ΤΕΧΝΟΛΟΓΙΚΟ ΕΚΠΑΙΔΕΥΤΙΚΟ ΙΔΡΥΜΑ ΔΥΤΙΚΗΣ ΕΛΛΑΔΑΣ

ΣΧΟΛΗ ΔΙΟΙΚΗΣΗ ΚΑΙ ΟΙΚΟΝΟΜΙΑΣ

ΤΜΗΜΑ ΔΙΟΙΚΗΣΗ ΕΠΙΧΕΙΡΗΣΕΩΝ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ ΑΠΟ ΤΟΝ ΠΑΓΚΟΣΜΙΟ ΙΣΤΟ

Μελέτη

Αθανάσιος Χαλάλης

Αλέξανδρος Φουντουλάκης

Επίβλεψη

Κωνσταντίνος Χαλκιόπουλος

ΠΑΤΡΑ 2015

ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ ΑΠΟ ΤΟΝ ΠΑΓΚΟΣΜΙΟ ΙΣΤΟ

Περίληψη

Η παρούσα πτυχιακή εργασία πραγματεύεται την εξόρυξη γνώσης από τον παγκόσμιο ιστό. Είναι δεδομένο ότι το διαδίκτυο συγκεντρώνει τεράστιες ποσότητες πληροφοριών, οι οποίες λόγω του όγκου τους πολλές φορές μένουν αναξιοποίητες. Οι τεχνικές εξόρυξης γνώσης μπορούν να αυτοματοποιήσουν τη διαδικασία της ανάκτησης χρήσιμων δεδομένων από τον ιστό και να συνδυαστούν με συστήματα που μπορούν να αξιοποιήσουν τη γνώση αυτή. Απώτερος στόχος της εξόρυξης γνώσης είναι η κατανοητή δομή της πληροφορίας που θα εξαχθεί και τα πρότυπα που θα προκύψουν προς το άτομο, προκειμένου να συμβάλλουν στη λήψη ορθών αποφάσεων. Τα είδη των βάσεων δεδομένων, όπως κατατάσσονται σύμφωνα με τους αριθμούς κλειδιών που τα χαρακτηρίζουν, διακρίνονται στις σχεσιακές βάσεις δεδομένων, στις βάσεις δεδομένων συναλλαγών, στις χωρικές και χρονικές βάσεις δεδομένων, στις βάσεις κειμένων και, τέλος, στις πολυμεσικές βάσεις. Είναι ευρέως γνωστό ότι οι αλγόριθμοι εξόρυξης δεδομένων ποικίλουν και στοχεύουν στην αυτόματη ή ημιαυτόματη ανάλυση μεγάλων ποσοτήτων δεδομένων για την εξαγωγή ενδιαφέροντος προτύπου που ήταν άγνωστο μέχρι εκείνη τη στιγμή. Η διπλωματική εργασία αναλύει τη συνεργατική διήθηση των δεδομένων, η οποία αποτελεί μια διαδικασία απόρριψης ή αποδοχής ορισμένων δεδομένων, καθώς και εξηγεί τα παραδείγματα εφαρμογών εξόρυξης γνώσης στον παγκόσμιο ιστό.

Abstract

This project deals with the extraction of knowledge from the Web. It is assumed that the internet brings huge amounts of information, which, because of their volume often left untapped. The data mining techniques can automate the process of recovering useful data from the web and combined with systems that can leverage this knowledge. The ultimate goal of data mining is the intelligible structure of information to be exported and standards that will arise for the individual, in order to contribute to sound decision making. The types of databases, such as classified according to the number of keys that characterize them, are divided into relational databases, databases on trade in spatial and temporal databases, the text bases and finally to multimedia databases. It is well known that data mining algorithms are varied and aim to automatic or semi-automatic analysis of large amounts of data to extract interest standard that was unknown until then. The thesis analyzes the collaborative filtering data, which is a method of disposal or acceptance of certain data, and explains examples mining applications on the web.

ΠΕΡΙΕΧΟΜΕΝΑ

ΕΥΡΕΤΗΡΙΟ ΕΙΚΟΝΩΝ

ΕΥΡΕΤΗΡΙΟ ΠΙΝΑΚΩΝ

ΕΙΣΑΓΩΓΗ

Η παρούσα πτυχιακή εργασία πραγματεύεται το αντικείμενο της εξόρυξης γνώσης από τον παγκόσμιο ιστό. Είναι γεγονός πως η υπολογιστική ισχύς των υπολογιστών διπλασιάζεται κάθε 18 μήνες και πλέον είναι ευρέως γνωστή η αδυναμία υπολογισμού του μεγέθους των δεδομένων του παγκόσμιου ιστού. Ο παγκόσμιος ιστός δύναται να θεωρηθεί ως η μεγαλύτερη βάση δεδομένων που είναι διαθέσιμη σε κάθε χρήστη και αντιμετωπίζει καθημερινά τις προκλήσεις `σε θέματα παρουσίασης και ποιότητας δεδομένων. Γενικά, τα δεδομένα του διαδικτύου διακρίνονται στο περιεχόμενο των ιστοσελίδων, την ενδοπληροφορία, εσωτερική δομή των ιστοσελίδων, στα δεδομένα χρήσης που περιγράφουν τον τρόπο που οι επισκέπτες τις προσπελαύνουν και το προφίλ των χρηστών που εμπεριέχουν πληροφορίες και δημογραφικλά δεδομένα.

Αντίστοιχα όσον αφορά τον τομέα εξόρυξης δεδομένων από τα πιο ενδιαφέροντα ερευνητικά πεδία του γενικού είναι οι εφαρμογές εξόρυξης δεδομένων στον παγκόσμιο ιστό. Ο όρος βάση δεδομένων εδώ χρησιμοποιείται με θεωρητικό τρόπο, καθώς στην πραγματικότητα δεν υπάρχει πρακτικά δομή ή σχήμα στον παγκόσμιο ιστό. Αυτό κάνει ακόμα πιο επιτακτική την ανάγκη για εξόρυξη δεδομένων στον παγκόσμιο ιστό, παρέχοντας τεράστια βοήθεια σε κάθε είδους χρήστη. Με τον όρο εξόρυξη γνώσης στον παγκόσμιο ιστό δεν αναφερόμαστε μόνο σε δεδομένα που περιέχονται σε ιστοσελίδες αλλά και σε δεδομένα που έχουν να κάνουν με τη δραστηριότητα ενός χρήστη σε αυτό.

Το πρώτο μέρος της μελέτης εστιάζει στις βασικές αρχές και τις εφαρμογές εξόρυξης δεδομένων. Γίνεται αναφορά σε δεδομένα που έχουν να κάνουν με τη δραστηριότητα ενός χρήστη και δεν επικεντρωνόμαστε μόνο σε δεδομένα που εμπεριέχονται σε ιστοσελίδες. Παρουσιάζεται το υλοποιημένο σύστημα εξόρυξης γνώσης και επισημαίνονται τα βήματα και τα μέρη του συστήματος.

Σε δεύτερο επίπεδο μελετούνται οι αλγόριθμοι εξόρυξης δεδομένων. Ο πραγματικός στόχος της εξόρυξης δεδομένων είναι η αυτόματη ή ημιαυτόματη ανάλυση μεγάλων ποσοτήτων με δεδομένα για την εξαγωγή κάποιου ενδιαφέροντος προτύπου που ήταν άγνωστο μέχρι εκείνη τη στιγμή.Πολυάριθμοι είναι οι αλγόριθμοι εξόρυξης δεδομένων και σε αυτή την ενότητα θα παρουσιαστούν οι πιο σημαντικοί από αυτούς. Οι κατηγορίες αφορούν την κατηγοριοποίηση, τη συσταδοποίηση, τους κανόνες συσχέτισης, τα πρότυπα καολουθιών, την παλινδρόμηση και τα δέντρα απόφασης.

Αυτές αναπαριστούν όλη την περιοχή των αλγορίθμων που εφαρμόζονται στον τομέα αυτό.

Ακολούθως, το τρίτο μέρος της εργασίας επικεντρώνεται στη συνεργατική διήθηση των δεδομένων. Είναι χρήσιμο να σημειωθεί ότι η Συνεργατική Διήθηση αποτελεί τη διαδικασία της απόρριψης ή αποδοχής ορισμένων δεδομένων συγκριτικά με ορισμένα άλλα. Η διαδικασία πραγματοποιείται μέσω υπολογιστή και με χρήση τεχνικών που απαιτούν την συνεργασία παραγόντων όπως είναι τα αποθηκευτικά μέσα, οι απόψεις των χρηστών, οι πηγές πληροφόρησης κλπ.. Επιπρόσθετα, για την αποτελεσματικότερη προσέγγιση των πελατών αναλύθηκαν οι μέθοδοι φιλτραρίσματος των πληροφοριών, καθώς και τα χαρακτηριστικά τους. Συνάμα, αναλύθηκε η αναγκαιότητα, ο στόχος και οι επιδιώξεις των συστημάτων.

Το τελευταίο μέρος της εργασίας αναλύει τα παραδείγματα εφαρμογών εξόρυξης γνώσης στον παγκόσμιο ιστό. Η εξόρυξη γνώσης αποτελεί την τεχνολογία αιχμής για την αποτελεσματική ανάλυση των δεδομένων και την αποκάλυψη νέων σχέσεων, όπως για παράδειγμα τα πρότυπα συμπεριφοράς των πελατών. Συγχρόνως, στοχεύει στην υποστήριξη στρατηγικών αποφάσεων και εφαρμόζει τεχνικές μηχανικής μάθησης, επαγωγικής εξαγωγής συμπερασμάτων διαχείρισης μεγάλων και ετερογενών βάσεων δεδομένων.

1. ΒΑΣΙΚΕΣ ΑΡΧΕΣ ΕΦΑΡΜΟΓΩΝ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ

1.1 Ορισμός της Έννοιας Εξόρυξη Δεδομένων (Data Mining)

Εξόρυξη δεδομένων ορίζεται ως η εξεύρεση μιας πληροφορίας από μια βάση δεδομένων με χρήση αλγορίθμου ομαδοποίησης ή κατηγοριοποίησης και των αρχών της στατιστικής, της τεχνητής νοημοσύνης, της μηχανικής μάθησης και των συστημάτων βάσεων δεδομένων. Στόχος της εξόρυξης δεδομένων είναι η πληροφορία που θα εξαχθεί και τα πρότυπα που θα προκύψουν να έχουν δομή κατανοητή προς τον άνθρωπο έτσι ώστε να τον βοηθήσουν να πάρει τις κατάλληλες αποφάσεις. (Κουρής, 2006)

Εικόνα 1 Οι κυριότεροι τομείς αλληλεπίδρασης του data mining

Ο όρος εξόρυξη δεδομένων είναι μία έννοια η οποία παραπέμπει σε κάθε είδος φόρμας με μεγάλη ποσότητα δεδομένων ή επεξεργασία δεδομένων (συλλογή, εξαγωγή δεδομένων, warehouse, ανάλυση δεδομένων και στατιστικής) αλλά επίσης γενικεύεται σε κάθε είδος συστήματος υποστήριξης αποφάσεων συμπεριλαμβανομένου της τεχνητής νοημοσύνης, της εκμάθησης μηχανής και της επιχειρηματικής ευφυΐας. Στην ορθή χρήση του όρου η λέξη κλειδί είναι η ανακάλυψη, που ορίζεται ως η ανίχνευση κάτι καινούριου.

Ο πραγματικός στόχος της εξόρυξης δεδομένων είναι η αυτόματη ή ημιαυτόματη ανάλυση μεγάλων ποσοτήτων δεδομένα για την εξαγωγή κάποιου ενδιαφέροντος προτύπου που ήταν άγνωστο μέχρι εκείνη τη στιγμή, όπως ομάδες από εγγραφές δεδομένων (συσταδοποίηση), ασυνήθιστες εγγραφές (anomaly detection) και εξαρτήσεις (κανόνες συσχετίσεων). Αυτό συνήθως συμπεριλαμβάνει τη χρήση βάσης δεδομένων όπως χωρικά ευρετήρια. Αυτά τα πρότυπα ύστερα μπορούν να θεωρηθούν ως μία περιγραφή των δεδομένων εισαγωγής και να χρησιμοποιηθούν για περαιτέρω ανάλυση ή για παράδειγμα στην εκμάθηση μηχανής και στην προγνωστική ανάλυση. Για παράδειγμα, η εξόρυξη δεδομένων θα μπορούσε να προσδιορίσει πολλαπλά σύνολα στα δεδομένα, τα οποία μπορούν να χρησιμοποιηθούν μετά για να εξασφαλίσουν περισσότερο ακριβή αποτελέσματα από ένα σύστημα υποστήριξης αποφάσεων. Παρότι η συλλογή δεδομένων και η προετοιμασία δεδομένων, αλλά και η ερμηνεία των αποτελεσμάτων και εκθέσεων δεν αποτελούν μέρος της εξόρυξης δεδομένων, παρ' όλα αυτά ανήκουν στην ανακάλυψη γνώσης από βάσεις δεδομένων σαν κάποια επιπρόσθετα βήματα.

Άλλοι σχετικοί όροι της εξόρυξης δεδομένων είναι οι data dredging, data fishing και data snooping, που αναφέρονται στην χρήση μεθόδων της εξόρυξης δεδομένων για να πάρουν δείγματα από μεγαλύτερη συλλογή δεδομένων που είναι (ή μπορεί να είναι) πολύ μικρά για αξιόπιστα στατιστικά συμπεράσματα που έγιναν σχετικά με τη εγκυρότητα των προτύπων που ανακαλύφθηκαν. Αυτές οι μέθοδοι, επίσης, μπορούν να χρησιμοποιηθούν για την δημιουργία νέων υποθέσεων προς εξέταση έναντι μεγαλύτερων συλλογών δεδομένων.

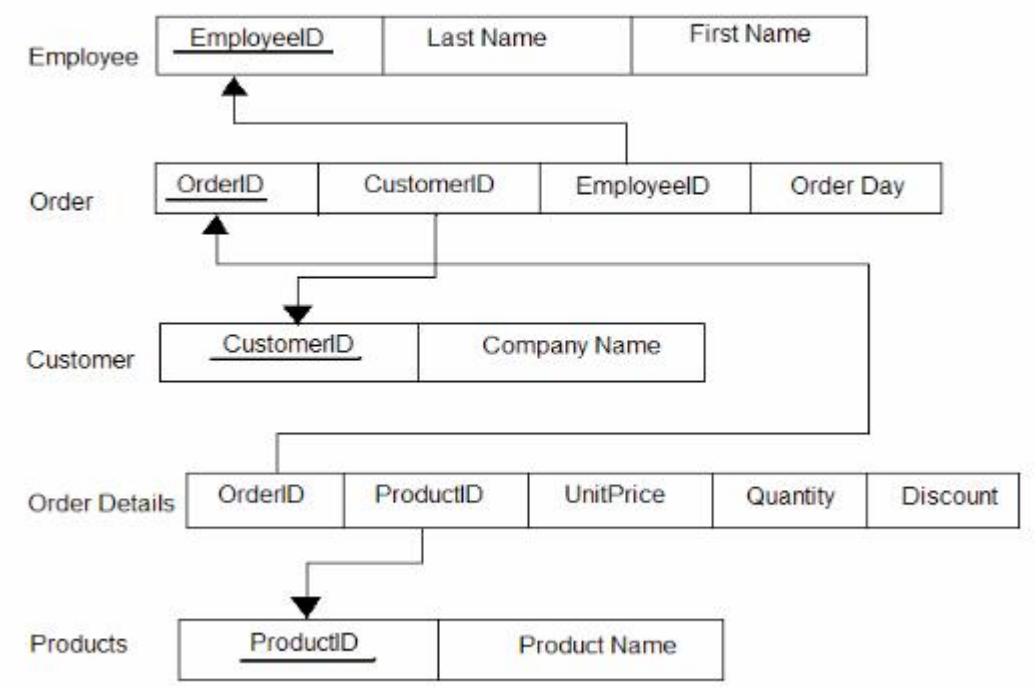
1.2 Είδη Βάσεων Δεδομένων όπου επιτρέπουν την Εξόρυξη Γνώσης

Η εφαρμογή τεχνικών εξόρυξης δεδομένων μπορεί θεωρητικά να εφαρμοστεί σε οποιαδήποτε είδος δεδομένων. Στην πράξη όμως ορισμένα είδη παρουσιάζουν και το μεγαλύτερο ενδιαφέρον από τους χρήστες. Στα επόμενα υποκεφάλαια παρουσιάζεται τα είδη των βάσεων δεδομένων όπως κατηγοριοποιούνται βάσει των αριθμών κλειδιών που τα χαρακτηρίζουν.

