση Δειγματοληψίας: Burn-in Phase
Αφελής Ταξινομητής Bayes: Naïve Bayes Classifier
Βαθμός Ανάλυσης: Granularity
Βεβαιότητα: Confidence
Γενετικό Μοντέλο: Generative Model
Γενετικός Αλγόριθμος: Genetic Algorithm
Γκαουσιανή Γραμμική Αναδρομή: Gaussian Linear Regression
Γκαουσιανός: Gaussian
Γνωσιακή Δομή: Knowledge Structure
Δειγματοληψία Gibbs: Gibbs Sampling
Δένδρο Αποφάσεων: Decision Tree
Διαδικασία Chinese Restaurant: Chinese Restaurant Process, CRP
Διαδικασίες Απόφασης Markov: Markov Decision Processes
Διαμέριση (ακεραίων): Partition
Διασταυρωμένη Επικύρωση: Cross Validation
Διαχωριστικό Μοντέλο: Discriminative Model
Διεπαφή Χρήστη: User Interface
Δυαδική Λογιστική Παλινδρόμηση: Binary Logistic Regression
Εκ των Υστέρων Διήθηση: Post-filtering
Εκ των Υστέρων Πιθανοτική Κατανομή: Posterior Distribution
Ελεύθερο Λογισμικό / Λογισμικό Ανοιχτού Κώδικα: Free / Open Source Software

Εμπιστοσύνη (για κανόνα συσχέτισης): Confidence
Ενιαίο Αναγνωριστικό Πόρου: Uniform Resource Identifier
Εξάπλωση Ενεργοποίησης: Spreading Activation
Εξατομίκευση: Personalization
Εξόρυξη Δεδομένων: Data Mining
Επαγωγή: Inference
Επαυξημένη Πραγματικότητα: Augmented Reality
Επέκταση Ερωτημάτων: Query Expansion
Επισημείωση: Tag / Annotation (αλλιώς ***Ετικέτα***)
Επιχείρηση 2.0: Enterprise 2.0
Επιχειρησιακή Έρευνα: Operational Research
Εύρεση Λήμματος: Stemming
Ευρετηρίαση: Indexing
Ευρετική Μέθοδος: Heuristic Method
Θεωρία Γράφων: Graph Theory
Ιστός 2.0: Web 2.0
Ιστός Δεδομένων: Web of Data
Ιστότοπος Συζητήσεων: Discussion Forum
Κανόνες Συσχέτισης: Association Rules
Κίνηση Brown: Brownian Motion
Κ-Κορυφαίες Προβλέψεις: Top K Predictions
Κλάδεμα (δεδομένων) : Pruning
Κοινωνικό Σημασιολογικό Νέφος Επισημειώσεων: Social Semantic Tag Cloud
Κόστος Συναλλαγών: Transaction Cost
Κ-Πλησιέστεροι Γείτονες: K-Nearest Neighbors
Λανθάνουσα Κατανομή Dirichlet: Latent Dirichlet Allocation
Λανθάνουσα Σημασιολογική Ανάλυση: Latent Semantic Analysis
Λεξιλόγιο: Vocabulary
Λογαριθμική Πιθανότητα: Log Likelihood
Λογικό Μοντέλο: Boolean Model (Information Retrieval)
Λογιστική Κανονική Κατανομή: Logistic Normal Distribution
Μαζικά Παιχνίδια Ρόλων μέσω Δικτύου με Πολλούς Παίκτες: Massive Multi-Player Online Role Playing Games
Μακρά Ουρά (φαινόμενο): Long-tail Effect
Με Βάση Περιπτώσεις: Case Based
Μεγιστοποίηση Αναμονής: Expectation Maximization
Μέθοδοι Εξαγωγής Ωφελείας: Utility Elicitation Methods
Μέθοδοι με Βάση Μοντέλα: Model-based Methods
Μέθοδοι με Βάση τη Μνήμη: Memory-based Methods

Μέθοδος Προσέγγισης Μέσου Πεδίου: Mean Field Variational Method
Μείωση Διαστάσεων: Dimensionality Reduction
Μέση Αμοιβαία Κατάταξη: Mean Reciprocal Rank
Μεταβολική Μέθοδος: Variational Method
Μηχανή Διανυσμάτων Υποστήριξης: Support Vector Machine
Μηχανική Γνώσης: Knowledge Engineering
Μηχανική Μάθηση: Machine Learning
Μοντέλο Διανυσματικού Χώρου: Vector Space Model
Μοντέλο Όψεων: Aspect Model
Μοντέλο Χώρου Κατάστασης: State Space Model
Μπεϋζιανά Δίκτυα: Bayesian Networks
Μπεϋζιανός Ταξινομητής: Bayesian Classifier
Νευρωνικό Δίκτυο: Neural Network
Νέφος Επισημειώσεων: Tag Cloud
Οικονομική Ψυχολογία: Behavioral Economics
Ομαδοποίηση Κ-Μέσων (αλγόριθμος): K-means Clustering
Ομαδοποίηση: Clustering
Ομοιότητα: Similarity
Οντολογία: Ontology
Περιηγητής: Browser
Πηγαίος Κώδικας: Source Code
Πιθανότητα Εμφάνισης Κρατημένων Εγγράφων: Held-out Probability
Πιθανοτικά Μοντέλα Θεμάτων: Probabilistic Topic Models
Πλαίσιο: Context
Πλήρως Ταξινομημένο Σύνολο: Totally Ordered Set
Πολλαπλός Πολλαπλασιαστικός Παράγοντας: Multiple Multiplicative Factor
Πράκτορας (Λογισμικού): (Software) Agent
Προσέγγιση από Πάνω προς τα Κάτω: Top-Down Approach
Προτάσεις με Βάση το Περιεχόμενο: Content-based Recommendation
Σημείο Ενδιαφέροντος: Place-Of-Interest
Στατιστική Επαγωγή: Statistical Inference
Στοιχείο Πληροφορίας: Information Resource
Στοιχειοσύνολο: Itemset
Συζυγής: Conjugate
Συλλογική Νοημοσύνη: Collective Intelligence
Σύμβουλος Πράκτορας: Counselor Agent
Συνάρτηση Πιθανοφάνειας: Likelihood Function
Συνάρτηση Πυκνότητας Πιθανότητας: Probability Density Function
Συνάφεια: Similarity

Συνεμφάνιση: Co-occurence
Συνεργατική Διήθηση με Επιρροές Κλίκας: Clique-effects Collaborative Filtering
Συνεργατική Διήθηση: Collaborative Filtering
Συνήθεις Λέξεις: Stopwords
Σύνολο Λέξεων (προσέγγιση): Bag of Words
Σύστημα Διαχείρισης Πηγαίου Κώδικα: Source Code Management System
Σύστημα Καταγραφής Προβλημάτων: Issue Tracking System
Σύστημα Προτάσεων: Recommender System (αλλιώς ***Σύστημα Συστάσεων***)
Ταξονομία: Taxonomy
Υλικό: Hardware
Υπερκείμενο: Hypertext
Υπερπροσαρμογή: Overfitting
Υπερφόρτωση Πληροφορίας: Information Overload
Υπηρεσίες Νέφους: Cloud Services
Υποστήριξη (για κανόνα συσχέτισης): Support
Φολκσονομία: Folksonomy
Χρονοσφραγίδα: Timestamp
Ψυχρή Έναρξη: Cold Start
Ωφέλεια: Utility

Σημείωση: Η ορολογία που χρησιμοποιήθηκε μεταφράστηκε στα ελληνικά με εξαιρέσεις όπου η αγγλική ορολογία έχει ενταχθεί στην ελληνική γλώσσα (για παράδειγμα με τη χρήση της λέξης «online»). Βάση για την αντιστοίχιση με ελληνικούς όρους αποτέλεσε η υπάρχουσα βιβλιογραφία στα αντίστοιχα πεδία, βιβλία και ερευνητικές εργασίες, καθώς και η βάση όρων της Ελληνικής Εταιρείας Ορολογίας (ΕΛΕΤΟ)[31].

