Πανεπιστήμιο Πειραιώς – Τμήμα Πληροφορικής

Πρόγραμμα Μεταπτυχιακών Σπουδών

«Πληροφορική»

Μεταπτυχιακή Διατριβή

Τίτλος Διατριβής	**(Ελληνικά)** Θεωρητική μελέτη των συνειρμικών τεχνικών ανάκτησης προκειμένου να υπάρξει βελτίωση στο πρόβλημα ανεπάρκειας του συνεργάσιμου φιλτραρίσματος **(Αγγλικά)** **Theoretical study of associative retrieval techniques in order to improve the sparsity problem of collaborative filtering**
Ονοματεπώνυμο Φοιτητή	ΛΑΙΟΣ ΑΘΑΝΑΣΙΟΣ
Πατρώνυμο	ΓΕΩΡΓΙΟΣ
Αριθμός Μητρώου	ΜΠΠΛ/**13047**
Επιβλέπων	ΤΣΙΧΡΙΝΤΖΗΣ ΓΕΩΡΓΙΟΣ,ΚΑΘΗΓΗΤΗΣ

Ημερομηνία Παράδοσης Νοέμβριος **2016**

Μεταπτυχιακή Διατριβή Αθανάσιος Λάιος

(υπογραφή)	(υπογραφή)	(υπογραφή)
Γεώργιος Τσιχριντζής Καθηγητής	Ευθύμιος Αλέπης Επίκουρος Καθηγητής	Διονύσιος Σωτηρόπουλος Διδάκτωρ

Ε Υ Χ Α Ρ Ι Σ Τ Ι Ε Σ

Για την εκπόνηση της μεταπτυχιακής μου διατριβής θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή Κύριο Γεώργιο Τσιχριντζή , για την εμπιστοσύνη που έδειξε στο πρόσωπο μου με την ανάθεση ενός πολύ ενδιαφέροντος θέματος , ως πνευματικό επιστέγασμα ενός πολύ όμορφου κύκλου που ολοκληρώνεται, του κύκλου μεταπτυχιακών σπουδών στην Πληροφορική του τμήματος Πληροφορικής του Πανεπιστημίου Πειραιώς.

Ο καθηγητής Κύριος Τσιχριντζής πρόσφερε αμέριστη βοήθεια στην προσπάθεια μου , έδωσε πολύτιμες οδηγίες και συμβουλές , μελέτησε τα κείμενα και πρότεινε βελτιώσεις και διορθώσεις προκειμένου να έχουμε το τελικό αποτέλεσμα. Θα ήθελα επίσης να ευχαριστήσω την οικογένεια μου , τους καθηγητές μου , τους φίλους μου που με στήριξαν και με βοήθησαν σε αυτή την δύσκολη προσπάθεια.

Αφιερώνεται στην μνήμη της αδερφής μου.

ΠΕΡΙΛΗΨΗ

Στην παρούσα διατριβή θα ασχοληθούμε σε θεωρητικό επίπεδο με τη μελέτη των συστημάτων σύστασης, που έχουν μεγάλη εφαρμογή στις τεχνικές τους, προκειμένου να προσφέρουν στα προϊόντα και στις υπηρεσίες τις καλύτερες πληροφορίες για τους δυνητικούς καταναλωτές. Το συνεργατικό φιλτράρισμα, που αποτελεί την πιο επιτυχημένη προσέγγιση της σύστασης, διατυπώνει προτάσεις και συστάσεις που βασίζονται στο ιστορικό των συναλλαγών. Ένα από τα σημαντικότερα προβλήματα, που περιορίζει την δημιουργία εμποδίου στο φιλτράρισμα είναι το πρόβλημα της ανεπάρκειας, που αναφέρετε σε μια κατάσταση που οι συναλλαγές ή οι ανατροφοδοτήσεις των δεδομένων είναι ανεπαρκείς για να αντιμετωπιστούν οι ομοιότητες που υπάρχουν στα συμφέροντά των καταναλωτών. Στην παρούσα διατριβή προτείνουμε να ασχοληθούμε με το πρόβλημα της ανεπάρκειας, αφού εφαρμόσουμε ένα συνεταιριστικό πλαίσιο των σχετικών αλγορίθμων ανάκτησης, για να μπορέσουμε να κάνουμε περαιτέρω εξερεύνηση των μεταβατικών ενώσεων που υπάρχουν ανάμεσα στους καταναλωτές βασισμένοι στο προηγούμενο ιστορικό συναλλαγών και την ανατροφοδότηση τους. Οι μεταβατικές ενώσεις, που αποτελούν πολύτιμη πηγή πληροφορίας, με αυτόν τον τρόπο βοηθούν στην διερεύνηση του ενδιαφέροντος των καταναλωτών και επιπλέον είναι απαραίτητα για την εξερεύνηση και αντιμετώπιση του προβλήματος της ανεπάρκειας.

Επομένως για να μπορέσουμε να έχουμε μία καλύτερη αξιολόγηση της αποτελεσματικότητας, που υπάρχει στην προσέγγιση μας, πραγματοποιήθηκε μία πειραματική μελέτη, αφού γίνει χρήση ενός συνόλου δεδομένων από ένα ηλεκτρονικό βιβλιοπωλείο. Πειραματιστήκαμε σε θεωρητικό επίπεδο τρεις επεκτάσιμους αλγόριθμους ενεργοποίησης και πιο συγκεκριμένα έναν περιορισμένο Μοντέλο Πυκνωτή Leaky, έναν αλγόριθμο Branch-and-Bound και έναν αλγόριθμο αναζήτησης δικτύου Hopfield. Επίσης γίνεται μία αναφορά στον αλγόριθμο του πλησιέστερου γείτονα αλλά και στον αλγόριθμο του συνεργατικού φιλτραρίσματος με εστίαση στα γραφήματα τους

Αυτοί οι αλγόριθμοι που συγκρίθηκαν με αρκετές προσεγγίσεις συνεργατικού φιλτραρίσματος δεν υπολογίζουν τις μεταβατικές ενώσεις. Παραδείγματα που μπορούμε να παραθέσουμε είναι, μια απλή προσέγγιση, δύο παραλλαγές της προσέγγισης με βάση το χρήστη, καθώς επίσης και μία προσέγγιση με βάση τα στοιχεία. Τα πειραματικά αποτελέσματα που προκύπτουν, από τη θεωρητική μας μελέτη δείχνουν ότι η διασπορά της ενεργοποίησης της προσέγγισης έχουν ξεπεράσει σε σημαντικό βαθμό τις άλλες συνεργατικές μεθόδους φιλτραρίσματος, όπως για παράδειγμα το μέτρο F και το σκορ κατάταξης. Επίσης παρατηρήσαμε, ότι το αποτέλεσμα της ενεργοποίησης της διάδοσης δεν είναι ανεπαρκές, ενώ στα δεδομένα που θα χρησιμοποιηθούν θα υπάρξει υποβιβασμός της παρουσίασης της σύστασης

ABSTRACT

The aim of this master thesis is the study and application of associative retrieval techniques that result is to have an improvement in sparsity problem. The recommendation systems are widely applied many installed applications. So as to offer to potential customers, products, services and information they need. Collaborative filtering, which is perhaps the most successful recommendation approach classifying them based on past experience and therefore no feedback from consumers who shared the same interests.A major drawback limiting the use of the co filtering is sparsity problem refers to a situation in which the transactions or data feedbacks are sparse and insufficient to have identification of similarities in consumer interests. n this thesis we propose considering the sparsity having problems posting using the application framework associative retrieval relatively disseminated in order to investigate the transitional links between consumers based on their previous transactions and then place the appropriate feedback.

Such transitional compounds are a valuable source of information is a valuable source of information that will help us to conclude consumer interests to be consideration of having problems posting sparsity. To make the evaluation of effectiveness of our approach, we will be carrying out an experimental study he used a data set from an on-line bookstore. Research theoretically done in three widespread activation algorithms including a capacitor leakage algorithm, a connected partial symbolic algorithm, and a Net parallel hop plantations search algorithm. Also made a report to the nearest neighbor algorithm and the algorithm of collaborative filtering to focus on the graphics The plantation relaxation search algorithm are compared with different filtering approaches where not review the transitional Union. A simple graph search approach,two variants of the use of comparison with the approach based a comparison variation element based approach. The experimental results help us to show that a comparison of disseminated activations with approaches based surpassed, to a very large extent, other filtering methods are measured based on the recommendation accuracy, reflection, the measure F and dense result. In addition we observe the effect of the trigger longitudinally, which means that after making integration transitional compounds were included element that was used in the past and not dilute the possibility of dilution. These data are used, so as to draw conclusions relating to user preferences, and ultimately lead to performance degradation recommendation.

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

ΚΕΦΑΛΑΙΟ 1-ΕΙΣΑΓΩΓΗ

1.1 Γενικά Στοιχεία

Σαν κοινωνική διαδικασία η σύσταση διαδραματίζει ένα σημαντικό ρόλο σε πολλές εφαρμογές που χρησιμοποιούνται από τους καταναλωτές , επειδή έχει υπερβολικό κόστος για κάθε καταναλωτή να μάθει ανεξάρτητα όλες τις εναλλακτικές λύσεις . Ανάλογα με την ρύθμιση της συγκεκριμένης εφαρμογής, ο καταναλωτής θα μπορούσε να είναι αγοραστής, για παράδειγμα σε online αγορές, ο αιτούμενος στις πληροφορίες, για παράδειγμα στην ανάκτηση πληροφοριών, ή ακόμα και ένας οργανισμός που προχωρά στην αναζήτηση κάποιας τεχνογνωσίας. Επιπλέον η σύσταση, η οποία αποτελείται από ένα εξατομικευμένο μηχανισμό μάρκετινγκ, έχει προσελκύσει πρόσφατα σημαντικό ενδιαφέρον από βιομηχανίες, όπως για παράδειγμα τις αγορές ηλεκτρονικού εμπορίου αλλά και τη διαφήμιση.

1.2 Στόχος Εργασίας

Σαν κύριο στόχο της παρούσας διατριβής θα είναι η όσο το δυνατόν μεγαλύτερη προσέγγιση των συστημάτων σύστασης τα οποία έχουν αναπτυχθεί, για να αυτοματοποιήσουν την διαδικασία της σύστασης. Τέτοια παραδείγματα ερευνητικών προτύπων συστημάτων σύστασης είναι τα παρακάτω :

1.PHOAKS από τον Terveen και την ερευνητική του ομάδα το 1997.
2 Syskils από τον Webert σε συνεργασία με τους Pazzani και Billsus το 1997
3.Fab από τους Bahanovic και Shaham το 1997
4.Grouplens από τον Konstan και την ερευνητική του ομάδα το 1997, συμπληρωμένη από τον Sarwar και την ερευνητική του ομάδα το 1998.

Αυτά τα συστήματα συνιστούν διάφορους τύπους διαδικτυακών πόρων , σαν παράδειγμα μπορούμε να αναφέρουμε τις απευθείας συνδέσεις για παρακολούθηση ειδήσεων, ταινιών που πρόκειται να παιχτούν, μεταξύ άλλων, σε δυνητικά ενδιαφερόμενους καταναλωτές. Η μεγάλη κλίμακα που παρέχουν οι συγκεκριμένες εφαρμογές είναι δυνατόν υπάρχουν σε πολλές δημοφιλείς, που μερικές από αυτές παραθέτονται παρακάτω:

1.Το Amazon
2.Το CDNOW
3.Το Pharmacy
4.Το Moviefinder

Τα συγκεκριμένα εμπορικά συστήματα προτείνουν τα προϊόντα στους δυνητικούς καταναλωτές βασισμένη στο ιστορικό των προηγούμενων συναλλαγών, σε συνδυασμό με την ανατροφοδότηση που υπάρχει. Γίνονται μέρος του προτύπου της τεχνολογίας της ηλεκτρονικής επιχείρησης και είναι πολύ πιθανό να αυξηθούν οι πωλήσεις του ηλεκτρονικού εμπορίου

μετατρέποντας παράλληλα του ερευνητές σε αγοραστές, αυξάνοντας παράλληλα την πίστη της πελατείας, όπως υποστηρίζει ο Schafer και η ερευνητική του ομάδα σε μελέτη που δημοσιεύθηκε το 2001.

Μία από τις πιο συχνά χρησιμοποιημένες και επιτυχημένες προσεγγίσεις σύστασης είναι η προσέγγιση του συνεργατικού φιλτραρίσματος ,σύμφωνα πάντα με τις έρευνες που δημοσιεύτηκαν από τους Hill και της ομάδας του το 1995, του Resnick και της ομάδας του το 1994 και των Shardamand και Mayes το 1995. Πραγματοποιείται μία πρόβλεψη των πιθανών συμφερόντων σε ένα συγκεκριμένο καταναλωτή βάσει της προηγούμενης συναλλαγής, γίνεται ανατροφοδότηση πληροφοριών και τέλος γίνεται μία πρόβλεψη, που βασίζεται στην παρατήρηση συμπεριφορών, από παρόμοιους καταναλωτές. Παρ'όλη την ευρεία υιοθέτηση και εξάπλωση του, το συνεργατικό φιλτράρισμα, παρουσιάζει σημαντικά μειονεκτήματα όπως για παράδειγμα την ανεπάρκεια στις αναφορές, την διαδικασία επέκτασης του συστήματος και τέλος την συνωνυμία μεταξύ πελατών που ενδεχομένως υπάρχει σύμφωνα πάντα που δημοσιεύτηκε από τον Sarwar και την ομάδα του το 2000.

1.3 Δομή Διατριβής

Στην παρούσα διατριβή θα επικεντρώσουμε την έρευνα μας στο πρόβλημα της ανεπάρκειας, στο οποίο γίνεται αναφορά στην στην έλλειψη ιστορικού, άρα και ανατροφοδότησης των δεδομένων με κύρια συνέπεια να είναι από δύσκολή έως και αδύνατη οποιαδήποτε πρόβλεψη για το ποιοι καταναλωτές έχουν παρόμοια συμπεριφορά με έναν δεδομένο καταναλωτή. Σαν παράδειγμα μπορούμε να αναφέρουμε ότι τα συστήματα συστάσεων που χρησιμοποιούνται από το ιστορικό των αγορών από ομάδες καταναλωτών, στη συνέχεια κάνουν συστάσεις σε μεμονωμένους καταναλωτές και τέλος απευθύνουν συστάσεις σε έναν μεμονωμένο καταναλωτή από την ίδια ομάδα που ήδη το έχει αγοράσει Όταν τέτοια συστήματα έχουν πρόσβαση σε περιορισμένο αριθμό ιστορικού συναλλαγών, σε σχέση πάντα με το συνολικό αριθμό των βιβλίων των καταναλωτών, παρ'όλα αυτά μπορούμε να αναφέρουμε ότι υπάρχει διάσπαρτος προσδιορισμός και σαν φυσική συνέπεια που προκύπτει είναι να καταστούν δύσκολα τα θεμελιώδη συμφέροντα των όμοιων καταναλωτών.

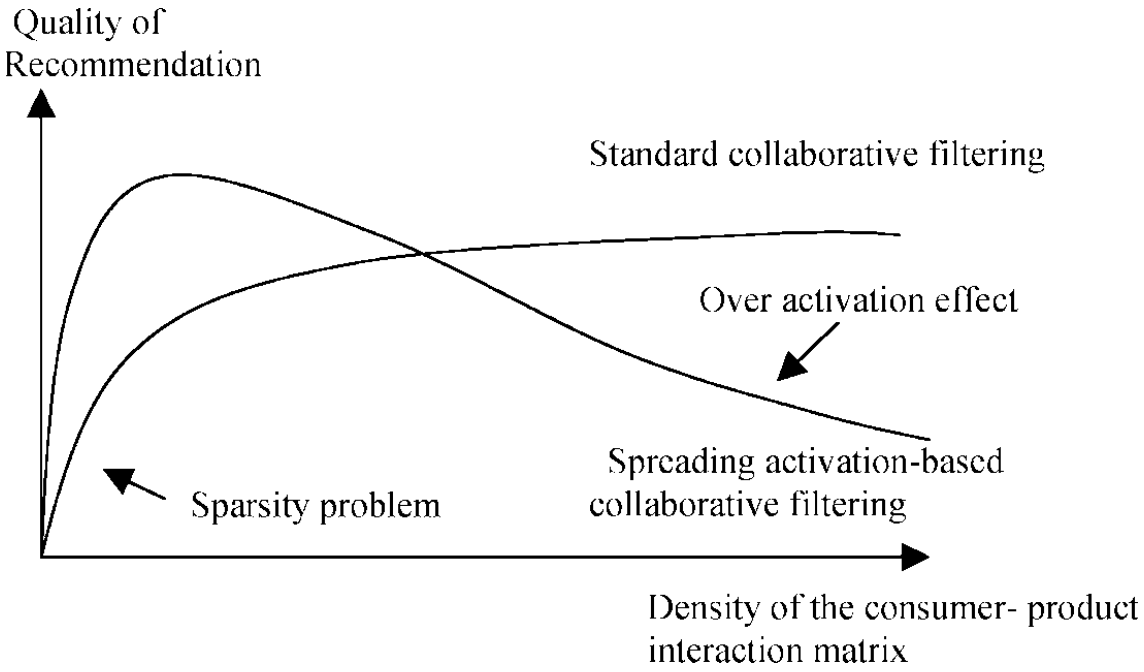
Επομένως σε αυτή την διατριβή θα γίνει μία νέα προσέγγιση στην εξέταση του προβλήματος της, ανεπάρκειας μέσα στα πλαίσια του συνεργατικού φιλτραρίσματος. Στην προσέγγιση αυτή το συνεργατικό φιλτράρισμα έχει μελετηθεί σε διμερή γραφήματα. Αποτελείται από ένα σύνολο σημειώσεων και αντιπροσωπεύει τα προϊόντα των υπηρεσιών καθώς επίσης και πληροφοριακά στοιχεία για πιθανές καταναλώσεις . Η άλλη ομάδα αντιπροσωπεύει τους καταναλωτές των χρηστών. Οι συναλλαγές και οι ανατροφοδοτήσεις έχουν μοντελοποιηθεί ως δεσμοί που συνδέουν τους κόμβους που υπάρχουν μεταξύ αυτών των δύο συνόλων. Στο πλαίσιο της βάσης του γραφήματος εφαρμόζουμε συνειρμικές τεχνικές ανάκτησης, συμπεριλαμβάνοντας αρκετούς αλγόριθμους ενεργοποίησης, προκειμένου να πραγματοποιηθούν ρητά οι μεταβατικές ενώσεις, που με τη σειρά τους βασίζονται στο συνεργατικό φιλτράρισμα. Τα αρχικά πειραματικά αποτελέσματα μας δείχνουν ότι η συνειρμική προσέγγιση, που είναι βασισμένη στην ανάκτηση, είναι δυνατόν να προκαλέσει σημαντική βελτίωση στην αποτελεσματικότητα ενός συνεργατικού φιλτραρίσματος παρ'όλο που δημιουργείται θέμα από την ύπαρξη του προβλήματος της ανεπάρκειας.

Στο παρόν κεφάλαιο μελετάμε τρεις αντιπροσωπευτικούς αλγόριθμούς διαδιδόμενης ενεργοποίησης απαραίτητα για την έρευνα μας.

1.Έναν περιορισμένης επέκτασης αλγόριθμος διαδιδόμενης ενεργοποίησης βασισμένος στο μοντέλο πυκνωτών Leaky, όπως προτάθηκε το 1983 από τον Anderson.