1.2.1 Σχεσιακές Βάσεις Δεδομένων

Ένα Σύστημα διαχείρισης βάσεων δεδομένων (ΣΔΒΔ) (database management system-DBMS) είναι μια συλλογή προγραμμάτων που επιτρέπουν στους χρήστες να δημιουργούν και να συντηρούν μια βάση δεδομένων. Τα ΣΔΒΔ συχνά απεικονίζουν τα δεδομένα σε μια δομή τύπου πίνακα. Αυτή είναι η αφορμή για την εισαγωγή του σχεσιακού μοντέλου (relational model), όπου τα δεδομένα απεικονίζονται να αποτελούνται από σχέσεις. Η πρόσβαση σε μια βάση δεδομένων συνήθως επιτυγχάνεται μέσω μιας γλώσσας ερωτήσεων (query language). Η πιο διαδεδομένη, που χρησιμοποιείται από τα περισσότερα ΣΔΒΔ, είναι η SQL. (Ξένος & Χριστοδουλάκης, 2000)

Μια σχεσιακή βάση δεδομένων είναι ουσιαστικά μια συλλογή από πίνακες, κάθε ένας από τους οποίους έχει ένα μοναδικό όνομα. Κάθε πίνακας αποτελείται από ένα σύνολο πεδίων (συνήθως στήλες) και σε αυτόν βρίσκονται αποθηκευμένα ένας μεγάλος αριθμός δεδομένων (εγγραφών). Κάθε εγγραφή σε έναν σχεσιακό πίνακα αναπαριστά ένα αντικείμενο και χαρακτηρίζεται από ένα μοναδικό “κλειδί”. Τα σχεσιακά δεδομένα μπορούν να επεξεργαστούν ή να αναλυθούν μέχρι κάποιο βαθμό με χρήση ερωτημάτων γραμμένων σε γλώσσα SQL ή με χρήση γραφικών περιβαλλόντων. Παράδειγμα τέτοιων ερωτημάτων θα μπορούσαν να ήταν “Δώσε μου τις πωλήσεις των τελευταίων 2 μηνών ανά κατάστημα” ή “ποια μετοχή είχε τη μεγαλύτερη μεταβολή το τελευταίο έτος”. Με την χρήση τεχνικών εξόρυξης δεδομένων τώρα κάποιος μπορεί να εισχωρήσει βαθύτερα στα δεδομένα και να ψάξει για μοτίβα ή τάσεις σε αυτά. Για παράδειγμα, ένα τέτοιο σύστημα μπορεί να αναλύσει τα δεδομένα των πελατών και να προβλέψει μελλοντικές συμπεριφορές βασισμένο σε προηγούμενα δεδομένα.



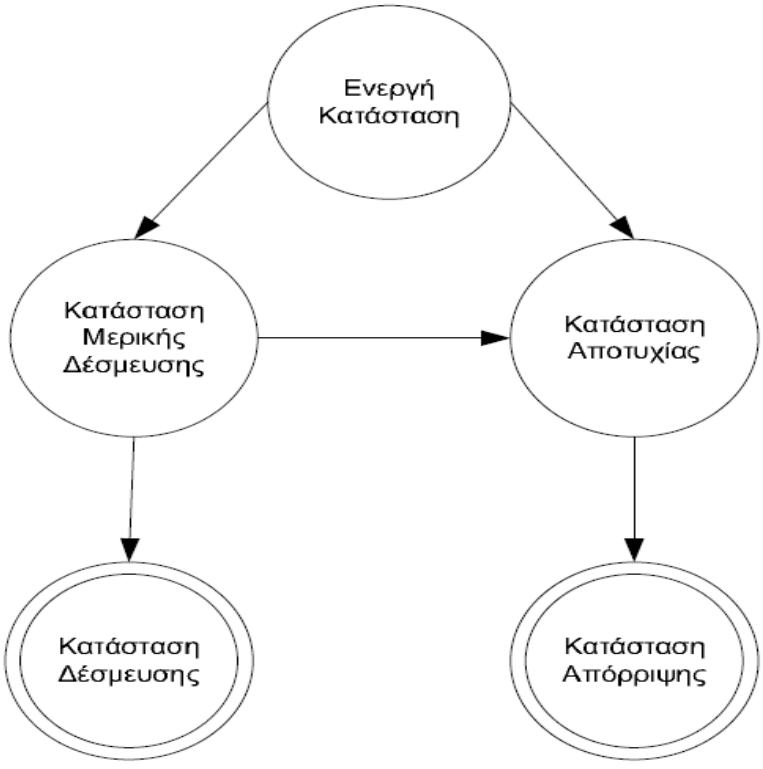
Εικόνα 2 Παράδειγμα του σχεσιακού μοντέλου

1.2.2 Βάσεις Δεδομένων Συναλλαγών

Ένα σύνολο λειτουργιών το οποίο αποτελεί μία λογική λειτουργική μονάδα καλείται συναλλαγή (transaction). Μια συναλλαγή περιέχει εντολές ανάγνωσης, εντολές εισαγωγής, εντολές διαγραφής, ή εντολές ενημέρωσης των δεδομένων της βάσης δεδομένων. Η πολυπλοκότητα των σύγχρονων βάσεων δεδομένων οδηγεί στην ανάπτυξη αξιόπιστων μηχανισμών ελέγχου της ακεραιότητας των δεδομένων πολλές συναλλαγές μπορεί να βρίσκονται σε εξέλιξη, προσπελάζοντας κοινά δεδομένα. Ο έλεγχος ταυτοχρονισμού (concurrency control) διαχειρίζεται τις συναλλαγές έτσι ώστε να αποφεύγονται οι παθολογικές καταστάσεις. Μια βάση δεδομένων μπορεί να βρεθεί σε ασταθή κατάσταση από μία βλάβη του συστήματος (βλάβη του φυσικού μέσου αποθήκευσης) ή (άλλου μέρους του συστήματος), μετά την αποκατάσταση της βλάβης πραγματοποιείται έλεγχος των δεδομένων. Υπάρχει ανάγκη ύπαρξης μηχανισμών επανάκτησης (recovery) δεδομένων, ώστε να επανέλθει η βάση δεδομένων στην κανονική της κατάσταση όπου πληρούνται όλοι οι περιορισμοί ακεραιότητας.

Γενικά μια βάση δεδομένων συναλλαγών αποτελείται από ένα αρχείο όπου κάθε εγγραφή αναπαριστά μια συναλλαγή. Μια συναλλαγή συνήθως περιλαμβάνει έναν μοναδικό αριθμό, και μια λίστα των αντικειμένων που αποτελούν την συναλλαγή (όπως τα προϊόντα που αγοράζονται σε ένα κατάστημα). Μια τέτοια βάση μπορεί να έχει επιπρόσθετους πίνακες συσχετισμένους με αυτή, οι οποίοι περιέχουν πρόσθετες πληροφορίες σχετικά με την πώληση, όπως την ημερομηνία και την ώρα πραγματοποίησης της συναλλαγής, τον κωδικό αριθμό του πελάτη, τον κωδικό αριθμό του πωλητή, το κατάστημα στο οποίο πραγματοποιήθηκε η συναλλαγή, κοκ. Με χρήση ενός συστήματος εξόρυξης δεδομένων μπορούμε να πραγματοποιούμε ανάλυση δεδομένων και να βρίσκουμε στοιχεία όπως π.χ. ποια προϊόντα αγοράζονται συνήθως μαζί και να προγραμματίζουμε έτσι καλύτερα την προώθηση τους. Αντίθετα ένα απλό σύστημα ανάκτησης πληροφορίας δεν είναι σε θέση να προχωρήσει σε τέτοιο βάθος στην ανάλυση και εξόρυξη των δεδομένων από τέτοιες βάσεις.

Μία συναλλαγή είναι συνήθως το αποτέλεσμα της εκτέλεσης ενός προγράμματος που είναι γραμμένο σε μία γλώσσα προγραμματισμού υψηλού επιπέδου. Οι εντολές που προσδιορίζουν μια συναλλαγή περικλείονται μεταξύ των εκφράσεων «begin transaction» και «end transaction». Μια συναλλαγή μπορεί να έχει τις εξής καταστάσεις:



Εικόνα 3 Καταστάσεις συναλλαγής

- Ενεργή (active) Κατάσταση. Η συναλλαγή εισέρχεται στην ενεργή κατάσταση κατά την αρχή της επεξεργασίας της και παραμένει σε αυτή ενόσω εκτελείται.
- Κατάσταση Μερικής Δέσμευσης (partial commit). Η συναλλαγή θεωρείται μερικώς δεσμευμένη όταν έχει ολοκληρωθεί και η τελευταία εντολή της συναλλαγής.
- Κατάσταση Αποτυχίας (failed). Η συναλλαγή αποτυγχάνει όταν το ΣΔΒΔ αντιληφθεί ότι δεν μπορεί να συνεχίσει την ομαλή επεξεργασία της.
- Κατάσταση Απόρριψης (aborted). Η συναλλαγή βρίσκεται στην κατάσταση αυτή όταν τα δεδομένα έχουν επανέλθει στην προηγούμενη σταθερή κατάσταση, πριν την αρχή της εκτέλεσης της συναλλαγής. Στην κατάσταση απόρριψης το ΣΔΒΔ έχει δύο επιλογές:
  (α) την επανεκκίνηση της συναλλαγής από την αρχή ή
  (β) την καταστροφή της συναλλαγής.
- Κατάσταση Δέσμευσης (committed). Η συναλλαγή έχει ολοκληρώσει επιτυχώς την εκτέλεσή της.

Οι Ιδιότητες Συναλλαγών (ACID) είναι οι εξής:

- Ατομικότητα (atomicity). Αν υπάρχει έστω και μία εντολή της συναλλαγής, η οποία αποτυγχάνει να εκτελεσθεί, τότε ολόκληρη η συναλλαγή αποτυγχάνει επίσης.
- Απομόνωση (isolation). Κάθε συναλλαγή πρέπει να εκτελείται ανεξάρτητα από άλλες συναλλαγές.
- Μονιμότητα (durability). Αν μία συναλλαγή ολοκληρωθεί με επιτυχία, τότε οι αλλαγές που έχει επιφέρει καταγράφονται μόνιμα στη βάση δεδομένων και δεν μπορούν να ανακληθούν.
- Συνέπεια (consistency). Η συναλλαγή πρέπει να μετατρέπει τη βάση δεδομένων από μία συνεπή κατάσταση σε μία άλλη συνεπή κατάσταση (τα δεδομένα πρέπει να είναι ορθά). Οι μηχανισμοί ακεραιότητας δεδομένων δεν επαρκούν για την εγγύηση της συνέπειας.

Θεωρείστε τη μεταφορά ενός ποσού από έναν τραπεζικό λογαριασμό σε άλλον. Η αφαίρεση του ποσού από τον πρώτο πρέπει να συνοδεύεται από την πρόσθεση του

ποσού στο δεύτερο. Αν το αφαιρούμενο ποσό διαφέρει από το προστιθέμενο, τότε τα δεδομένα της βάσης δεν έχουν συνέπεια (δεν είναι ορθά).

Οι τρόποι επεξεργασίας έχουν ως εξής:

- Ακολουθιακή ή σειριακή εκτέλεση. Οι συναλλαγές εκτελούνται η μία μετά την άλλη αλλά εμφανίζονται σημαντικές καθυστερήσεις με αποτέλεσμα να μειώνεται η γενική απόδοση του συστήματος.
- Ταυτόχρονη εκτέλεση πολλών συναλλαγών. Ο δίσκος ή η CPU μπορούν να απασχολούνται με άλλη συναλλαγή, οπότε αυξάνεται ο αριθμός των συναλλαγών που ολοκληρώνονται στη μονάδα του χρόνου (throughput). Το πλεονέκτημα είναι ότι μειώνονται οι καθυστερήσεις (waiting time) και ο μέσος χρόνος εκτέλεσης των συναλλαγών (mean response time), ενώ το μειονέκτημα είναι ότι ενδέχεται να επιφέρει προβλήματα στην ακεραιότητα και συνέπεια των δεδομένων της βάσης δεδομένων.

1.2.3 Χωρικές και Χρονικές Βάσεις Δεδομένων

Οι χωρικές βάσεις δεδομένων περιέχουν δεδομένα που καθορίζονται – περιλαμβάνουν μια χωρική διάσταση. Τέτοιες βάσεις είναι οι γεωγραφικές βάσεις (χάρτες), βάσεις σχετικά με την σχεδίαση VLSI κυκλωμάτων, ιατρικές εικόνες καθώς και εικόνες δορυφόρων. Για παράδειγμα σε μια βάση που έχει καταχωρημένη την κατανομή πλούτου σε σχέση με μια γεωγραφική περιοχή μπορούμε να ανακαλύψουμε τάσεις συγκεντρώσεων ή αραιώσεων πληθυσμών. Οι συγκεκριμένες βάσεις έχουν μια πληθώρα εφαρμογών όπως οικολογία, logistics, χωροταξία κ.α. Αναφορικά ορισμένες εργασίες σχετικές και με την εξόρυξη δεδομένων που χρησιμοποιούν χωρικά δεδομένα μπορούν να βρεθούν στα.

Οι χρονικές βάσεις έχουν, όπως εύκολα μπορεί να γίνει κατανοητό, δεδομένα τα οποία περιέχουν και τη χρονική διάσταση. Η διάσταση αυτή μπορεί να είναι απλά η ημερομηνία ή ώρα πραγματοποίησης ενός γεγονότος ή η καταχώρηση πολλαπλών τιμών χρονικής καταγραφής κάποιων παραμέτρων. Σε αυτού του είδους τις βάσεις οι τεχνικές εξόρυξης δεδομένων μπορούν να χρησιμοποιηθούν προκειμένου να βρουν μεταβολές σε σχέση με το χρόνο, ή τάσεις μεταβολής διαφόρων αντικειμένων. Τέτοιες πληροφορίες μπορεί να είναι ιδιαιτέρως χρήσιμες στην λήψη αποφάσεων ή

στην χάραξη στρατηγικής σε επιχειρήσεις. Για παράδειγμα οι μεταβολές των τιμών μετοχών σε σχέση με το χρόνο μπορεί να μας αποκαλύψουν πότε είναι η κατάλληλη περίοδος για αγορά ή πώληση μιας μετοχής. Στον τομέα της εξόρυξης δεδομένων έχουν γίνει κάποιες εργασίες πάνω σε χρονικά δεδομένα. (Silberschatz & Korth & Sudarshan, 2011)

1.2.4 Βάσεις Κειμένων

Οι βάσεις κειμένων είναι βάσεις οι οποίες περιέχουν λέξεις ή ολόκληρα κείμενα, ή εναλλακτικά που περιέχουν λεκτικές περιγραφές αντικειμένων. Αυτές οι περιγραφές μπορούν να κυμαίνονται από απλές λέξεις κλειδιά, μέχρι ολόκληρες προτάσεις, όπως για παράδειγμα περιγραφές προϊόντων, απαντήσεις σε ερωτήματα – παράπονα χρηστών σε ένα call-center κ.α. Η πληροφορία που μπορεί να ανακαλύψει κάποιος από τέτοιες βάσεις είναι ανεξάντλητη. Παράδειγμα από μια βάση κειμένων μπορεί να δημιουργήσει έναν θησαυρό λέξεων, ή μια λίστα συνωνύμων. Από μια βάση παραπόνων – απαντήσεων χρηστών σε ένα call – center μπορεί πάλι να δημιουργήσει μια λίστα σχετικών αυτοματοποιημένων απαντήσεων σε αντίστοιχα ερωτήματα.