[31] http://www.eleto.gr/

Δημοσιεύσεις και Ανακοινώσεις

Στα πλαίσια της παρούσας εργασίας έχουν πραγματοποιηθεί οι παρακάτω δημοσιεύσεις και ανακοινώσεις.

Δημοσιεύσεις σε Περιοδικά

[J1] K. Christidis and G.Mentzas, "A Topic-based Recommender System for Electronic Marketplace Platforms", Expert Systems with Applications, in print, 2013, Impact Factor 1.924

[J2] K. Christidis, G. Mentzas, and D. Apostolou, "Using latent topics to enhance search and recommendation in Enterprise Social Software," Expert Systems with Applications, vol. 39, no. 10, pp. 9297–9307, Aug. 2012, Impact Factor 1.924

[J3] K. Christidis, G. Mentzas, and D. Apostolou, "Supercharging Enterprise 2.0," IEEE IT Professional, vol. 13, no. 4, pp. 29–35, 2011.

[J4] K. Christidis, N. Papailiou, D. Apostolou, and G. Mentzas, "Semantic Interfaces for Personal and Social Knowledge Work," International Journal of Knowledge-Based Organizations, vol. 1, no. 1, pp. 61–77, 2011.

Ανακοινώσεις σε Συνέδρια

[C1] K. Christidis and G. Mentzas. "A Topic-based Recommender System for Electronic Marketplace Platforms", In Proceedings of the 24st IEEE International Conference on Tools with Artificial Intelligence (ICTAI '12), IEEE, Athens, Greece

[C2] K. Christidis, G. Mentzas, and D. Apostolou, "A Socially Intelligent Approach for Enterprise Information Search and Recommendation," In Proceedings of 18th International ICE-Conference on Engineering, Technology and Innovation, 18 - 20 June 2012, Munich

[C3] K. Christidis, F. Paraskevopoulos, D. Panagiotou and G. Mentzas. "Combining Activity Metrics and Contribution Topics for Software Recommendations". In Proceedings of the 3rd International Workshop on Recommendation Systems for Software Engineering (RSSE '12). ACM, Zurich, Switzerland

[C4] E. Bothos, K. Christidis, D. Apostolou, and G. Mentzas, "Information market based recommender systems fusion," in Proceedings of the 2nd International

Workshop on Information Heterogeneity and Fusion in Recommender Systems, New York, NY, USA, 2011, pp. 1–8.

[C5] K. Christidis, D. Apostolou, and G. Mentzas, “Exploring Customer Preferences with Probabilistic Topics Models.” In Proceedings of Preference Learning workshop, European Conference of Machine Learning 2010, Barcelona, Spain

[C6] K. Christidis and G. Mentzas, “Using Probabilistic Topic Models in Enterprise Social Software,” in Business Information Systems, 2010, pp. 23–34.

[C7] K. Christidis, N. Papailiou, G. Mentzas, and D. Apostolou, “Exploring Gadget-Based Interfaces for the Social Semantic Desktop,” in 2009 13th Panhellenic Conference on Informatics, 2009, pp. 215–219.

[C8] N. Papailiou, K. Christidis, D. Apostolou, G. Mentzas, R. Gudjonsdottir, ”Personal and Group Knowledge Management with the Social Semantic Desktop”, in O. Cunnigham and M. Cunnigham (eds) Collaboration and the Knowledge Economy: issues, Applications and Case Studies, pp. 1475-1482, eChallenges e-2008 Conference, 22 - 24 October 2008, Stockholm, Sweden

Βιβλιογραφία

[1] M. D. Ekstrand, “Collaborative Filtering Recommender Systems,” *Foundations and Trends® in Human–Computer Interaction*, vol. 4, no. 2, pp. 81–173, 2010.

[2] B. Schwartz, *The paradox of choice: Why more is less*. Harper Perennial, 2005.

[3] H.-J. Happel and W. Maalej, “Potentials and challenges of recommendation systems for software development,” in *Proceedings of the 2008 international workshop on Recommendation systems for software engineering*, New York, NY, USA, 2008, pp. 11–15.

[4] E. Rich, “User modeling via stereotypes,” *Cognitive Science*, vol. 3, no. 4, pp. 329–354, Oct. 1979.

[5] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry, “Using collaborative filtering to weave an information tapestry,” *Commun. ACM*, vol. 35, no. 12, pp. 61–70, Dec. 1992.

[6] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl, “GroupLens: an open architecture for collaborative filtering of netnews,” in *Proceedings of the 1994 ACM conference on Computer supported cooperative work*, New York, NY, USA, 1994, pp. 175–186.

[7] U. Shardanand and P. Maes, “Social information filtering: algorithms for automating “word of mouth”,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 1995, pp. 210–217.

[8] W. Hill, L. Stead, M. Rosenstein, and G. Furnas, “Recommending and evaluating choices in a virtual community of use,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 1995, pp. 194–201.

[9] K. Goldberg, T. Roeder, D. Gupta, and C. Perkins, “Eigentaste: A Constant Time Collaborative Filtering Algorithm,” *Information Retrieval*, vol. 4, no. 2, pp. 133–151, 2001.

[10] P. M. West, D. Ariely, S. Bellman, E. Bradlow, J. Huber, E. Johnson, B. Kahn, J. Little, and D. Schkade, “Agents to the Rescue?,” *Marketing Letters*, vol. 10, no. 3, pp. 285–300, 1999.

[11] A. Ansari, S. Essegaier, and R. Kohli, “Internet Recommendation Systems,” *Journal of Marketing Research*, vol. 37, no. 3, pp. 363–375, Aug. 2000.

[12] G. Linden, B. Smith, and J. York, “Amazon.com recommendations: item-to-item collaborative filtering,” *Internet Computing, IEEE*, vol. 7, no. 1, pp. 76–80, 2003.

[13] R. M. Bell, Y. Koren, and C. Volinsky, “The BellKor solution to the Netflix prize,” *KorBell Team’s Report to Netflix*, 2007.

[14] G. Adomavicius and A. Tuzhilin, “Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions,” *IEEE transactions on knowledge and data engineering*, vol. 17, no. 6, pp. 734–749, 2005.

[15] N. J. Belkin and W. B. Croft, “Information filtering and information retrieval: two sides of the same coin?,” *Commun. ACM*, vol. 35, no. 12, pp. 29–38, Dec. 1992.