2.Έναν Branch-and-Bound σειριακό, συμβολικό ερευνητικό αλγόριθμο και

3.Έναν παράλληλης χαλάρωσης και έρευνας δικτύου αλγόριθμο Hopfield.
Σε αυτό το κεφάλαιο συνοψίζουμε την έρευνα μας πάνω σε αυτούς τους αλγόριθμους και πραγματοποιούμε μία συζήτηση σχετικά με την υλοποίηση τους



Σχήμα 1ο : Το πρόβλημα της ανεπάρκειας και επιδράσεις της υπερενργοποίησης

1.4 Περιορισμένης επέκτασης Αλγόριθμος Διαδιδόμενης Ενεργοποίησης βασισμένος στα πρότυπα πυκνωτών Leaky

Χρησιμοποιώντας το μοντέλο πυκνωτών Leaky, που στη διεθνή βιβλιογραφία ονομάζεται και σαν αλγόριθμος LCM και προτάθηκε από τον Anderson το 1983, οι καταναλωτές και τα προϊόντα θεωρούνται ως γενικοί κόμβοι. Ένας κόμβος σύνδεσης, που συμβολίζεται με R, ορίζεται παρακάτω προκειμένου να συμπεριλάβει τις ενώσεις που προκύπτουν από τους κόμβους.

$$R(r*r)=\begin{bmatrix} |P|*|P| & A^T(|C|*|P|) \\ \\ \lambda(|P|*|C| & I(|P|*|C|) \end{bmatrix} \quad (1)$$

Στον παραπάνω ορισμό, η τιμή του |P| υποδηλώνει τον αριθμό των προϊόντων, το |C| τον αριθμό των καταναλωτών, το άθροισμα r=|P|+|C| και το γινόμενο A(|P|*|C| αντιπροσωπεύει τη μήτρα αλληλεπίδρασης μεταξύ καταναλωτή και προϊόντος. Ο κόμβος R είναι η μήτρα γειτνίασης για το διμερές μας γράφημα και αντιστοιχίζεται με τη μήτρα αλληλεπίδρασης καταναλωτών και προϊόντων A. Επειδή οι συνδέσεις των ομοιοτήτων που προκύπτουν ανάμεσα στα προϊόντα και στους καταναλωτές απουσιάζουν από το μοντέλο γραφημάτων τα αντίστοιχα στοιχεία και οι ενώσεις των καταναλωτών εκπροσωπούνται από μήτρες ταυτοτήτων. Τα κύρια στάδια της υλοποίησης του περιορισμένης επέκτασης αλγόριθμου LCM συνοψίζονται παρακάτω

- **Βήμα 1ο –Έναρξη** Αρχικά ένας κόμβος διανύσματος V δημιουργείται για να παρουσιάσει το χρήστη στόχων. Αυτό το διάνυσμα περιλαμβάνει r στοιχεία από τα οποία μόνο το ένα αντιστοιχίζεται στο χρήστη στόχων και αποδίδεται σε αυτό τιμή ίση με τη μονάδα. Σε όλα τα υπόλοιπα στοιχεία εκχωρούμε τιμή ίση με το μηδέν. Ένα διάνυσμα ενεργοποίησης D δημιουργείται για να συλλάβει όλα τα επίπεδα ενεργοποίησης από όλους τους κόμβους του συγκεκριμένου μοντέλου. Όλα τα στοιχεία του D(0) αρχικοποιούνται με την τιμή μηδέν.
- **Βήμα 2ο- Ενεργοποίηση και υπολογισμός επιπέδου ενεργοποίησης.** Κατά τη διάρκεια της επανάληψης t ο αλγόριθμός υπολογίζει το διάνυσμα ενεργοποίησης D(t) ως εξής:

  $$D(t)=V+M'D(t-1) \text{ και } M=(1-\gamma)I + aR \quad (2)$$

  Όπου η παράσταση (1-γ) υπολογίζει την ταχύτητα της αποσύνθεσης στο επίπεδο ενεργοποίησης των ενεργών κόμβων και η σταθερά α περιγράφει την αποδοτικότητα με την οποία οι κόμβοι μετατρέπουν την ενεργοποίηση που λήφθηκαν άμεσα από τους ενεργούς τους κόμβους με βάση τα δικά τους επίπεδα ενεργοποίησης. Μόνο ένας συγκεκριμένος αριθμός από κόμβους με το υψηλότερο επίπεδο ενεργοποίησης διατηρούν τα επίπεδα ενεργοποίησης τους σε τιμή A(t). Όλοι τα υπόλοιπα στοιχεία του A(t) αρχικοποιούνται με την τιμή μηδέν. Οι παράμετροι ελέγχου λαμβάνουν, στα πειράματά μας τιμές από 0.2 έως 0.8, μετά την ολοκλήρωση της διαδικασίας εκτέλεσης διάφορων αλγορίθμων.

- **Βήμα 3ο-Παύση Υπολογισμού.** Ο αλγόριθμος τερματίζει έπειτα από ένα συγκεκριμένο αριθμό επαναλήψεων. Το όριο αυτών των επαναλήψεων ορίζεται σε 10 για την τρέχουσα υλοποίηση της εφαρμογής. Τα 50 κορυφαία αντικείμενα κόμβων έχουν το υψηλότερο επίπεδο ενεργοποίησης μέσα στο διάνυσμα ενεργοποίησης του τελικού σταδίου A(10) και για αυτό δεν έχουν προηγουμένως αγοραστεί από τη σύσταση για τον στοχευμένο καταναλωτή.

1.5 Αλγόριθμος Branch-and-Bound

Η υλοποίηση του αλγόριθμου Branch-and-Bound, που αλλιώς ονομάζεται και αλγόριθμος BNB, ακολουθεί την χρήση που την χρήση που πρότεινε ο Chen και NG το 1995, αναπτύχθηκε στο πλαίσιο της χρήσης της έννοιας. Η υλοποίηση μας ξεκινά με τον χρήστη κόμβων που

αντιστοιχίζεται με το χρήστη στόχων. Οι γειτονικοί κόμβοι, που είναι οι κόμβοι των αντικειμένων και αντιστοιχίζονται με το ιστορικό αγορών των χρηστών στόχων, ενεργοποιούνται. Οι ενεργοί κόμβοι τοποθετούνται μέσα σε μία ουρά προτεραιότητας, η οποία είναι βασισμένη στα επίπεδα ενεργοποίησης τους και στην υψηλή προτεραιότητα των κόμβων, που είναι απαραίτητα για την ενεργοποίησης τους. Τα κύρια βήματα για την υλοποίηση του αλγόριθμου Branch-and-Bound συνοψίζονται ως εξής

- **Βήμα 1ο-Έναρξη**. Ο κόμβος, που αντιστοιχίζεται με το χρήστη στόχων, αρχικοποιείται για να λάβει το επίπεδο ενεργοποίησης 1. Όλοι οι υπόλοιποι κόμβοι αρχικοποιούνται στο επίπεδο μηδέν. Μία ουρά προτεραιότητας $Q_{priority}$, δημιουργείται μόνο με τον χρήστη κόμβων σαν αρχικό αριθμό τους. Μία άδεια αρχικοποιημένη ουρά εξόδου Q_{output} δημιουργείται προκειμένου να αποθηκεύσει τους ενεργούς κόμβους.
- **Βήμα 2ο –Ενεργοποίηση και υπολογισμός της ενεργοποίησης.** Κατά τη διάρκεια κάθε ενεργοποίησης, ο αλγόριθμος διαγράφει τον αμέσως επόμενο κόμβο από τον $Q_{priority}$, ο οποίος κόμβος έχει το υψηλότερο επίπεδο ενεργοποίησης, ενεργοποιεί τους γειτονικούς κόμβους, ενώ παράλληλα υπολογίζει τα γειτονικά επίπεδα ενεργοποίησης, βάσει του τύπου $\mu_j(t+1)=\mu_i(t)*t_{ij}$ όπου $\mu_i(t)$ είναι η μεταβλητή που παρουσιάζει το επίπεδο ενεργοποίησης από τον αμέσως επόμενο κόμβο που έχει ήδη διαγραφεί από τον $Q_{priority}$, η μεταβλητή t_{ij} παρουσιάζει το βάρος του συνδέσμου που συνδέεται με τον αμέσως επόμενο κόμβο και με τον γειτονικό κόμβο, ενώ παράλληλα ορίζουμε στην εφαρμογή μας κάθε σύνδεσμο να έχει τιμή 0.5, και η μεταβλητή $\mu_j(t+1)$ παρουσιάζει τον νέο τρόπο υπολογισμού του επιπέδου ενεργοποίησης για τους γειτονικούς κόμβους. Οι ενεργοί κόμβοι, οι οποίοι δεν είχαν εγγραφεί νωρίτερα στο Q_{output}, καταχωρούνται στην ουρά εξόδου. Ενώ αν είναι ήδη διαθέσιμες στο Q_{output} το επίπεδο των ενεργοποιήσεων θα αυξηθεί κατά $\mu_j(t+1)$
- **Βήμα 3ο Παύση Υπόθεσης** Η παραπάνω διαδικασία ενεργοποίησης επαναλαμβάνεται για ένα σταθερό αριθμό επαναλήψεων προτού ο αλγόριθμος τερματιστεί και εξάγει τα 50 κορυφαία αντικείμενα των κόμβων από τον Q_{output}. Στα πειράματα μας θα ορίσουμε οι επαναλήψεις του αλγόριθμου να είναι 70.

1.6 Αλγόριθμος Αναζήτησης Δικτύου Hopfield

Ο αλγόριθμος αναζήτησης δικτύου Hopfield, ο οποίος έπειτα για συντομία ονομάστηκε αλγόριθμος Hopfield, εκτελεί μία παράλληλη διαδικασία χαλαρής έρευνας για να υποστηρίξει την διαδιδόμενη ενεργοποίηση. Σε αυτό το πλαίσιο, τα μοντέλα γραφημάτων του συνεργατικού φιλτραρίσματος σχεδιάζονται για να υπερσυνδέσουν νευρώνες και συνάψεις μέσα στο δίκτυο Hopfield, με τους νευρώνες να εκπροσωπούν τους χρήστες και τα αντικείμενα, ενώ οι συνάψεις εκπροσωπούν τον υπερσύνδεσμο μεταξύ χρηστών και αντικειμένων. Η υλοποίηση του αλγορίθμου ενεργοποίησης δικτύου Hopfield περιγράφεται με τα ακόλουθα βήματα

- **Βήμα 1ο Προετοιμασία** Ο κόμβος των χρηστών που αντιστοιχίζεται με τον χρήστη στόχων, προετοιμάζεται για να λάβει το επίπεδο ενεργοποίησης 1. Όλοι οι υπόλοιποι κόμβοι προετοιμάζονται για το επίπεδο 0
- **Βήμα 2ο Ενεργοποίηση Αλγορίθμου και Ενεργοποίηση Επιπέδου Υπολογισμών.** Όπως αναφέραμε και στον αλγόριθμο LCM, πραγματοποιούμε ενεργοποίηση σε ένα συγκεκριμένο αριθμό κόμβων τα οποία έχουν υψηλό επίπεδο ενεργοποίησης. Το επίπεδο της ενεργοποίησης, για κάθε κόμβο, υπολογίζεται από τον ακόλουθο τύπο:

$$\mu_j(t+1)=f_s\sum_{i=0}^{n-1} t_{ij}\mu_i(t),\ 0\leq j\geq 1 \qquad (3)$$

όπου η f_s είναι μία συνεχόμενη σιγμοειδές συνάρτηση μετασχηματισμού. Ο Knight το 1990 όρισε την συνάρτηση fs με τον ακόλουθο τύπο:

$$f_s=1/1+\exp((\theta 1-x)/\theta 2) \quad (4)$$

όπου $\mu_j(t+1)$ είναι το επίπεδο ενεργοποίησης του κόμβου j με βάση την επανάληψη t+1, ενώ t_{ij} είναι το βάρος του συνδέσμου που συνδέει τον κόμβο I με τον κόμβο j, με παρόμοιο τρόπο

όπως και στον αλγόριθμο branch-and-bound, ορίζουμε το βάρος του συνδέσμου να παίρνει την τιμή 0.5. Σε συμφωνία υλοποίησης της παράστασης (6), κάθε νέος ενεργοποιημένος κόμβος υπολογίζει το επίπεδο ενεργοποίησης του με βάση το άθροισμα των προϊόντων από τα γειτονικά επίπεδα ενεργοποίησης και της συνάψεις τους. Ο έλεγχος των παραμέτρων θ_1 και θ_2 από την σιγμοειδή συνάρτηση με ευρετικό τρόπο ορίζονται με τιμές 10 και 0.8 αντίστοιχα στα πειράματα μας.

- **<u>Βήμα 3ο Παύση της Υπόθεσης</u>** Η παραπάνω μεθοδολογία επαναλαμβάνεται μέχρι η συνθήκη (8) ικανοποιείται και μας δείχνει ότι δεν υπάρχει σημαντική αλλαγή υπερύψωσης μεταξύ των δύο τελευταίων επαναλήψεων.

$$\sum_j \mu_j(t+1)-\sum_j \mu_j(t) < \varepsilon * t \quad (5)$$

Σε αυτή τη συνθήκη, το ε είναι ένας πολύ μικρός θετικός αριθμός. Σημειώνουμε εδώ ότι οι επιτρεπόμενες αλλαγές είναι ανάλογες ως προς τον αριθμό των επαναλήψεων που εκτελούνται για την επιτάχυνση της σύγκλισης. Όπως και σε όλες τις άλλες προσεγγίσεις, οι κορυφαίοι κόμβοι αντικειμένων που έχουν το υψηλότερο επίπεδο ενεργοποίησης στην τελική κατάσταση δικτύου, συνιστώνται στους καταναλωτές μετά τη διαδικασία διαγραφής των αντικειμένων που έχουν ήδη αγοραστεί από το χρήστη στόχων.

Το υπόλοιπο της διατριβής οργανώνεται ως εξής Στο 2ο κεφάλαιο πραγματοποιείται έρευνα, βασισμένη στις μελέτες που ήδη υπάρχουν σχετικά με το συνεργατικό φιλτράρισμα ενώ παράλληλα ασχολείται λεπτομερώς με το πρόβλημα της ανεπάρκειας. Στο 3ο κεφάλαιο συνοψίζουμε τα μοντέλα συνειρμικής ανάκτησης, παρουσιάζοντας παράλληλα τη γραφική παράσταση, που βασίζεται στο συνεργατικό φιλτράρισμα. Στο υποκεφάλαιο 3.1 πραγματοποιείται εισαγωγή της συνειρμικής ανάκτησης, ενώ ταυτόχρονα παρουσιάζουμε την αντίστοιχη γραφική παράσταση, που βασίζεται στα μοντέλα του συνεργατικού φιλτραρίσματος. Στο υποκεφάλαιο 3.2 γίνεται μια αναλυτική παρουσίαση του γενικού σχεδιασμού της προτεινόμενης προσέγγισης φιλτραρίσματος η οποία είναι βασισμένη στην συνειρμική ανάκτηση. Στο υποκεφάλαιο 3.3 πραγματοποιούμε εισαγωγή του αλγόριθμου ενεργοποίησης της διάδοσης ο οποίος παρέχει τον υπολογιστικό μηχανισμό για να χρησιμοποιηθεί σε περαιτέρω διερεύνηση, κάτω από το πλαίσιο της μεταβατικής ένωσης, των ειδικών ερευνητικών ερωτημάτων, τα οποία προκύπτουν από την ανάλυση μας, στο υποκεφάλαιο 3.4. Στο 4ο θα επιχειρήσουμε να αμβλύνουμε το πρόβλημα της ανεπάρκειας με χρήση μιας προσαρμοσμένης απόστασης και μίας μεθόδου που βασίζεται σε μοντέλα γραφημάτων. Στο 5ο κεφάλαιο πραγματοποιούμε μια αναλυτική παρουσίαση δύο ενδεικτικών αλγορίθμων και μία ενδεικτική

μελέτη από τη χρήση τους. Ολοκληρώνοντας τη διατριβή μας στο 6ο κεφάλαιο συνοψίζουμε την έρευνα και τα αποτελέσματά μας ενώ επισημαίνουμε τις μελλοντικές μας κατευθύνσεις.

ΚΕΦΑΛΑΙΟ 2-ΣΥΝΕΡΓΑΤΙΚΟ ΦΙΛΤΡΑΡΙΣΜΑ ΚΑΙ ΤΟ ΠΡΟΒΛΗΜΑ ΤΗΣ ΑΝΕΠΑΡΚΕΙΑΣ

Σε αυτό το κεφάλαιο θα εξετάσουμε την προηγούμενη έρευνα μας, καθώς επίσης και την ανάπτυξη του συστήματος του συνεργατικού φιλτραρίσματος, ενώ θα εισάγουμε το πρόβλημα της ανεπάρκειάς, που έχει αναγνωριστεί σαν μία από τις σημαντικότερες τεχνικές που παρεμποδίζουν την περαιτέρω ανάπτυξη και υιοθέτηση των συνεργατικών συστημάτων φιλτραρίσματος

2.1 Συνεργατικό Φιλτράρισμα

Στο συνεργατικό φιλτράρισμα πραγματοποιείται παραγωγή εξατομικευμένων συστάσεων οι οποίες προκύπτουν από το άθροισμα των εμπειριών παρόμοιων χρηστών στο σύστημα. Σαν έννοια μπορεί να αναφερθεί ότι η προσέγγιση αυτοματοποιεί την διαδικασία της σύστασης αρκεί να γίνει διάδοση «από στόμα σε στόμα». Μία άλλη πτυχή που προκύπτει από τη χρήση του συνεργατικού φιλτραρίσματος, είναι ότι ο προσδιορισμός των καταναλωτών ή των χρηστών είναι παρόμοιος με αυτόν που είναι απαραίτητος στη σύσταση. Τα μοντέλα Cluster, τα μοντέλα δικτύου Bayesian, που είναι κατά κανόνα εξειδικευμένοι αλγόριθμοι συσχέτισης, σε σχέση με τις άλλες τεχνικές που έχουν χρησιμοποιηθεί για τον σκοπό της ταυτοποίησης, σύμφωνα με την έρευνα που δημοσιεύτηκε από τον Brees και την ομάδα του το 1998 και ξαναδημοσιεύτηκε με βετιώσεις από τον Lin και την ομάδα του το 2002. Οι μέθοδοι που βασίζονται στην συμπεριφορά των παρόμοιων καταναλωτών, όπως επίσης και των γειτόνων τους, αποτελούν το ποιο συχνό της προσέγγισης και είναι δυνατόν να προκαλέσουν συστάσεις, όπως αναφέρεται στη μελέτη που δημοσίευσε ο Sarwar και η ομάδα του το 2002.

Το συνεργατικό φιλτράρισμα αποτέλεσε την πιο επιτυχημένη προσέγγιση του συστήματος και έχει εφαρμοστεί ευρέως σε πολλές εφαρμογές όπως προκύπτει από τις μελέτες που δημοσιεύτηκαν, από τον Burke το 2000, τον Claypool και την ομάδα του το 2000, τον Mobasher και την ομάδα του το 2000, τον Nasraoui και την ομάδα του το 1999, τον Pazzani το 1999 και τον Sarwar και την ομάδα του το 1998. Παρά την επιτυχία που υπάρχει σε πολλές ρυθμίσεις των εφαρμογών, από την χρήση της προσέγγισης του συνεργατικού φιλτραρίσματος, έχει αναφερθεί η ύπαρξη αρκετών και μεγάλων περιορισμών μεταξύ αυτών, τα προβλήματα ανεπάρκειας, επεκτασιμότητας και συνωνυμίας, όπως αναφέρεται στη μελέτη του Sarwar και

της ομάδας που δημοσιεύτηκε το 2000. Το πρόβλημα της ανεπάρκειας προκύπτει από το γεγονός ότι οι συναλλαγές ή οι ανατροφοδοτήσεις δεν είναι αρκετές, προκειμένου να υπάρξει αναζήτηση γειτόνων, ενώ παράλληλα αποτελεί περιορισμό, σε γενικές γραμμές, της χρήσης του φιλτραρίσματος στην ποιότητα των συστάσεων. Η μελέτη μας εστιάστηκε στην ανάπτυξη μιας αποτελεσματικής προσέγγισης προκειμένου να υπάρξει σύσταση υψηλής ποιότητας. Το επόμενο υποκεφάλαιο αναφέρει το πρόβλημα τις ανεπάρκειας με λεπτομέρειες.

2.2 Το πρόβλημα Ανεπάρκειας

Στα συνεργατικά συστήματα φιλτραρίσματος, υπάρχει τυπική εκπροσώπηση χρηστών ή καταναλωτών από τα αντικείμενα που έχουν αγοραστεί ονομαστικά. Σαν παράδειγμα μπορούμε να αναφέρουμε, ότι σε ένα online βιβλιοπωλείο με πωλήσεις 2 εκατομμυρίων βιβλίων, κάθε καταναλωτής αντιπροσωπεύεται από ένα Boooolean διάνυσμα χαρακτηριστικών γνωρισμάτων που αποτελείται από 2 εκατομμύρια στοιχεία. Η αξία του κάθε στοιχείου καθορίζεται από το αν ο πελάτης έχει αγοράσει το αντίστοιχο βιβλίο, το οποίο φαίνεται από το ιστορικό των συναλλαγών. Συνήθως όταν έχει πραγματοποιηθεί η αγορά η τιμή που λαμβάνει είναι ίση με τη μονάδα, ενώ όταν η τιμή ισούται με το 0 σημαίνει ότι η αγορά δεν έχει πραγματοποιηθεί. Κατά τη διάρκεια της εξέτασης πολλών καταναλωτών, μία μήτρα που περιλαμβάνει όλους τους φορείς και παράλληλα εκπροσωπείται από τους καταναλωτές, μπορεί να χρησιμοποιηθεί και σαν μήτρα δράσεων. Καλούμε λοιπόν αυτήν τη μήτρα, ως μήτρα αλληλεπίδρασης καταναλωτικών προϊόντων. Η χρήση, γενικά του όρου της αλληλεπίδρασης, μας παραπέμπει στη μήτρα, στην οποία υπάρχει αντίθεση σε συγκεκριμένη «αγορά» ή «συναλλαγή», επειδή υπάρχουν και άλλα είδη σχέσεων, όπως για παράδειγμα οι άμεσες και έμμεσες αξιολογήσεις μεταξύ των καταναλωτών και των προϊόντων, γενικά για τα συστήματα των συστάσεων.