1.2.5 Πολυμεσικές Βάσεις

Οι βάσεις πολυμέσων αποθηκεύουν ήχο, στατική και κινούμενη εικόνα, καθώς και κείμενο και έχουν ποικίλες εφαρμογές. Οι συγκεκριμένες βάσεις είναι συνήθως πάρα πολύ μεγάλες σε μέγεθος λόγω της φύσεως των δεδομένων που αποθηκεύουν. Η χρήση των τεχνικών εξόρυξης δεδομένων μπορεί να μας απαλλάξει από διάφορα προβλήματα και δυσκολίες που συναντάμε στις συγκεκριμένες εφαρμογές, όπως την εύρεση και την εξαγωγή πολλαπλών χαρακτηριστικών από τα πολυμεσικά δεδομένα, εύρεση με βάση κάποια μετρική ομοιότητας κ.α.

Τα πολυμεσικά δεδομένα αποτελούνται από την περιγραφική πληροφορία (π.χ., τίτλος ταινίας) και την πληροφορία περιεχομένου (content). Ένας τρόπος διαχείρισης των πολυμεσικών τύπων δεδομένων χρησιμοποιώντας ένα παραδοσιακό ΣΔΒΔ είναι να αποθηκεύσουμε την περιγραφική πληροφορία στη βάση δεδομένων του ΣΔΒΔ και να χρησιμοποιήσουμε εξωτερικά αρχεία για την αποθήκευση του περιεχομένου. Το βασικό μειονέκτημα αυτής της προσέγγισης είναι ότι δεν μπορούμε να

χρησιμοποιήσουμε τη λειτουργικότητα του ΣΔΒΔ για το περιεχόμενο των τύπων δεδομένων (π.χ. την κατασκευή δομών καταλόγων / ευρετηρίων - indexes). Μπορεί επίσης να καταλήξει σε ασυνέπειες, όπως ένα αρχείο που είναι σημειωμένο στην βάση δεδομένων, αλλά του οποίου τα περιεχόμενα λείπουν ή το αντίστροφο. Συνεπώς είναι επιθυμητό να αποθηκεύονται στην βάση δεδομένων τα ίδια τα δεδομένα. (Μανωλόπουλος & Παπαδόπουλος, 2009)

Οι περισσότερο γνωστοί τύποι δεδομένων πολυμέσων που είναι διαθέσιμοι στις Βάσεις Δεδομένων Πολυμέσων είναι οι παρακάτω. (Elmasri & Navathe, 2007)

· Κείμενο: Μπορεί να είναι ή να μην είναι μορφοποιημένο. Για ευκολία ανάλυσης δομημένων εγγράφων, χρησιμοποιούνται πρότυπα όπως η SGML και παραλλαγές όπως η HTML.

· Γραφικά: Παραδείγματα περιλαμβάνουν σχέδια και εικονογραφήσεις που κωδικοποιούνται με χρήση κάποιου πρότυπου (πχ., CGM, PICT, postscript).

· Εικόνες: Περιλαμβάνουν σχέδια, φωτογραφίες, κοκ., κωδικοποιημένα σε τυπικές μορφοποιήσεις όπως bitmap, JPEG, και MPEG. Στα JPEG, και MPEG υπάρχει συμπίεση. Οι εικόνες αυτές δεν διαιρούνται σε επί μέρους στοιχεία. Επομένως τα ερωτήματα με βάση το περιεχόμενο (πχ., βρες όλες τις εικόνες που περιέχουν κύκλους) δεν είναι εύκολες.

· Κινούμενες Εικόνες: Χρονικές ακολουθίες από δεδομένα εικόνων ή γραφικών.

· Βίντεο: Ένα σύνολο από φωτογραφικά δεδομένα σε χρονική ακολουθία με καθορισμένο ρυθμό - για παράδειγμα 30 καρέ το δευτερόλεπτο.

· Δομημένος Ήχος: Μια ακολουθία από στοιχεία ήχου που περιλαμβάνουν νότες, τόνο, διάρκεια, κοκ.

· Ήχος: Δειγματοληπτικά δεδομένα από ηχητικές ηχογραφήσεις σαν συμβολοσειρές από bits σε ψηφιακή μορφή. Τυπικά οι αναλογικές ηχογραφήσεις μετατρέπονται σε ψηφιακή μορφή πριν την αποθήκευση.

· Σύνθετα Δεδομένα Πολυμέσων: Ένας συνδυασμός από τύπους δεδομένων πολυμέσων όπως ήχος και βίντεο που μπορεί να αναμειγνύονται φυσικά για να δώσουν ένα νέο τύπο μορφοποίησης αποθήκευσης ή λογική ανάμειξη ενώ διατηρούν τους αρχικούς τύπους και τις μορφοποιήσεις.

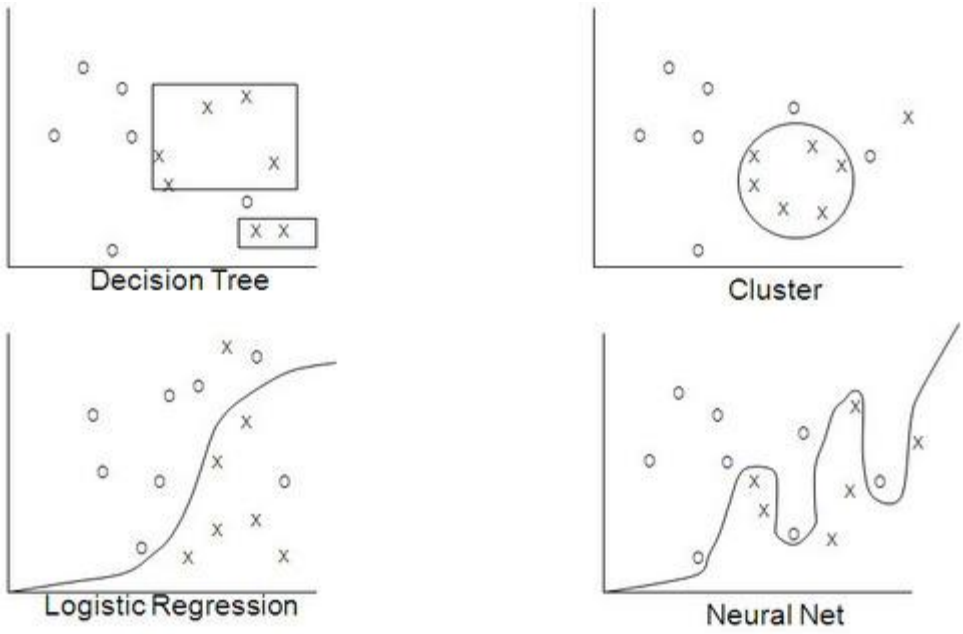
1.3 Τεχνικές Εξόρυξης Γνώσεις

Οι πιο σημαντικές κατηγορίες τεχνικών εξόρυξης δεδομένων είναι οι ακόλουθες :

- Κατηγοριοποίηση - Classification
- Συσταδοποίηση - Clustering
- Κανόνες Συσχέτισης - Association Rules

1.3.1 Τεχνική Κατηγοριοποίησης

Η κατηγοριοποίηση είναι μία από τις πιο σημαντικές κατηγορίες τεχνικών εξόρυξης δεδομένων. Απώτερος στόχος της είναι η ανάθεση ενός νέου αντικειμένου σε μία από τις προϋπάρχουσες κλάσεις (κατηγορίες). Η κατηγοριοποίηση προϋποθέτει πρότερη γνώση για τα δεδομένα, καθώς οι προϋπάρχουσες κλάσεις είναι από πριν δεδομένες και ταυτόχρονα η δημιουργία και εκπαίδευση του μοντέλου κατηγοριοποίησης στηρίζεται σε ένα σύνολο από προ-κατηγοριοποιημένα αντικείμενα. Το μοντέλο κατηγοριοποίησης εξετάζει κάθε νέο αντικείμενο και, ανάλογα με τις τιμές των χαρακτηριστικών του, εφαρμόζεται σε μία από τις υπάρχουσες κλάσεις.



Εικόνα 4 Διαγραμματική απεικόνιση Αλγόριθμων κατηγοριοποίησης

Στη βιβλιογραφία υπάρχει ποικιλία αλγορίθμων κατηγοριοποίησης, με τον καθένα να καταρτίζεται συνήθως σε έναν συγκεκριμένο τομέα δεδομένων. Σε καθολικό επίπεδο, οι αλγόριθμοι κατηγοριοποίησης μπορούν να διαχωριστούν σε δύο μεγάλες κατηγορίες, όπως εστιάζονται στον ακόλουθο πίνακα.

Πίνακας 1 Κατηγορίες αλγορίθμων

Δένδρα απόφασης (decision trees)	Οι αλγόριθμοι της κατηγορίας αυτής δημιουργούν ένα δενδρικό μοντέλο κατηγοριοποίησης, το οποίο βασίζεται στην τεχνική «διαίρει και βασίλευε». Με τον τρόπο αυτό ο συνολικός χώρος αναζήτησης χωρίζεται σε ορθογώνια υποσύνολα, με βάση συνθήκες ελέγχου των τιμών συγκεκριμένων χαρακτηριστικών.
Νευρωνικά δίκτυα (neural networks)	Οι αλγόριθμοι αυτοί μοντελοποιούνται με βάση τις λειτουργίες του ανθρώπινου εγκεφάλου. Ένα νευρωνικό δίκτυο μπορεί να θεωρηθεί σαν ένας κατευθυνόμενος γράφος με πηγή (είσοδος), καταβόθρα (έξοδος) και εσωτερικούς (κρυμμένους) κόμβους. Πρέπει να ανεφερθεί ότι τα νευρωνικά δίκτυα δεν είναι κατάλληλα για εφαρμογές πραγματικού χρόνου αφού αυτά απαιτούν μακρύ χρόνο εκπαίδευσης.

1.3.2 Τεχνική Συσταδοποίησης

Συσταδοποίηση[1] (Clustering) είναι η διαδικασία διαίρεσης ενός συνόλου δεδομένων σε αμοιβαία αποκλειόμενες ομάδες, τέτοιες ώστε τα μέλη κάθε ομάδας να είναι όσο κοντά γίνεται το ένα με το άλλο, ενώ οι διαφορετικές ομάδες να είναι όσο το δυνατόν πιο μακριά η μια από την άλλη, όπου η απόσταση μετράται σε σχέση με τις διαθέσιμες μεταβλητές. Σίγουρα, με την αναπαράσταση των δεδομένων με λιγότερες συστάδες χάνονται ορισμένες μικρολεπτομέρειες, αλλά επιτελείται απλοποίηση (μια συστάδα είναι μια ταξινομημένη λίστα αντικειμένων, τα αντικείμενα της οποίας έχουν κάποια κοινά χαρακτηριστικά).

[1] Η συσταδοποίηση μπορεί να βρεθεί με διαφορετικά ονόματα σε διαφορετικά πεδία, όπως *μη εποπτευόμενη μάθηση* (*unsupervised learning*) στην αναγνώριση προτύπων, *αριθμητική ταξονομία* (numerical taxonomy) στην βιολογία, οικολογία, *τυπολογία*, στις κοινωνικές επιστήμες και *τμηματοποίηση* (segmentation, partitioning) στη θεωρία των γράφων και στις Βάσεις Δεδομένων

Η συσταδοποίηση διαδραματίζει σπουδαίο ρόλο σε εφαρμογές εξόρυξης δεδομένων, όπως είναι η διερεύνηση επιστημονικών δεδομένων, η ανάκτηση πληροφορίας και η εξόρυξη κειμένου, εφαρμογές χωρικών βάσεων δεδομένων, web ανάλυση, CRM, μάρκετινγκ, ιατρική διάγνωση, υπολογιστική βιολογία και πολλά άλλα.

Η θεμελιώδης διάκριση μεταξύ της συσταδοποίησης και της κατηγοριοποίησης είναι ότι στη δεύτερη, ένα σύνολο από προ-ομαδοποιημένα αντικείμενα είναι διαθέσιμο, και αυτό που απαιτείται είναι να τοποθετήσουμε και να εφαρμόσουμε ένα νέο αντικείμενο σε κάποια από τις τρέχουσες ομάδες. Απεναντίας, στη συσταδοποίηση δε διαθέτουμε καμία πρότερη γνώση για τις ομάδες/συστάδες στις οποίες είναι απαραίτητο να διαχωρισθούν τα δεδομένα. Συνεπώς, η συσταδοποίηση παράγεται μόνο από τα δεδομένα και είναι απόλυτα οδηγούμενη από αυτά.

Η συσταδοποίηση είναι πολύ ωφέλιμη σε ένα σύνολο γνήσιων εφαρμογών, όπως η ανάλυση προτύπων (pattern analysis), η λήψη αποφάσεων (decision making), η ανάκτηση πληροφορίας (information retrieval) κ.α. Στην πλειοψηφία των περιπτώσεων που υλοποιείται η συσταδοποίηση, ανακύπτει μικρή ή καθόλου γνώση για την δομή και το είδος των δεδομένων. Σε ανάλογες περιπτώσεις, η συσταδοποίηση των δεδομένων είναι αρμόδια για την ανεύρεση αλληλοσχετισμών μεταξύ των δεδομένων με σκοπό να κατανοηθεί η δομή τους, το οποίο είναι και ο απώτερος στόχος.

Η συσταδοποίηση αποτελείται από 4 διαφορετικές φάσεις:

Στη φάση της αναπαράστασης, επιλέγεται η βέλτιστη δομή αναπαράστασης των δεδομένων, για να επικυρωθούν τα χαρακτηριστικά εκείνα που είναι αναγκαία και χρήσιμα για τη συσταδοποίηση. Ακολούθως, κατά τη φάση της μοντελοποίησης, επιλέγεται το μέτρο ομοιότητας μεταξύ των αναπαράστασεων των αντικειμένων, καθώς και οι αρχικές συστάδες. Η συνάρτηση απόστασης καθορίζει το μέτρο ομοιότητας μεταξύ των αντικειμένων. Στην επόμενη φάση της συσταδοποίησης, οι τελικές συστάδες αναπτύσσεται με την υλοποίηση του καθεαυτού αλγορίθμου

συσταδοποίησης. Τέλος, οι παραχθέντες συστάδες βελτιστοποιούνται με βάση ορισμένες μετρικές αξιολόγησης της ποιότητάς τους: ορισμένες συστάδες είναι πιθανόν να συγχωνευθούν μεταξύ τους ή να διαιρεθούν σε αντίστοιχες συστάδες.
Στη βιβλιογραφία υπάρχει ποικιλία αλγορίθμων συσταδοποίησης, με τον καθένα να εξειδικεύεται κατά κανόνα σε ένα συγκεκριμένο είδος δεδομένων. Στον ακόλουθο πίνακα παρουσιάζονται δυο από τις πιο σημαντικές κατηγορίες αλγορίθμων συσταδοποίησης με τα χαρακτηριστικά τους.