[16] C. D. Manning, P. Raghavan, and H. Sch\ütze, “Introduction to information retrieval,” 2008.

[17] R. Baeza-Yates, B. Ribeiro-Neto, and others, *Modern information retrieval*, vol. 463. 1999.

[18] T. Berners-Lee and R. Cailliau, “WorldWideWeb: Proposal for a HyperText Project.” 1990.

[19] T. OReilly, “What is Web 2.0: Design patterns and business models for the next generation of software.”

[20] W. Kim, O.-R. Jeong, and S.-W. Lee, “On social Web sites,” *Information Systems*, vol. 35, no. 2, pp. 215–236, Apr. 2010.

[21] J. Porter, *Designing for the social web*, First. Thousand Oaks, CA, USA: New Riders Publishing, 2008.

[22] C. Anderson, *The long tail*. Business Books, 2004.

[23] I. Guy and D. Carmel, “Social recommender systems,” in *Proceedings of the 20th international conference companion on World wide web*, New York, NY, USA, 2011, pp. 283–284.

[24] A. I. Schein, A. Popescul, L. H. Ungar, and D. M. Pennock, “Methods and metrics for cold-start recommendations,” in *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, 2002, pp. 253–260.

[25] G. Adomavicius and A. Tuzhilin, “Context-aware recommender systems,” *Recommender Systems Handbook*, pp. 217–253, 2011.

[26] S. Abbar, “Context-Aware Recommender Systems : A Service-Oriented Approach,” *PRism*, pp. 1–6.

[27] K. Church, B. Smyth, P. Cotter, and K. Bradley, “Mobile information access: A study of emerging search behavior on the mobile Internet,” *ACM Trans. Web*, vol. 1, no. 1, May 2007.

[28] M.-H. Park, J.-H. Hong, and S.-B. Cho, “Location-Based Recommendation System Using Bayesian User’s Preference Model in Mobile Devices,” in *Ubiquitous Intelligence and Computing*, vol. 4611, J. Indulska, J. Ma, L. Yang, T. Ungerer, and J. Cao, Eds. Springer Berlin / Heidelberg, 2007, pp. 1130–1139.

[29] Federica Cena, Luca Console, Cristina Gena, Anna Goy, Guido Levi, Sonia Modeo, and I. Torre, “An adaptive tourist guide in mobile context *,” CiteSeerX, 2008.

[30] I. Cantador, P. Brusilovsky, and T. Kuflik, “Second workshop on information heterogeneity and fusion in recommender systems (HetRec2011),” in *Proceedings of the fifth ACM conference on Recommender systems*, New York, NY, USA, 2011, pp. 387–388.

[31] M. J. Pazzani and D. Billsus, “Content-Based Recommendation Systems,” in *The Adaptive Web*, vol. 4321, P. Brusilovsky, A. Kobsa, and W. Nejdl, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 325–341.

[32] M. F. Porter, “An algorithm for suffix stripping,” *Program: electronic library and information systems*, vol. 14, no. 3, pp. 130–137, Dec. 1980.

[33] G. Salton, A. Wong, and C. S. Yang, “A vector space model for automatic indexing,” *Communications of the ACM*, vol. 18, no. 11, p. 620, 1975.

[34] W. Cohen, “Learning Rules that Classify E-Mail,” in *In Papers from the AAAI Spring Symposium on Machine Learning in Information Access*, pp. 18–25.

[35] D. Billsus, M. J. Pazzani, and J. Chen, “A learning agent for wireless news access,” in *Proceedings of the 5th international conference on Intelligent user interfaces*, New York, NY, USA, 2000, pp. 33–36.

[36] M. Pazzani and D. Billsus, “Learning and Revising User Profiles: The Identification of Interesting Web Sites,” *Machine Learning*, vol. 27, no. 3, pp. 313–331, 1997.

[37] Z. Huang, D. Zeng, and H. Chen, “A Comparison of Collaborative-Filtering Recommendation Algorithms for E-commerce,” *IEEE Intelligent Systems*, vol. 22, no. 5, pp. 68–78, 2007.

[38] B. M. Sarwar, G. Karypis, J. Konstan, and J. Riedl, “Recommender systems for large-scale e-commerce: Scalable neighborhood formation using clustering,” in *Proceedings of the Fifth International Conference on Computer and Information Technology*, 2002, pp. 158–167.

[39] R. Burke, “Knowledge-based recommender systems,” *Encyclopedia of Library and Information Systems*, vol. 69, no. Supplement 32, pp. 175–186, 2000.

[40] R. Burke, “Hybrid recommender systems: Survey and experiments,” *User Modeling and User-Adapted Interaction*, vol. 12, no. 4, pp. 331–370, 2002.

[41] G. Shani and A. Gunawardana, “Evaluating Recommender Systems,” *Microsoft Research Report MSR-TR-2009-159, Nov*, 2009.

[42] G. Salton and M. E. Lesk, “The SMART automatic document retrieval systems—an illustration,” *Commun. ACM*, vol. 8, no. 6, pp. 391–398, Jun. 1965.

[43] M. Uschold and M. Gruninger, “Ontologies and semantics for seamless connectivity,” *SIGMOD Rec.*, vol. 33, no. 4, pp. 58–64, 2004.

[44] T. Berners-Lee and J. Hendler, “Scientific publishing on the semantic web,” *Nature*, vol. 410, pp. 1023–1024, 2001.

[45] C. Bizer, T. Heath, and T. Berners-Lee, “Linked Data - The Story So Far,” *International Journal on Semantic Web and Information Systems*, vol. 5, no. 3, pp. 1–22, 33 2009.

[46] G. Klyne and J. J. Carroll, “Resource Description Framework (RDF): Concepts and Abstract Syntax,” vol. 10, no. February, pp. 1–20, 2004.

[47] T. Berners-Lee, Y. Chen, L. Chilton, D. Connolly, R. Dhanaraj, J. Hollenbach, A. Lerer, and D. Sheets, “Tabulator: Exploring and analyzing linked data on the semantic web,” in *Proceedings of the 3rd International Semantic Web User Interaction Workshop*, 2006, vol. 2006.

[48] A. Halevy, P. Norvig, and F. Pereira, “The unreasonable effectiveness of data,” *Intelligent Systems, IEEE*, vol. 24, no. 2, pp. 8–12, 2009.

[49] S. T. Dumais, “Latent Semantic Analysis.,” *Annual Review of Information Science and Technology (ARIST)*, vol. 38, pp. 189–230, 2004.

[50] P. Wiemer-Hastings, K. Wiemer-Hastings, and A. Graesser, “Latent semantic analysis,” in *Proceedings of the 16th international joint conference on Artificial intelligence*, 2004, pp. 1–14.

[51] T. Hofmann, “Probabilistic latent semantic indexing,” in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, Berkeley, California, United States, 1999, pp. 50–57.

[52] A. Y. Ng and M. I. Jordan, “On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes,” in *Neural Information Processing Systems*, 2002, vol. 2, pp. 841–848.