Σε αυτό το σημείο, θα γίνει εισαγωγή μερικών σημειώσεων που είναι απαραίτητες για τη διατριβή μας. Ορίζουμε σαν μεταβλητή C το σύνολο των καταναλωτών και σαν P το σύνολο των καταναλωτών. Συμβολίζουμε την μήτρα αλληλεπίδρασης καταναλωτικών προϊόντων με τον ακόλουθο τύπο |C|X|P| matrix $A=(a_{ij})$, ενώ το πεδίο τιμών που λαμβάνει η μήτρα είναι το ακόλουθο:

$$a_{ij=} \begin{cases} 1, \text{Εάν ο χρήστης έχει αγορασμένο το στοιχείο j} \\ 0, \text{Σε διαφορετική περίπτωση} \end{cases} \quad (6)$$

Σε αυτό το σημείο μπορούμε να σημειώσουμε στην μελέτη μας ότι εστιάσαμε στις πραγματικές συναλλαγές που έλαβαν χώρα έτσι ώστε η τιμή της μήτρας a_{ij} είναι δυαδική, δηλαδή να λαμβάνει τιμές 0 και 1. Σε άλλα σενάρια σύστασης, όπως για παράδειγμα εκείνα που εμπλέκουν και αξιολογήσεις, οι τιμές που θα λάβει η μήτρα a_{ij} μπορεί να λάβει κατηγορηματικές οι συνεχείς τιμές, όπως για παράδειγμα βαθμολογία πέντε επιπέδων και πιθανότητες των συμφερόντων.

Σε πολλές εφαρμογές μεγάλης κλίμακας, όπως μεγάλες ιστοσελίδες ηλεκτρονικού εμπορίου, τόσο ο αριθμός των αντικειμένων |P| όσο και των καταναλωτών |C| είναι τεράστιος. Σε τέτοιες περιπτώσεις, ακόμα και όταν έχουν καταγραφεί πολλές συναλλαγές, η μήτρα αλληλεπίδρασης καταναλωτικών προϊόντων μπορεί ακόμα να είναι εξαιρετικά ανεπαρκής, δηλαδή να υπάρχουν πολύ λίγα στοιχεία στο A που η αξία τους να είναι ίση με τη μονάδα. Αυτό το πρόβλημα που, συνήθως το αναφέρουμε και σαν πρόβλημα ανεπάρκειας, θα έχει σημαντικές αρνητικές επιπτώσεις στην αποτελεσματικότητα της προσέγγισης του συνεργατικού φιλτραρίσματος.

Λόγω της ανεπάρκειας είναι πολύ πιθανό η ομοιότητα της συσχέτισης μεταξύ δύο χρηστών να είναι μηδέν, τα συνεργατικά συστήματα φιλτραρίσματος καθίστανται άχρηστα, σύμφωνα με την έρευνα των Billsus και Pazzani που δημοσιεύτηκε το 1998. Ακόμα και αν τα ζεύγη των χρηστών που συσχετίζονται θετικά, δεν μπορούν να θεωρούνται αξιοποιήσιμα.

Το πρόβλημα της κρύας εκκίνησης μας δείχνει περαιτέρω τη σημασία με την οποία αντιμετωπίζουμε το πρόβλημα της ανεπάρκειας. Το συγκεκριμένο πρόβλημα αναφέρεται στην κατάσταση κατά την οποία ένας χρήστης ή ένα στοιχείο μόλις έχει εισαχθεί στο σύστημα, σύμφωνα με την μελέτη του Schein και τη ερευνητική του ομάδα το 2002. Το συνεργατικό φιλτράρισμα δεν μπορεί να δημιουργήσει χρήσιμες προτάσεις για το νέο χρήστη λόγω της έλλειψης επάρκειας προηγούμενων αξιολογήσεων ή αγορών. Με τον ίδιο τρόπο, όταν ένα νέο στοιχείο εισέρχεται στο σύστημα, είναι απίθανο το συνεργατικό φιλτράρισμα να προβεί σε συστάσεις αγοράς σε πολλούς χρήστες, επειδή είναι πολύ λίγοι αυτοί που έχουν αξιολογηθεί ή ακόμα και να το αγοράσουν. Εννοιολογικά το πρόβλημα της κρύας εκκίνησης μπορεί να θεωρηθεί σαν ειδικό παράδειγμα του προβλήματος της ανεπάρκειας, όπου τα περισσότερα στοιχεία που υπάρχουν στις γραμμές ή της στήλες, της μήτρας αλληλεπίδρασης ανάμεσα στον καταναλωτών και προϊόντων, λαμβάνουν την τιμή 0.

Πολλοί ερευνητές έχουν προσπαθήσει να ανακουφίσουν το πρόβλημα της ανεπάρκειας. Ο Sarwar και η ομάδα του, σε μελέτη που δημοσίευσαν το 2001, πρότειναν μία προσέγγιση του στοιχείου που βασίζεται στην αντιμετώπιση, τόσο της παρουσίας, όσο και της κλιμάκωσης του προβλήματος της ανεπάρκειας. Με βάση τα δεδομένα των συναλλαγών αλλά και της ανατροφοδότησης, Τα στοιχεία που είναι παρόμοια με εκείνα που αγοράστηκαν στο παρελθόν από το χρήστη του στόχου και είναι αναγνωρισμένα, εν συνεχεία συνίστανται. Οι ομοιότητες του στοιχείου υπολογίζονται όπως η συσχετίσεις των φορέων που εντοπίζονται μεταξύ των αντίστοιχων στηλών του στοιχείου. Έχει αναφερθεί ότι σε ορισμένες εφαρμογές ότι η προσέγγιση βασισμένη στο στοιχείο επιτυγχάνει καλύτερη ποιότητα σύστασης, σε σύγκριση με την προσέγγιση βασισμένη στον χρήστη, ενώ η κυρίαρχη προσέγγιση, που χρησιμοποιεί τα συστήματα σύστασης, στηρίζεται σε συσχετίσεις που προκύπτουν από τα διανύσματα σειρών και χρηστών.

Μία άλλη προσέγγιση που προτείνουμε, αφορά στη μείωση της διάστασης, ενώ ο στόχος της είναι η απευθείας μείωση της διάστασης της μήτρας αλληλεπίδρασης καταναλωτικών προϊόντων. Μία απλή στρατηγική είναι ο σχηματισμός ομάδων στοιχείων ή χρηστών ενώ στη συνέχεια οι δεσμοί που προκύπτουν χρησιμοποιούνται, σαν βασικές μονάδες, για τη διατύπωση συστάσεων. Υπάρχουν πιο προηγμένες τεχνικές που μπορούν να εφαρμοστούν προκειμένου να επιτευχθεί η μείωση της διάστασης. Τέτοια παραδείγματα είναι οι στατιστικές τεχνικές, όπως η ανάλυση της βασικής συνιστώσας (Principle Component Analysis-PCA), που προτάθηκε από τον Golberg και την ομάδα του το 2001, ενώ μπορούμε να αναφέρουμε και τις τεχνικές ανάκτησης των πληροφοριών με συγκεκριμένη περίπτωση την Λανθάνουσα Σημασιολογική Ευρετηρίαση (Latent Semantic Indexing-LSI), η οποία προτάθηκε από τους Billsus και Pazzani το 1998 και τον Sarwar και την ομάδα του το 2000. Εμπειρικές μελέτες αναφέρουν ότι η μείωση της διάστασης είναι δυνατόν να προκαλέσει σημαντική βελτίωση της ποιότητας της σύστασης σε μερικές εφαρμογές, ενώ στις υπόλοιπες απλά να παρουσιάζει καλή απόδοση, όπως υποστηρίζει σε μελέτη που δημοσίευσε ο Sarwar και η ομάδα του το 2000. Η προσέγγιση της μείωσης της διάστασης αντιμετωπίζει το πρόβλημα του προβλήματος της ανεπάρκειας, προβαίνοντας σε διαγραφή αντιπροσωπευτικά ασήμαντων καταναλωτών ή προϊόντων συμπυκνώνοντας με αυτόν τον τρόπο τη μήτρα αλληλεπίδρασης καταναλωτικών προϊόντων. Ωστόσο είναι δυνατόν κατά τη διάρκεια της μείωσης της διάστασης να χαθούν δυνητικά απαραίτητες πληροφορίες. Αυτό, εν μέρει, μπορεί να εξηγήσει τα ανάμεικτα αποτελέσματα των

αναφορών, σχετικά με την απόδοση της μείωσης της διάστασης που βασίζεται στις προσεγγίσεις συνεργατικού φιλτραρίσματος.

Οι ερευνητές έχουν προσπαθήσει να συνδυάσουν το συνεργατικό φιλτράρισμα με βάση τις προσεγγίσεις του περιεχομένου της σύστασης προκειμένου να αντιμετωπίσουμε το πρόβλημα τις ανεπάρκειας, σύμφωνα με τις έρευνες που δημοσιεύτηκαν από τους Balabanovic και Shoham το 1997, τον Basu και την ομάδα του το 1998, τον Condiff και την ομάδα του το 1999, τον Good και την ομάδα του το 1999, τον Huang και την ομάδα του το 2002, τον Pazzani το 1999 και τον Sarwar και την ομάδα το 1998. Μία τέτοια προσέγγιση δεν θεωρεί μόνο τις αλληλεπιδράσεις των καταναλωτικών προϊόντων που αγοράστηκαν στο παρελθόν, αλλά και τις ομοιότητες μεταξύ των προϊόντων ή των αντικειμένων που προέρχονται από τις εγγενείς ιδιότητες ή τα χαρακτηριστικά τους. Γίνεται αναφορά σε αυτήν την προσέγγιση ως υβριδική προσέγγιση. Οι περισσότερες από τις προηγούμενες μελέτες που χρησιμοποιούν την υβριδική προσέγγιση έδειξαν σημαντική βελτίωση της ποιότητας της σύστασης πάνω από την προσέγγιση βασισμένη στους χρήστες, όπως συζητήθηκε παραπάνω. Ωστόσο η υβριδική προσέγγιση απαιτεί πρόσθετες πληροφορίες σχετικά με τα προϊόντα και μία μέτρηση για τον υπολογισμό του νοήματος των ομοιοτήτων μεταξύ τους. Στην πράξη τέτοιες πληροφορίες για το προϊόν ίσως είναι δύσκολο ή έχει μεγάλο κόστος να δοθούν και επομένως η σχετικός υπολογισμός μπορεί να μην είναι στην πραγματικότητα άμεσα διαθέσιμος.

Η έρευνα μας ασχολήθηκε σε διαφορετικό, από τα προηγούμενα, πλαίσιο με το πρόβλημα της ανεπάρκειας. Αντί να μειώσουμε την διάσταση της μήτρας της αλληλεπίδρασης καταναλωτών προϊόντων Α, με φυσικό επακόλουθο να γίνει ανεπαρκής, προτείναμε να εξερευνήσουμε τις μεταβατικές διεθνείς δράσεις μεταξύ των καταναλωτών και των στοιχείων για την αύξηση της μήτρας Α και σαν φυσική συνέπεια είναι η μήτρα να γίνει πυκνή για τους σκοπούς της σύστασης. Η διαίσθηση που υπάρχει πίσω από τις μεταβατικές αλληλεπιδράσεις μπορεί να εξηγηθεί από το ακόλουθο παράδειγμα. Ας υποθέσουμε ότι οι χρήστες C1 και C2 έχουν αγοράσει το βιβλίο P1 και οι χρήστες C2 και C3 έχουν αγοράσει το βιβλίο P2. Βασισμένη στο πρότυπο των προσεγγίσεων του συνεργατικού φιλτραρίσματος τα οποία δεν θεωρούν τις μεταβατικές αλληλεπιδράσεις, θα πραγματοποιηθεί σύνδεση του χρήστη C1 με τον χρήστη C2, όπως επίσης το ίδιο ισχύει και για τους χρήστες C2 και C3. Ωστόσο δεν θα πραγματοποιηθεί σύνδεση μεταξύ των χρηστών C1 και C3. Μία προσέγγιση που θα συμπεριλαμβάνει τιις μεταβατικές αλληλεπιδράσεις, ωστόσο θα αναγνωρίσει τη συνειρμική σχέση που υπάρχει ανάμεσα στους χρήστες C1 και C3, επομένως θα εισάγει τις μεταβατικές αλληλεπιδράσεις ώστε η μήτρα αλληλεπίδρασης καταναλωτικών προϊόντων Α να προβεί σε συστάσεις.

Η παρούσα έρευνα εστιάζεται στην ανάπτυξη μίας υπολογιστικής προσέγγισης για την διερεύνηση ομοιοτήτων ανάμεσα στο μεταβατικό χρήστη και το στοιχείο με την αντιμετώπιση του προβλήματος της ανεπάρκειας μέσα στα πλαίσια του συνεργατικού φιλτραρίσματος.

Στο επόμενο κεφάλαιο πραγματοποιείται μία γενική παρουσίαση μοντέλων πλαισίου και γίνεται μία συζήτηση πάνω στις υπάρχουσες έρευνες που σχετίζονται με τον υπολογισμό και την εφαρμογή των μεταβατικών ενώσεων πάνω στην ανάκτηση πληροφοριών και στα συστήματα σύστασης.

Κεφάλαιο 3- Η ΣΥΣΤΗΣΗ ΤΗΣ ΔΙΑΜΟΡΦΩΣΗΣ ΩΣ ΣΥΝΕΙΡΜΙΚΟ ΠΡΟΒΛΗΜΑ ΤΗΣ ΑΝΑΚΤΗΣΗΣ

3.1 Η Συνειρμική Ανάκτηση που βασίζεται στα Μοντέλα των Γραφημάτων

Στο συγκεκριμένο υποκεφάλαιο θα αναφέρουμε ότι η δυνητική αξία των μεταβατικών ενώσεων έχει αναγνωριστεί από τους ερευνητές οι οποίοι εργάζονται πάνω στον τομέα των συστημάτων σύστασης όπως είναι για παράδειγμα οι μελέτες που δημοσιεύτηκαν από τους Billsus και Pazzani το 1998 και τον Sarwar και την ομάδα του το 2002.Η εξερεύνηση των μεταβατικών ενώσεων διεξάγεται τυπικά πάνω σε ένα μοντέλο γραφήματος σύστασης που βασίζεται σε δύο λόγους. Ο πρώτος λόγος είναι ότι μία γραφική παράσταση ή ένα μοντέλο που βασίζεται σε δίκτυο είναι εύκολο να ερμηνεύσει και να παρέχει ένα φυσικό και γενικό πλαίσιο για πολλούς και διαφορετικούς τύπους εφαρμογών όπως για παράδειγμα τα συστήματα σύστασης. Ενώ ο δεύτερος λόγος, αφορά στην ύπαρξη ενός πλούσιου συνόλου γραφικών παραστάσεων που βασίζονται στους αλγορίθμους έχουν άμεση εφαρμογή όταν το έργο της σύστασης διαμορφώνεται θεωρητικά ως γραφική παράσταση του προβλήματος.

Παρακάτω θα γίνε μία σύντομη έρευνα γραφικών παραστάσεων τριών αντιπροσωπευτικών μοντέλων που βασίζονται στην εξερεύνηση των μεταβατικών σχέσεων. Ο Aggrawal και η ομάδα του το 1999 εισήγαγε ένα μοντέλο σύστασης βασισμένο σε μία γραφική παράσταση με κατεύθυνση προς τους χρήστες. Στο συγκεκριμένο μοντέλο, ένας κατευθυνόμενος σύνδεσμος πραγματοποιεί εκκίνηση από τον χρήστη C1 και τερματίζει στον χρήστη C2, που αυτό σημαίνει ότι η συμπεριφορά του χρήστη C2 είναι ισχυρά προβλέψιμη έναντι της συμπεριφοράς του C1. Οι συστάσεις δημιουργήθηκαν από μια μικρή εξερεύνηση υποδεικνύουν ισχυρή πρόβλεψη στις διαδρομές που ενώνουν πολλούς χρήστες. Ο Mirza σε ατομική έρευνα το 2001,αλλά και με την ομάδα του το 2003, μία γραφική παράσταση κοινωνικού δικτύου των χρηστών προκειμένου να υπάρξει παροχή συστάσεων. Οι σύνδεσμοι που προκύπτουν από τη γραφική παράσταση κοινωνικού δικτύου, ορίζονται, μεταξύ δύο χρηστών οι οποίοι έχουν συμφωνήσει σε αξιολογήσεις για τουλάχιστον ένα δεδομένο αριθμό αντικειμένων. Τόσο τα μοντέλα που εισήγαγε ο Aggraval όσο και αυτά του Mirza τονίζουν την σημασία της χρήσης των γραφικών παραστάσεων, οι οποίες χρησιμοποιούνται από τις ενώσεις των χρηστών, προκειμένου να γίνει εξερεύνηση των μεταβατικών ενώσεων. Ο Huang και η ομάδα σε προηγούμενη έρευνα που δημοσίευσε το 2003, πρότεινε την ανάπτυξη ενός άλλου μοντέλου γραφικής παράστασης, που βασίζεται στο συνεργατικό φιλτράρισμα και να περιλαμβάνει, τόσο τους χρήστες όσο και τα σημεία τους στη γραφική παράσταση. Το μοντέλο αυτό έχει σαν κύριο στόχο την δημιουργία συστάσεων και την ανάληψη πρόσθετων τύπων εισροών σε ένα μη προσδιορισμένο πλαίσιο.

Τα μοντέλα των γραφικών παραστάσεων που αναφέρθηκαν παραπάνω βασίζονται στην παροχή της βασικής αναπαράστασης και του εκσυγχρονισμού του πλαισίου, που είναι απαραίτητο για την έρευνα μας σχετικά με το πρόβλημα της ανεπάρκειας, ενώ μας δίνουν τη δυνατότητα να σχεδιάσουμε μία αναλογία ανάμεσα στα συστήματα σύστασης και στα συστήματα συνεργατικής ανάκτησης. Αυτή η αναλογία, με τη σειρά της υποδηλώνει ότι το πρόβλημα της ανεπάρκειας, είναι δυνατό δυνατόν δυνητικά να αντιμετωπιστεί με αποτελεσματικότητα αρκεί να χρησιμοποιηθούν υπολογιστικά συστήματα, όπως για παράδειγμα ένας επεκτάσιμος αλγόριθμος ενεργοποίησης, ο οποίος εφαρμόστηκε με επιτυχία στη συνειρμική ανάκτηση.

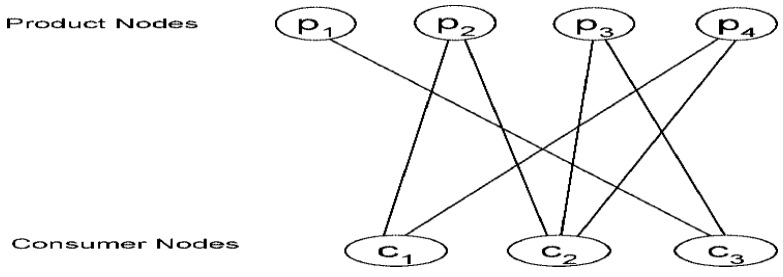
Σε αυτό το κεφάλαιο, θα συζητήσουμε με λεπτομέρειες, πως η διαδικασία της σύστασης μπορεί να διατυπωθεί σαν συνειρμικό πρόβλημα ανάκτησης, αλλά και πως οι επεκτάσιμοι αλγόριθμοι ενεργοποίησης θα χρησιμοποιηθούν για να βοηθήσουμε στην επίλυση του προβλήματος της ανεπάρκειας. Με αυτόν τον τρόπο καταλήγουμε στο συμπέρασμα ότι παρουσιάζοντας μία έρευνα ερωτήσεων σαν κύριο στόχο έχουμε την εφαρμογή των επεκτάσιμων αλγόριθμων ενεργοποίησης μέσα στα πλαίσια των συστημάτων σύστασης.

3.2 Το Συνεργατικό Φιλτράρισμα με εφαρμογή της Συνειρμικής Ανάκτησης

Η συνειρμική ανάκτηση πληροφοριών έχει τις ρίζες της στις στατιστικές μελέτες των ενώσεων ανάμεσα στους όρους και τα έγγραφα που υπάρχουν σε μία συλλογή κειμένων. Η βασική ιδέα πίσω από τη συνειρμική ανάκτηση είναι η δημιουργία μίας γραφικής παράστασης ή ενός μοντέλου δικτύου των εγγράφων, που περιλαμβάνει όρους δεικτών και ερωτήματα, και στη συνέχεια να πραγματοποιήσει διερεύνηση των μεταβατικών ενώσεων μεταξύ των όρων και των εγγράφων, που χρησιμοποιεί η συγκεκριμένη γραφική παράσταση, έτσι ώστε να επιτευχθεί η βελτίωση ποιότητας της ανάκτησης πληροφοριών. Για παράδειγμα, το γενικευμένο μοντέλο διανυσματικού μοντέλου, όπως το περιέγραψε ο Wonk και η ομάδα του σε μελέτη που δημοσιεύτηκε το 1985 παρουσιάζει το έγγραφο από ένα διάνυσμα και τις ομοιότητες που προκύπτουν σε όλα τα άλλα έγγραφα που υπάρχουν μέσα στο σώμα. Οι ομοιότητες των ενώσεων που υπάρχουν μεταξύ των εγγράφων, ορίζονται από τις μεταβατικές ενώσεις, που προκύπτουν από τους όρους των δεικτών, κατασκευάζονται και χρησιμοποιούνται κατευθείαν για την υποστήριξη της ανάκτησης των πληροφοριών. Ένας αριθμός τεχνικών έχει προταθεί για την κατασκευή και χρήση αυτών των ενώσεων των δικτύων προκειμένου να επιτευχθεί η ανάκτηση των πληροφοριών. Παραδείγματα τέτοιων τεχνικών είναι οι διάφορες στατιστικές προσεγγίσεις, οι οποίες προτάθηκαν από τους Crouch και Yang το 1992, τα νευρωνικά δίκτυα, που προτάθηκαν από τους Jung και Raghavan το 1990, οι γενετικοί αλγόριθμοι, που προτάθηκαν από τον Gordon το 1998, όπως επίσης και οι επεκτάσιμες προσεγγίσεις των ενεργοποιήσεων, που προτάθηκαν από τους Cohen και Kjeldsen το 1987 αλλά και από τους Salton και Buckley το 1988.