Πίνακας 2 Κατηγορίες αλγορίθμων συσταδοποίησης

Διαχωριστικοί (partitional)	Οι επαναληπτικοί αλγόριθμοι συσταδοποίησης βασίζονται σε μια αρχική εκτίμηση των συστάδων (είτε τυχαία ή βασισμένη σε κάποιες παραδοχές ή γνώση των δεδομένων) και στη συνέχεια παράγουν ένα διαμερισμό των αντικειμένων σε συστάδες.
Ιεραρχικοί (hierarchical)	Οι αλγόριθμοι της κατηγορίας αυτής δημιουργούν ένα ιεραρχικό δένδρο από συστάδες, επιτρέποντας σε μία συστάδα να έχει ένα σύνολο από υποσυστάδες. Ανάλογα με τον τρόπο κατασκευής του δένδρου ιεραρχίας, οι αλγόριθμοι διακρίνονται σε bottom-up (ή agglomerative) και top-down. Οι bottom-up αλγόριθμοι θεωρούν κάθε αντικείμενο ως μια ξεχωριστή συστάδα και σε κάθε βήμα συγχωνεύουν τις δύο πιο κοντινές συστάδες, μέχρις ότου απομείνει μόνο μία συστάδα η οποία περιλαμβάνει όλα τα αντικείμενα. Αντίθετα, οι top-down αλγόριθμοι ξεκινούν από μια μοναδική συστάδα, η οποία περιέχει όλα τα αντικείμενα, και σε κάθε βήμα χωρίζουν τις υπάρχουσες συστάδες σε επιμέρους υποσυστάδες.

Μία αποτελεσματική συσταδοποίηση έχει ως πρωταρχικούς στόχους:

- Την ελαχιστοποίηση του σφάλματος που παρουσιάζεται και,

- Τον περιορισμό, όσο αυτό είναι δυνατόν, του συνολικού αριθμού των διαφορετικών ομάδων που θα εξαχθούν ,κατά την τελειοποίηση της διαδικασίας.

Όπως αποδεικνύεται, τα δύο παραπάνω κριτήρια αποτελεσματικότητας είναι αντικρουόμενα. Αυτό συμβαίνει διότι η δημιουργία όλο και λιγότερων ομάδων αυξάνει την εισαγωγή σφάλματος και σημαντικό αντικείμενο έρευνας οφείλει να αποτελέσει ο σωστός συνδυασμός αυτών των δύο.

Η επιτυχία ή η αποτυχία στη διαδικασία συσταδοποίησης, δεν μπορεί να αποφανθεί μετά την ολοκλήρωση της. Και τούτο, διότι δεν μπορεί να οριστεί ένα απόλυτα «καλό κριτήριο», το οποίο θα ήταν αυτάρκες από τον τελικό στόχο του Clustering. Ένα διαφορετικό κριτήριο, μία διαφορετική μέθοδος υλοποίησης ή ακόμα και μία διαφορετική επιλογή παραμέτρων, ορισμένων αναλόγως με το επιθυμητό αποτέλεσμα, είναι δυνατόν να κατευθυνθεί σε ανόμοια αποτελέσματα συσταδοποίησης. Παραδείγματος χάρη, αντικείμενο ενδιαφέροντος είναι δυνατόν να οριστεί ο εντοπισμός αντιπροσωπευτικών ομάδων, «φυσικών» συστάδων, που να περιγράφουν άγνωστα για το αντικείμενο χαρακτηριστικά, ή ο εντοπισμός ασυνήθιστων και περίεργων αντικειμένων που ανήκουν στα δεδομένα.

1.3.3 Τεχνική Κανόνων Συσχέτισης

Οι κανόνες συσχέτισης αποτελούν μια σύγχρονη μέθοδο για την εξαγωγή γνώσης από μεγάλες βάσεις δεδομένων. Η εμφάνισή τους οφείλεται στις ανάγκες ανάλυσης του «καλαθιού αγοράς» (market basket analysis). Οι κανόνες συσχέτισης εφαρμόστηκαν για να αποκαλύπτουν ενδιαφέρουσες σχέσεις μεταξύ των δεδομένων. Οι σχέσεις αυτές παρουσιάζονται στη μορφή A->B, όπου το A και το B αναφέρονται στα σύνολα γνωρισμάτων που υπάρχουν στα δεδομένα υπό ανάλυση.

Για την αξιολόγηση της σποδαιότητας του κάθε κανόνα συσχέτισης, έχουν προταθεί ορισμένα μέτρα, όπως είναι η υποστήριξη, η εμπιστοσύνη, η κάλυψη. Στον ακόλουθο πίνακα διακρίνονται αναλυτικά τα χαρακτηριστικά αυτών των μέτρων :

Πίνακας 3

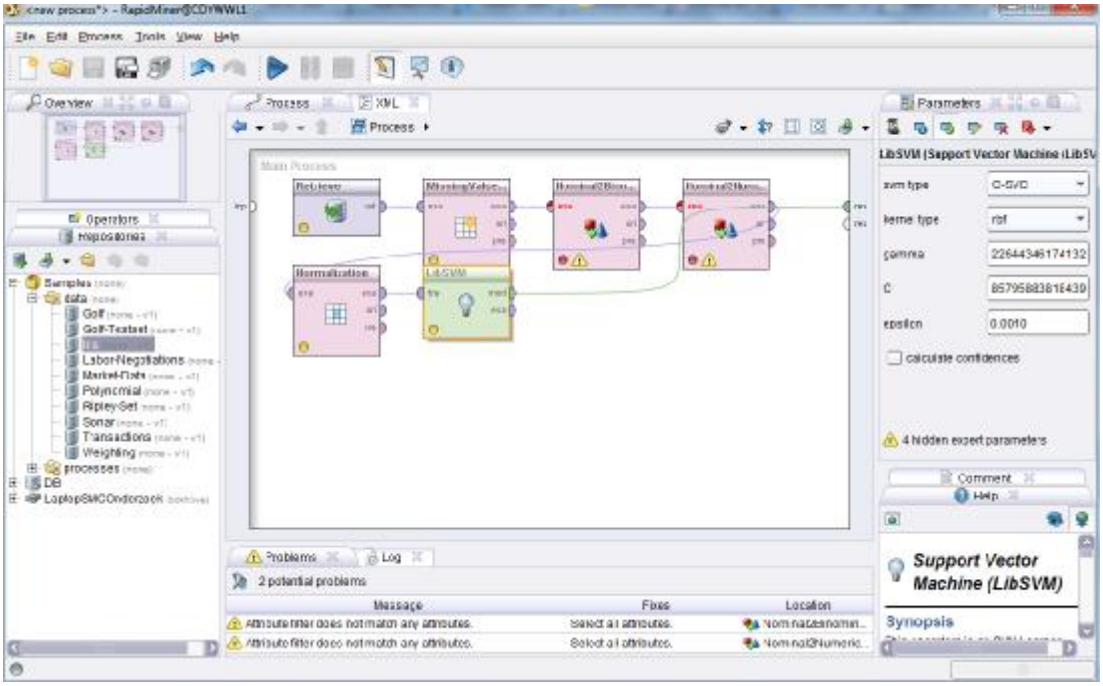
Υποστήριξη (Support)	Η υποστήριξη ενός κανόνα συσχέτισης είναι το ποσοστό όλων των περιπτώσεων στο σύνολο δεδομένων που ικανοποιούν έναν κανόνα, δηλαδή ικανοποιούν το αριστερό και το δεξιό μέλος του κανόνα.
Εμπιστοσύνη (Confidence)	Η εμπιστοσύνη ενός κανόνα συσχέτισης είναι το πηλίκο της υποστήριξης του κανόνα προς το ποσοστό των περιπτώσεων που καλύπτονται από το αριστερό μέλος του κανόνα.
Κάλυψη (Coverage)	Η κάλυψη ενός κανόνα συσχέτισης είναι το ποσοστό των περιπτώσεων των δεδομένων που έχουν τις τιμές των γνωρισμάτων που ορίζονται στο αριστερό μέλος του κανόνα. Ένας κανόνας συσχέτισης με τιμή κάλυψης κοντά στο 1, μπορεί να θεωρηθεί ως κανόνας με ενδιαφέρον.
Lift.	Το lift ορίζεται ως η εμπιστοσύνη του κανόνα διαιρούμενη με το ποσοστό όλων των περιπτώσεων που καλύπτονται από το δεξιό μέλος του κανόνα. Είναι ένα μέτρο της σπουδαιότητας της συσχέτισης και είναι ανεξάρτητο από την κάλυψη.

1.4 Λογισμικά Εξόρυξης Δεδομένων

Απαραίτητη κρίνεται η ύπαρξη κατάλληλου λογισμικού και υπολογιστικών συστημάτων. Αυτό οφείλεται στον αυξανόμενο όγκο των δεδομένων που αποθηκεύονται διαρκώς ψηφιακά, προκειμένου να διατηρηθεί και να αξιοποιηθεί η ωφέλιμη πληροφορία. Ακολουθεί συνοπτική παρουσίαση των σημαντικότερων προγραμμάτων που εφαρμόζουν τεχνικές εξόρυξης δεδομένων, έτσι όπως παρουσιάζονται στο διαδίκτυο.

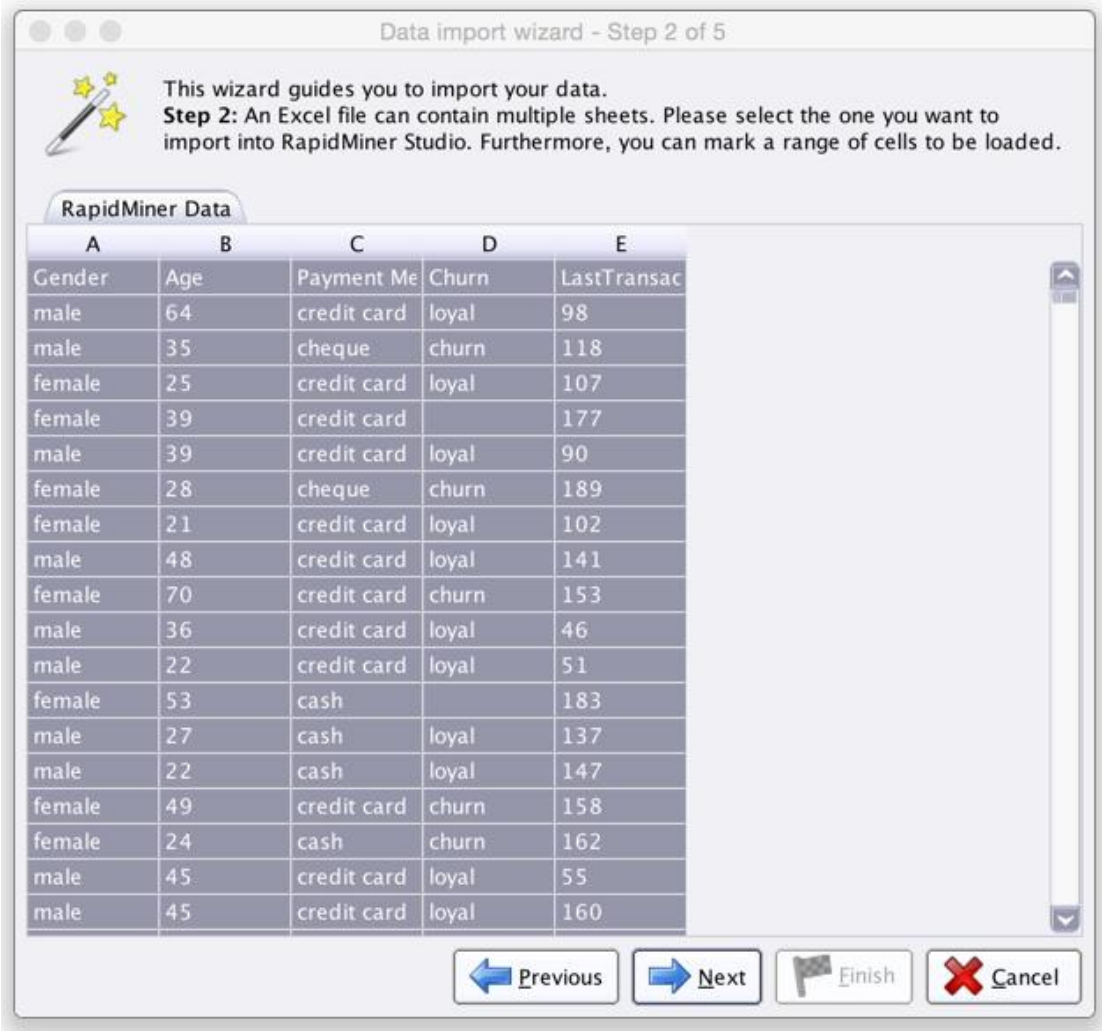
1.4.1 RapidMiner

Το RapidMiner, πρώην YALE (Yet Another Learning Environment), αποτελεί ένα περιβάλλον για μηχανική εκμάθηση, εξόρυξη δεδομένων, προβλεπτική και επιχειρησιακή ανάλυση. Υλοποιείται στην έρευνα, εκπαίδευση, ανάπτυξη εφαρμογών, γρήγορη προτυποποίηση και σε βιομηχανικές εφαρμογές. Διανέμεται υπό την άδεια χρήσης λογισμικών ανοιχτού κώδικα AGPL, και στεγάζεται στο SourceForge από το 2004. Είναι υλοποιημένο σε Java.



Εικόνα 5 Φύλλο εργασίας Rapidminer

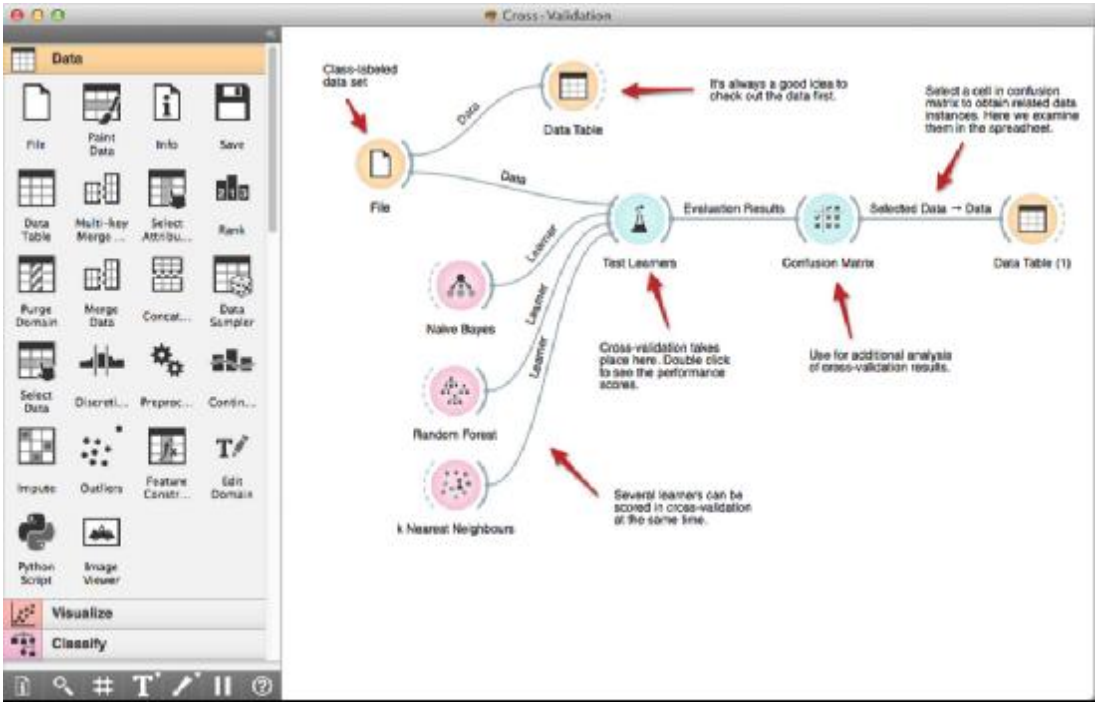
Το RapidMiner είναι δυνατόν να χρησιμοποιηθεί σε εξόρυξη κειμένων, εξόρυξη πολυμέσων, μηχανική χαρακτηριστικών, εξόρυξη ροής δεδομένων και ανίχνευση εννοιών παρέκκλισης, ανάπτυξη μεθόδων συνόλου και επιμεριστική εξόρυξης δεδομένων. Το RapidMiner συναντάται στο τομέα της ηλεκτρονικής, ενέργειας, πληροφορικής, φαρμακευτική και αυτοκινητιστική βιομηχανία, εμπόριο, αεροπλοΐα, τηλεπικοινωνίες, τραπεζικό και ασφαλιστικό κλάδο, στη παραγωγική διαδικασία, έρευνα αγοράς και άλλα πολλά πεδία.