[53] M. Steyvers and T. Griffiths, “Probabilistic topic models,” *Handbook of Latent Semantic Analysis*, pp. 424–440, 2007.

[54] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” *The Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[55] T. L. Griffiths and M. Steyvers, “Finding scientific topics.” National Academy of Sciences, 2004.

[56] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning, “Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora,” in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, 2009, pp. 248–256.

[57] D. M. B. J. . Lafferty, “Correlated Topic Models,” in *Advances in Neural Information Processing Systems 18: Proceedings of the 2005 Conference*, 2006, p. 147.

[58] J. Aitchison, *The statistical analysis of compositional data*. London, UK, UK: Chapman & Hall, Ltd., 1986.

[59] D. M. Blei and J. D. Lafferty, “Dynamic topic models,” in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 113–120.

[60] C. Wang, “Continuous Time Dynamic Topic Models,” 2009.

[61] X. Wang and A. McCallum, “Topics over time: a non-Markov continuous-time model of topical trends,” in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA, 2006, pp. 424–433.

[62] X. Wei, J. Sun, and X. Wang, “Dynamic mixture models for multiple time series,” in *Proceedings of the 20th international joint conference on Artifical intelligence*, San Francisco, CA, USA, 2007, pp. 2909–2914.

[63] D. Harman, “Overview of the first TREC conference,” in *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, 1993, pp. 36–47.

[64] D. M. Blei, T. L. Griffiths, M. I. Jordan, and J. B. Tenenbaum, “Hierarchical Topic Models and the Nested Chinese Restaurant Process,” 2004.

[65] R. Schirru, S. Baumann, M. Memmel, and A. Dengel, “Topic-Based Recommendations for Enterprise 2.0 Resource Sharing Platforms,” in *Knowledge-Based and Intelligent Information and Engineering Systems*, vol. 6881, A. König, A. Dengel, K. Hinkelmann, K. Kise, R. J. Howlett, and L. C. Jain, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 495–504.

[66] X. Xie, W. Zhang, Y. Yang, and Q. Wang, "DRETOM: developer recommendation based on topic models for bug resolution," in *Proceedings of the 8th International Conference on Predictive Models in Software Engineering*, New York, NY, USA, 2012, pp. 19–28.

[67] C.-Y. Tsai and S.-H. Chung, "A personalized route recommendation service for theme parks using RFID information and tourist behavior," *Decision Support Systems*.

[68] W. Y. Chen, J. Luan, H. Bai, Y. Wang, and E. Y. Chang, "Collaborative filtering for orkut communities: discovery of user latent behavior," in *Proceedings of the 18th international conference on World wide web*, 2009, pp. 681–690.

[69] A. S. Das, M. Datar, A. Garg, and S. Rajaram, "Google news personalization: scalable online collaborative filtering," in *Proceedings of the 16th international conference on World Wide Web*, New York, NY, USA, 2007, pp. 271–280.

[70] T. Kurashima, T. Iwata, G. Irie, and K. Fujimura, "Travel route recommendation using geotags in photo sharing sites," in *Proceedings of the 19th ACM international conference on Information and knowledge management*, New York, NY, USA, 2010, pp. 579–588.

[71] B. Fields, C. Rhodes, and M. Inverno, "Using Song Social Tags and Topic Models to Describe and Compare Playlists," *Computing*, 2010.

[72] C. Haruechaiyasak and C. Damrongrat, "Article Recommendation Based on a Topic Model for Wikipedia Selection for Schools," *Digital Libraries: Universal and Ubiquitous Access to Information*, pp. 339–342, 2008.

[73] R. Krestel and P. Fankhauser, "Personalized topic-based tag recommendation," *Neurocomputing*, vol. In Press, Corrected Proof.

[74] E. Diaz-Aviles, M. Georgescu, A. Stewart, and W. Nejdl, "LDA for on-the-fly auto tagging," in *Proceedings of the fourth ACM conference on Recommender systems*, 2010, pp. 309–312.

[75] Z. Zhao, S. Feng, Q. Wang, J. Z. Huang, G. J. Williams, and J. Fan, "Topic oriented community detection through social objects and link analysis in social networks," *Knowledge-Based Systems*, vol. In Press, Uncorrected Proof.

[76] T. Hofmann, "Latent semantic models for collaborative filtering," *ACM Trans. Inf. Syst.*, vol. 22, no. 1, pp. 89–115, Jan. 2004.

[77] J. Wang, A. P. de Vries, and M. J. T. Reinders, "Unifying user-based and item-based collaborative filtering approaches by similarity fusion," in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, 2006, pp. 501–508.

[78] Timothy N. Rubin and M. Steyvers, "A Topic Model For Movie Choices and Ratings," CiteSeerX, 2009.

[79] A. Popescul, L. H. Ungar, D. M. Pennock, and S. Lawrence, "Probabilistic models for unified collaborative and content-based recommendation in sparse-data environments," in *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, 2001, pp. 437–444.

[80] D. Agarwal and B.-C. Chen, "fLDA: matrix factorization through latent dirichlet allocation," in *Proceedings of the third ACM international conference on Web search and data mining*, New York, NY, USA, 2010, pp. 91–100.

[81] C. Wang and D. M. Blei, "Collaborative topic modeling for recommending scientific articles," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA, 2011, pp. 448–456.

[82] J. Bughin and M. Chui, "The rise of the networked enterprise: Web 2.0 finds its payday," *McKinsey Quarterly*, vol. 4, pp. 3–8, 2010.

[83] F. A. Hayek, "The Use of Knowledge in Society," *SSRN eLibrary*, 1945.

[84] A. P. McAfee, "Enterprise 2.0: The dawn of emergent collaboration," *MIT Sloan Management Review*, vol. 47, no. 3, p. 21, 2006.

[85] I. Guy, N. Zwerdling, I. Ronen, D. Carmel, and E. Uziel, "Social media recommendation based on people and tags," in *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, 2010, pp. 194–201.

[86] P. A. Dmitriev, N. Eiron, M. Fontoura, and E. Shekita, "Using annotations in enterprise search," in *Proceedings of the 15th international conference on World Wide Web*, Edinburgh, Scotland, 2006, pp. 811–817.

[87] A. Passant, P. Laublet, J. G. Breslin, and S. Decker, "SemSLATES: Improving enterprise 2.0 information systems using semantic Web technologies," in *5th International Conference on Collaborative Computing: Networking, Applications and Worksharing, 2009. CollaborateCom 2009*, 2009, pp. 1–10.

[88] A. Mockus and J. D. Herbsleb, "Expertise browser: a quantitative approach to identifying expertise," in *Proceedings of the 24th International Conference on Software Engineering*, 2002, pp. 503–512.

[89] H. Kagdi, M. Hammad, and J. I. Maletic, "Who can help me with this source code change?," in *Software Maintenance, 2008. ICSM 2008. IEEE International Conference on*, 2008, pp. 157–166.

[90] T. Fritz, G. C. Murphy, and E. Hill, "Does a programmer's activity indicate knowledge of code?," in *Proceedings of the the 6th joint meeting of the European software engineering conference and the ACM SIGSOFT symposium on The foundations of software engineering*, 2007, pp. 341–350.