Η ομοιότητα που υπάρχει ανάμεσα στη συνεταιριστική ανάλυση και το συνεργατικό φιλτράρισμα έχει αναγνωριστεί από ορισμένες πρόσφατες μελέτες, όπως η συγκεκριμένη που δημοσιεύτηκε από τους Soborof και Nicolas το 2000. Στη συνειρμική ανάκτηση, τα έγγραφα αντιπροσωπεύονται από όρους δεικτών. Ταυτόχρονα η σημασιολογία του όρου δείκτη μπορεί επίσης να αντιπροσωπεύεται από το σύνολο των εγγράφων που το περιλαμβάνουν. Με τον ίδιο τρόπο, στα συνεργατικά φιλτραρίσματα, οι προτιμήσεις των χρηστών είναι δυνατόν να αντιπροσωπεύονται από τα στοιχεία και τις αλληλεπιδράσεις που προκύπτουν ανάμεσα τους. Τα εγγενής χαρακτηριστικά ενός στοιχείου μπορούν επίσης να αντιπροσωπευτούν από τους χρήστες και τις αλληλεπιδράσεις του.

Το ακόλουθο παράδειγμα απεικονίζει την ιδέα της εξερεύνησης των μεταβατικών στα συστήματα συστάσεων. Χρησιμοποιώντας το συμβολισμό που αναπτύξαμε στο υποκεφάλαιο 2.2 οι τελευταίες συναλλαγές μπορούν να χρησιμοποιηθούν στην κάτωθι μήτρα αλληλεπιδράσεων καταναλωτικών προϊόντων.



Σχήμα 2: Ένα απλό παράδειγμα μεταβατικών ενώσεων στο συνεργατικό φιλτράρισμα

$$\begin{array}{c} \\ c1 \\ c2 \\ c3 \end{array} \begin{array}{c} \begin{array}{cccc} p1 & p2 & p3 & p4 \end{array} \\ \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix} \end{array} \quad (7)$$

Σε αυτό το σημείο μπορούμε να σημειώσουμε, ότι στην παρούσα διατριβή, υποθέτουμε ότι οι μόνες πληροφορίες που διαθέτουμε για το σύστημα συστάσεων είναι η πιο πάνω μήτρα. Επομένως το γράφημα που φαίνεται στο Σχήμα 1 είναι ένα διμερές γράφημα. Παρατηρούμε επίσης ότι σε ένα διμερές γράφημα, οι κόμβοι διαχωρίζονται σε δύο διακριτές τιμές. Οι δεσμοί που προκύπτουν ανάμεσα από τα ζεύγη των διαφορετικών συνόλων κόμβων θεωρείται παραδεκτή, ενώ συνδέσεις ανάμεσα στους κόμβους από το ίδιο σύνολο δεν επιτρέπονται.

Υποθέτουμε τώρα ότι το σύστημα προβαίνει σε σύσταση των προϊόντων στον καταναλωτή C1. Ο πρότυπος αλγόριθμος συνεργατικού φιλτραρίσματος θα προχωρήσει σε επιβραβεύσεις, οι οποίες θα είναι βασισμένες στις ομοιότητες, μεταξύ των καταναλωτών C1, C2 και C3. Οι ομοιότητες μεταξύ των καταναλωτών C1 και C2 είναι προφανείς, λόγω του προηγούμενου κοινού ιστορικού αγορών των στοιχείων P2 και P4. Σαν αποτέλεσμα, το στοιχείο P3 συνίσταται για τον καταναλωτή C1, επειδή έχει ήδη αγοραστεί από τον C2. Όμως δεν μπορεί να βρεθεί ισχυρή ομοιότητα ανάμεσα στους C1 και C3. Επομένως, το στοιχείο P1, το οποίο έχει αγοραστεί από τον C3, δεν πρέπει να συσταθεί στον C1.

Η παραπάνω προσέγγιση της σύστασης μπορεί εύκολα να υλοποιηθεί σε μία γραφική παράσταση βασισμένη στους υπολογισμούς των συσχετίσεων μεταξύ των κόμβων των καταναλωτών και των κόμβων των καταναλωτών. Σε αυτό το πλαίσιο, σύνδεση μεταξύ δύο κόμβων καθορίζεται από την ύπαρξη και το μήκος της διαδρομής , ή των διαδρομών, που τα συνδέει. Στο πρότυπο των προσεγγίσεων του συνεργατικού φιλτραρίσματος, θα θεωρήσουμε σαν δεδομένες τις διαδρομές που έχουν μήκος ίσο με 3. Σαν παράδειγμα, η συσχέτιση μεταξύ του καταναλωτή C1 και του στοιχείου P3 καθορίζεται από το μήκος του που είναι ίσο με 3 και συνδέει τον καταναλωτή C1 με το στοιχείο P3. Είναι εύκολο να δούμε από το Σχήμα 1 ότι υπάρχουν δύο μονοπάτια που συνδέουν το C1 με το P3 και είναι τα ακόλουθα C1-P2-C2-P3 και C1-P4-C2-P3. Από την ισχυρή συσχέτιση που προκύπτει καταλήγουμε το στοιχείο P3 να δώσει συστάσεις για αγορά από τον καταναλωτή C1. Η σύνδεση ανάμεσα στον C1 και το P1 δεν υφίσταται γιατί δεν υπάρχει μήκος διαδρομής που να ισούται με 3. Διαισθητικά, όσο υψηλότερος είναι ο αριθμός των διαδρομών που συνδέουν διακριτικά έναν κόμβο καταναλωτών και ένα κόμβο προϊόντων τόσο υψηλότερη είναι η συσχέτιση μεταξύ αυτών των δύο κόμβων. Συνεπώς το προϊόν είναι πολύ πιθανό να συσταθεί στον καταναλωτή.

Αν επεκτείνουμε την πιο πάνω προσέγγιση μας για να εξερευνήσουμε και να ενσωματώσουμε τις μεταβατικές ενώσεις, αρκεί να κατασκευαστεί μία απλή γραφική παράσταση. Εξετάζοντας τα μονοπάτια, τα οποία το μήκος υπερβαίνει το 3, το μοντέλο θα είναι σε θέση να ερευνήσει μεταβατικές ενώσεις που προκύπτουν. Για παράδειγμα υπάρχουν δύο μονοπάτια τον καταναλωτή C1 με το στοιχείο P1 και είναι τα ακόλουθα. C1-P2-C3-P3-C1 και C1-P4-C2-C3-

P3-P4. Έτσι, το στοιχείο P1 μπορεί επίσης να συσταθεί στον καταναλωτή C1 έχοντας σαν απαραίτητη προϋπόθεση να λάβουμε σαν δεδομένες τις μεταβατικές ενώσεις.

Σε αυτό το σημείο θα γίνει μία παρουσίαση των κύριων βημάτων μίας νέας προσέγγισης συνεργατικού φιλτραρίσματος που λαμβάνει ρητώς τις μεταβατικές ενώσεις που προκύπτουν για να αντιμετωπιστεί το πρόβλημα της ανεπάρκειας.

Η προσέγγιση μας παίρνει σαν είσοδο την μήτρα αλληλεπίδρασης καταναλωτών και προϊόντων A. Στη συνέχεια κατασκευάζεται μία διμερής γραφική παράσταση. Οι συστάσεις κατασκευάζονται βασισμένες στα ζεύγη των κόμβων καταναλωτών αλλά και στοιχειών. Λαμβάνοντας υπ'οψιν έναν κόμβο καταναλωτών c_t και ένα κόμβο των στοιχείων p_j, η συσχέτιση που υπάρχει μεταξύ τους, η οποία συμβολίζεται με $a(c_t,p_j)$, αποτελεί το άθροισμα των βαρών όλων των διακριτικών μονοπατιών που συνδέουν το c_t και το p_j. Στον συγκεκριμένο υπολογισμό, θα λάβουμε υπ'όψιν μόνο τα μονοπάτια εκείνα, που το μέγιστο επιτρεπόμενο μήκος είναι μικρότερο ή ίσο με το M. Το όριο M είναι μία παράμετρος, η οποία μπορεί να ελεγχθεί από τον σχεδιαστή συστημάτων σύστασης. Σαν παράδειγμα μπορούμε να αναφέρουμε ότι το όριο M είναι ίσο με 3 σε πολλές προσεγγίσεις, όπως αναφέρεται στις μελέτες που δημοσιεύτηκαν από τον Breese και την ομάδα του το 1998, τον Resnik και την ομάδα του το 1994, όπως και τον Sarwar και την ομάδα του το 2001. Είναι εύκολο να παρατηρήσουμε ότι το M πρέπει να είναι περιττός προκειμένου οι μεταβατικές ενώσεις να παρουσιαστούν σε μία διμερής γραφική παράσταση. Για μία δεδομένη διαδρομή μήκους x, όπου x≤M, το βάρος της διαδρομής υπολογίζεται ως a^x, όπου a είναι μία σταθερά μεταξύ 0 και 1, εξασφαλίζοντας ότι οι μακρύτερες διαδρομές έχουν μικρότερο αντίκτυπο. Η ιδιαίτερη αξία που προκύπτει για την σταθερά a είναι δυνατόν να προσδιοριστεί από τον σχεδιαστή του συστήματος με βάση τα χαρακτηριστικά του στο υποκείμενο πεδίο εφαρμογής του, Σε εφαρμογές όπου οι μεταβατικές ενώσεις μπορούν να έχουν ισχυρή επίδραση από τα ενδιαφέροντα των καταναλωτών, η σταθερά a πρέπει να πάρει τιμή κοντά στη μονάδα, ενώ στις εφαρμογές, όπου οι μεταβατικέςενώσεις τείνουν να λάβουν λίγες πληροφορίες, η σταθερά a πρέπει να πάρει τιμή κοντά στο 0. Χρησιμοποιούμε το παράδειγμα που φαίνεται στο σχήμα 1 για να απεικονίσουμε τον παραπάνω υπολογισμό. Όταν ορίζουμε το μήκος M=3, που θεωρείται σταθερό συνεργατικό φιλτράρισμα, έχουμε τα ακόλουθα αποτελέσματα $a(c1,p3)=0,5^3+0.5^3=0.25$ και $a(c1,p1)=0$. Ενώ όταν το M=5 τότε $a(c1,p3)=0.5^3+0.5^3=0.25$ και $a(c1,p1)=0.5^5+0.5^5=0.0625$.

Για τον καταναλωτή c_t, ο παραπάνω υπολογισμός συσχετίσεων επαναλαμβάνεται για όλα τα στοιχεία p_jE P. Τα στοιχεία του συνόλου P ταξινομούνται κατά φθίνουσα σειρά με βάση τον κόμβο $a(c_t, p_j)$. Τα πρώτα k αντικείμενα, περιλαμβάνοντας και τα αντικείμενα που έχουν αγοραστεί από τον c_t στο παρελθόν, από την φθίνουσα λίστα συστήνονται στον καταναλωτή c_t.

Σε αυτό το σημείο θα γίνει περιγραφή της παραπάνω διαδικασίας κάνοντας χρήση του συμβολισμού της μήτρας που συμπεριλάβαμε στο υποκεφάλαιο 2.2. Λαμβάνοντας σαν δεδομένο ότι έχουμε τη μήτρα αλληλεπίδρασης καταναλωτικών προϊόντων A, η παράμετρος του βάρους του μονοπατιού που συμβολίζεται με a, και το μέγιστο επιτρεπόμενο μήκος του μονοπατιού που το ορίζουμε με M, τότε οι μεταβατικές ενώσεις που προκύπτουν από τα αντικείμενα και τους καταναλωτές δίνεται από τη μήτρα A_a^M ενώ το πεδίο τιμών που λαμβάνει καθορίζεται από την ακόλουθη σχέση

$$A^M_a = \begin{cases} aA, & \text{εάν } M=1 \\ a^2 * A*A^T * A_a^{M-2}, & \text{εάν } M=3,5,7\ldots \end{cases} \quad (8)$$

Με βάση το παραπάνω αριθμητικό παράδειγμα και θεωρώντας ότι η σταθερά a ισούται με 0.5 οι μεταβατικές ενώσεις που προκύπτουν όταν M=3 και M=5 δίδονται παρακάτω

$$A^3_{0.5}=\begin{pmatrix} 0 & 0.5 & 0.25 & 0.5 \\ 0.125 & 0.625 & 0.5 & 0.625 \\ 0.25 & 0.125 & 0.375 & 0.125 \end{pmatrix}$$

$$A^5_{0.5}=\begin{pmatrix} 0.0625 & 0.5625 & 0.375 & 0.5625 \\ 0.15625 & 0.75 & 0.59375 & 0.75 \\ 0.15625 & 0.21875 & 0.3125 & 0.21875 \end{pmatrix}$$

Μία βασική πρόκληση της εφαρμογής της παραπάνω προσέγγισης είναι ότι ο υπολογισμός του A* απαιτεί εκτεταμένη χρήση τέτοιων πόρων, ειδικά όταν υπάρχουν πολλοί κόμβοι τόσο για τον καταναλωτή όσο και για το προϊόν, όπως για παράδειγμα είναι το τυπικό των μεγάλων ιστοσελίδων επικοινωνίας ή όταν το μονοπάτι M είναι μεγάλο. Αυτή η εξέταση αποτελεί κίνητρο στη διατριβή μας για εκ νέου εφαρμογή των σχετικών συνειρμικών, αλλά και των επεκτάσιμων αλγορίθμων ενεργοποίησης για να εκτελεστεί ο υπολογισμός της ένωσης. Στο επόμενο υποκεφάλαιο παρουσιάζουμε την συνειρμική ανάκτηση που είναι βασισμένη στην προσέγγιση της σύστασης.

3.3 Διάδοση της ενεργοποίησης με παράλληλη αναζήτηση γραφημάτων

Οι τεχνικές της διάδοσης της ενεργοποίησης έχουν μεγάλη εξάπλωση και εφαρμόζονται σε συνειρμικές ανακτήσεις τόσο στο μοντέλο της ανθρώπινης γνωστικής λειτουργίας και επεξεργασίας πληροφοριών, όπως αναφέρουν οι Collins και Loftus σε μελέτη που δημοσιεύτηκε το 1975, όσο και σαν ένα υπολογιστικό μηχανισμό για την διαδικασία της εξερεύνησης των συλλογικών δικτύων. Οι τεχνικές της διαδιδόμενης ενεργοποίησης, προσφάτως έχουν επίσης εφαρμογή στη διερεύνηση διάφορων τύπων δικτύων, συμπεριλαμβανομένων του διαδικτύου, των δικτύων παραπομπής, όπως και παρόμοιων δικτύων περιεχομένου, σύμφωνα με μελέτες που δημοσιεύτηκαν από τον Bollen και την ομάδα του το 1999, τους Crestani και Lee το 2000, άλλα και τον Pirolli και την ομάδα του το 1996. Στην έρευνα μας θα δώσουμε έμφαση της διαδιδόμενης ενεργοποίησης σαν υπολογιστική μέθοδος η οποία σκοπεύει σε μία αποτελεσματικότερη διερεύνηση των μεταβατικών ενώσεων, που προκύπτουν μεταξύ των καταναλωτών και των προϊόντων, στο συνεργατικό φιλτράρισμα.

Γενικά, θα μπορούσαμε να αναφέρουμε ότι στην προσέγγιση γραφήματος εξερεύνησης, η διαδιδόμενη ενεργοποίηση σε πρώτο στάδιο ενεργοποιεί ένα επιλεγμένο υποσύνολο των κόμβων μέσα από μία δεδομένη γραφική παράσταση τους οποίους τους ορίζει σαν εναρκτήριους και σε δεύτερο στάδιο ακολουθώντας τους συνδέσμους των επαναλήψεων ενεργοποιεί τους κόμβους που μπορούν να επιτευχθούν άμεσα από τους κόμβους που είναι

σχεδόν ενεργοποιημένοι. Χρησιμοποιώντας το παράδειγμά μας, ο κόμβος c_2 , ο οποίος εκπροσωπεί τον στοχευμένο καταναλωτή και χρειάζεται συστάσεις, αποτελεί τον εναρκτήριο κόμβο της διαδιδόμενης ενεργοποίησης και είναι ο πρώτος που ενεργοποιείται. Μετά την πρώτη επανάληψη, οι απευθείας ενεργοποιημένοι κόμβοι ο p_2 και ο p_4 ενεργοποιούνται. Κατά τη διάρκεια της δεύτερης επανάληψης και οι τρεις ενεργοί κόμβοι c_1, p_2 και p_4 ενεργοποιούν τους απευθείας γειτονικούς κόμβους. Έτσι τα επίπεδα ενεργοποίησης των p_2 και p_4 αναβαθμίζονται και ένας επιπρόσθετος κόμβος, συγκεκριμένα ο c_2 ενεργοποιείται. Η διαδικασία της ενεργοποίησης επαναλαμβάνει και το επίπεδο ενεργοποίησης που εξαπλώνεται σταδιακά, από τον εναρκτήριο κόμβο στους απευθείας ή μη συνδεδεμένους κόμβους περιλαμβάνοντας και τον κόμβο αντικειμένων p_1.

Οι υπο-κατασκευή υλοποιήσεις της διαδιδόμενης ενεργοποίησης, έχουν πρόσβαση σε όλους τους κόμβους και θα έχουν σα τελικό στόχο την ενεργοποίηση σε ορισμένο επίπεδο. Σε άλλα σχήματα της διαδιδόμενης ενεργοποίησης, η συγκεκριμένη διαδικασία συνεχίζεται μέχρις ότου να εκπληρωθούν ορισμένα προκαθορισμένα κριτήρια. Ο Salton και ο Buckley το 1988 σχεδίασαν και αξιολόγησαν, χρησιμοποιώντας διάφορες τεχνικές διαδιδόμενης ενεργοποίησης, την ανάκτηση πληροφοριών σαν έννοια της επέκτασης της έρευνας του λεξιλογίου αλλά και την συμπλήρωση στα ανακτημένα κείμενα. Οι περιορισμένες μέθοδοι διαδιδόμενων ενεργοποιήσεων επισημάνθηκαν από τους Cohen και Kjeldsen το 1987, έχοντας σαν κύριο στόχο τη βελτίωση της υπολογιστικής απόδοσης με παράλληλη διατήρηση των επιδόσεων εξερεύνησης αφού περιορίσουμε την διαδικασία της, σε κάθε ένα από τα στάδια ενεργοποίησης της διασποράς, έτσι ώστε μόνο ένα υποσύνολο των ενεργών κόμβων να ενεργοποιηθούν. Ο Chen και Dhar το 1991 πρότειναν μία έρευνα του αλγορίθμου Branch-and-bound για τις διαδιδόμενες ενεργοποιήσεις, η οποία αντιμετωπίζει την επέκταση της ενεργοποίησης σαν μία παραλλαγή της διαδικασίας διάσχισης της κατάστασης. Ο Chen και η ομάδα του το 1993, αλλά και αργότερα με την εταιρεία Ng το 1995. Εισήγαγε άλλες διαδιδόμενες τεχνικές ενεργοποίησης χρησιμοποιώντας τον αλγόριθμο αναζήτησης δικτύου Hopfield. Αυτό το νευρωνικό δίκτυο, το οποίο βασίζεται στην ενεργοποίηση των παράλληλων κόμβων, ολοκληρώνει την διαδικασία της επέκτασης όταν δίκτυο καταλήξει σε μία σταθερή κατάσταση. Τόσο οι προσεγγίσεις των αλγορίθμων Branch-and-Bound αλλά και του αλγορίθμου Hopfield εφαρμόστηκαν στην εξερεύνηση της έννοιας της διαδιδόμενης ενεργοποίησης που βασίζεται στους χώρους δικτύων, όπως υποστηρίζει ο Chen και η ομάδα του το 1993 αλλά και ο ίδιος με την εταιρεία Ng το 1995.

Στο επόμενο υποκεφάλαιο, παρουσιάζουμε τα ερευνητικά ερωτήματα που προκύπτουν από την εφαρμογή των διαδιδόμενων τεχνικών ενεργοποίησης του συνεργατικού φιλτραρίσματος.