Εικόνα 6 Εισαγωγή της βάσης δεδομένων στο λογισμικό rapidminer

1.4.2 Orange

Το Orange αποτελεί μία βιβλιοθήκη αντικειμένων πυρήνα και ρουτινών της C++ και περιλαμβάνει μία μεγάλη ποικιλία ορισμένων κύριων και ορισμένων όχι τόσο βασικών αλγορίθμων μηχανικής εκμάθησης και εξόρυξης δεδομένων. Συνάμα, περιέχει ρουτίνες για εισαγωγή και χειρισμό δεδομένων. Επιπλέον, το περιβάλλον της επιτρέπει τη δημιουργία κώδικα για γρήγορη προτυποποίηση νέων αλγορίθμων και έλεγχο συστημάτων. Πρόκειται για μια συλλογή από υπό-προγράμματα (modules) Python, τα οποία περιλαμβάνονται στη βασική βιβλιοθήκη και εφαρμόζουν κάποια λειτουργία για την οποία ο χρόνος εκτέλεσης δεν είναι σημαντικός και που γίνεται πιο εύκολα με Python παρά με C++.



Εικόνα 7 Στιγμιότυπο από το λογισμικό orange

1.5 Πεδία Εφαρμογής Μεθόδων Εξόρυξης Δεδομένων

Από τα πιο ενδιαφέροντα ερευνητικά πεδία του γενικού τομέα εξόρυξης δεδομένων είναι οι εφαρμογές εξόρυξης δεδομένων στον παγκόσμιο ιστό. Ο όρος βάση δεδομένων εδώ χρησιμοποιείται με θεωρητικό τρόπο, καθώς στην πραγματικότητα δεν υπάρχει πρακτικά δομή ή σχήμα στον παγκόσμιο ιστό. Αυτό κάνει ακόμα πιο επιτακτική την ανάγκη για εξόρυξη δεδομένων στον παγκόσμιο ιστό, παρέχοντας τεράστια βοήθεια σε κάθε είδους χρήστη. Με τον όρο εξόρυξη γνώσης στον παγκόσμιο ιστό δεν αναφερόμαστε μόνο σε δεδομένα που περιέχονται σε ιστοσελίδες αλλά και σε δεδομένα που έχουν να κάνουν με τη δραστηριότητα ενός χρήστη σε αυτό.

1.5.1 Μάρκετινγκ

Μια κατηγορία πολύ γνωστών εφαρμογών εξόρυξης δεδομένων είναι αυτές του μάρκετινγκ. Αυτό είναι αναμενόμενο μιας και μεγάλες εταιρίες χρησιμοποιούν μεγάλα συστήματα διαχείρισης δεδομένων για να διαχειρίζονται μεγάλο αριθμό πελατών και οικονομικών στοιχείων. (Κολλάρας, 2007)Τα τελευταία χρόνια οι τάσεις του μάρκετινγκ ορίζουν μια πολιτική έρευνας των αναγκών των πελατών. Αναζητούν απαντήσεις σε ερωτήματα όπως, τι είναι αυτό που θέλουν οι πελάτες, ποιες είναι οι ανάγκες τους κ.α. Ο τομέας της εξόρυξης δεδομένων έχει συνεισφέρει σημαντικά σε αυτή την κατεύθυνση από την ανάλυση δεδομένων μια επιχείρησης και την εξαγωγή χρήσιμων συμπερασμάτων για την συμπεριφορά των πελατών. Ένας αρκετά γνωστός αλγόριθμος εξόρυξης δεδομένων είναι ο A-Priori. Ο αλγόριθμος αυτός κάνει ανάλυση δεδομένων αγοράς, όπου υπάρχουν δεδομένα σχετικά με πελάτες ή αγορών σε καταστήματα. Ο A-Priori μπορεί αποδοτικά να δώσει συμπεράσματα όπως «κάθε πελάτης που αγοράζει βαμβακερά υφάσματα θα αγοράσει και μπίρα με μεγάλη πιθανότητα». (Τσιράκης, 2006)

Άλλα παραδείγματα εξόρυξης δεδομένων στο μάρκετινγκ είναι η ανάλυση της συμπεριφοράς των πελατών ηλεκτρονικών καταστημάτων χρησιμοποιώντας τα log αρχεία ή η πρόβλεψη εάν ένας πελάτης θα αγοράσει ένα συγκεκριμένο προϊόν χρησιμοποιώντας παρελθοντικές του κινήσεις.

1.5.2 Επενδύσεις

Πολυάριθμες χρηματιστηριακές εταιρίες χρησιμοποιούν τεχνικές εξόρυξης δεδομένων έτσι ώστε να μπορούν να γνωρίζουν που να επενδύσουν. Στην πραγματικότητα μια μεγάλη μερίδα έρευνας στο τομέα εξόρυξης δεδομένων έχει γίνει έχοντας ως αφετηρία χρηματιστηριακές εφαρμογές. Μια άλλη χρήση των τεχνικών εξόρυξης δεδομένων είναι οι εφαρμογές εξόρυξης δεδομένων από κείμενα. Για παράδειγμα αλγόριθμοι που εξάγουν χρήσιμη πληροφορία από μη δομημένα κείμενα, έτσι ώστε να προβλεφθούν οι τάσεις σε μετοχές. (Μαρκέλλου, 2005)

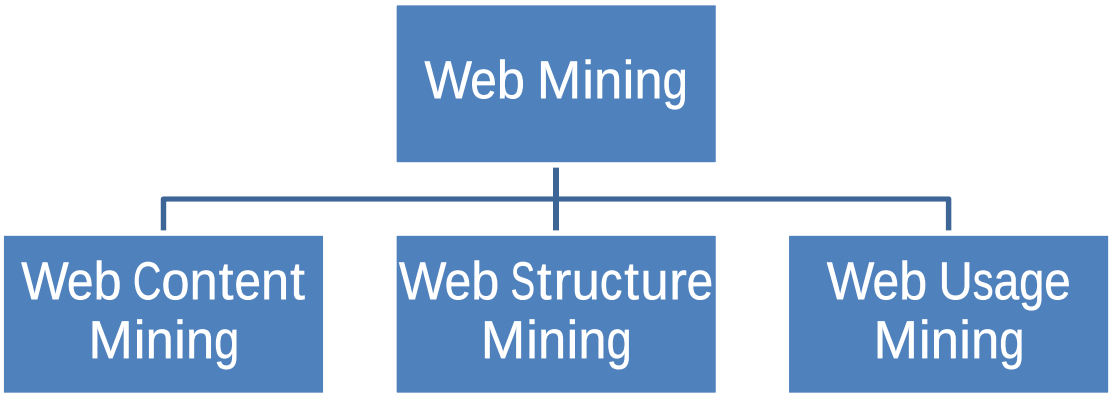
1.5.3 Πρόληψη και Ασφάλεια

Η εξόρυξη δεδομένων έχει με επιτυχία εφαρμοστεί και στην πρόληψη και αποφυγή διάφορων τύπων απάτης. Από την αναγνώριση κακόβουλων ενεργειών σε συναλλαγές κάποιος μπορεί να αντιληφθεί συναλλαγές που μπορεί να σχετίζονται με οικονομικές παρανομίες ή άλλου είδους απάτες. Ένα παράδειγμα συστήματος είναι το FAIS. Ωστόσο τα τελευταία χρόνια, όπως βλέπουμε και ακούμε, υπάρχει μια τάση για πρόληψη σε κακόβουλες ενέργειες. (Κουρής, 2006)Οι κινήσεις μας σε δημόσιους χώρους καταγράφεται όπως και αυτές που έχουν να κάνουν με τον παγκόσμιο ιστό. Για παράδειγμα μια πρόσφατη εφαρμογή μπορούσε να αναγνωρίζει ανώμαλα πρότυπα χρησιμοποιώντας κανόνες σε δεδομένα νοσοκομείων έτσι ώστε να αναγνωρίζει, σε πραγματικό χρόνο, εμφάνιση ασθενειών.

1.5.4 Παγκόσμιος Ιστός

Ο τομέας της εξόρυξης δεδομένων είχε άμεση εφαρμογή με επιτυχία στο Διαδίκτυο. Το πιο δημοφιλές παράδειγμα εξόρυξης δεδομένων στο διαδίκτυο είναι η Google. Για να γίνει πιο κατανοητή η σημαντικότητα της συνεισφοράς αυτής θα πρέπει να αντιληφθούμε πως ο όγκος της πληροφορίας που υπάρχει μέχρι τώρα στο διαδίκτυο είναι αδύνατο να μετρηθεί με ακρίβεια. Οι σελίδες που κάθε φορά ερευνά η Google δηλώνεται πως είναι περίπου 4,285,199,774. Κάθε ερώτημα στην μηχανή αναζήτησης δεν ξεπερνά σε χρόνο τα δυο δευτερόλεπτα.

Η Google και γενικά ο τομέας της εξόρυξης δεδομένων στο Διαδίκτυο έχουν σήμερα τεράστια επιτυχία γιατί έχουν εκπληρώσει δυο σημαντικούς στόχους. (Πλέγας, 2013) Πρώτα, μπορούν να κάνουν αναζήτηση (με κάθε ερώτημα) σε τόσα πολλά δεδομένα σε πολύ σύντομο χρόνο. Δεύτερον, μπορούν να επιστρέψουν σε κάθε ερώτημα τα πρώτα αποτελέσματα που είναι πιο χρήσιμα. Έτσι τελικά ο χρήστης λαμβάνει γρήγορα και εύκολα μόνο της ουσιώδη πληροφορία που θέλει.



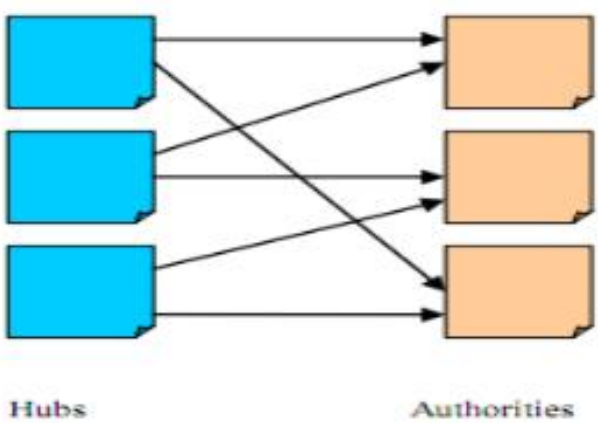
Εικόνα 8 Ταξινόμηση του τομέα εξόρυξης δεδομένων διαδικτύου

1.5.4.1 Web Content Mining

Το web content mining, εξετάζει τα περιεχόμενα των ιστοσελίδων καθώς και τα αποτελέσματα αναζητήσεων. Το περιεχόμενο αυτό μπορεί να περιέχει τόσο κείμενο όσο και γραφικά. Οι προκλήσεις αυτού του τομέα της εξόρυξης δεδομένων είναι πολλές μιας και το μέγεθος των ιστοσελίδων είναι απροσδιόριστα μεγάλο και η δομή τους δεν είναι ομοιόμορφη. Επίσης υπάρχει πληθώρα κειμένων σε πολλαπλές εκδόσεις καθώς και λανθασμένη και ατελής πληροφορία. Αυτό κάνει ακόμα πιο επιτακτική την ανάγκη για χρήση τεχνικών ώστε τα αποτελέσματα από αναζητήσεις να είναι ορθό και ακριβές. (Κάπρος, 2006) Εκτός από αυτό, υπάρχει ένα τμήμα του διαδικτύου γνωστό και ως «βαθύς ιστός (deep web)» το οποίο δεν μπορεί εύκολα να ευρετηριοποιηθεί από μηχανές αναζήτησης. Ο «βαθύς ιστός» περιέχει βάσεις δεδομένων, βιβλιοθήκες, γενετικά δεδομένα και ευρετήρια. Μεγάλο μέρος του «βαθύ ιστού» είναι δομημένο ή ημι-δομημένο και έτσι είναι ευκολότερο να αναλυθεί και να ενοποιηθεί, το δύσκολο είναι να βρεθούν τεχνικές να ευρετηριοποιηθεί. Το web content mining χωρίζεται επιμέρους στο web page content mining και στο search result mining. Το πρώτο είναι η παραδοσιακή αναζήτηση ιστοσελίδων σύμφωνα με το περιεχόμενο τους, ενώ το δεύτερο είναι περαιτέρω αναζήτηση σε ιστοσελίδες που είναι αποτέλεσμα προηγούμενης αναζήτησης. (Πλέγας, 2008)

1.5.4.2 Web Structure Mining

Το web structure mining, είναι ο ερευνητικός τομέας που εστιάζει στη χρήση της ανάλυσης της δομής των συνδέσμων του διαδικτύου, και ένας βασικός σκοπός του είναι η ανακάλυψη των πιο προτιμητέων κειμένων. Ο παγκόσμιος ιστός θεωρείται σαν ένας κατευθυνόμενος γράφος όπου οι ιστοσελίδες είναι οι κόμβοι του και οι σύνδεσμοι είναι οι πλευρές που τους ενώνουν. (Φαλιάγκα, 2012) Η βασική ιδέα εδώ είναι πως ένας υπερσύνδεσμος από ένα κείμενο Α σε ένα κείμενο Β υποδηλώνει πως ο συγγραφέας του κειμένου Α θεωρεί το περιεχόμενο του κειμένου Β αξιοσημείωτο. Οι Αλγόριθμοι και Τεχνικές Εξόρυξης Δεδομένων από Ροές Δεδομένων στον Παγκόσμιο Ιστό υπερσύνδεσμοι χρησιμοποιούνται ευρέως στις μηχανές αναζήτησης για να αναγνωρίσουν σχέσεις συσχέτισης μεταξύ κειμένων, να ομαδοποιήσουν κείμενα ανάλογα τη σημαντικότητα τους και τελευταία για να βρουν κοινότητες στον παγκόσμιο ιστό από τις παραπομπές ή την μη ύπαρξη παραπομπών. (Ζώτος, 2007)

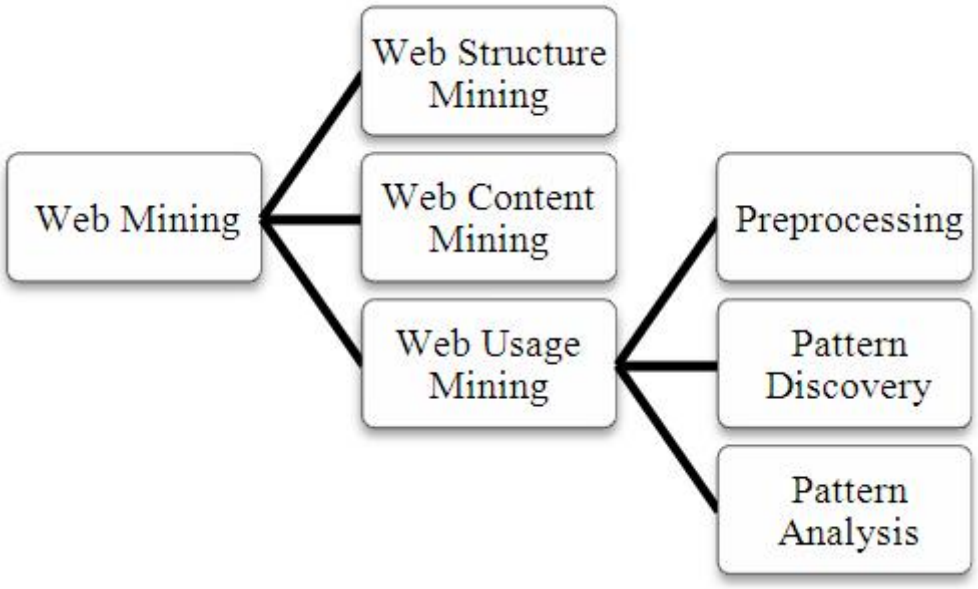


http://www.upet.ro/annals/economics/pdf/2011/part4/Dinuca.pdf

1.5.4.3 Web Usage Mining

Στο web usage mining γνωστό και ως web log mining, γίνεται επεξεργασία των log αρχείων σχετικά με τις προσβάσεις χρηστών στις διάφορες ιστοσελίδες. Με τη βοήθεια τεχνικών αυτού του τομέα γίνεται κατανοητή η συμπεριφορά ενός χρήστη αλλά και η δομή της πληροφορίας. Τα δεδομένα των click-stream, τα cookies, τα ερωτήματα των χρηστών, και κάθε είδους δεδομένα σχετικά με τα αποτελέσματα της αλληλεπίδρασης μεταξύ ανθρώπου και διαδικτύου χρησιμοποιούνται επίσης για να

τονιστούν οι ανάγκες των πελατών και να βελτιωθεί η ποιότητα των υπηρεσιών τους. (Μεττούρης, 2008)Το general access pattern tracking είναι ένας τύπος του web usage mining ο οποίος εξετάζει στο ιστορικό επισκέψεων των ιστοσελίδων.