[91] J. Geldenhuys, "Finding the Core Developers," in *2010 36th EUROMICRO Conference on Software Engineering and Advanced Applications (SEAA)*, 2010, pp. 447–450.

[92] D. Matter, A. Kuhn, and O. Nierstrasz, "Assigning bug reports using a vocabulary-based expertise model of developers," in *Mining Software Repositories, 2009. MSR'09. 6th IEEE International Working Conference on*, 2009, pp. 131–140.

[93] W. Maalej and A. Sahm, "Assisting engineers in switching artifacts by using task semantic and interaction history," in *Proceedings of the 2nd International Workshop on Recommendation Systems for Software Engineering*, New York, NY, USA, 2010, pp. 59–63.

[94] M. Gethers, T. Savage, M. Di Penta, R. Oliveto, D. Poshyvanyk, and A. De Lucia, "CodeTopics: which topic am I coding now?," in *Software Engineering (ICSE), 2011 33rd International Conference on*, 2011, pp. 1034 –1036.

[95] D. Pagano and W. Maalej, “How do developers blog?: an exploratory study,” in *Proceedings of the 8th Working Conference on Mining Software Repositories*, New York, NY, USA, 2011, pp. 123–132.

[96] A. Hindle, N. A. Ernst, M. W. Godfrey, and J. Mylopoulos, “Automated topic naming to support cross-project analysis of software maintenance activities,” in *Proceedings of the 8th Working Conference on Mining Software Repositories*, New York, NY, USA, 2011, pp. 163–172.

[97] E. Linstead, P. Rigor, S. Bajracharya, C. Lopes, and P. Baldi, “Mining Eclipse Developer Contributions via Author-Topic Models,” in *Mining Software Repositories, 2007. ICSE Workshops MSR ’07. Fourth International Workshop on*, 2007, p. 30.

[98] Y. Ohsawa and K. Yada, *Data Mining for Design and Marketing*. CRC Press, 2009.

[99] J. Fürnkranz and E. Hüllermeier, *Preference Learning*. Springer, 2010.

[100] J. B. Schafer, J. Konstan, and J. Riedi, “Recommender systems in e-commerce,” in *Proceedings of the 1st ACM conference on Electronic commerce*, 1999, pp. 158–166.

[101] T. Calders and B. Goethals, “Non-derivable itemset mining,” *Data Mining and Knowledge Discovery*, vol. 14, no. 1, pp. 171–206, 2007.

[102] J.-S. Lee, C.-H. Jun, J. Lee, and S. Kim, “Classification-based collaborative filtering using market basket data,” *Expert Systems with Applications*, vol. 29, no. 3, pp. 700–704, Oct. 2005.

[103] A. Mild and T. Reutterer, “An improved collaborative filtering approach for predicting cross-category purchases based on binary market basket data,” *Journal of Retailing and Consumer Services*, vol. 10, no. 3, pp. 123–133, May 2003.

[104] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, “Analysis of recommendation algorithms for e-commerce,” in *Proceedings of the 2nd ACM conference on Electronic commerce*, Minneapolis, Minnesota, United States, 2000, pp. 158–167.

[105] Y. H. Cho, J. K. Kim, and S. H. Kim, “A personalized recommender system based on web usage mining and decision tree induction,” *Expert Systems with Applications*, vol. 23, no. 3, pp. 329–342, 2002.

[106] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques, Third Edition*, 3rd ed. Morgan Kaufmann, 2011.

[107] B. Xu, M. Zhang, Z. Pan, and H. Yang, “Content-Based Recommendation in E-Commerce,” in *Computational Science and Its Applications – ICCSA 2005*, vol. 3481, O. Gervasi, M. L. Gavrilova, V. Kumar, A. Laganà, H. P. Lee, Y. Mun, D. Taniar, and C. J. K. Tan, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 946–955.

[108] K. Kim and H. Ahn, “A recommender system using GA K-means clustering in an online shopping market,” *Expert systems with applications*, vol. 34, no. 2, pp. 1200–1209, 2008.

[109] C.-W. Chen and P.-J. Cheng, “Title-Based Product Search – Exemplified in a Chinese E-commerce Portal,” in *Information Retrieval Technology*, vol. 6458,

P.-J. Cheng, M.-Y. Kan, W. Lam, and P. Nakov, Eds. Springer Berlin / Heidelberg, 2010, pp. 25–36.

[110] M. Ovsjanikov and Y. Chen, "Topic modeling for personalized recommendation of volatile items," *Machine Learning and Knowledge Discovery in Databases*, pp. 483–498, 2010.

[111] H. Wenxing, Y. Weng, L. Xie, and L. Maoqing, "Design and implementation of web-based DSS for online shopping mall," in *IEEE International Conference on Control and Automation, 2009. ICCA 2009*, 2009, pp. 1308 –1313.

[112] H.-F. Wang and C.-T. Wu, "A mathematical model for product selection strategies in a recommender system," *Expert Systems with Applications*, vol. 36, no. 3, Part 2, pp. 7299–7308, Apr. 2009.

[113] J. J. Castro-Schez, R. Miguel, D. Vallejo, and L. M. López-López, "A highly adaptive recommender system based on fuzzy logic for B2C e-commerce portals," *Expert Systems with Applications*, vol. 38, no. 3, pp. 2441–2454, Mar. 2011.

[114] H. F. Wang and C. T. Wu, "A strategy-oriented operation module for recommender systems in E-commerce," *Computers & Operations Research*, 2010.

[115] S. Kaiser, S. Kansy, G. Mueller-Seitz, and M. Ringlstetter, "Weblogs for organizational knowledge sharing and creation: a comparative case study," *Knowledge Management Research & Practice*, vol. 7, no. 2, pp. 120–130, Jun. 2009.

[116] J. Bughin, A. H. Byers, and M. Chui, "How social technologies are extending the organization," *MacKinsey Quaterly, November*, 2011.

[117] E. Amitay, D. Carmel, N. Har'El, S. Ofek-Koifman, A. Soffer, S. Yogev, and N. Golbandi, "Social search and discovery using a unified approach," 2009, p. 199.

[118] C. Dugan, M. Muller, D. R. Millen, W. Geyer, B. Brownholtz, and M. Moore, "The dogear game: a social bookmark recommender system," in *Proceedings of the 2007 international ACM conference on Supporting group work*, Sanibel Island, Florida, USA, 2007, pp. 387–390.

[119] S. J. Andriole, "Business impact of Web 2.0 technologies," *Communications of the ACM*, vol. 53, no. 12, pp. 67–79, 2010.

[120] I. Guy, S. Ur, I. Ronen, A. Perer, and M. Jacovi, "Do you want to know?: recommending strangers in the enterprise," in *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, Hangzhou, China, 2011, pp. 285–294.

[121] X. Jin, Y. Zhou, and B. Mobasher, "Web usage mining based on probabilistic latent semantic analysis," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, Seattle, WA, USA, 2004, pp. 197–205.

[122] A. Abecker and L. van Elst, "Ontologies for Knowledge Management," in *Handbook on Ontologies*, S. Staab and R. Studer, Eds. Springer Berlin Heidelberg, 2009, pp. 713–734.