3.4 Ερευνητικές ερωτήσεις

Το κεντρικό θέμα της έρευνας μας είναι η αποδοχή της διαδιδόμενης ενεργοποίησης για να επιλυθεί το πρόβλημα της ανεπάρκειας που προκύπτει από την χρήση των συστημάτων σύστασης. Στοχεύουμε να ερευνήσουμε πόσο σημαντική βελτίωση μπορεί να επιτευχθεί, στην ποιότητα της σύστασης, από την αποδοχή των τεχνικών της διαδιδόμενης ενεργοποίησης, προκειμένου να γίνει εξερεύνηση των μεταβατικών ενώσεων ανάμεσα στους χρήστες και τα αντικείμενα μέσω των συστημάτων του συνεργατικού φιλτραρίσματος. Επίσης στοχεύουμε να αποκομίσουμε ένα κέρδος κατανοώντας την συμπεριφορά των συστημάτων σύστασης που κάνουν χρήση των μεταβατικών ενώσεων σε σχέση με την ποσότητα των δεδομένων συναλλαγής που διατίθενται σε αυτά τα συστήματα. Αυτό σημαίνει ότι διαισθητικά, όταν η μήτρα αλληλεπίδρασης καταναλωτικών προϊόντων είναι ανεπαρκής η προσέγγιση που βασίζεται στις μεταβατικές ενώσεις αναμένεται να ξεπεράσει τις προσεγγίσεις του συνεργατικού

φιλτραρίσματος που δεν χρησιμοποιούν μεταβατικές ενώσεις, επειδή οι χρήσιμες πληροφορίες είναι ήδη διαθέσιμες στις υπάρχουσες μεταβατικές ενώσεις. Όταν η μήτρα γίνεται πολύ πυκνή, δηλαδή όταν πολλά δεδομένα συναλλαγών είναι διαθέσιμα, ωστόσο, αναμένουμε ότι οι μεταβατικές ενώσεις θα έχουν περιορισμένη ή αρνητική επίδραση πάνω στην παρουσίαση των συστημάτων σύστασης.

Για τις υπάρχουσες προσεγγίσεις συνεργατικού φιλτραρίσματος οι οποίες δεν εξερευνούν τις μεταβατικές ενώσεις, στο σημείο αυτό μπορούμε να αναφέρουμε το σταθερό συνεργατικό φιλτράρισμα, η πυκνότητα της μήτρας αλληλεπίδρασης καταναλωτών και προϊόντων καθορίζεται από το ύψος της συνολικής ποιότητας της σύστασης με βάση τα όσα αναφέρει ο Sarwar και η ομάδα του σε έρευνα που δημοσίευσε το 2000. Ωστόσο, για τις προσεγγίσεις που βασίζονται στις μεταβατικές ενώσεις, η υπέρθεση των μεταβατικών ενώσεων σε ένα γράφημα καταναλωτών και υπολογισμού να μην προκαλέσει ανεπάρκεια δεδομένων, αλλά η ανεπάρκεια να προκληθεί όταν τα δεδομένα χρησιμοποιηθούν για να δείξουν τις προτιμήσεις των χρηστών. Το σχήμα 2 δείχνει την αναμενόμενη εκτέλεση των διάφορων ειδών συνεργατικού φιλτραρίσματος όταν διαφέρει η πυκνότητα στα καταναλωτών και προϊόντων.

Σε αντίθεση με τα προηγούμενα, ενδιαφερθήκαμε να εξερευνήσουμε σχετικά πλεονεκτήματα και μειονεκτήματα από διάφορους τύπους αλγορίθμων ενεργοποίησης σε σχέση με την ποιότητα των παραγόμενων συστάσεων και την υπολογιστική τους απόδοση. Στο επόμενο κεφάλαιο θα προβούμε σε υλοποίηση δύο αλγορίθμων και θα καταγράψουμε τα συμπεράσματα μας

Κεφάλαιο 4- ΑΜΒΛΥΝΣΗ ΤΟΥ ΠΡΒΛΗΜΑΤΟΣ ΤΗΣ ΑΝΕΠΑΡΚΕΙΑΣ ΣΤΟ ΣΥΝΕΡΓΑΤΙΚΟ ΜΕ ΧΡΗΣΗ ΜΙΑΣ ΠΡΟΣΑΡΜΟΣΜΕΝΗΣ ΑΠΟΣΤΑΣΗΣ ΚΑΙ ΜΙΑΣ ΜΕΘΟΔΟΥ ΠΟΥ ΒΑΣΙΖΕΤΑΙ ΣΤΟ ΓΡΑΦΗΜΑ

Το συνεργατικό φιλτράρισμα είναι η διαδικασία που προβλέπει τα ενδιαφέροντα των καταναλωτών σε διάφορα αντικείμενα, όπως για παράδειγμα σε βιβλία ή ταινίες, βασίζονται σε προτιμήσεις πληροφοριών, η οποίες εξειδικεύονται σε μία φόρμα κατάταξης, που προκύπτει από πολλούς άλλους χρήστες. Ένα από τα σημαντικότερα κλειδιά που περιλαμβάνει το συνεργατικό φιλτράρισμα είναι πως διαπραγματεύεται με ανεπάρκεια δεδομένων από τη στιγμή που οι περισσότεροι χρήστες βαθμολογούν μόνο ένα μικρό αριθμό από αντικείμενα. Σαν πρώτη συνεισφορά σε αυτό το κεφάλαιο είναι η υπολογιστική μέτρηση.

Η μέτρηση της απόστασης βασίζεται στην πιθανότητα και είναι προσαρμοσμένη στο χρήστη που έχει ανεπάρκεια δεδομένων. Μπορεί να χρησιμοποιηθεί στιγμιαία για την συντομότερη γειτονική μέθοδο ή για ένα γράφημα που βασίζεται στις μεθόδους της καταγραφής των άκρων του γραφήματος. Η δεύτερη συνεισφορά είναι ένα λογοτεχνικό και ταυτόχρονα πιθανολογικό γράφημα που είναι βασισμένο στον αλγόριθμο συνεργατικού φιλτραρίσματος και απασχολείται με τον υπολογισμό της απόστασης. Χρησιμοποιώντας τον πολλαπλασιαστικό πιθανολογικό αλγόριθμο του συνεργατικού φιλτραρίσματος στα γραφήματα στόχων, δεν θα χρησιμοποιήθει η βαθμολογία από τους συνδεδεμένους γείτονες αλλά θα χρησιμοποιήσει και την πληροφορία που διατίθεται και για τους μη χρησιμοποιημένους γείτονες.

Τα πειράματα μας δείχνουν ότι και η προσαρμογή των μετρήσεων της απόστασης αλλά και το γράφημα που βασίζεται στο συνεργατικό φιλτράρισμα μας καθοδηγούν σε περισσότερες και ακριβείς προβλέψεις.

4.1 Εισαγωγή

Τα συστήματα σύστασης είναι συστήματα λογισμικού σχεδιασμένα να προτείνουν, έχοντας μια τεράστια κατηγορία από διαθέσιμα αντικείμενα, όπως για παράδειγμα βιβλία, ταινίες, μουσικά κομμάτια, νέα άρθρα και πολλά άλλα αντικείμενα τα οποία είναι τα πιο δημοφιλή σύμφωνα με τα ενδιαφέροντα του χρήστη. Δύο κοινές κλάσεις από αλγορίθμους χρησιμοποιούνται στα συστήματα σύστασης και είναι ο αλγόριθμος που βασίζεται στο περιεχόμενο και το συνεργατικό φιλτράρισμα. Ο αλγόριθμος που βασίζεται στο περιεχόμενο χρησιμοποιεί σαν βάση τις συστάσεις από το περιεχόμενο των αντικειμένων, όπως για παράδειγμα το κείμενο από ένα νέο άρθρο, ενώ ο αλγόριθμος του συνεργατικού φιλτραρίσματος χρησιμοποιεί σαν βάση τις συστάσεις από τις πληροφορίες από τις πληροφορίες, οι οποίες εγγράφονται από το σύστημα, σχετικά με τις προτιμήσεις των διαφόρων χρηστών.

Στα συστήματα συνεργατικού φιλτραρίσματος ίσως χρειαστεί να γίνει εκτίμηση στους χρήστες οι οποίοι έχουν καταχωρηθεί ρητώς, όλοι ή ένα υποσύνολο από αυτούς, στα αντικείμενα, ή στα δυαδικά δεδομένα, που εγγράφει κάθε χρήστη στα αντικείμενα, με τέτοιο τρόπο ώστε αυτός ή αυτή έχει προηγουμένως αγοράσει. Σαν παράδειγμα μπορούμε να αναφέρουμε τα συμφραζόμενα από ένα σύστημα ηλεκτρονικής διαφήμισης. Αυτό το κεφάλαιο θεωρεί την προηγούμενη ρύθμιση. Σε αυτό το σημείο η εισαγωγή δεδομένων μπορεί να αντιπροσωπεύονται με όρους, που προκύπτουν από τη μήτρα εκτίμησης, η οποία έχει μία γραμμή για κάθε χρήστη και μία στήλη για κάθε αντικείμενο, και η οποία συμπεριλαμβάνει τα στοιχεία στις εκτιμήσεις που δίδονται από τους χρήστες στα διάφορα αντικείμενα. Υποθέτουμε ότι η εκτίμηση είναι ένας φυσικός αριθμός και ορίζεται από ένα σταθερό εύρος, παράδειγμα από το ένα έως το 5. Εάν υπάρχουν m χρήστες και n αντικείμενα, τότε η μήτρα εκτίμησης A είναι μία μήτρα m*n, επομένως το στοιχείο $A_{i\ j}$ είτε ανατίθεται να αξιολογηθεί από το χρήστη i για το αντικείμενο j, ή αλλιώς είναι ίσο με την ένδειξη της απώλειας της αξίας εάν ο χρήστης i δεν έχει κάνει εκτίμηση για το αντικείμενο j.

Δεδομένης της μήτρας εκτιμήσεων, ένας αλγόριθμος συνεργατικού φιλτραρίσματος μπορεί να συστήσει έναν αριθμό από αντικείμενα σε ένα συγκεκριμένο χρήστη κατά την διάρκεια της πρώτης πρόβλεψης για όλα τα αντικείμενα που δεν έχει εκτιμήσει ακόμα, συνεπώς να προτείνει τα αντικείμενα εκείνα τα οποία έχουν υψηλή πρόβλεψη εκτίμησης. Για παράδειγμα, ο προβλεπόμενος χρήστης i ο οποίος κάνει εκτίμηση για το αντικείμενο j, θα θεωρήσει ένας από τους κοντινότερους γειτονικούς κόμβους k που βασίζεται στον αλγόριθμο CI, την ομάδα όλων των άλλων χρηστών που εκτιμούν το αντικείμενο j, μεταξύ αυτής της έρευνας για τις πιο παρόμοιες τιμές του k, με όρους από την εκτίμηση τους, οπότε παίρνει την πιο συχνά εμφανίσιμη τιμή, μεταξύ των εκτιμήσεων των γειτονικών κόμβων για το αντικείμενο j, σαν πρόβλεψη. Εάν η πρόβλεψη εκτίμησης επιτρέπει την εισαγωγή πραγματικών αριθμών, είναι πιθανό εναλλακτικά να υπολογιστεί ο μέσος όρος που προκύπτει από τις συχνά εμφανιζόμενες τιμές.

Μία από τις πιο σημαντικές επιδράσεις του συνεργατικού φιλτραρίσματος είναι πως συμφωνεί με την απώλεια δεδομένων. Τυπικά οι χρήστες εκτιμούν μόνο ένα πολύ μικρό ποσοστό αντικειμένων από τη μήτρα επομένως η βαθμολογία της μήτρας να έχει απώλειες, συμπεριλαμβανομένων πολλών χαμένων τιμών. Η ανεπάρκεια των δεδομένων ίσως είναι επιζήμια για την ακρίβεια για τους ακόλουθους δύο λόγους. Ο πρώτος λόγος είναι, ότι η αξιοπιστία της εκτίμησης ανάμεσα σε παρόμοιους χρήστες γίνεται δύσκολη, επειδή, λόγω των πολλών τιμών που λείπουν, ίσως πολλά αντικείμενα να εκτιμηθούν από πολλούς χρήστες, οπότε ως εκ τούτου να περιέχει κάποιες πληροφορίες σχετικά με την ομοιότητα τους. Ο

δεύτερος λόγος είναι, ότι για να γίνει πρόβλεψη εκτίμησης του $A_{i\ j}$, πρέπει να υπάρχει μία επάρκεια σε σχέση με το χρήστη i οι οποίοι επίσης να εκτιμούν το αντικείμενο j. Και αυτό γιατί, λόγω της ανεπάρκειας δεδομένων, θα υπάρξει ένας σχετικά μικρός αριθμός τέτοιων χρηστών με αποτέλεσμα το σύνολο εισόδου για τον πλησιέστερο γείτονα να είναι μικρό.

Σε αυτό το κεφάλαιο προτείνεται μία προσέγγιση που ασχολείται με άγνωστες τιμές οι οποίες προκύπτουν από τη μετάβαση σε μία πιθανολογική αναπαράσταση του προβλήματος. Αυτό σημαίνει ότι, ο αλγόριθμος μας αντί να εργάζεται πάνω στις εκτιμήσεις, εργάζεται πάνω στην κατανομή των πιθανοτήτων οι οποίες εξάγονται από το πεδίο των θετικών αξιολογήσεων. Εάν η εκτίμηση ενός χρήστη για ένα αντικείμενο είναι δεδομένη, τότε η αντίστοιχη κατανομή της πιθανότητας θα έχει μηδενική διακύμανση, ενώ για την συγκεκριμένη εκτίμηση, θα τοποθετηθεί όλη η μάζα πιθανότητας. Εάν αυτό δεν είναι γνωστό, τότε οι αλγόριθμοι κάνουν χρήση μίας ενιαίας, παγκόσμιας, ή προβλεπόμενης κατανομής.

Στην κορυφή της πιθανολογικής αναπαράστασης, θα προτείνουμε μία αναμενόμενη μέτρηση απόστασης για να υπολογίσουμε τις ομοιότητες μεταξύ του προφίλ αξιολογήσεων από δύο χρήστες. Αυτή η μέτρηση μπορεί να χρειαστεί για τον πλησιέστερο γείτονα που ο τύπος του είναι ο αλγόριθμος συνεργατικού φιλτραρίσματος. Όπως θα δούμε, αυτό βελτιώνει σημαντικά την απόδοση της προσέγγισης του πλησιέστερου γείτονα.

Η δεύτερη προσέγγιση μας για να βελτιώσουμε την κατάσταση στην προσέγγιση του πλησιέστερου γείτονα αποτελείται από έναν αλγόριθμο βασισμένο σε γράφημα, ο οποίος καλείται πιθανολογικό συνεργατικό φιλτράρισμα βασισμένο σε γράφημα. Ο αλγόριθμος αυτός δουλεύει πάνω σε ένα γράφημα του χρήστη, στο οποίο οι χρήστες είναι οι κόμβοι, ενώ τα άκρα επισημαίνονται από την μέτρηση της απόστασης. Αυτός ο αλγόριθμος επιδιώκει να βελτιώσει την κλασσική μέθοδο του πλησιέστερου γείτονα από τη χρησιμοποιώντας τις πληροφορίες που υπάρχουν στους έμμεσους γείτονες. Το γράφημα που βασίζεται στο συνεργατικό φιλτράρισμα επιτυγχάνει με πολλαπλάσιο τρόπο τις πιθανολογικές προβλέψεις που προκύπτουν, κατά μήκος των ακμών, στο γράφημα του χρήστη.

Το υπόλοιπο του κεφαλαίου οργανώνεται με τον ακόλουθο τρόπο. Στο υποκεφάλαιο 4.2 ξεκινάμε μία συζήτηση που αφορά σχετικές προσεγγίσεις. Στο υποκεφάλαιο 4.3, συζητάμε παραλλαγές τις μέτρησης απόστασης Manhattan και προτείνουμε την πιθανολογική αναπαράσταση.

4.2 Σχετικά με το Κεφάλαιο

Ένας αριθμός από διαφορές τεχνικές του συνεργατικού φιλτραρίσματος περιγράφονται και συγκρίνονται στο παρόν κεφάλαιο, συμπεριλαμβανομένης και της προσέγγισης του πλησιέστερου γείτονα, η οποία θα χρησιμοποιηθεί για την αξιολόγηση της πρότασης που εξάγεται από την μέτρηση της απόστασης.

Οι μέθοδοι σχεδιάζονται με ειδικό τρόπο, για να αμβλύνουν τα προβλήματα ανεπάρκειας και παγωμένης έναρξης, ενώ συνήθως για την επίλυση τους περιλαμβάνουν μία υβριδική μέθοδο ανάμεσα σε ένα συνεργατικό φίλτρο και μία σύσταση που βασίζεται στο περιεχόμενο. Οι προσεγγίσεις μας διαφέρουν στο ότι παραμένουν σε ένα συνεργατικό φίλτρο.

Οι μέθοδοι που προτείνουμε χρησιμοποιούν μία πιθανολογική αναπαράσταση. Αυτές οι πιθανολογικές αναπαραστάσεις προηγουμένως έχουν επιτυχώς χρησιμοποιηθεί σε προσεγγίσεις Bayesian.

Ένας αριθμός από διαφορετικά γραφήματα που βασίζονται στις προσεγγίσεις έχουν προταθεί στο παρελθόν ως επαρκή, ενώ μερικά από αυτά έχουν σχεδιαστεί με ειδικό τρόπο για την άμβλυνση του προβλήματος ανεπάρκειας.

Ο Huang και η ομάδα του κατασκεύασε ένα διμερές γράφημα, που στο ένα τμήμα περιλαμβάνει χρήστες ενώ στο άλλο αντικείμενα. Μία κορυφή προστίθεται ανάμεσα σε ένα χρήστη και ένα αντικείμενο όταν το αντικείμενο αρέσει στο χρήστη. Τότε ένα αντικείμενο κατατάσσεται για ένα χρήστη βασισμένο στα μονοπάτια που προκύπτουν ανάμεσα στο χρήστη και το αντικείμενο. Αν και το σύστημα δίνει καλά αποτελέσματα, ωστόσο είναι περιορισμένα για χρήση επειδή μπορεί να χειριστεί μόνο δυαδικά δεδομένα. Ένα τώρα ένας χρήστης δείξει ότι του αρέσει ή όχι το αντικείμενο, οι κορυφές δεν θα έχουν βάρος .Ο Papangelis προτείνει έναν άλλο αλγόριθμο που βασίζεται σε γράφημα το οποίο περιέχει μόνο χρήστες. Το σύστημα χρησιμοποιεί μόνο δυαδικά δεδομένα. Οι κορυφές σταθμίζονται με βάση τον αριθμό των συνεχόμενων αξιολογήσεων μεταξύ των δύο χρηστών.

Μία άλλη μέθοδος που προτείνει ο Alexandros περιγράφει την εκμετάλλευση των μεταβατικών συσχετίσεων μεταξύ των αντικειμένων. Σαν παράδειγμα μπορούμε να αναφέρουμε ότι μερικά αντικείμενα μπορεί να συσχετίζονται σε μεγάλο βαθμό, καθώς επίσης μία δεδομένη αξιολόγηση για ένα στοιχείο θα μπορούσε, ως εκ τούτου, να μας πει κάτι που σχετίζεται με την αξιολόγηση των άλλων στοιχείων. Ο αλγόριθμος συστήνει τα κορυφαία αντικείμενα βάσει των μεταβατικών συσχετίσεων. Αυτό μπορεί να φανεί και σαν υβριδική προσέγγιση του προβλήματος της ανεπάρκειας.

Ο Huang και η ομάδα του παρουσιάζει ένα σύστημα σύστασης που βασίζεται στο γράφημα. Σε αντίθεση με το σύστημα μας, το σύστημα συστήνει μόνο ένα αριθμό από αντικείμενα. Δεν δημιουργεί προβλέψεις για άλλα αντικείμενα. Αυτό το σύστημα χρησιμοποιεί ένα γράφημα δύο επιπέδων και περιλαμβάνει χρήστη με χρήστη, αντικείμενο με αντικείμενο και τις κορυφές από το χρήστη με το αντικείμενο. Αυτό το γράφημα αποτελεί την αναζήτηση ενδεδειγμένων στοιχείων.

Εδώ μπορούμε να σημειώσουμε ότι καμία από τις μεθόδους που βασίζονται στο γράφημα δεν μπορούν να συγκριθούν άμεσα με τη μέθοδο μας, είτε επειδή προβλέπουν κατάταξη, αντί της απόλυτης τιμής ή επειδή μπορεί να χειριστεί μόνο δυαδικά δεδομένα.

4.3 Μετρήσεις Απόστασης και Αναπαράσταση Αποτελεσμάτων

Σε αυτό το υποκεφάλαιο, συζητάμε πρώτα μία δημοφιλής μέθοδο που συμφωνεί για τις ελλιπείς εκτιμήσεις κατά τον υπολογισμό των ομοιοτήτων μεταξύ των προφίλ δύο χρηστών. Στη συνέχεια ορίζουμε ένα προφίλ πιθανολογικής αναπαράστασης και ορίζουμε την έννοια της αναμενόμενης απόστασης που βασίζεται στην αναπαράσταση με διάφορους φυσικούς τρόπους έτσι ώστε να έχουμε καλύτερο χειρισμό στις ελλείψεις αξιολογήσεων.

4.3.1 Μετρήσεις Απόστασης

Η ομοιότητα ανάμεσα στο προφίλ των δύο χρηστών, που αυτό σημαίνει ότι στη μήτρα εκτίμησης είναι στην ίδια σειρά, μπορεί να εκτιμηθεί με όρους μέσω μίας μέτρησης ομοιοτήτων ή μίας μέτρησης απόστασης. Αυτή η έρευνα θεωρεί ότι ο συντελεστής συσχέτισης, η αλλιώς μία μέτρηση ομοιοτήτων, αλλά και η απόσταση Manhattan χρησιμοποιούνται συνήθως στο συνεργατικό φιλτράρισμα. Για να απλοποιήσουμε τους χειρισμούς υποθέτουμε πρώτα ότι καμία εκτίμηση δεν χάθηκε και μετά ορίζουμε ότι και δύο επεκτάθηκαν προκειμένου να χειριστούν την έλλειψη αξιολογήσεων.