Εικόνα 9

Αυτή η χρήση (usage) μπορεί να είναι γενική ή μπορεί να στοχεύει σε συγκεκριμένη χρήση ή χρήστες. Επίσης αναγνωρίζοντας τα πρότυπα της κίνησης, το usage mining γίνεται εξόρυξη αυτών ακολουθιακών προτύπων (sequential patterns). Για παράδειγμα τα πρότυπα μπορούν να συσταδοποιηθούν βάση των ομοιοτήτων τους. Αυτό στη συνέχεια μπορεί να χρησιμοποιηθεί ώστε να γίνει συσταδοποίηση των χρηστών σε ομάδες βασιζόμενοι σε ομοιότητες των προσβάσεων τους σε ιστοσελίδες. Τέλος ένας άλλος τύπος του web usage mining είναι το customized usage tracking το οποίο αναλύει μεμονωμένες τάσεις έτσι ώστε οι ιστοσελίδες να προσδίδονται σε συγκεκριμένου χρήστες. Βασιζόμενοι σε πρότυπα προσβάσεων, μια ιστοσελίδα μπορεί δυναμικά να τροποποιηθεί για ένα χρήστη όσον αφορά την πληροφορία που παρουσιάζει, το βάθος της δομής του και τη μορφή των πηγών που παρουσιάζονται. (Ταράτσα, 2011) Πολύ σημαντικό είναι και το πρόβλημα της εξαγωγής κοινοτήτων διαδικτύου πραγματικού χρόνου (online web communities). Μια κοινότητα στον παγκόσμιο ιστό είναι μια ομάδα σελίδων που έχουν κάποιο κοινό αντικείμενο (π.χ. σελίδες που γράφουν για αθλητικά). Τελικός σκοπός της παρούσας εργασίας είναι η παρουσίαση ενός μοντέλου για την ομαδοποίηση χρηστών και κοινοτήτων βάση χαρακτηριστικών ομοιότητας με τεχνικές συσταδοποίησης. Εξαιτίας της ποικιλίας των θεμάτων που υπάρχουν στον παγκόσμιο ιστό, το πρόβλημα της εξαγωγής

κοινοτήτων έχει γίνει πολύ σημαντικό και δύσκολο. Έτσι με τη δημιουργία προτύπων χρηστών μπορούμε να βρούμε μεγάλα υποσύνολα σελίδων από συγκεκριμένες κοινότητες. Ιδιαίτερο ενδιαφέρον στον τομέα του web mining παρουσιάζει ο τομέας του web usage mining ή αλλιώς web log mining και του web usage mining. Τα τελευταία χρόνια έχουν υλοποιηθεί πολλές εφαρμογές και αλγόριθμοι για την εξαγωγή συμπερασμάτων από δεδομένα διαδικτύου. Ο παγκόσμιος ιστός είναι μια απέραντη πηγή δεδομένων που προέρχονται είτε από το περιεχόμενο διαδικτύου (web content), δηλαδή τα δισεκατομμύρια σελίδων που είναι διαθέσιμες, είτε από τη χρήση διαδικτύου (web usage), δηλαδή από τα log αρχεία δεδομένων που συλλέγονται καθημερινά από τους διακομιστές. Ο τομέας του web mining είναι η περιοχή της εξόρυξης δεδομένων η οποία έχει να κάνει με την εξαγωγή ενδιαφέρουσας γνώσης από τον παγκόσμιο ιστό. Η έρευνα στην συγκεκριμένη περιοχή παρουσιάζει μεγάλη άνθιση με πολλές δημοσιευμένες εργασίες σε παγκόσμια συνέδρια. (Παπανικολάου, 2011)

2. ΑΛΓΟΡΙΘΜΟΙ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ

Οι αλγόριθμοι εξόρυξης δεδομένων είναι πολλοί και σε αυτή την ενότητα θα παρουσιαστούν σε κατηγορίες οι πιο σημαντικοί από αυτούς. Οι κατηγορίες στις οποίες θα τους συναντήσουμε είναι οι παρακάτω:

- Κατηγοριοποίηση
- Συσταδοποίηση
- Κανόνες Συσχέτισης
- Πρότυπα Ακολουθιών
- Παλινδρόμηση
- Δέντρα Απόφασης

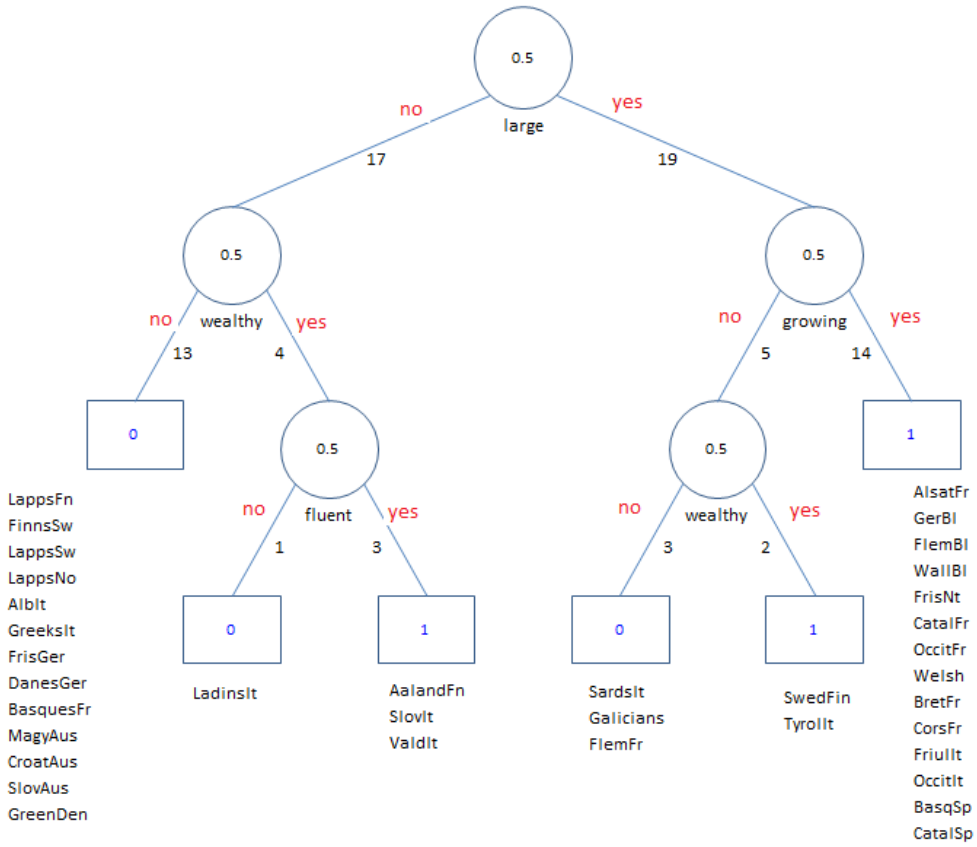
Οι παραπάνω κατηγορίες χωρίς αμφιβολία αναπαριστούν όλη την περιοχή των αλγορίθμων που χρησιμοποιούνται στον τομέα αυτό. Τα τελευταία χρόνια η ερευνητική κοινότητα δίνει πολύ βάση στη βελτίωση υπαρχόντων τεχνικών και δημιουργία νέων για να αντιμετωπιστούν τα προβλήματα που τείθονται σε αυτές τις κατηγορίες που θα αναλύσουμε παρακάτω.

2.1 Κατηγοριοποίηση

Η κατηγοριοποίηση (classification) αποτελεί μια από τις βασικές εργασίες (tasks) εξόρυξης δεδομένων. Βασίζεται στην εξέταση των χαρακτηριστικών ενός νέου αντικειμένου το οποίο με βάση τα χαρακτηριστικά αυτά αντιστοιχίζεται σε ένα προκαθορισμένο σύνολο κλάσεων. Τα αντικείμενα που πρόκειται να κατηγοριοποιηθούν αναπαριστάνονται γενικά από τις εγγραφές της βάσης δεδομένων και η διαδικασία της κατηγοριοποίησης αποτελείται από την ανάθεση κάθε εγγραφής σε κάποιες από τις προκαθορισμένες κατηγορίες.

Η εργασία της κατηγοριοποίησης χαρακτηρίζεται από έναν καλά καθορισμένο ορισμό των κατηγοριών και το σύνολο που χρησιμοποιείται για την εκπαίδευση του μοντέλου αποτελείται από προκατηγοριοποιημένα παραδείγματα. Η βασική εργασία είναι να δημιουργηθεί ένα μοντέλο το οποίο θα μπορούσε να εφαρμοστεί για να κατηγοριοποιήσει δεδομένα που δεν έχουν ακόμα κατηγοριοποιηθεί (να ανατεθεί σε κάποια από τις κατηγορίες).

Στις περισσότερες περιπτώσεις, υπάρχει ένα περιορισμένος αριθμός κατηγοριών και εμείς θα πρέπει να αναθέσουμε κάθε εγγραφή στην κατάλληλη κατηγορία. Για αυτό το σκοπό χρησιμοποιούνται κάποιες τεχνικές, τις οποίες μπορούμε να κατατάξουμε σε δύο κατηγορίες. Η πρώτη χρησιμοποιεί δέντρα απόφασης (decision trees) και η δεύτερη νευρωνικά δίκτυα (neural networks).



Εικόνα 10 Διαγραμματική απεικόνιση Αλγόριθμων κατηγοριοποίησης

2.2 Συσταδοποιήση

Συσταδοποίηση[2] (Clustering) είναι η διαδικασία διαίρεσης ενός συνόλου δεδομένων σε αμοιβαία αποκλειόμενες ομάδες, τέτοιες ώστε τα μέλη κάθε ομάδας να είναι όσο

[2] Η συσταδοποίηση μπορεί να βρεθεί με διαφορετικά ονόματα σε διαφορετικά πεδία, όπως *μη εποπτευόμενη μάθηση* (*unsupervised learning*) στην αναγνώριση προτύπων, *αριθμητική ταξονομία* (numerical taxonomy) στην βιολογία, οικολογία, *τυπολογία*, στις κοινωνικές επιστήμες και *τμηματοποίηση* (segmentation, partitioning) στη θεωρία των γράφων και στις Βάσεις Δεδομένων

κοντά γίνεται το ένα με το άλλο, ενώ οι διαφορετικές ομάδες να είναι όσο το δυνατόν πιο μακριά η μια από την άλλη, όπου η απόσταση μετράται σε σχέση με τις διαθέσιμες μεταβλητές. Σίγουρα, με την αναπαράσταση των δεδομένων με λιγότερες συστάδες χάνονται ορισμένες μικρολεπτομέρειες, αλλά επιτελείται απλοποίηση (μια συστάδα είναι μια ταξινομημένη λίστα αντικειμένων, τα αντικείμενα της οποίας έχουν κάποια κοινά χαρακτηριστικά).

Η συσταδοποίηση διαδραματίζει σπουδαίο ρόλο σε εφαρμογές εξόρυξης δεδομένων, όπως είναι η διερεύνηση επιστημονικών δεδομένων, η ανάκτηση πληροφορίας και η εξόρυξη κειμένου, εφαρμογές χωρικών βάσεων δεδομένων, web ανάλυση, CRM, μάρκετινγκ, ιατρική διάγνωση, υπολογιστική βιολογία και πολλά άλλα.

Η θεμελιώδης διάκριση μεταξύ της συσταδοποίησης και της κατηγοριοποίησης είναι ότι στη δεύτερη, ένα σύνολο από προ-ομαδοποιημένα αντικείμενα είναι διαθέσιμο, και αυτό που απαιτείται είναι να τοποθετήσουμε και να εφαρμόσουμε ένα νέο αντικείμενο σε κάποια από τις τρέχουσες ομάδες. Απεναντίας, στη συσταδοποίηση δε διαθέτουμε καμία πρότερη γνώση για τις ομάδες/συστάδες στις οποίες είναι απαραίτητο να διαχωρισθούν τα δεδομένα. Συνεπώς, η συσταδοποίηση παράγεται μόνο από τα δεδομένα και είναι απόλυτα οδηγούμενη από αυτά.

Η συσταδοποίηση είναι πολύ ωφέλιμη σε ένα σύνολο γνήσιων εφαρμογών, όπως η ανάλυση προτύπων (pattern analysis), η λήψη αποφάσεων (decision making), η ανάκτηση πληροφορίας (information retrieval) κ.α. Στην πλειοψηφία των περιπτώσεων που υλοποιείται η συσταδοποίηση, ανακύπτει μικρή ή καθόλου γνώση για την δομή και το είδος των δεδομένων. Σε ανάλογες περιπτώσεις, η συσταδοποίηση των δεδομένων είναι αρμόδια για την ανεύρεση αλληλοσχετισμών μεταξύ των δεδομένων με σκοπό να κατανοηθεί η δομή τους, το οποίο είναι και ο απώτερος στόχος.

Η συσταδοποίηση αποτελείται από 4 διαφορετικές φάσεις:

Στη φάση της αναπαράστασης, επιλέγεται η βέλτιστη δομή αναπαράστασης των δεδομένων, για να επικυρωθούν τα χαρακτηριστικά εκείνα που είναι αναγκαία και χρήσιμα για τη συσταδοποίηση. Ακολούθως, κατά τη φάση της μοντελοποίησης, επιλέγεται το μέτρο ομοιότητας μεταξύ των αναπαράστασεων των αντικειμένων, καθώς και οι αρχικές συστάδες. Η συνάρτηση απόστασης καθορίζει το μέτρο ομοιότητας μεταξύ των αντικειμένων. Στην επόμενη φάση της συσταδοποίησης, οι τελικές συστάδες αναπτύσσεται με την υλοποίηση του καθεαυτού αλγορίθμου συσταδοποίησης. Τέλος, οι παραχθέντες συστάδες βελτιστοποιούνται με βάση ορισμένες μετρικές αξιολόγησης της ποιότητάς τους: ορισμένες συστάδες είναι πιθανόν να συγχωνευθούν μεταξύ τους ή να διαιρεθούν σε αντίστοιχες συστάδες.

Στη βιβλιογραφία υπάρχει ποικιλία αλγορίθμων συσταδοποίησης, με τον καθένα να εξειδικεύεται κατά κανόνα σε ένα συγκεκριμένο είδος δεδομένων. Στον ακόλουθο πίνακα παρουσιάζονται δυο από τις πιο σημαντικές κατηγορίες αλγορίθμων συσταδοποίησης με τα χαρακτηριστικά τους.