[123] A. Gilchrist, "Corporate taxonomies: report on a survey of current practice," *Online Information Review*, vol. 25, no. 2, pp. 94–103, 2001.

[124] F. Dotsika, “Uniting formal and informal descriptive power: Reconciling ontologies with folksonomies,” *International Journal of Information Management*, vol. 29, no. 5, pp. 407–415, 2009.

[125] G. Mansingh, K. M. Osei-Bryson, and H. Reichgelt, “Building ontology-based knowledge maps to assist knowledge process outsourcing decisions,” *Knowledge Management Research & Practice*, vol. 7, no. 1, pp. 37–51, 2009.

[126] C.-C. Kiu and E. Tsui, “TaxoFolk: a hybrid taxonomy–folksonomy classification for enhanced knowledge navigation,” *Knowledge Management Research & Practice*, vol. 8, no. 1, pp. 24–32, Mar. 2010.

[127] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne, “Finding high-quality content in social media,” in *Proceedings of the international conference on Web search and web data mining*, New York, NY, USA, 2008, pp. 183–194.

[128] C. Mangold, “A survey and classification of semantic search approaches,” *IJMSO*, vol. 2, no. 1, p. 23, 2007.

[129] L. Park and K. Ramamohanarao, “Efficient storage and retrieval of probabilistic latent semantic information for information retrieval,” *The VLDB Journal*, vol. 18, no. 1, pp. 141–155, Jan. 2009.

[130] A. Hotho, R. J\äschke, C. Schmitz, and G. Stumme, “Information retrieval in folksonomies: Search and ranking,” *The Semantic Web: Research and Applications*, pp. 411–426, 2006.

[131] K. Christidis, G. Mentzas, and D. Apostolou, “Using latent topics to enhance search and recommendation in Enterprise Social Software,” *Expert Systems with Applications*, vol. 39, no. 10, pp. 9297–9307, Aug. 2012.

[132] H. M. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno, “Evaluation methods for topic models,” in *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009, pp. 1105–1112.

[133] H. J. Happel and W. Maalej, “Potentials and challenges of recommendation systems for software development,” in *Proceedings of the 2008 international workshop on Recommendation systems for software engineering*, 2008, pp. 11–15.

[134] D. Schuler and T. Zimmermann, “Mining usage expertise from version archives,” in *Proceedings of the 2008 international working conference on Mining software repositories*, 2008, pp. 121–124.

[135] O. Baysal, M. W. Godfrey, and R. Cohen, “A bug you like: A framework for automated assignment of bugs,” in *Program Comprehension, 2009. ICPC ’09. IEEE 17th International Conference on*, 2009, pp. 297 –298.

[136] G. Gousios, E. Kalliamvakou, and D. Spinellis, “Measuring developer contribution from software repository data,” in *Proceedings of the 2008 international working conference on Mining software repositories*, New York, NY, USA, 2008, pp. 129–132.

[137] H. Kautz, B. Selman, and M. Shah, “Referral Web: combining social networks and collaborative filtering,” *Commun. ACM*, vol. 40, no. 3, pp. 63–65, Mar. 1997.

[138] D. L. Olson and D. Delen, *Advanced data mining techniques*. Springer, 2008.

[139] I. Song and P. K. Chintagunta, “Measuring Cross-Category Price Effects with Aggregate Store Data,” *Manage. Sci.*, vol. 52, no. 10, pp. 1594–1609, Oct. 2006.

[140] Y.-L. Chen, K. Tang, R.-J. Shen, and Y.-H. Hu, “Market basket analysis in a multiple store environment,” *Decis. Support Syst.*, vol. 40, no. 2, pp. 339–354, Aug. 2005.

[141] Z. Huang, W. Chung, and H. Chen, “A graph model for E-commerce recommender systems,” *Journal of the American Society for Information Science and Technology*, vol. 55, no. 3, pp. 259–274, 2004.

[142] Y.-J. Park and K.-N. Chang, “Individual and group behavior-based customer profile model for personalized product recommendation,” *Expert Systems with Applications*, vol. 36, no. 2, Part 1, pp. 1932–1939, Mar. 2009.

[143] R. Agrawal, T. Imieliński, and A. Swami, “Mining association rules between sets of items in large databases,” *SIGMOD Rec.*, vol. 22, no. 2, pp. 207–216, Jun. 1993.

[144] A. Savasere, E. Omiccinski, and S. Navathe, “An efficient algorithm for mining association rules in large databases,” in *Proceedings of the International Conference on Very Large Data Bases*, 1995, pp. 432–444.

[145] L. Cavique, “A scalable algorithm for the market basket analysis,” *Journal of Retailing and Consumer Services*, vol. 14, no. 6, pp. 400–407, 2007.

[146] W.-P. Lee, “Towards agent-based decision making in the electronic marketplace: interactive recommendation and automated negotiation,” *Expert Systems with Applications*, vol. 27, no. 4, pp. 665–679, Nov. 2004.

[147] D. G. Gregg and S. Walczak, “Auction Advisor: an agent-based online-auction decision support system,” *Decision Support Systems*, vol. 41, no. 2, pp. 449–471, Jan. 2006.

[148] T. Iwata, S. Watanabe, T. Yamada, and N. Ueda, “Topic tracking model for analyzing consumer purchase behavior,” in *Proceedings of the 21st international jont conference on Artifical intelligence*, 2009, pp. 1427–1432.

[149] K. Tang, Y.-L. Chen, and H.-W. Hu, “Context-based market basket analysis in a multiple-store environment,” *Decis. Support Syst.*, vol. 45, no. 1, pp. 150–163, Apr. 2008.

[150] A. K. McCallum, “MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu,” 2002.

[151] J. Han, J. Pei, Y. Yin, and R. Mao, “Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach,” *Data Mining and Knowledge Discovery*, vol. 8, no. 1, pp. 53–87, Jan. 2004.

[152] *Rapidminer*. .

[153] P. Baum, “A new track for technology: Can ICT take care for healthier lifestyles?,” 2011.

[154] R. Thaler, C. Sunstein, and J. Balz, “Choice architecture,” *Available at SSRN 1583509*, 2010.

[155] Y. Bakos, “The emerging role of electronic marketplaces on the Internet,” *Communications of the ACM*, vol. 41, no. 8, pp. 35–42, 1998.

[156] M. Grieger, “Electronic marketplaces: A literature review and a call for supply chain management research,” *European Journal of Operational Research*, vol. 144, no. 2, pp. 280–294, Jan. 2003.

[157] M. A. Abramson and G. Means, *E-Government 2001*. Rowman & Littlefield, 2001.

[158] E. Turban, R. K. Rainer, and R. E. Potter, *Introduction to information technology*. John Wiley & Sons, 2005.

[159] H. Sharifi, D. F. Kehoe, and J. Hopkins, “A classification and selection model of e-marketplaces for better alignment of supply chains,” *Journal of Enterprise Information Management*, vol. 19, no. 5, pp. 483–503, Jan. 2006.

[160] D. Fensel, D. L. McGuiness, E. Schulten, W. K. Ng, G. P. Lim, and G. Yan, “Ontologies and electronic commerce,” *IEEE Intelligent Systems*, vol. 16, no. 1, pp. 8–14, 2001.