Δίνοντας δύο τιμές στα προφίλ των χρηστών **p** και **q**, σαν παράδειγμα αναφέρουμε τα n-διαστατα διανύσματα, τα οποία περιέχουν τα συστατικά των αξιολογήσεων, που λαμβάνονται από το σύνολο των πιθανών αξιολογήσεων R, με πεδίο τιμών R={1,2,3,4,5}, τότε η συσχέτιση Pearson ορίζεται ως εξής:

$$pc(\mathbf{p},\mathbf{q})=\sum_i(p_i - \vec{p})* (q_i-\vec{q})/\sqrt{\sum_I (p_i-\vec{p})^{2*})\sum_I (q_i-\vec{q})^2} \quad (8)$$

Στην οποία το διάνυσμα u υποδηλώνει τη μέση τιμή από τα περιεχόμενα u. Η απόσταση Manhattan ορίζεται ως εξής

$$d_{MD}(\mathbf{p},\mathbf{q})=||\mathbf{p-q}||_1=\sum_{i=1}^{n} |p_i-q_i| \quad (9)$$

Με $||-||_1$ να αποτελεί τον πρώτο κανόνα.

Με την παρουσία των ελλιπών εκτιμήσεων, η συσχέτιση υπολογίζεται μόνο πάνω από τα περιεχόμενα και τα οποία είναι γνωστά τόσο για τον χρήστη p όσο και για τον q. Αυτή η προσέγγιση δεν αποδέχεται την απόσταση Manhattan επειδή το εύρος αυτού του μέτρου εξαρτάται από τον αριθμό των περιεχομένων. Υποθέτουμε την ίδια απόσταση κατά μήκος της κάθε διάστασης, που αυτό σημαίνει ότι θα κατέληγε σε μικρότερη αξία, με την προϋπόθεση ότι οι περισσότερες εκτιμήσεις είναι ελλιπείς και αποτελεί τον κύριο σκοπό μας. Ως εκ τούτου, υπολογίζουμε, με κλιμακωτό τρόπο την απόσταση Manhattan με τον ακόλουθο τύπο:

$$d_{MD\text{-}Scaled}(\mathbf{p},\mathbf{q})=n/|K(\mathbf{p})\cap K(\mathbf{q}) \quad * \sum_{i\in(k(\mathbf{p})\cap K(\mathbf{q}))} |p_i - q_i| \quad (10)$$

Με $K(\mathbf{u})=\{i|u_i\neq ?\}$ να αποτελούν την ομάδα των δεικτών για τους οποίους οι τιμές u είναι γνωστές. Σαν παράδειγμα που μπορούμε να αναφέρουμε είναι ότι, υποθέτουμε ότι ο χρήστης **p** παίρνει τιμές **p**=(3,1,?,0,?) ενώ ο χρήστης **q** παίρνει τιμές **q**=(1,2,?,?,?), επομένως η τιμή του K**(p)** είναι K(**p**)={1,2,4} ενώ η τιμή του K(**q**) είναι η K(**q**)={1,2}. Με βάση τα όσα αναφέραμε η κλιμακωτή απόσταση Manhattan υπολογίζεται ως εξής: $d_{MD\text{-}Scaled}(\mathbf{p},\mathbf{q})=5/2*(|3-1|+|1-2|)=5$.

Όπως θα δούμε και στο 5ο κεφάλαιο, οι παραπάνω επιδόσεις των μετρήσεων δεν έχουν καλή κλιμάκωση όταν υπάρχει το πρόβλημα της ανεπάρκειας. Εάν έχουμε το πρόβλημα της ανεπάρκειας σε ποσοστό 50%, που αυτό σημαίνει για παράδειγμα ότι η πιθανότητα ορίζεται σαν p=0,5, τότε η πιθανότητα της μέσης κλασματικής απόδοσης της επικάλυψης $|K(\mathbf{p})\cap K(\mathbf{q})|/n$ ανάμεσα στα δύο διανύσματα **p** και **q** να είναι μόνο $p^2=0,25$. Τόσο οι συσχετίσεις όσο και η κλιμακωτή μέθοδος Manhattan είναι απροσδιόριστες όταν $K(\mathbf{p})\cap K(\mathbf{q})=0$, για παράδειγμα όταν οι χρήστες δεν έχουν εκτίμηση για τα συνεργαζόμενα αντικείμενα.

4.3.2 Πιθανολογική Αναπαράσταση

Για να μπορέσουμε να αντιμετωπίσουμε την προαναφερθείσα αστάθεια, η οποία προκύπτει από τις ομοιότητες και τις μετρήσεις των αποστάσεων, προτείνουμε να δημιουργήσουμε μία πιθανολογική αναπαράσταση με βάση τα προφίλ των χρηστών, στην οποία οι ασταθείς

αξιολογήσεις αντικαθίστανται από κατανομές πιθανοτήτων, οι οποίες βασίζονται στο πεδίο ορισμού πιθανών αξιολογήσεων. Αυτό σημαίνει ότι αντικαθιστούμε κάθε περιεχόμενο του διανύσματος p_i με μία οριακή κατανομή της πιθανότητας που προκύπτει από τον τύπο

$P_r(p_i)$ ($\sum_{uER} Pr(p_i=u)=1$).

Μπορούμε να υπολογίσουμε την αναμενόμενη απόσταση Manhattan ανάμεσα σε δύο πιθανολογικούς τύπους ως εξής

$$d_{EMD}(\mathbf{p},\mathbf{q})=E[||\mathbf{p}-q||_1=\sum_{i=1}^{n} \sum_{u_E R} \sum_{u_q E r} |u_p - u_q| P_r(p_i=u_p)= P_r(q_i=u_q) \quad (11)$$

Σε αυτό το σημείο υποθέτουμε ότι οι τυχαίες μεταβλητές που αντιστοιχίζονται με τις διάφορές αξιολογήσεις είναι ανεξάρτητες.

Για τις δεδομένες αξιολογήσεις $p_i=p\neq?$ Η οριακή κατανομή της πιθανότητας P_r ορίζεται ως εξής

$$P_r=\begin{cases} 1 \text{ εάν } u=p \\ 0 \text{ εάν } u\neq p \end{cases} \quad (12)$$

Η συνεισφορά για τα συστατικά του φορέα που αντιστοιχούν σε άγνωστες αξιολογήσεις θα είναι διαφορετική. Θα θεωρήσουμε διάφορους τρόπους που ορίζουν μία αντιστοιχία για τη συγκεκριμένη περίπτωση.

<u>Κατανομή των αγνώστων αξιολογήσεων</u>

Σε αυτό το σημείο θέτουμε το ερώτημα ποια διανομή θα χρησιμοποιήσουμε για να υπολογίσουμε την αναμενόμενη απόσταση Manhattan στην περίπτωση που μία αξιολόγηση για ένα αντικείμενο είναι άγνωστη. Θεωρούμε λοιπόν τις ακόλουθες δύο υποθέσεις:

- Υποθέτουμε λοιπόν, ότι η αξιολόγηση του αντικειμένου ακολουθεί την ομοιόμορφη κατανομή πιθανότητας, η οποία υπολογίζεται από τον τύπο $P_r(p_i=u)=1/|R|$ με την μεταβλητή uER. Καλούμε λοιπόν την αναμενόμενή απόσταση Manhattan, που ο υπολογισμός της βασίζεται σε αυτήν την παραδοχή, ως αναμενόμενη και ομοιόμορφη απόσταση Manhattan.
- Υποστηρίζουμε, ότι η αξιολόγηση για το αντικείμενο i από το χρήστη j είναι άγνωστη. Σε αυτήν την περίπτωση υποστηρίζουμε ότι ο χρήστης j αξιολογεί με παρόμοιο τρόπο όλους τους υπόλοιπους χρήστες, που χρησιμοποιούν την διεθνή κατανομή από όλες τις γνωστές αξιολογήσεις που αφορούν το αντικείμενο i, αντί της ομοιόμορφης κατανομής. Θα αναφέρουμε λοιπόν το αποτέλεσμα της απόστασης ως αναμενόμενη και διεθνώς ομοιόμορφη απόσταση Manhattan.

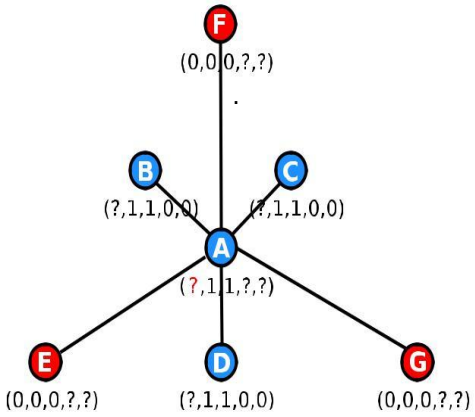
Θεωρούμε την πρώτη άποψη. Αν και οι δύο χρήστες δεν αξιολόγησαν το δεδομένο αντικείμενο, τότε ο όρος που προκύπτει από τον τύπο d_{EMD} , ο οποίος αντιστοιχίζεται με το συγκεκριμένο αντικείμενο εξαρτάται από το εύρος αξιολόγησης R. Δεν εξαρτάται όμως από το χρήστη ή από το ενεργό αντικείμενο. Για παράδειγμα, εάν το R λαμβάνει τιμές R={1,….,r} τότε η τιμή της σταθεράς c είναι η ακόλουθη:

$$c=\sum_{up} \sum_{uq} |u_p-u_q| P_r(p_i=u_p) P_r(q_i=u_q)=1/r^2 \sum_{up} \sum_{uq} |u_p-u_q|=r(r^2-1)/3r^2 \quad (13)$$

Η τιμή που προκύπτει, φαίνεται να είναι ένα είδος ποινής, επειδή προστέθηκε η μέτρηση Manhattan σε κάθε αντικείμενο που δεν έχει αξιολογηθεί και από τους δύο χρήστες. Αυτό μας παρακινεί να ορίσουμε σαν μέτρο την ακόλουθη μέτρηση. Υπολογίζουμε την απόσταση Manhattan πάνω από τα αντικείμενα, που είναι γνωστά και για τους δύο χρήστες, ενώ παράλληλα προσθέτουμε την σταθερά ποινής c σε κάθε αντικείμενο που δεν έχει αξιολογηθεί από κανέναν από τους δύο χρήστες. Καλούμε λοιπόν αυτήν την απόσταση ως απόσταση ποινής Manhattan. Αυτή λοιπόν είναι η κύρια διαφορά ανάμεσα στην αναμενόμενη και ομοιόμορφη απόσταση Manhattan με την αντίστοιχη απόσταση ποινής Manhattan, ενώ έρχεται σε πλήρη διάψευση ότι τα αντικείμενα αξιολογούνται μόνο από τον ένα χρήστη. Για αυτά τα αντικείμενα η απόσταση ποινής Manhattan, αποδίδει μία σταθερά ποινής c, ενώ η αναμενόμενη και ομοιόμορφη απόσταση Manhattan υπολογίζει την αναμενόμενη απόσταση βασισμένη σε μία δεδομένη κατανομή. Η απόσταση ποινής Manhattan μπορεί να υπολογίσει έναν παράγοντα |R| πιο γρήγορα σε σύγκριση με την αναμενόμενη ομοιόμορφη απόσταση Manhattan αλλά και από τη διεθνή απόσταση κατανομής Manhattan.

ΚΕΦΑΛΑΙΟ 5- ΘΕΩΡΗΤΙΚΗ ΠΕΡΙΓΡΑΦΗ ΑΛΓΟΡΙΘΜΩΝ ΚΑΙ ΤΩΝ ΑΝΤΙΣΤΟΙΧΩΝ ΓΡΑΦΗΜΑΤΩΝ ΤΟΥΣ

Σε αυτό το κεφάλαιο θα γίνει μία θεωρητική περιγραφή των αλγορίθμων. Θα εξηγήσουμε επίσης πως ένα γράφημα χρησιμοποιείται σαν βάση για το συνεργατικό φιλτράρισμα



Σχήμα 3: Το πρόβλημα της ανεπάρκειας. Οι κόκκινοι χρήστες έχουν μία γνωστή αξιολόγηση για το αντικείμενο 1. Ωστόσο ο χρήστης A έχει τρεις κλειστούς γείτονες, τους B, C και D, που κανένας από αυτούς δεν έχει μία γνωστή αξιολόγηση για το αντικείμενο 1. Για να δημιουργήσουμε μία πρόβλεψη για το αντικείμενο 1, ο χρήστης A βασίζεται στους πιο μακρινούς χρήστες E,F και G, επομένως τα αποτελέσματα δεν θα δίνουν τόσο ακριβές προβλέψεις. Το μήκος των κορυφών είναι ανάλογα με την απόσταση που υπάρχει ανάμεσα στους χρήστες.

5.1 Το γράφημα του χρήστη

Παρουσιάζουμε τα δεδομένα που στο εξής θα καλούνται γράφημα χρηστών G=(V,E). Κάθε κόμβος **v**EV αντιπροσωπεύει ένα χρήστη. Εάν η απόσταση d ανάμεσα σε δύο χρήστες/κόμβους μπορεί να υπολογιστεί, ένα βάρος eEE που υπάρχει ανάμεσα στους κόμβους προστίθεται στο

γράφημα. Το βάρος w που προκύπτει από αυτές τις κορυφές είναι αντιστρόφως ανάλογο από την απόσταση που υπάρχει ανάμεσα στους συνδεδεμένους κόμβους υπολογίζεται από τον ακόλουθο τύπο:

$$w_{e(v1,v2)}=1/d(v1,v2) \qquad (14)$$

Το γράφημα δεν θα περιλαμβάνει κορυφές ανάμεσα στους χρήστες που συναξιολογούν αντικείμενα όταν χρησιμοποιούν τη συσχέτιση Pearson ή την αναμενόμενη και ομοιόμορφη κατανομή Manhattan, ή πρόκειται να χρησιμοποιήσουν την αναμενόμενη απόσταση Manhattan ή την απόσταση ποινής Manhattan, αυτή η διαδικασία θα έχει σαν αποτέλεσμα ένα πλήρες συνδεδεμένο γράφημα.

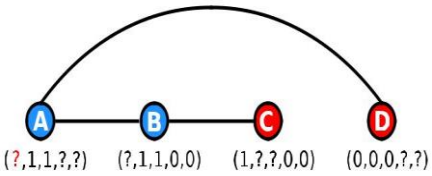
Για να χρησιμοποιήσουμε σε ένα γράφημα την μέθοδο του συνεργατικού φιλτραρίσματος θα πρέπει να έχουμε ολοκληρώσει τις πληροφορίες για κάθε κόμβο, οι οποίες βασίζονται στην παρουσία πληροφοριών για τους συνδεδεμένους κόμβους.

Αρχίζουμε τη μετάφραση του της κλασσικής προσέγγισης του πλησιέστερου γείτονα που βασίζεται στη μέθοδο του γραφήματος. Αυτή η μέθοδος είναι επεκτάσιμη για να δημιουργήσει σαφής χρήση στη δομή του γραφήματος.

5.2 Παρουσίαση και Υλοποίηση του Αλγόριθμου του Πλησιέστερου Γείτονα και του γραφήματος ολοκλήρωσης του.

Υποστηρίζουμε ότι έχουμε έναν κόμβο **v**=(3,1,?,0,?). Ο στόχος μας περιλαμβάνει μία πρόβλεψη των ελλιπών τιμών u_3 και u_4. Μπορούμε να χρησιμοποιήσουμε το ήδη κατασκευασμένο γράφημα του σχήματος 2 προκειμένου να δημιουργήσουμε αυτές τις προβλέψεις. Στην περίπτωση που υπάρχει μία υψηλότερη συσχέτιση ανάμεσα στην κορυφή και το βάρος της, πρακτικά σημαίνει ότι είναι ο πλησιέστερος γείτονας, επομένως αυτό σημαίνει ότι οι ελλιπείς τιμές τείνουν να γίνουν ίσες με τις αντίστοιχες τιμές των συνδεδεμένων κόμβων για τις οποίες η συσχέτιση ανάμεσα στο βάρος της κορυφής με την συνδεδεμένη κορυφή είναι υψηλή. Μία ελλιπής τιμή μπορεί να προβλεφθεί με τον ακόλουθο τρόπο. Επιλέγουμε τους k συνδεδεμένους κόμβους με το υψηλότερη συσχέτιση βάρους και κορυφής, για τους οποίους η τιμή πρόβλεψης είναι γνωστή, ενώ μετά λαμβάνουμε το μέσο όρο από αυτές τις γνωστές τιμές. Αυτό αντιστοιχίζεται με την k πρόβλεψη του πλησιέστερου γείτονα.

Ενώ αυτή η μέθοδος σε πολλές περιπτώσεις δίνει καλά αποτελέσματα, ωστόσο το κύριο μειονέκτημα της είναι η ύπαρξη του προβλήματος της παγωμένης εκκίνησης. Οι χρήστες που έχουν αξιολογήσει λίγα αντικείμενα θα δώσουν άσχημες αξιολογήσεις, ενώ νέα αντικείμενα που ήδη αξιολογήθηκαν από μερικούς χρήστες σπάνια θα συνίστανται.



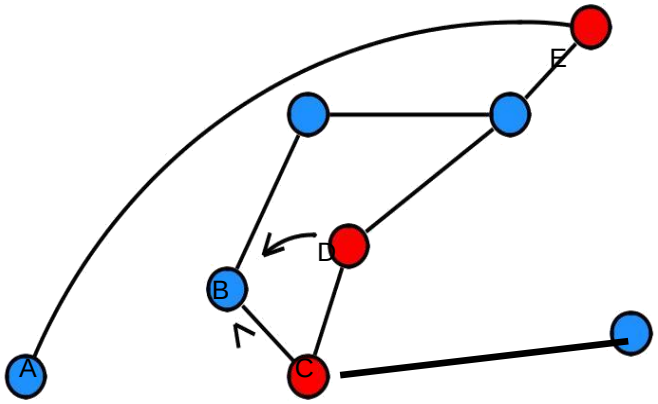
Σχήμα 4: Εδώ χρησιμοποιούμε έμμεσους κόμβους. Οι χρήστες C και D δίνουν μία αξιολόγηση για το αντικείμενο 1. Αλλά μόνο ο χρήστης D είναι ένας απομακρυσμένος γείτονας από τον χρήστη A.

5.3 Πιθανολογικό Γράφημα που βασίζεται στο Συνεργατικό φιλτράρισμα

Για να αμβλύνουμε τα προβλήματα της παγωμένης εκκίνησης αλλά και της ανεπάρκειας, προτείνουμε να χρησιμοποιήσουμε κλειστούς άλλα έμμεσους κόμβους αντί των μη σχετικών γειτόνων. Αυτό απεικονίζεται στις εικόνες 2 και 3. Στην εικόνα 2 ο στόχος μας είναι να προβλέψουμε μία εκτίμηση για το αντικείμενο 1, που συμβολίζεται με το κόκκινο ερωτηματικό, για τον χρήστη A. Ο μόνος άμεσος γείτονας που εκτιμά επίσης αυτό το αντικείμενο είναι ο χρήστης D. Η αποτελεσματική πρόβλεψη θα είναι 0 όταν ο χρήστης D δώσει στο πρώτο αντικείμενο μία εκτίμηση που θα είναι ίση με το 0. Εάν παρατηρήσουμε με πιο κλειστό τρόπο, σχετικά με τις κοινές εκτιμήσεις που υπάρχουν ανάμεσα στα αντικείμενα A και D, μπορούμε να σημειώσουμε ότι οι εκτιμήσεις που δίνουν είναι εντελώς αντίθετες μεταξύ τους. Ο χρήστης A εκτιμά τα αντικείμενα 2 και 3 με μονάδα, ενώ ο D εκτιμά και τα δύο με μηδέν. Επομένως είναι απίθανο ότι η πρόβλεψη που δημιουργήσαμε να ήταν ακριβής. Εάν δούμε τις κοινές εκτιμήσεις, ανάμεσα στους χρήστες A και D, παρατηρούμε ότι είναι πανομοιότυπες ενώ μπορούμε να υποθέσουμε ότι και τα άλλα αντικείμενα θα βαθμολογηθούν με τον ίδιο τρόπο. Τα παραπάνω ισχύουν και για τους χρήστες B και C. Έτσι μπορούμε να υποστηρίξουμε ότι ο χρήστης B θα εκτιμήσει το αντικείμενο με τιμή που ισούται με την μονάδα όπως επίσης με τον ίδιο τρόπο θα πραγματοποιήσει και ο χρήστης C, ενώ εάν ο χρήστης B θα πραγματοποιήσει εκτίμηση για το αντικείμενο 1 και θα όριζε τιμή ίση με την μονάδα με παρόμοιο τρόπο θα όριζε και ο χρήστης A. Δημιουργώντας πρόβλεψη για τον χρήστη B, σε πρώτο στάδιο και αφού συμπεριλάβουμε την πρόβλεψη που έγινε από το χρήστη A, μπορούμε να πούμε ότι αποφεύγουμε σε πολύ μεγάλο βαθμό το πρόβλημα της ανεπάρκειας αποκτώντας παράλληλα πιο ακριβείς προβλέψεις.