Πίνακας 4 Κατηγορίες αλγορίθμων συσταδοποίησης

Διαχωριστικοί (partitional)	Οι επαναληπτικοί αλγόριθμοι συσταδοποίησης βασίζονται σε μια αρχική εκτίμηση των συστάδων (είτε τυχαία ή βασισμένη σε κάποιες παραδοχές ή γνώση των δεδομένων) και στη συνέχεια παράγουν ένα διαμερισμό των αντικειμένων σε συστάδες.
Ιεραρχικοί (hierarchical)	Οι αλγόριθμοι της κατηγορίας αυτής δημιουργούν ένα ιεραρχικό δένδρο από συστάδες, επιτρέποντας σε μία συστάδα να έχει ένα σύνολο από υποσυστάδες. Ανάλογα με τον τρόπο κατασκευής του δένδρου ιεραρχίας, οι αλγόριθμοι διακρίνονται σε bottom-up (ή agglomerative) και top-down. Οι bottom-up αλγόριθμοι θεωρούν κάθε αντικείμενο ως μια ξεχωριστή συστάδα και σε κάθε βήμα συγχωνεύουν τις δύο πιο κοντινές συστάδες, μέχρις ότου απομείνει μόνο μία συστάδα η οποία περιλαμβάνει όλα τα αντικείμενα. Αντίθετα, οι top-down αλγόριθμοι ξεκινούν από μια μοναδική συστάδα, η οποία περιέχει όλα τα αντικείμενα, και σε κάθε βήμα χωρίζουν τις υπάρχουσες συστάδες σε επιμέρους υποσυστάδες.

Μία αποτελεσματική συσταδοποίηση έχει ως πρωταρχικούς στόχους:

- Την ελαχιστοποίηση του σφάλματος που παρουσιάζεται και,
- Τον περιορισμό, όσο αυτό είναι δυνατόν, του συνολικού αριθμού των διαφορετικών ομάδων που θα εξαχθούν ,κατά την τελειοποίηση της διαδικασίας.

Όπως αποδεικνύεται, τα δύο παραπάνω κριτήρια αποτελεσματικότητας είναι αντικρουόμενα. Αυτό συμβαίνει διότι η δημιουργία όλο και λιγότερων ομάδων αυξάνει την εισαγωγή σφάλματος και σημαντικό αντικείμενο έρευνας οφείλει να αποτελέσει ο σωστός συνδυασμός αυτών των δύο.

Η επιτυχία ή η αποτυχία στη διαδικασία συσταδοποίησης, δεν μπορεί να αποφανθεί μετά την ολοκλήρωση της. Και τούτο, διότι δεν μπορεί να οριστεί ένα απόλυτα «καλό κριτήριο», το οποίο θα ήταν αυτάρκες από τον τελικό στόχο του Clustering. Ένα διαφορετικό κριτήριο, μία διαφορετική μέθοδος υλοποίησης ή ακόμα και μία διαφορετική επιλογή παραμέτρων, ορισμένων αναλόγως με το επιθυμητό αποτέλεσμα, είναι δυνατόν να κατευθυνθεί σε ανόμοια αποτελέσματα συσταδοποίησης. Παραδείγματος χάρη, αντικείμενο ενδιαφέροντος είναι δυνατόν να οριστεί ο εντοπισμός αντιπροσωπευτικών ομάδων, «φυσικών» συστάδων, που να περιγράφουν άγνωστα για το αντικείμενο χαρακτηριστικά, ή ο εντοπισμός ασυνήθιστων και περίεργων αντικειμένων που ανήκουν στα δεδομένα.

2.3 Κανόνες Συσχέτισης

Οι κανόνες συσχέτισης αποτελούν μια σύγχρονη μέθοδο για την εξαγωγή γνώσης από μεγάλες βάσεις δεδομένων. Η εμφάνισή τους οφείλεται στις ανάγκες ανάλυσης του «καλαθιού αγοράς» (market basket analysis). Οι κανόνες συσχέτισης εφαρμόστηκαν για να αποκαλύπτουν ενδιαφέρουσες σχέσεις μεταξύ των δεδομένων. Οι σχέσεις αυτές παρουσιάζονται στη μορφή A->B, όπου το A και το B αναφέρονται στα σύνολα γνωρισμάτων που υπάρχουν στα δεδομένα υπό ανάλυση.

Για την αξιολόγηση της σποδαιότητας του κάθε κανόνα συσχέτισης, έχουν προταθεί ορισμένα μέτρα, όπως είναι η υποστήριξη, η εμπιστοσύνη, η κάλυψη. Στον ακόλουθο πίνακα διακρίνονται αναλυτικά τα χαρακτηριστικά αυτών των μέτρων :

Υποστήριξη (Support)	Η υποστήριξη ενός κανόνα συσχέτισης είναι το ποσοστό όλων των περιπτώσεων στο σύνολο δεδομένων που ικανοποιούν έναν κανόνα, δηλαδή ικανοποιούν το αριστερό και το δεξιό μέλος του κανόνα.
Εμπιστοσύνη (Confidence)	Η εμπιστοσύνη ενός κανόνα συσχέτισης είναι το πηλίκο της υποστήριξης του κανόνα προς το ποσοστό των περιπτώσεων που καλύπτονται από το αριστερό μέλος του κανόνα.
Κάλυψη (Coverage)	Η κάλυψη ενός κανόνα συσχέτισης είναι το ποσοστό των περιπτώσεων των δεδομένων που έχουν τις τιμές των γνωρισμάτων που ορίζονται στο αριστερό μέλος του κανόνα. Ένας κανόνας συσχέτισης με τιμή κάλυψης κοντά στο 1, μπορεί να θεωρηθεί ως κανόνας με ενδιαφέρον.
Lift.	Το lift ορίζεται ως η εμπιστοσύνη του κανόνα διαιρούμενη με το ποσοστό όλων των περιπτώσεων που καλύπτονται από το δεξιό μέλος του κανόνα. Είναι ένα μέτρο της σπουδαιότητας της συσχέτισης και είναι ανεξάρτητο από την κάλυψη.

2.4 Πρότυπα Ακολουθιών

Η εξόρυξη πρότυπων ακολουθιών (sequential patterns) είναι η εξόρυξη των συχνά εμφανιζόμενων προτύπων σχετικών με το χρόνο ή άλλες ακολουθίες. Οι περισσότερες μελέτες στα πρότυπα ακολουθιών επικεντρώνονται στα συμβολικά πρότυπα. Ο χρήστης εδώ μπορεί να προσδιορίσει τους περιορισμούς στα είδη των προτύπων ακολουθιών που εξάγονται με την παροχή των προσχεδίων προτύπων (template patterns) υπό μορφή σειριακών επεισοδίων, παράλληλων επεισοδίων ή κανονικών εκφράσεων. Παραδείγματα προτύπων ακολουθιών έχουμε στην καθημερινή μας ζωή όπως τα κείμενα, οι μουσικές νότες, τα δεδομένα του καιρού και οι ακολουθίες του DNA.

2.5 Παλινδρόμηση

Η παλινδρόμηση (regression) είναι θέμα το οποίο έχει μελετηθεί πολύ στην στατιστική και στα νευρωνικά δίκτυα. Κύριος σκοπός εδώ είναι η πρόβλεψη της τιμής μιας μεταβλητής μελετώντας τις τιμές που είχε στο παρελθόν. Συνήθως χρησιμοποιούμε ένα μοντέλο για την μεταβλητή. Η παλινδρόμηση καλύπτει ένα μεγάλο τμήμα του τομέα της εξόρυξης δεδομένων που έχει να κάνει με προβλέψεις.

2.6 Δέντρα Απόφασης

Τα δέντρα απόφασης (decision trees) έχουν μελετηθεί αρκετά σαν ένα ζήτημα μηχανικής μάθησης. Για να γίνει κατανοητό, ας υποθέσουμε ότι έχουμε ένα σύνολο εγγραφών και καθεμία από αυτές έχει μια λίστα χαρακτηριστικών. Ένα δέντρο απόφασης στο σύνολο των εγγραφών είναι ένα δέντρο όπου σε κάθε κόμβο του (που δεν είναι φύλλο) υπάρχει ένα ερώτημα που αναφέρεται στα χαρακτηριστικά των εγγραφών και κάθε ερώτημα καταλήγει σε ένα συγκεκριμένο παιδί ενός κόμβου. Τα φύλλα του δηλώνουν τις κλάσεις. Έτσι ένα δέντρο απόφασης εκτελεί κατηγοριοποίηση χρησιμοποιώντας ερωτήματα σχετικά με τα χαρακτηριστικά των εγγραφών. Οι εφαρμογές που χρησιμοποιούν δέντρα απόφασης είναι παρόμοιες με αυτές που κάνουν κατηγοριοποίηση.

2.7 Συνεργατική Διήθηση Δεδομένων

Με τον όρο συνεργατική διήθηση δεδομένων, περιγράφεται η διαδικασία της απόρριψης ή αποδοχής κάποιων δεδομένων σε σχέση με κάποια άλλα. Η διαδικασία πραγματοποιείται μέσω υπολογιστή και με χρήση τεχνικών που απαιτούν την συνεργασία παραγόντων όπως είναι τα αποθηκευτικά μέσα, οι απόψεις των χρηστών, οι πηγές πληροφόρησης κλπ.

Ο όρος αναφέρεται κυρίως στα δεδομένα χρηστών του διαδικτύου, καθώς σ' αυτό υπάρχουν υπερβολικά μεγάλες συγκεντρώσεις πηγών πληροφορίας.

Για να γίνει περισσότερο κατανοητή η σημασία του όρου, πρέπει να σημειωθεί ότι στην καθημερινή ζωή, οι άνθρωποι προκειμένου να κάνουν μια επιλογή, βασίζονται

σε συστάσεις ή προτροπές άλλων ανθρώπων μέσω του προφορικού λόγου, συστατικών επιστολών, ειδησεογραφικών αναφορών από τα μέσα ενημέρωσης, γενικών ερευνών, τουριστικών οδηγών και ούτω καθεξής. Έτσι, έχουν αναπτυχθεί παρόμοια ηλεκτρονικά συστήματα συστάσεων, τα οποία προτείνουν την καταλληλότερη για το χρήστη πληροφορία. Τα ηλεκτρονικά συστήματα συστάσεων, ενισχύουν και αυξάνουν αυτή τη φυσική κοινωνική διαδικασία βοηθώντας τους ανθρώπους να διακρίνουν ανάμεσα στα διαθέσιμα βιβλία, άρθρα, ιστοσελίδες, ταινίες, μουσική, εστιατόρια ή λίστα ανεκδότων ώστε να επιλέξουν τις πιο ενδιαφέρουσες και αξιόλογες για τους ίδιους πληροφορίες.

Ουσιαστικά, λοιπόν, πρόκειται για την μέθοδο παραγωγής μιας αυτόματης πρόβλεψης (φιλτράρισμα, διήθηση) προς όφελος του χρήστη με συλλογή πληροφοριών για τις προτιμήσεις άλλων χρηστών (συνεργασία).

Να σημειωθεί ότι παρ' όλο που αυτές οι προβλέψεις απευθύνονται στον συγκεκριμένο χρήστη χρησιμοποιούν πληροφορίες προερχόμενες από πολλούς άλλους. Αυτό διαφέρει από την πιο απλή προσέγγιση της παροχής μιας μέσης βαθμολογίας για το αντικείμενο ενδιαφέροντος, όπως για παράδειγμα με βάση τον αριθμό των θετικών ψήφων που παρέχονται σε ένα ερωτηματολόγιο.

Ο όρος "collaborating filtering" επινοήθηκε από τους προγραμματιστές ενός από πρώτα συστήματα συστάσεων, του Tapestry και έκτοτε έχει υιοθετηθεί ευρέως, ανεξαρτήτως του ότι οι "συστήνοντες" δεν "συνεργάζονται" στην πραγματικότητα ούτε μεταξύ τους ούτε με τους αποδέκτες. Από την άλλη, ενώ τα αποτελέσματα μπορεί να υποδεικνύουν εξαιρετικά ενδιαφέρουσες επιλογές για το χρήστη, μπορεί όμως να περιέχουν και προτάσεις που θα έπρεπε τελικά να φιλτράρονται και να μην εμφανίζονται.

2.8 Εφαρμογές Εξόρυξης Δεδομένων

Η διαδικασία της ανακάλυψης Γνώσης σε Βάσεις Δεδομένων και η εξόρυξη δεδομένων διαθέτει ποικιλία εφαρμογών, όπως προαναφέρθηκε. Στον παρακάτω πίνακα εστιάζονται ορισμένα ενδεικτικά παραδείγματα.

Πίνακας 5 Εφαρμογές εξόρυξης δεδομένων

Έρευνα Αγοράς	Σε εταιρείες όπου υπάρχει αυξημένος όγκος δεδομένων λόγω του μεγάλου αριθμού πελατών και οικονομικών στοιχείων, γίνεται χρήση συστημάτων διαχείρισης δεδομένων για τη βελτιστοποίηση της ανάλυσης και της χρήσης των δεδομένων αυτών. Είναι γεγονός πως η επιστήμη του marketing προσανατολίζεται στην έρευνα των αναγκών των πελατών με σκοπό τη μεγιστοποίησης του κέρδους. Αυτό επιτυγχάνεται με την εξόρυξη δεδομένων καθώς αναλύονται τα δεδομένα μιας επιχείρησης και με τον τρόπο αυτό εξάγονται χρήσιμα συμπεράσματα για τη συμπεριφορά των πελατών. Στο marketing επίσης χρησιμοποιείται η εξόρυξη δεδομένων με τη μορφή της ανάλυσης των πελατών ηλεκτρονικών καταστημάτων κάνοντας χρήση Log αρχείων, αλλά και της πρόβλεψης των μελλοντικών αγορών ενός πελάτη, με βάση παρελθοντικές του κινήσεις.
Χρηματοοικονομικά	Έρευνες στον τομέα της εξόρυξης δεδομένων έχουν πραγματοποιηθεί και για την κάλυψη αναγκών των χρηματιστηριακών εφαρμογών. Με βάση τις τεχνικές που αναπτύσσονται από τις έρευνες αυτές, οι χρηματιστηριακές εταιρείες επιλέγουν τις επενδύσεις τους. Βέβαια σε αυτή την περίπτωση, η εξόρυξη δεδομένων γίνεται από κείμενα και τεχνικές αναφορές επιχειρήσεων με στόχο την επίτευξη μίας πρόβλεψης της τάσης των μετοχών.
Ασφάλεια Συστημάτων	Μία από τις πιο σημαντικές και επιτυχημένες εφαρμογές της εξόρυξης δεδομένων αποτελεί η πρόληψη και η αποφυγή διαφόρων τύπων απάτης. Τέτοιες μπορεί να είναι οι απάτες

	στο διαδίκτυο με περίεργες συναλλαγές, ή οι οικονομικές απάτες των οποίων η πρόληψη επιτυγχάνεται με την χρήση συστημάτων αναγνώρισης ανωμαλιών που ανιχνεύονται στις συναλλαγές, όπως είναι το FAIS.
Παγκόσμιος Ιστός	Η Google είναι το μεγαλύτερο και πιο εμφανές παράδειγμα εξόρυξης δεδομένων στο διαδίκτυο. Ενώ ο όγκος των δεδομένων είναι τεράστιος, η κάθε αναζήτησης παράγει αποτελέσματα η παρουσίαση των οποίων δεν ξεπερνάει χρονικά τα δύο δευτερόλεπτα. Μέσα από τη μηχανή αναζήτησης της Google γίνεται εύκολα αντιληπτή η επιτυχία της εφαρμογής της διαδικασίας της εξόρυξης δεδομένων στο διαδίκτυο, καθώς ο τελικός χρήστης λαμβάνει εύκολα και γρήγορα μόνο την προς αναζήτηση πληροφορία.