[161] W.-Y. Liang, C.-C. Huang, T.-L. (Bill) Tseng, Y.-C. Lin, and J. Tseng, “The evaluation of intelligent agent performance - An example of B2C e-commerce negotiation,” *Comput. Stand. Interfaces*, vol. 34, no. 5, pp. 439–446, Sep. 2012.

[162] H. Li and H. Wang, “A multi-agent-based model for a negotiation support system in electronic commerce,” *Enterprise Information Systems*, vol. 1, no. 4, pp. 457–472, 2007.

[163] D. Rosaci and G. Sarné, “A multi-agent recommender system for supporting device adaptivity in e-Commerce,” *Journal of Intelligent Information Systems*, vol. 38, no. 2, pp. 393–418, 2012.

[164] S. Huang, “Designing utility-based recommender systems for e-commerce: Evaluation of preference-elicitation methods,” *Electronic Commerce Research and Applications*, vol. 10, no. 4, pp. 398–407, Jul. 2011.

[165] S. J. García-Dastugue and D. M. Lambert, “Internet-enabled coordination in the supply chain,” *Industrial Marketing Management*, vol. 32, no. 3, pp. 251–263, Apr. 2003.

[166] L. Chen and P. Pu, “A cross-cultural user evaluation of product recommender interfaces,” in *Proceedings of the 2008 ACM conference on Recommender systems*, New York, NY, USA, 2008, pp. 75–82.

[167] R. Kohavi, R. Longbotham, D. Sommerfield, and R. M. Henne, “Controlled experiments on the web: survey and practical guide,” *Data Min Knowl Disc*, vol. 18, no. 1, pp. 140–181, Feb. 2009.

[168] P. Buitelaar, P. Cimiano, and B. Magnini, *Ontology Learning from Text: Methods, Evaluation And Applications*. IOS Press, 2005.

[169] G. Aastrand, R. Celebi, and L. Sauermann, “Using linked open data to bootstrap corporate knowledge management in the OrganiK project,” in *Proceedings of the 6th International Conference on Semantic Systems*, New York, NY, USA, 2010, pp. 18:1–18:8.

[170] F. Monaghan, G. Bordea, K. Samp, and P. Buitelaar, “Exploring Your Research: Sprinkling some Saffron on Semantic Web Dog Food,” in *Semantic Web Challenge at the International Semantic Web Conference*, 2010.

[171] M. Ge, C. Delgado-Battenfeld, and D. Jannach, “Beyond accuracy: evaluating recommender systems by coverage and serendipity,” in *Proceedings of the*

fourth ACM conference on Recommender systems, New York, NY, USA, 2010, pp. 257–260.

[172] T. Murakami, K. Mori, and R. Orihara, "Metrics for Evaluating the Serendipity of Recommendation Lists," in *New Frontiers in Artificial Intelligence*, K. Satoh, A. Inokuchi, K. Nagao, and T. Kawamura, Eds. Springer Berlin Heidelberg, 2008, pp. 40–46.

[173] A. Foster and N. Ford, "Serendipity and information seeking: an empirical study," *Journal of Documentation*, vol. 59, no. 3, pp. 321–340, Jun. 2003.

[174] R. Krestel and N. Dokoohaki, "Diversifying Product Review Rankings: Getting the Full Picture," in *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2011 IEEE/WIC/ACM International Conference on*, 2011, vol. 1, pp. 138 – 145.

[175] B. Ludwig, F. Ricci, and Z. Yumak, "1st workshop on recommendation technologies for lifestyle change 2012," in *Proceedings of the sixth ACM conference on Recommender systems*, New York, NY, USA, 2012, pp. 357–358.

Παράρτημα 1: Ερωτηματολόγιο Αξιολόγησης Συστήματος Entasis

Αριθμός Ερώτησης	Διατύπωση Ερώτησης	Τύπος Ερώτησης
1	I have used the OrganiK search functionality.	Ναι/Όχι
2	I was able to locate the items that I was looking for in OrganiK.	Κλίμακα Likert
3	I found the search results relevant to what I was looking for.	Κλίμακα Likert
4	I have noticed the related tags suggested by the system.	Ναι/Όχι
5	I find that the suggested tags were suitable for annotating the document.	Κλίμακα Likert
6	The system suggested in general all suitable tags I could use for annotating the specific document.	Κλίμακα Likert
7	I have noticed the related content suggested by the system.	Ναι/Όχι
8	I find that the suggested content was relevant to the document I was initially reading.	Κλίμακα Likert
9	The system suggested all related content to the document I was initially reading.	Κλίμακα Likert
10	I have created a blog post of my own in OrganiK.	Ναι/Όχι
11	I find it useful to share information in blog posts and receive comments from my colleagues.	Κλίμακα Likert
12	I found important or helpful information in blogs.	Κλίμακα Likert
13	I have shared my status/ thoughts using microblogging (Shout Box).	Ναι/Όχι
14	I find it useful to quickly communicate with your colleagues using microblogging (Shout Box).	Κλίμακα Likert
15	I have inserted the OrganiK bookmarklet into my web browser and used it to share bookmarks	Ναι/Όχι
16	I think its usage improves the discovery of useful resources in the web.	Κλίμακα Likert
17	I have added new information or made a change to a wiki page.	Ναι/Όχι
18	I have found information in wiki pages that was of immediate relevance to my work.	Κλίμακα Likert
19	I have used wiki pages in order to keep knowledge updated or disseminate my work.	Κλίμακα Likert
20	I use the notification features in order to keep myself informed.	Κλίμακα Likert
21	I think that automated notification in OrganiK has helped me keep in touch with developments in my organization.	Κλίμακα Likert
22	OrganiK helped me to earn visibility in my organization	Κλίμακα Likert
23	Do you consider that OrganiK could help you be more informed and aware about activities in your organization	Ανοιχτού Τύπου

	concerning your work?	
24	Do you consider OrganiK could help you get a better social connection with your colleagues?	Ανοιχτού Τύπου
25	Do you consider that OrganiK could help in capturing and organizing the knowledge and expertise of your organization?	Ανοιχτού Τύπου
26	Do you consider OrganiK could help you or your team discovers items helpful in your work activities?	Ανοιχτού Τύπου
27	OrganiK helped me to improve reputation in company	Κλίμακα Likert
28	OrganiK has helped my organization increase knowledge sharing	Κλίμακα Likert
29	OrganiK has helped my organization increase knowledge reuse	Κλίμακα Likert
30	OrganiK has helped my organization identify new business opportunities	Κλίμακα Likert
31	Do you consider that OrganiK could help you be more informed and aware about activities in your organization concerning your work?	Ανοιχτού Τύπου
32	Do you consider OrganiK could help you get a better social connection with your colleagues?	Ανοιχτού Τύπου
33	Do you consider that OrganiK could help in capturing and organizing the knowledge and expertise of your organization?	Ανοιχτού Τύπου
34	Do you consider OrganiK could help you or your team discovers items helpful in your work activities?	Ανοιχτού Τύπου
35	OrganiK has helped my organization increase collaboration efficiency	Κλίμακα Likert
36	I found the search results relevant to what I was looking for.	Κλίμακα Likert