Αυτή η μέθοδος των πολλαπλασιαστικών προβλέψεων θα αποτελέσει τη βάση για τα γραφήματα που βασίζονται στον αλγόριθμο του συνεργατικού φιλτραρίσματος. Χρησιμοποιώντας όμως αυτή την προσέγγιση το πρόβλημα μας είναι το ακόλουθο. Αφού δημιουργήσουμε προβλέψεις, οι οποίες σχετίζονται με τις γνωστές τιμές, οι τιμές αυτές είναι επιρρεπείς σε αποκλίσεις. Τα σφάλματα θα πολλαπλασιαστούν με φυσικό επακόλουθο την συσσώρευση τους. Για να μπορέσουμε να επιλύσουμε τα σφάλματα θα χρειαστεί να χρησιμοποιήσουμε δύο μετρήσεις

- Η πρώτη μέτρηση αφορά την δημιουργία πιο συγκεκριμένων προβλέψεων.
- Ενώ η δεύτερη λαμβάνει υπ'όψιν την αβεβαιότητα που προκύπτει από τις προηγούμενες προβλέψεις προκειμένου να μπορέσει να χρησιμοποιηθεί για άλλες προβλέψεις. Ε



Σχήμα 5: Το γράφημα του χρήστη. Κάθε κόμβος αντιπροσωπεύεται από ένα χρήστη στην ομάδα των δεδομένων. Όταν η απόσταση ανάμεσα σε δύο χρήστες μπορεί να υπολογιστεί η σταθμισμένη κορυφή τοποθετείται ανάμεσα στους χρήστες. Το τοξοειδές μήκος αποτελείται από χρήστες. Οι γνωστές τιμές για τους κόμβους C, D και E μπορεί να διαδοθεί παράδειγμα από τον C και D και το από B στον A ακολούθως. Χωρίς διάδοση, η προβλεπόμενη τιμή για το χρήστη A θα βασίζονταν μόνο στο χρήστη E, αρά θα υπήρχε μία κακή πρόβλεψη κατά μήκος της τεράστιας απόστασης ανάμεσα από τους χρήστες A και E.

5.3.1 Η σειρά πρόβλεψης

Ο στόχος μας σε αυτήν την ενότητα είναι να συμπληρώσουμε, σε πρώτο στάδιο τις προβλέψεις μας, οι οποίες μπορούν να γίνουν με μεγαλύτερη σιγουριά. Για το τέλος θα χρειαστούμε έναν τρόπο για να αποφασίσουμε, σχετικά με τη αξιοπιστία που έχουμε για μία πρόβλεψη. Υποστηρίζουμε ότι δημιουργούμε μία πρόβλεψη που βασίζεται πάνω στους k γείτονες και όλους εκείνους τους γείτονες που κάνουν εκτίμηση για το αντικείμενο για το οποίο όλες οι εκτιμήσεις είναι ακριβώς οι ίδιες. Αυτό θα μας δώσει μία πρόβλεψη με την μέγιστη σιγουριά. Στην αντίθετη περίπτωση, εάν όλοι οι γείτονες εκτιμήσουν το αντικείμενο με διαφορετικό τρόπο, δεν θα μπορούμε να δημιουργήσουμε μία σίγουρη πρόβλεψη.

Η διακύμανση της συσχέτισης ανάμεσα στους γείτονες και τις εκτιμήσεις τους, η οποία χρησιμοποιείται για την δημιουργία προβλέψεων, μας δίνει ένα μέτρο σχετικά με την σιγουριά της πρόβλεψης μας. Μία χαμηλή διακύμανση μας λέει ότι η πρόβλεψη μας είναι πιθανότατα σωστή. Ενώ μία υψηλή διακύμανση μας λέει ότι έχουμε χαμηλή σιγουριά για την πρόβλεψη.

Μπορούμε λοιπόν να χρησιμοποιήσουμε την διακύμανση για να ορίσουμε τη σειρά με την οποία θα συμπληρωθούν οι προβλέψεις μας στο γράφημα.

5.3.2 Πολλαπλασιαστική Αβεβαιότητα

Υποστηρίζουμε ότι οι χρήστες μας μπορούν να εκτιμήσουν μόνο ένα αντικείμενο, θετικά ή αρνητικά, δημιουργώντας με αυτόν τον τρόπο μία πρόβλεψη, για κάποια αντικείμενα που σχετίζονται με το χρήστη A, που βασίζονται πάνω σε τρεις γείτονες. Οι δύο από τους τρεις γείτονες εκτιμούν το αντικείμενο θετικά ενώ ο τρίτος δίνει στο αντικείμενο αρνητική εκτίμηση. Σε αυτήν την περίπτωση χρήστης A θα μπορούσε να εκτιμήσει το αντικείμενο θετικά αλλά δεν θα μπορούσε να είναι σίγουρος για την πρόβλεψη του. Μόνο τα 2/3 των γειτόνων θα μπορούσαν

να δώσουν θετική εκτίμηση. Σαν παράδειγμα μπορούμε να αναφέρουμε ότι θα ορίσουμε ένα επίπεδο εμπιστοσύνης που αφορά τα 2/3 των προβλέψεων.

Στον αλγόριθμό μας, οι προβλεπόμενες τιμές χρησιμοποιούνται στις προβλέψεις που θα πραγματοποιηθούν αργότερα. Για παράδειγμα όταν δημιουργηθεί μία πρόβλεψη για το χρήστη Β, αυτός ο χρήστης θα χρησιμοποιήσει το χρήστη Α σαν έναν από τους γείτονές του. Εάν οι γείτονες του χρήστη Β, εκτός από τον χρήστη Α, εκτιμούν το αντικείμενο θετικά, οπότε η πρόβλεψη μας για το χρήστη Α είναι θετική, η πρόβλεψη για το χρήστη Β είναι επίσης θετική.

Καταχωρώντας μία σίγουρη τιμή η συγκεκριμένη πρόβλεψη γίνεται λίγο πιο ισχυρή. Αν θεωρήσουμε την προβλεπόμενη τιμή για τον χρήστη Α σαν γνωστή τιμή, όλοι οι γείτονες του χρήστη Β, θα εκτιμήσουν το αντικείμενο θετικά οπότε με βεβαιότητα το αποτέλεσμα της εκτίμησης θα είναι ίσο με τη μονάδα. Κάτι τέτοιο όμως δεν μπορεί να ληφθεί σαν δεδομένο και αφορά το γεγονός ότι δεν ήμασταν απολύτως σίγουροι στην πρόβλεψη που πραγματοποιήθηκε για το χρήστη Α.

Σαν λύση που μπορούμε να προτείνουμε σε αυτό το κενό που προκύπτει, υποστηρίζουμε την αποθήκευση των συσχετίσεων των αξιολογήσεων ανάμεσα στους γείτονες αντί της πρόβλεψης που προκύπτει από αυτή την κατανομή με κύριο παράδειγμα την αποθήκευση μίας θετικής αξιολόγησης που θα μπορούσαμε να αποθηκεύσουμε και αποτελείται από 2/3 θετικών εκτιμήσεων και 1/3 αρνητικών. Όταν δημιουργήσουμε την πρόβλεψη για τον χρήστη Β, χρησιμοποιούμε την συσχέτιση αντί για την πρόβλεψη. Αυτό μπορεί να επιτευχθεί με διάσπαση της ψήφου για το χρήστη Α στην πρόβλεψη για το χρήστη Β σύμφωνα με την πρόβλεψη. Σαν παράδειγμα μπορούμε να αναφέρουμε ότι εάν ο χρήστης Β έχει δύο γείτονες που δίνουν θετική ψήφο και μία συσχέτιση για τον χρήστη Α, τότε θα έχουμε την ακόλουθη πρόβλεψη συσχέτισης για τον χρήστη Β, η οποία θα αποτελείται από 1+1+2/3/3=8/9 θετικές αξιολογήσεις και 1/3/3=1/9 αρνητικές αξιολογήσεις. Στην περίπτωση που δεν θα έχουμε ολοκληρωμένες προβλέψεις για τον χρήστη Α τότε μπορούμε να λάβουμε υπόψιν ότι όταν κάνουμε μία πρόβλεψη για τον χρήστη Β, αντανακλάται ένα είδος εμπιστοσύνης για την συγκεκριμένη πρόβλεψη. Σαν δεύτερο κέρδος από τη χρήση αυτής της μεθόδου είναι ότι, θα υπάρχει μικρότερη επιρροή σε μία επόμενη πρόβλεψη, άρα θα υπάρχει και μία πιο ομοιόμορφος διαχωρισμός στο δικαίωμα ψήφου.

Ο συγκεκριμένος αλγόριθμος δουλεύει πάνω σε μη δυαδικά δεδομένα, ενώ η μέθοδος που περιγράφεται παραπάνω, παρατείνεται σε διακριτές συσχετίσεις. Σαν παράδειγμα μπορούμε να αναφέρουμε αξιολογήσεις ανάμεσα στο 1 και στο 5, πέντε γείτονες να αξιολογούν με 3,4,3,5 και δύο αντίστοιχους να παίρνουν την ακόλουθη συσχέτιση πιθανότητας Pr=(0,1/5,2/5,1/5,1/5). Επομένως η αξιολόγηση, η οποία μπορεί να είναι το αποτέλεσμα μία πρόβλεψης, είναι πιθανό να παρουσιαστεί σαν μία συσχέτιση. Η συσχέτιση D, που βασίζεται πάνω στις συσχετίσεις Pr^i και αποτελείται από n γείτονες υπολογίζεται με τον ακόλουθο τρόπο:

$$Pr_j=\sum_{i=1..n}D^i_j / n \qquad (15)$$

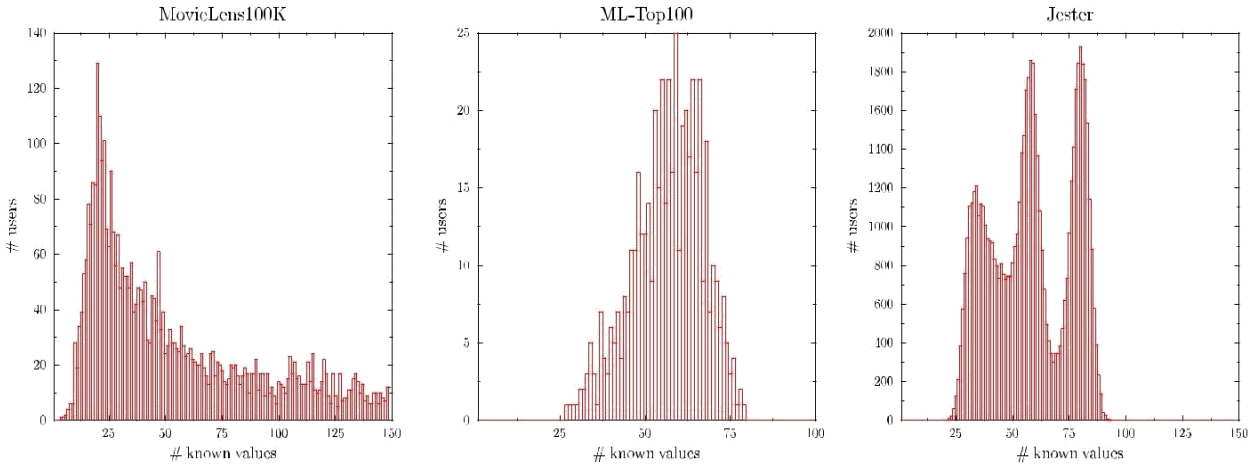
Η τελική πρόβλεψη δημιουργείται αφού λάβουμε υπ'όψιν την ενδιάμεση τιμή των συσχετίσεων. Προκειμένου να συγκρίνουμε την εμπιστοσύνη μας σε μία πρόβλεψη αρκεί να χρησιμοποιήσουμε την διαφορά των συσχετίσεων. Όσο χαμηλότερη είναι η διαφορά, τόσο περισσότερη εμπιστοσύνη έχουμε για μία πρόβλεψη. Σαν παράδειγμα μπορούμε να αναφέρουμε ότι ο χρήστης D=(0,1,0,0,0), ενώ η διαφορά είναι 0, όλοι οι γείτονες δίνουν την ίδια αξιολόγηση, που στην περίπτωση αυτή ισούται με δύο, επομένως η πρόβλεψη είναι ακριβής.

5.4 Πειραματικά Αποτελέσματα

Σε αυτήν την ενότητα θα αναφέρουμε την εκτέλεση ενός αριθμού πειραμάτων προκειμένου να αξιολογηθεί η απόδοση των προτεινόμενων μέτρων απόδοσης καθώς επίσης και τα γραφήματα που βασίζονται στους αλγόριθμους. Τα πειράματα που εκτελέσαμε χρησιμοποιούν τη γνωστή βάση δεδομένων MovieLens, ένα υποσύνολο του οποίου περιλαμβάνει τις 100 πιο εκτιμώμενες ταινίες και 100 κορυφαίους χρήστες που έδωσαν τις υψηλότερες εκτιμήσεις που έχουν χρησιμοποιηθεί. Το δεύτερο σύνολο δεδομένων είναι η βάση δεδομένων που αποτελείται από αστεία τύπου Jester και αποτελείται από έναν αριθμό εκτιμώμενων αστείων. Οι πληροφορίες για τα σύνολα των δεδομένων μπορούν να βρεθούν στον πίνακα 1 ενώ στο σχήμα 6 δείχνει την κατανομή των εκτιμώμενων αντικειμένων από τους χρήστες

	ML-100K	ML-TOP100	JESTER
#αντικείμενα	943	100	73.421
# χρήστες	1682	100	100
#εκτιμώμενα Αντικείμενα	100.000	2914	4.100.000
# Ανεπάρκεια	0,93	0.70	0.44

Πίνακας 1: Περιγραφή των συνόλων δεδομένων MovieLens, οι 100 κορυφαίες ταινίες και η βάση δεδομένων Jester



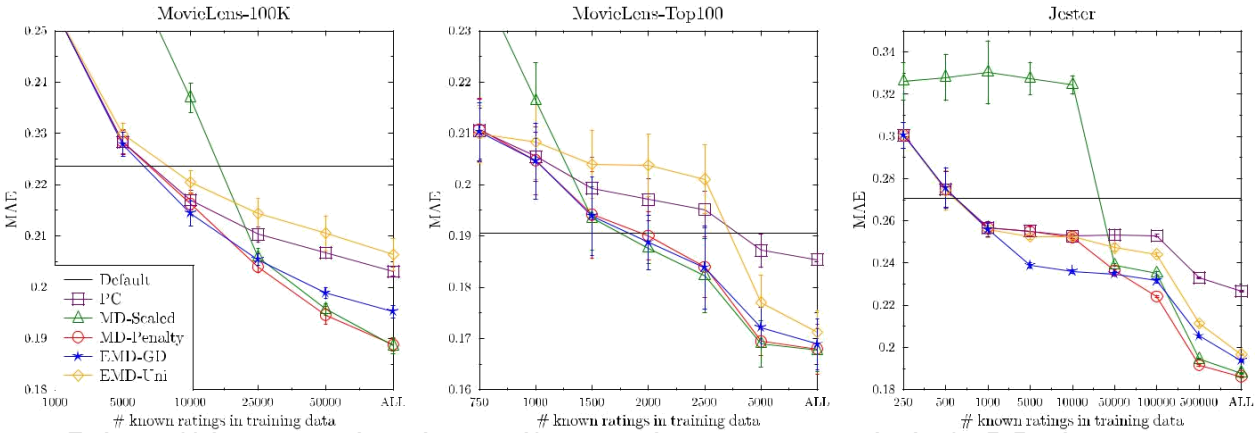
Σχήμα 6:Αριθμός των χρηστών για κάθε αριθμό των γνωστών εκτιμήσεων

Για την αξιολόγηση της απόδοσης, το σύνολο των δεδομένων χωρίζεται σε 5 ασυνεχείς εκπαιδεύσεις προς τις διασπώμενες δοκιμές, με ποσοστό 80% των δεδομένων για κάθε εκπαιδευόμενο σύνολο. Η απόδοση μετριέται με τη χρήση της έννοιας του μέσου όρου σφάλματος. Θα συζητήσουμε την απόδοση με βάση διαφορετικά επίπεδα ανεπάρκειας ενώ θα

αναλύσουμε σε βάθος την απόδοση του αλγορίθμου, όταν στην συγκεκριμένη επίδραση εμφανίζεται το πρόβλημα της παγωμένης έναρξης.

5.4.1 Αξιολόγηση Μέτρησης Απόδοσης

Το σχήμα 7 δείχνει ότι η έννοια της συνάρτησης του μέσου παράλληλου διαφοροποιημένου σφάλματος για το πακέτο των προβλημάτων ανεπάρκειας συνδυάζεται με την μέτρηση της απόστασης που αναλύσαμε στο 4ο κεφάλαιο. Επίσης περιλαμβάνονται η συσχέτιση Pearson, η κλιμακωτή απόσταση Manhattan, η απόσταση Manhattan με ποινή, αλλά και η αναμενόμενη απόσταση Manhattan βασισμένη στη διεθνή και ομοιόμορφη κατανομή. Το σχήμα επίσης υποδηλώνει την προεπιλεγμένη έννοια του μέσου όρου του σφάλματος, που αντιστοιχεί πάντα την πρόβλεψη της ολικής διαμέσου κατάταξης του συνόλου δεδομένων. Άλλες μέθοδοι άμβλυνσης του προβλήματος ανεπάρκειας δεν υπολογίζονται σε αυτή τη σύγκριση επειδή η



Σχήμα 7: Η έννοια του μέσου όρου σφάλματος ενάντια στα ανεπαρκή σύνολα δεδομένων

απόδοση τους έχει μικρή επίδραση στην χρησιμότητα του προτεινόμενου μας μέτρου απόστασης. Για παράδειγμα, χρησιμοποιώντας μία υβριδική μέθοδο για να αμβλύνουμε το πρόβλημα ανεπάρκειας είναι πολύ καλό να συνδυαστεί με τις μετρήσεις των αποστάσεων μας. Εδώ μπορούμε να σημειώσουμε ότι η προεπιλεγμένη έννοια του μέσου όρου σφάλματος δεν είναι χρήσιμη στην πράξη στην περίπτωση που δεν μπορεί να χρησιμοποιηθεί για την κατάταξη αντικειμένων.

Η συσχέτιση Pearson, η οποία συνήθως χρησιμοποιείται στο συνεργατικό φιλτράρισμα, καθώς επίσης και η ομοιόμορφη επεκτάσιμη κατανομή Manhatan αποτελείται από από τις καλύτερες επιδόσεις, συνέπεια της συνδυασμένης χρήσης της κατανομής Manhattan με ποινή αλλά και της επεκτάσιμης κατανομής Manhattan που στηρίζεται σε διεθνείς όρους, σε όλα τα σύνολα δεδομένων και τα επίπεδα ανεπάρκειας. Ωστόσο εάν η ανεπάρκεια αυξηθεί η απόδοση του θα επιδεινωθεί σημαντικά. Το κλιμακούμενο μέτρο απόδοσης αποτελεί το χειρότερο εκτελέσιμο μέτρο απόδοσης που χρησιμοποιείται πάνω σε ανεπαρκή σύνολα δεδομένων. Ο λόγος είναι ότι το μέτρο απόστασης γίνεται ασταθής εάν η επικάλυψη μεταξύ δύο προφίλ χρηστών είναι πολύ μικρή όπως αναφέραμε και στο υποκεφάλαιο 4.3 και συγκεκριμένα στην ενότητα 4,3.1. Η εκτέλεση του μέτρου απόστασης με ποινή αλλά και του επεκτάσιμου μέτρου απόστασης

βασισμένο στα διεθνή πρότυπα θα έχει συνολικά καλές επιδόσεις. Επιπλέον θα έχουν παρόμοια απόδοση και για τις κορυφαίες 100 εκτιμώμενες ταινίες από το MovieLens. Τα αποτελέσματα που προκύπτουν επί της βάσης δεδομένων Jester, αλλά και σε μικρότερο βαθμό από τα άτομα που επίσης χρησιμοποίησαν το σύνολο δεδομένων MovieLens-100k, δείχνουν ότι το μέτρο απόστασης με ποινή δουλεύει σε καλύτερο βαθμό πάνω σε χαμηλά επίπεδα ανεπάρκειας, ενώ το επεκτάσιμο μέτρο απόδοσης που βασίζεται στα διεθνή πρότυπα είναι καλύτερο για υψηλά επίπεδα ανεπάρκειας. Βασιζόμενοι λοιπόν πάνω σε αυτές τις παρατηρήσεις, μπορούμε να συστήσουμε λοιπόν το επεκτάσιμο μέτρο απόστασης που βασίζεται σε διεθνή πρότυπα για σύνολα δεδομένων με υψηλά επίπεδα ανεπάρκειας, ενώ το μέτρο απόστασης με ποινή συστήνεται για σύνολα δεδομένων με χαμηλά επίπεδα ανεπάρκειας.

Υπενθυμίζουμε επίσης ότι το μέτρο απόστασης με ποινή προσθέτει μία σταθερή ποινή για την απόσταση Manhattan για κάθε αντικείμενο το οποία δεν είναι εκτιμώμενο από τους δύο χρήστες. Σαν αποτέλεσμα το μέτρο απόστασης με ποινή ρητώς παίρνει των αριθμό των κοινών εκτιμήσεων μέσα σε ένα λογαριασμό. Ένας μικρότερος αριθμός από συνεκτιμώμενα αντικείμενα αποδίδει μία υψηλότερη απόδοση. Θα μπορούσε κανείς να υποστηρίξει ότι οι χρήστες με ένα υψηλό αριθμό από κοινές εκτιμήσεις είναι πιθανό να έχουν παρόμοια συμπεριφορά ανεξάρτητα από τις δεδομένες εκτιμήσεις. Αυτό αντικατοπτρίζεται στη μέτρηση του μέτρου απόστασης με ποινή.

5.4.2 Απόδοση της πιθανολογικής βάσης του Συνεργατικού Φιλτραρίσματος

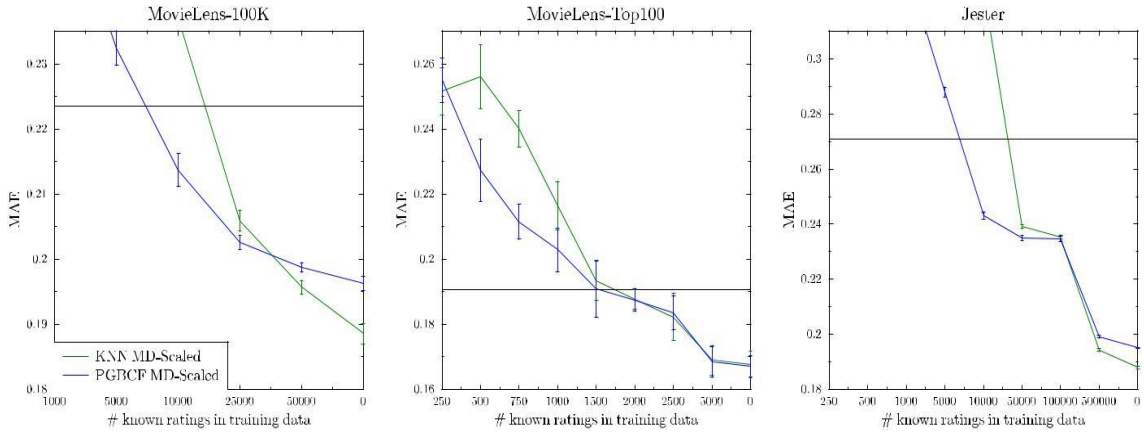
Το σχήμα 8 μας δείχνει την απόδοση του πιθανολογικού γραφήματος που βασίζεται στο συνεργατικό φιλτράρισμα ενώ επίσης γίνεται σύγκριση με το αντίστοιχο του αλγόριθμου του πλησιέστερου γείτονα. Και τα δύο χρησιμοποιούν σαν μέτρο απόστασης το διαβαθμισμένο μέτρο απόστασης. Συγκρίνοντας τα δύο γραφήματα που βασίζονται στους αλγόριθμους όπως τους περιγράψαμε στα υποκεφάλαια 5.2 και 5.3 αυτό δεν είναι δυνατόν να υφίσταται λόγω ότι οι συγκεκριμένοι αλγόριθμοι εργάζονται πάνω σε δυαδικά σύνολα δεδομένων η προβλέπουν απευθείας μία κατάταξη

Η απόδοση του πιθανολογικού γραφήματος που βασίζεται στο συνεργατικό φιλτράρισμα συγκρίνεται με το γράφημα του πλησιέστερου γείτονα στα περισσότερα από τα πειράματά μας. Αναμένουμε ότι η απόδοση του πιθανολογικού γραφήματος που βασίζεται στο συνεργατικό φιλτράρισμα να είναι περισσότερο χρήσιμη σε σύνολα δεδομένων που έχουν ανεπάρκεια. Εάν τα δεδομένα δεν είναι ανεπαρκή, τότε οι πολλαπλασιαστικές προβλέψεις είναι λιγότερο χρήσιμες και μπορούν να μειώσουν την ακρίβεια. Αυτό είναι γνωστό σαν επίδραση της υπέρ-ενεργοποίησης. Αυτή η υπόθεση επιβεβαιώνεται από τα αποτελέσματα. Στο σύνολο δεδομένων MovieLens 100k, η απόδοση του πιθανολογικού γραφήματος που βασίζεται στο συνεργατικό φιλτράρισμα υπερτερεί έναντι του αλγόριθμου του πλησιέστερου γείτονα σε σύνολα δεδομένων που υπάρχει ανεπάρκεια. Το ίδιο ισχύει και για τη βάση δεδομένων Jester. Επομένως, η απόδοση του πιθανολογικού γραφήματος που βασίζεται στο συνεργατικό φιλτράρισμα μπορεί να θεωρηθεί ως πρόβλεψη εκτιμήσεων για σύνολα δεδομένων με ανεπάρκεια.

Το κόστος υπολογισμού της απόδοσης του πιθανολογικού γραφήματος που βασίζεται στο συνεργατικό φιλτράρισμα είναι περίπου διπλάσιο σε σύγκριση με το γράφημα του πλησιέστερου γείτονα, εφόσον αγνοήσουμε το χρόνο διαλογής των άγνωστων εκτιμήσεων. Έτσι η επιβάρυνση στις πολλαπλασιαστικές προβλέψεις είναι σχετικά μικρή. Αυτό γίνεται μόνο στην περίπτωση που ο στόχος μας είναι να εκτιμήσεις αντικειμένων για άλλους χρήστες. Εάν, ενδιαφερόμαστε να δημιουργήσουμε προβλέψεις για μόνο ένα χρήστη, τότε αυτό μπορεί να πραγματοποιηθεί εύκολα με τη χρήση του γραφήματος του πλησιέστερου γείτονα. Δεν μπορεί να

πραγματοποιηθεί με την απόδοση του γραφήματος που βασίζεται στο συνεργατικό φιλτράρισμα επειδή δημιουργεί προβλέψεις για όλους τους χρήστες σε μια προκαθορισμένη σειρά.

Συνδυάζοντας το μέτρο απόστασης με ποινή με τη μέθοδο που βασίζεται στο γράφημα μας δίνει σαν αποτέλεσμα μία μικρή πρόσθετη βελτίωση πάνω στο σύνολο δεδομένων MovieLens με υψηλή ανεπάρκεια. Όταν το σύνολο των δεδομένων γίνει λιγότερο ανεπαρκές. Ο συνδυασμός των μεθόδων, του γραφήματος του πλησιέστερου γείτονα αλλά και του μέτρου απόστασης με ποινή, υπερτερεί έναντι μίας συνηθισμένης μεθόδου. Αυτό όμως έχει σαν αποτέλεσμα ότι η μέθοδος της απόδοσης του γραφήματος του συνεργατικού φιλτραρίσματος θα έχει ελαφρώς χειρότερα αποτελέσματα, σε σχέση με το συνδυασμό των μεθόδων του γραφήματος του πλησιέστερου γείτονα και του μέτρου απόστασης με ποινή, πάνω σε σύνολα δεδομένων με όχι τόσο μεγάλη ανεπάρκεια. Για το σύνολο δεδομένων Jester η μέθοδος που θα εκτελεστεί, είναι κοντά στο συνδυασμό των μεθόδων του γραφήματος του πλησιέστερου γείτονα και στο μέτρο απόστασης με ποινή, εκτός εάν η απόδοση του γραφήματος του συνεργατικού φιλτραρίσματος προκαλεί υπέρ-απόδοση στην περίπτωση που χρησιμοποιηθεί ο συνδυασμός των μεθόδων του γραφήματος του πλησιέστερου γείτονα αλλά και του μέτρου απόστασης με ποινή.



Σχήμα 8: Η έννοια του μέσου όρου σφάλματος στον αλγόριθμο συνεργατικού φιλτραρίσματος και στον αλγόριθμο του πλησιέστερου γείτονα με χρήση κλιμακωτού μέτρου απόστασης για διάφορα επίπεδα ανεπάρκειας

5.4.3 Ανθεκτικότητα Ενάντια στο πρόβλημα της παγωμένης έναρξης

Για να μπορέσουμε να αναλύσουμε την ανθεκτικότητα την ανθεκτικότητα των αλγορίθμων, σε σχέση με το πρόβλημα της παγωμένης έναρξης, υπολογίζουμε την έννοια του μέσου όρου σφάλματος επί των χρηστών, των οποίων ο αριθμός των εκτιμώμενων αντικειμένων έγκειται σε ένα δεδομένο χρονικό διάστημα. Για παράδειγμα η έννοια του μέσου όρου σφάλματος αφορά όλους τους χρήστες που εκτιμούν 2 με 3 αντικείμενα και δείχνει πως οι αλγόριθμοι μπορούν να χρησιμοποιηθούν πάνω σε νέους χρήστες. Το σχήμα δείχνει αυτά τα αποτελέσματα για διαφορετικά χρονικά διαστήματα. Μπορούμε να δείξουμε αυτά τα αποτελέσματα μόνο για το σύνολο δεδομένων MovieLens-100k. Για όλα τα υπόλοιπα σύνολα δεδομένων δεν υπάρχουν χρήστες που προέβησαν σε εκτίμηση αυτών των λίγων αντικειμένων.

Τα αποτελέσματα δείχνουν καθαρά ότι το νέο μέτρο απόστασης που χρησιμοποιήθηκε ξεπερνά την ήδη χρησιμοποιούμενη συσχέτιση Pearson. Αν και η διακύμανση των αποτελεσμάτων είναι υψηλή μπορούμε να παρατηρήσουμε ότι η χρήση της απόδοσης του γραφήματος που βασίζεται στο συνεργατικό φιλτράρισμα είναι καλύτερη από τη χρήση του γραφήματος του πλησιέστερου γείτονα που εκτιμούν ένα μικρό αριθμό από αντικείμενα. Μόνο στην περίπτωση που εκτιμώνται 2 με 3 αντικείμενα, η απόδοση του γραφήματος που βασίζεται στο συνεργατικό φιλτράρισμά μειονεκτεί έναντι του γραφήματος του πλησιέστερου γείτονα.

6. ΣΥΝΟΨΗ ΚΑΙ ΣΥΜΠΕΡΑΣΜΑΤΑ

Στην παρούσα διατριβή εστιάσαμε στην θεωρητική άμβλυνση του προβλήματος ανεπάρκειας που προκύπτει από την χρήση των συστημάτων συνεργατικού φιλτραρίσματος. Μοντελοποιήσαμε το πρόβλημα της σύστασης σαν ένα πρόβλημα συνειρμικής ανάκτησης. Οι διαδιδόμενοι αλγόριθμοι αναπτύσσονται μέσα στο πλαίσιο της συνειρμικής ανάκτησης πληροφοριών. Τα δεδομένα που χρησιμοποιήθηκαν ανακτήθηκαν μέσα από ένα ηλεκτρονικό βιβλιοπωλείο. Κατά τη θεωρητική προσέγγιση παρατηρήσαμε ότι:

α) Η διαδιδόμενη ενεργοποίηση που βασίζεται στο συνεργατικό φιλτράρισμα επιτυγχάνει σημαντικότερη ποιότητα της σύστασης σε σύγκριση με την προσέγγιση του απλού συνεργατικού φιλτραρίσματος ενώ δεν λαμβάνει υπόψην τις μεταβατικές ενώσεις.

β) Η προσέγγιση που βασίζεται στη διαδιδόμενη ενεργοποίηση μπορεί με αποτελεσματικό τρόπο να αμβλύνει το πρόβλημα της παγωμένης έναρξης δημιουργώντας υψηλής ποιότητας συστάσεις για τους νέους χρήστες.

Προτείναμε για την επίλυση των προβλημάτων ανεπάρκειας και παγωμένης έναρξης μία σειρά από μετρήσεις της απόστασης Manhattan, όπως για παράδειγμα το επεκτάσιμο μέτρο απόστασης βασισμένο στα διεθνή πρότυπα, ή το επεκτάσιμο ομοιόμορφο κατανεμημένο μέτρο απόστασης, αλλά και το μέτρο απόστασης με ποινή. Σε θεωρητικό επίπεδο παρουσιάσαμε την απόδοση του πιθανολογικού γραφήματος του συνεργατικού φιλτραρίσματος, ενός αλγόριθμου που κατατάσσει τις διαδιδόμενες πιθανολογικές προβλέψεις μέσω ενός γραφήματος χρηστών. Μέσω της θεωρητικής προσέγγισης δείχνουμε ότι με τη χρήση νέων μέτρων απόστασης, όπως για παράδειγμα το μέτρο απόστασης με ποινή και το επεκτάσιμο μέτρο απόστασης βασισμένο στα διεθνή πρότυπα, αποδίδουν σταθερά υψηλές επιδόσεις όπως με την συσχέτιση Pearson και με το κλιμακούμενο μέτρο απόδοσης Manhattan. Με βάση τη θεωρητική μας προσέγγιση παρατηρήσαμε ότι η απόδοση του γραφήματος που βασίζεται στο συνεργατικό φιλτράρισμα είναι δυνατόν να ξεπεράσει το γράφημα του πλησιέστερου γείτονα, σε σύνολα δεδομένων που έχουν ανεπάρκεια ή ακόμα και σε νέους χρήστες που εκτίμησαν λίγα αντικείμενα

ΠΑΡΑΡΤΗΜΑ-ΒΙΒΛΙΟΓΡΑΦΙΑ

1. AGGARWAL, C. C., WOLF, J. L., WU, K.-L., AND YU, P. S. 1999. Horting hatches an egg: A new graph theoretic approach to collaborative filtering. In Proceedings of the 5th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'99) (San Diego, Calif.). ACM, New York, 201– 212.
2.ALBERT, R. AND BARABASI, A.-L. 2002. Statistical mechanics of complex networks. Rev. Mod. Phys.74, 47–97.
ANDERSON, J. R. 1983. A spreading activation theory of memory. J. Verb. Learn. Verb. Behav. 22, 261–295.
3. BALABANOVIC, M. AND SHOHAM, Y. 1997. FAB: Content-based, collaborative recommendation. Com- mun. ACM 40, 3, 66–72.
4.BASU, C., HIRSH, H., AND COHEN, W. 1998. Recommendation as classification: Using social and content-based information in recommendation. In Proceedings of the 15th National Conference on Artificial Intelligence, 714–720.
5.BILLSUS, D. AND PAZZANI, M. J. 1998. Learning collaborative information filters. In Proceedings of the 15th International Conference on Machine Learning, 46–54.
6.BOLLEN, J., VANDESOMPEL, H., AND ROCHA, L. M. 1999. Mining associative relations from website logs and their application to context-dependent retrieval using spreading activation. In Proceedings of the Workshop on Organizing Web Space (WOWS). ACM Digital Libraries 99.
7.BREESE, J. S., HECKERMAN, D., AND KADIE, C. 1998. Empirical analysis of predictive algorithms for collaborative filtering. In Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence (Madison, Wisc.). Morgan-Kaufmann, Reading, Mass. 43–52.
8.BURKE, R. 2000. Semantic ratings and heuristic similarity for collaborative filtering. In Proceedings of the 17th National Conference on Artificial Intelligence.
9.CHEN, H. AND DHAR, V. 1991. Cognitive process as a basis for intelligent retrieval systems design Information Processing and Management 27, 5, 405–432.
10.CHEN, H., LYNCH, K. J., BASU, K., AND NG, D. T. 1993. Generating, integrating, and activating thesauri for concept-based document retrieval. IEEE Exp., Spec. Series Artif. Intell. Text-based Inf. Systems 8, 2, 25–34.
11.CHEN, H. AND NG, D. T. 1995. An algorithmic approach to concept exploration in a large knowledge network (automatic thesaurus consultation): Symbolic branch-and-bound search vs. Connection- ist Hopfield net activation. J. ASIS 46, 5, 348–369.
12.CLAYPOOL, M., GOKHALE, A., MIRANDA, T., MURNIKOV, P., NETES, D., AND SARTIN
, M. 1999. Combining content-based and collaborative filters in an online newspaper. In Proceedings of the ACM SIGIR Workshop on Recommender Systems. ACM, New York.
13.COHEN, P. R. AND KJELDSEN, R. 1987. Information retrieval by constrained spreading activation in semantic networks. Information Processing and Management 23, 4, 255–268.
14.COLLINS, A. M. AND LOFTUS, E. F. 1975. A spreading activation theory of semantic processing. Psych. Rev. 82, 6, 407–428.
15.CONDLIFF, M. K., LEWIS, D. D., MADIGAN, D., AND POSSE, C. 1999. Bayesian mixed-effects models for recommender systems. In Proceedings of the ACM SIGIR Workshop on

.

Recommender Systems. ACM, New York.

16.CRESTANI, F. AND LEE, P. L. 2000. Searching the web by constrained spreading activation. Inf.Proc. Manage. 36, 585–605.
17.CROUCH, C. AND YANG, B. 1992. Experiments in automatic statistical thesaurus construction. In Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, (Copenhagen, Denmark). ACM, New York, 77–88.
18.GOLDBERG, K., ROEDER, T., GUPTA, D., AND PERKINS, C. 2001. Eigentaste: A constant time collaborative filtering algorithm. Inf. Ret. 4, 2, 133–151.
19.GOOD, N., SCHAFER, J., KONSTAN, J., BORCHERS, A., SARWAR, B., HERLOCKER, J., AND RIEDL, J. 1999. Combining collaborative filtering with personal agents for better recommendations. In Proceedings of the 16th National Conference on Artificial Intelligence, 439–446.
20.GORDON, M. 1988. Probabilistic and genetic algorithm for document retrieval. Commun. ACM 31, 10, 1208–1218.
21.HILL, W., STEAD, L., ROSENSTEIN, M., AND FURNAS, G. 1995. Recommending and evaluating choices in a virtual community of use. In Proceedings of the ACM CHI'95 Conference on Human Factors in Computing Systems. ACM, New York, 194–201.
22.HUANG, Z., CHUNG, W., AND CHEN, H. 2003. A graph model for e-commerce recommender systems. J. ASIST, in press.
23.Gediminas Adomavicius and Alexander Tuzhilin. To-ward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions.
24. Laurent Candillier, Frank Meyer, and Fran_coise Fessant. Designing speci_c weighted similarity measures to improve collaborative _ltering systems. In ICDM'08: Proceedings of the 8th industrial conference on Advances in Data Mining, pages 242{255, Berlin, Heidelberg, 2008.Springer-Verlag.
25. Thomas Hofmann. Latent semantic models for collaborative filtering. ACM Transactions on InformationSystems, 22(1):89{115, 2004.
26. Zan Huang, Hsinchun Chen, and Daniel Zeng. Ap-plying associative retrieval techniques to alleviate thesparsity problem in collaborative filtering. ACM Trans. Inf. Syst., 22(1):116{142, January 2004.
27. Zan Huang, Wingyan Chung, Thian-Huat Ong, and Hsinchun Chen. A graph-based recommender system for digital library. In JCDL '02: Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries,pages 65{73, New York, NY, USA, 2002. ACM.
28. Zan Huang, D. Zeng, and Hsinchun Chen. A comparison of collaborative-_ltering recommendation algorithms for e-commerce. Intelligent Systems, IEEE,
22(5):68{78, 2007.
29. Koji Miyahara and Michael J. Pazzani. Collaborative filtering with the simple bayesian classi_er. In In: Pro-IEEE Transactions on Knowledge and Data Engineering, 17(6):734{749, June 2005.Proceedings of the 6th Pacificc Rim International Conference on Artiticial Intelligence, pages 679{689, 2000.
30. Alexandros Nanopoulos. Collaborative filtering based on transitive correlations between items. Advances in Information Retrieval, 2007.
31. M. Papagelis, D. Plexousakis, and T. Kutsuras. Alleviating the sparsity problem of collaborative filtering using trust inferences. In P. Herrmann, editor, iTrust, pages 224{239. Springer-Verlag Berlin Heidelberg, 2005.

32. Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. Grouplens: An open architecture for collaborative filtering of netnews. Pages 175{186. ACM Press, 1994.
33. Badrul M. Sarwar, George Karypis, Joseph A. Konstan, and John T. Riedl. Application of dimensionality reduction in recommender system - a case study. In ACM WebKDD Workshop, 2000.
34. Andrew I. Schein, Alexandrin Popescul, Lyle H., Rin Popescul, Lyle H. Ungar, and David 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, pages 253{260. ACM Press, 2002.
35. Andrew I. Schein, Alexandrin Popescul, Lyle H. Ungar, and David M. Pennock. Methods and metrics for coldstart recommendations. In SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval,pages 253{260, New York, NY, USA, 2002. ACM.
36.Nathan Srebro, Jason D. M. Rennie, and Tommi S. Jaakkola. Maximum-margin matrix factorization. In Advances in Neural Information Processing Systems 17, pages 1329{1336, 2005.
37. Xiaoyuan Su and Taghi M. Khoshgoftaar. Collaborative filtering for multi-class data using belief nets algorithms. In ICTAI '06: Proceedings of the 18th IEEE International Conference on Tools with Artificial Intel ligence, pages 497 {504, Washington, DC, USA, 2006.
IEEE Computer Society.