Παράρτημα 2: Σενάρια Αξιολόγησης Συστήματος TradingLink

Αριθμός Ερώτησης	Διατύπωση Ερώτησης
1	You are trying to sell a new camera, and you type the following words:" " The system proposes the use of the following words: " ", "", "", "",. How relevant are those words to your items?
2	How useful are they for describing your item?
3	When you type:"casio g-shock mens watch", the system additionally suggests "band", "time","measuring "
4	Suppose that you are trying to sell an item and you type the following words:" ". The system in order to assist you in not missing any important parts of the description presents the following words: " ", " "," ", " "," ". How useful do you find the appearance of these words?
5	Suppose that you are trying to sell an item and you type the following words:" ". The system in order to assist you in not missing any important parts of the description presents the following words: " ", " "," ", " "," ". How useful do you find the appearance of these words?
6	Suppose that you are trying to sell an item and you type the following words:" ". The system in order to assist you in not missing any important parts of the description presents the following words: " ", " "," ", " "," ". How useful do you find the appearance of these words?
7	Suppose that you are trying to sell an item and you type the following words:" ". The system in order to assist you in not missing any important parts of the description presents the following words: " ", " "," ", " "," ". How useful do you find the appearance of these words?
8	When you type "casio g-shock mens watch", the system suggests the following items: " Casio G-Shock Color Youth Culture Street Fashions Watch DW-6900CS-4DR: bid $0.73 buy $100 ", "Casio G-Shock Riseman Tough Solar Mens Watch bid $160.21 buy $164.11"," Casio Model: BGD-100-7A Color: White Condition: Brand new... bid $89.87 buy $92.22".
9	Suppose that you are trying to sell an item and you type the following words:" ". The system in order to assist you in filling in the starting price and the buy price presents the following items: " ", " "," ", " "," ". How useful do you find the appearance of these items?
10	Suppose that you are trying to sell an item and you type the following words:" ". The system in order to assist you in filling in the starting price and the buy price presents the following items: " ", " "," ", " "," ". How useful do you find the appearance of these items?
11	Suppose that you are trying to sell an item and you type the following words:" ". The system in order to assist you in filling in the starting price and the buy price presents the following items: " ", " "," ", " "," ". How useful do you find the appearance of these items?
12	Suppose that you are trying to sell an item and you type the following words:" ". The system in order to assist you in filling in the starting price and the buy price presents the following items: " ", " "," ", " "," ". How

	useful do you find the appearance of these items?
13	When you browse the item: " camera canon 660D..., bidding price 10.29$, buy price 14.02$ ". The system also suggests: "camera canon 550D bidding price 204.2$, buy price 20.39$ ", " "," ", " "," ".
14	Suppose that looking for what to buy you find the following item: " ". The system in order to assist you in finding what you need presents the following items: " ", " "," ", " "," ". How useful do you find the appearance of these items?
15	Suppose that looking for what to buy you find the following item: " ". The system in order to assist you in finding what you need presents the following items: " ", " "," ", " "," ". How useful do you find the appearance of these items?
16	Suppose that looking for what to buy you find the following item: " ". The system in order to assist you in finding what you need presents the following items: " ", " "," ", " "," ". How useful do you find the appearance of these items?
17	Suppose that looking for what to buy you find the following item: " ". The system in order to assist you in finding what you need presents the following items: " ", " "," ", " "," ". How useful do you find the appearance of these items?
18	When you type:"comics disney edition", the system additionally suggests "comic", "issue","item"
19	When you type:" zoom camera", the system additionally suggests " screen", "photo"," case", " working"
20	When you type:"quality art reprint", the system additionally suggests " pieces", "good ","picture ", "comic"
21	When you type:"used unisex sunglasses", the system additionally suggests "shipping", "brand "," body"
22	When you type "comics disney edition", the system suggests the following items:"GQ Gentlemen's quarterly issue: Cover Story - Robert Downey Jr - IRON MAN comic, bid $7.81"," This is a nice copy of Marvel Age Spiderman Free Comic Book Day #1 2004 bid $2.49, buy $2.49","2004 The new avengers illumminati comic book 1 book 1 of 5 February 2007 bid $2.0"
23	When you type "camera zoom", the system suggests the following items: "Diveset with a refurbished to new silver canon ixus 105IS 12.1MP digital camera bid $224.05", "lovely quality exclusive leather wallet case for apple iPhone 4/4G bid $3.12", "Pentax Zoom-70 35mm Film Camera, In very good condition, working, bid $7.83"
24	When you type "quality art reprint", the system suggests the following items:"Exact photographic replica of the original cult movie poster, "American Werewolf in London" bid $3.91","Exact photographic replica of the original cult movie poster, "Alligator", bid $3.91","Another of Mom's pins. Deep purple rhinestones, faux pearls and turtuise-colored stones. bid $7, buy $11"
25	When you type "used unisex sunglasses", the system suggests the following items: "Brand new Ray-Ban model RB3025 Unisex Sunglasses You are bidding on a brand new Italian made..., bid $65.99, buy $69.99","Gold CP30 Star Wars Booble Figure. Plastic body and plastic base. bid $0.99, buy $3.25","Ray Ban Wayferer sunglasses Original 2140 black. bid $125.13"

26	When you browse a zoom camera, the system suggests the following items: "Diveset with a refurbished to new silver canon ixus 105IS 12.1MP digital camera bid $224.05", "lovely quality exclusive leather wallet case for apple iPhone 4/4G bid $3.12", "Pentax Zoom-70 35mm Film Camera, In very good condition, working, bid $7.83"
27	When you browse a pair of used sunglasses, the system suggests the following items: "Brand new Ray-Ban model RB3025 Unisex Sunglasses You are bidding on a brand new Italian made..., bid $65.99, buy $69.99","Gold CP30 Star Wars Booble Figure. Plastic body and plastic base. bid $0.99, buy $3.25","Ray Ban Wayferer sunglasses Original 2140 black. bid $125.13"
28	When you browse an old zippo lighter the system suggests the following items: "Zippo 1996, camel joe new work,from timeless design to the unmistakable click, bid $68.63,buy $78.44","Zippo 2003, Wright Brothers anniversary lighter. bid $42.14","Brand new thailand elephant zippo lighter in a beautiful blue colour complete with zippo bid $15.64, buy $23.47"
29	When you browse an old vinyl record, the system suggests the following items: "1973 USA Vinyl LP, Tracks Sweet Music, Goodbye Line, Feather Bed,... bid $1.55","Wild Cherry 2 track 7" vinyl in a printed company sleeve, bid $1.1","The hybirds "the only ones (part two)" 3 track limited edition promo 10" comes in a sleeve, bid $6.96,buy $7.74"
30	When you browse a postcard from scotland, the system suggests the following items: "Inverness From thecastle scotland postcard. Printed postcard. Very good unused condition. bid $1.1, buy $1.17","Luss Bay and Ben Lomond Scotland RP. Real Photograph postcard. Very good unused condition. bid $1.1, buy $1.17","Modern postcard of War Poster - Wings over Amreica An images of War poscartd - c. 1990s, bid $3.87, buy $3.87"