3. ΣΥΝΕΡΓΑΤΙΚΗ ΔΙΗΘΗΣΗ ΔΕΔΟΜΕΝΩΝ

3.1 Φιλτράρισμα Πληροφοριών

Τα συστήματα ανάκτησης πληροφοριών επιτρέπουν στους χρήστες να εκφράσουν ερωτήσεις για να επιλέξουν στοιχεία που ταιριάζουν με ένα συγκεκριμένο θέμα και παράλληλα επιτρέπουν στις εταιρείες να συλλέξουν πληροφορίες σχετικά με το προφίλ του κάθε χρήστη. Οι τεχνικές ανάκτησης πληροφοριών ωστόσο δεν είναι αρκετά χρήσιμες στην πραγματική διαδικασία σύστασης, δεδομένου ότι δεν συλλαμβάνουν καμία πληροφορία για τις προτιμήσεις των χρηστών πέραν της συγκεκριμένης ερώτησης. Για την αποτελεσματικότερη προσέγγιση των πελατών εφαρμόζονται μέθοδοι φιλτραρίσματος των πληροφοριών, τα χαρακτηριστικά των οποίων παρουσιάζονται στον επόμενο πίνακα.

Πίνακας 6 Κατηγορίες φιλτραρίσματος πληροφοριών

Δημογραφικό Φιλτράρισμα	Οι δημογραφικές προσεγγίσεις φιλτραρίσματος χρησιμοποιούν τις περιγραφές των χρηστών για να μάθουν τη σχέση μεταξύ ενός στοιχείου και του τύπου ανθρώπων που τους αρέσει. Τα προφίλ χρηστών δημιουργούνται με την ταξινόμηση των χρηστών σε στερεοτυπικές περιγραφές, που αντιπροσωπεύουν τα χαρακτηριστικά γνωρίσματα των κατηγοριών των χρηστών. Τα προσωπικά στοιχεία του χρήστη είναι απαραίτητα και χρησιμοποιούνται για την ταξινόμηση. Οι ταξινομήσεις χρησιμοποιούνται ως γενικοί χαρακτηρισμοί για τους χρήστες και τα ενδιαφέροντά τους. Συνήθως, τα προσωπικά στοιχεία του χρήστη λαμβάνονται κατά την αίτηση εγγραφής στο σύστημα (φύλο, επάγγελμα, ηλικία, τρόπος ζωής κ.α.). (Χ. Χριστάκου, 2012)
Φιλτράρισμα Περιεχομένου	Το φιλτράρισμα περιεχομένου επιλέγει τις σωστές πληροφορίες για τους χρήστες με τη σύγκριση της αναπαράστασης της πληροφορίας αναζήτησης με την αναπαράσταση του περιεχομένου των παραμέτρων χρήστη που εκφράζει τα ενδιαφέροντά του (profile). Το

	φιλτράρισμα περιεχομένου πληροφοριών έχει αποδειχθεί αποτελεσματικό στην εντόπιση κειμενικών στοιχείων σχετικών με ένα θέμα χρησιμοποιώντας τεχνικές, όπως Boolean ερωτήσεις, ερωτήσεις διανυσματικού χώρου, το πιθανολογικό πρότυπο, τα νευρωνικά δίκτυα και το μοντέλο ασαφών συνόλων.
Φιλτράρισμα με βάση την Γνώση	Τα συστήματα αυτού του είδους στηρίζονται για τις συστάσεις που κάνουν σε συγκεκριμένη γνώση η οποία καθορίζει κατά πόσο τα χαρακτηριστικά ενός προϊόντος ανταποκρίνονται στις ανάγκες και τα ενδιαφέροντα του χρήστη, δηλαδή αν το προϊόν θα είναι χρήσιμο στον χρήστη ή όχι. Το σύστημα συγκεντρώνει τα αιτήματα του χρήστη και προτείνει και εξηγεί τις συστάσεις που βρίσκει ως λύση. Η συνάρτηση ομοιότητας στα Συστήματα Προτάσεων με βάση την γνώση εκτιμά πόσο οι ανάγκες του χρήστη συσχετίζονται με τις συστάσεις και έτσι τελικά δείχνει την χρησιμότητα της σύστασης για τον ενδιαφερόμενο.
Συνεργατικό Φιλτράρισμα (Διήθηση) Δεδομένων	Στην συνεργατική προσέγγιση αντί να συστηθούν στοιχεία επειδή είναι παρόμοια με τα στοιχεία που ένας χρήστης επιδοκίμασε στο παρελθόν, συστήνονται στοιχεία που άλλοι χρήστες με γειτονικό προφιλ έχουν συμπαθήσει. Αντί δηλαδή να υπολογιστεί η ομοιότητα των προϊόντων, υπολογίζεται η ομοιότητα των πελατών. Χαρακτηριστικά, για κάθε χρήστη βρίσκεται ένα σύνολο «πλησιέστερων χρηστών γειτόνων» με των οποίων τις μέχρι τώρα εκτιμήσεις υπάρχει ο ισχυρότερος συσχετισμός. Τα αποτελέσματα για τα άγνωστα στοιχεία προβλέπονται με βάση συνδυασμό αποτελεσμάτων που είναι γνωστά από τους «πλησιέστερους γείτονες».
Υβριδικό Φιλτράρισμα	Η κατηγορία αυτή συστημάτων χρησιμοποιεί ένα συνδυασμό των μεθόδων που αναφέραμε παραπάνω, εκμεταλλευόμενα τα προτερήματα τις μίας τεχνικής για να καλύψουν τα μειονεκτήματα της άλλης. Υπάρχουν πολλοί διαφορετικοί τρόποι με τους οποίους συνδυάζονται δύο ή και περισσότερες τεχνικές συστημάτων

	προτάσεων για να δημιουργηθεί ένα υβριδικό σύστημα. Στόχος του συνδυασμού διαφορετικών μεθόδων είναι η βελτίωση της απόδοσής

Σχετικά όμως με τις δύο πρώτες κατηγορίες φιλτραρίσματος εντοπίζονται κάποια σημαντικά μειονεκτήματα, τα όποια φαίνεται να υπερπηδήσει το συνδυαστικό φιλτράρισμα. Συγκεκριμένα το δημογραφικό σύστημα φιλτραρίσματος έχει δύο βασικά μειονεκτήματα:

- Είναι βασισμένο σε μια γενίκευση των ενδιαφερόντων του χρήστη, έτσι ώστε το σύστημα συστήνει τα ίδια στοιχεία στους ανθρώπους με παρόμοια δημογραφικά προφίλ. Δεδομένου ότι κάθε χρήστης είναι διαφορετικός, αυτές οι συστάσεις αποδεικνύονται πάρα πολύ γενικές.

- Οι δημογραφικές προσεγγίσεις δεν παρέχουν οποιαδήποτε μεμονωμένη προσαρμογή στις αλλαγές ενδιαφέροντος. Τα ενδιαφέροντα του χρήστη τείνουν να αλλάζουν με το πέρασμα του χρόνου, και έτσι οι παράμετροι του χρήστη πρέπει να προσαρμόζονται στην αλλαγή.

Αντίστοιχα για το φιλτράρισμα περιεχομένου ενώ ορθώς είναι βασισμένο στις αντικειμενικές πληροφορίες του προϊόντος πολλές φορές, η επιλογή κάποιου στοιχείου βασίζεται σε ένα μεγάλο ποσοστό στις υποκειμενικές ιδιότητες του στοιχείου. Για παράδειγμα στα έγγραφα κειμένων οι αντιπροσωπεύσεις συλλαμβάνουν μόνο ορισμένες πτυχές του περιεχομένου, ενώ υπάρχουν πολλές άλλες που θα επηρέαζαν την εμπειρία ενός χρήστη. Για ιστοσελίδες, παραδείγματος χάριν, οι τεχνικές φιλτραρίσματος περιχεόμενου αγνοούν εντελώς τις αισθητικές ιδιότητες, και τους παράγοντες δικτύων όπως ο χρόνος φόρτωσης. Ακόμα, είναι δύσκολο να προέρχονται οι συστάσεις από ένα ευρύ φάσμα θεμάτων, επειδή όλες οι πληροφορίες επιλέγονται και συστήνονται βασισμένες στο περιεχόμενο.

3.2 Ορισμός Συνεργατικής Διήθησης Δεδομένων

Με τον όρο collaborative filtering, ο οποίος στα Ελληνικά αποδίδεται ως «συνεργατική διήθηση δεδομένων», περιγράφεται η διαδικασία της απόρριψης ή αποδοχής κάποιων δεδομένων σε σχέση με κάποια άλλα. Η διαδικασία πραγματοποιείται μέσω υπολογιστή και με χρήση τεχνικών που απαιτούν την συνεργασία παραγόντων όπως είναι τα αποθηκευτικά μέσα, οι απόψεις των χρηστών, οι πηγές πληροφόρησης κλπ.

Ο όρος αναφέρεται κυρίως στα δεδομένα χρηστών του διαδικτύου, καθώς σ' αυτό υπάρχουν υπερβολικά μεγάλες συγκεντρώσεις πηγών πληροφορίας.

Για να γίνει περισσότερο κατανοητή η σημασία του όρου, πρέπει να σημειωθεί ότι στην καθημερινή ζωή, οι άνθρωποι προκειμένου να κάνουν μια επιλογή, βασίζονται σε συστάσεις ή προτροπές άλλων ανθρώπων μέσω του προφορικού λόγου, συστατικών επιστολών, ειδησεογραφικών αναφορών από τα μέσα ενημέρωσης, γενικών ερευνών, τουριστικών οδηγών και ούτω καθεξής. Έτσι, έχουν αναπτυχθεί παρόμοια ηλεκτρονικά συστήματα συστάσεων, τα οποία προτείνουν την καταλληλότερη για το χρήστη πληροφορία. Τα ηλεκτρονικά συστήματα συστάσεων, ενισχύουν και αυξάνουν αυτή τη φυσική κοινωνική διαδικασία βοηθώντας τους ανθρώπους να διακρίνουν ανάμεσα στα διαθέσιμα βιβλία, άρθρα, ιστοσελίδες, ταινίες, μουσική, εστιατόρια ή λίστα ανεκδότων ώστε να επιλέξουν τις πιο ενδιαφέρουσες και αξιόλογες για τους ίδιους πληροφορίες.

Ουσιαστικά, λοιπόν, πρόκειται για την μέθοδο παραγωγής μιας αυτόματης πρόβλεψης (φιλτράρισμα, διήθηση) προς όφελος του χρήστη με συλλογή πληροφοριών για τις προτιμήσεις άλλων χρηστών (συνεργασία).

Να σημειωθεί ότι παρ' όλο που αυτές οι προβλέψεις απευθύνονται στον συγκεκριμένο χρήστη χρησιμοποιούν πληροφορίες προερχόμενες από πολλούς άλλους. Αυτό διαφέρει από την πιο απλή προσέγγιση της παροχής μιας μέσης βαθμολογίας για το αντικείμενο ενδιαφέροντος, όπως για παράδειγμα με βάση τον αριθμό των θετικών ψήφων που παρέχονται σε ένα ερωτηματολόγιο.

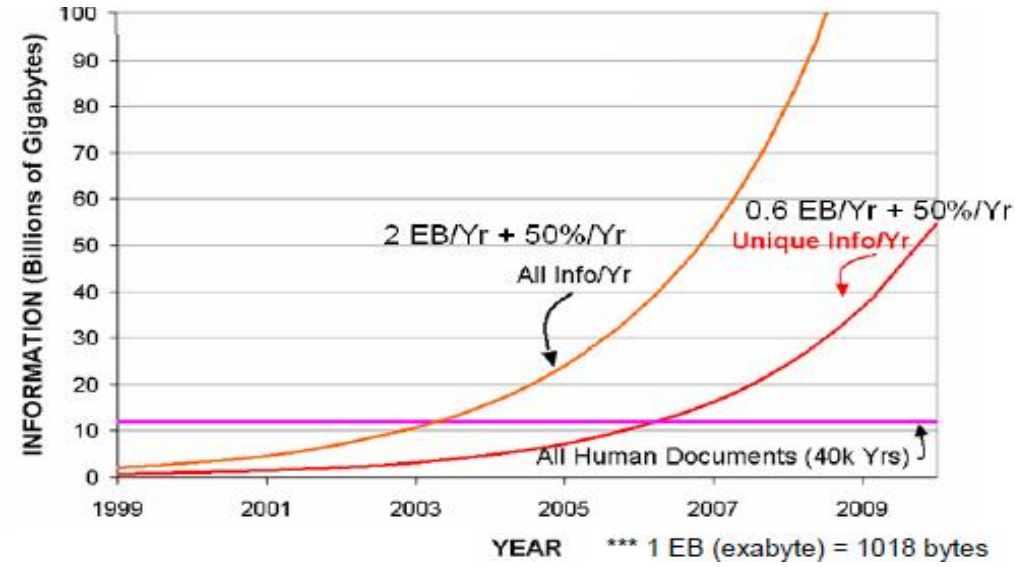
Ο όρος "collaborating filtering" επινοήθηκε από τους προγραμματιστές ενός από πρώτα συστήματα συστάσεων, του Tapestry και έκτοτε έχει υιοθετηθεί ευρέως, ανεξαρτήτως του ότι οι "συστήνοντες" δεν "συνεργάζονται" στην πραγματικότητα ούτε μεταξύ τους ούτε με τους αποδέκτες. Από την άλλη, ενώ τα αποτελέσματα μπορεί να υποδεικνύουν εξαιρετικά ενδιαφέρουσες επιλογές για το χρήστη, μπορεί

όμως να περιέχουν και προτάσεις που θα έπρεπε τελικά να φιλτράρονται και να μην εμφανίζονται.

3.3 Αναγκαιότητα Συστημάτων

Πλέον τα συστήματα καταγραφής και η διαχείρισης δεδομένων από ηλεκτρονικές εφαρμογές, καλούνται να εφαρμόσουν νέες τεχνικές καθώς τα δεδομένα αυξάνονται ραγδαία και σε πολλές περιπτώσεις καταρρίπτονται καθιερωμένες μέθοδοι επεξεργασίας τους. Παράλληλα η αυξημένη χρήση των web υπηρεσιών δημιούργησε την ανάγκη εμφάνισης προσωποποιημένων συστάσεων στους χρήστες Αποτέλεσμα αυτής της εξέλιξης είναι η αναζήτηση νέων μεθόδων διαχείρισης του μεγάλου όγκου πληροφοριών στα πλαίσια παροχής εξατομικευμένων υπηρεσιών στους χρήστες.

Πίνακας 7 Ποσότητες παραγόμενων πληροφοριών σε παγκόσμιο επίπεδο



Το Collaborative Filtering είναι μια τεχνική που χρησιμοποιείται κυρίως από συστήματα που κάνουν στον χρήστη προτάσεις επιλογών. Συστήματα που έχουν επεξεργαστεί τις επιλογές και τις συνήθειες των προηγούμενων χρηστών και με τη μορφή προτιμήσεων ή βαθμολογήσεων που δίνουν οι χρήστες και μέσω μιας μηχανής, προσεγγίζουν τις προτιμήσεις των νέων χρηστών. Λόγω του ότι στηρίζεται στην εμπειρία που αποκτά από τη συμπεριφορά των χρηστών, συνεχώς βελτιώνονται

τα αποτελέσματα του. Το σύστημα ταιριάζει τις συμπεριφορές των χρηστών και μόλις εντοπίσει κοινά στοιχεία των νέων χρηστών με τους παλιότερους κάνει σύσταση για το πιθανώς ο νέος πελάτης χρειάζεται.
Η πολυπλοκότητα του συστήματος έγκειται στο γεγονός ότι συγκεντρώνει φαινομενικά ασύνδετες πληροφορίες και από τον συνδυασμό τους πιθανολογεί το αποτέλεσμα. Αν βρει άλλους χρήστες με τις ίδιες καταναλωτικές συμπεριφορές τότε ελέγχει τι δεν καταναλώθηκε και γίνεται σύσταση.
Επίσης χαρακτηριστικό στοιχείο του συστήματος είναι ότι όσο περισσότερο το χρησιμοποιεί ο ίδιος χρήστης τόσο περισσότερο πιο εξατομικευμένη γίνεται η σελίδα προς αυτόν. Σύμφωνα με τα ως άνω προκύπτει πως η εφαρμογή ισχυροποιείται όταν :

- Οι χρήστες που την χρησιμοποιούν γίνονται περισσότεροι

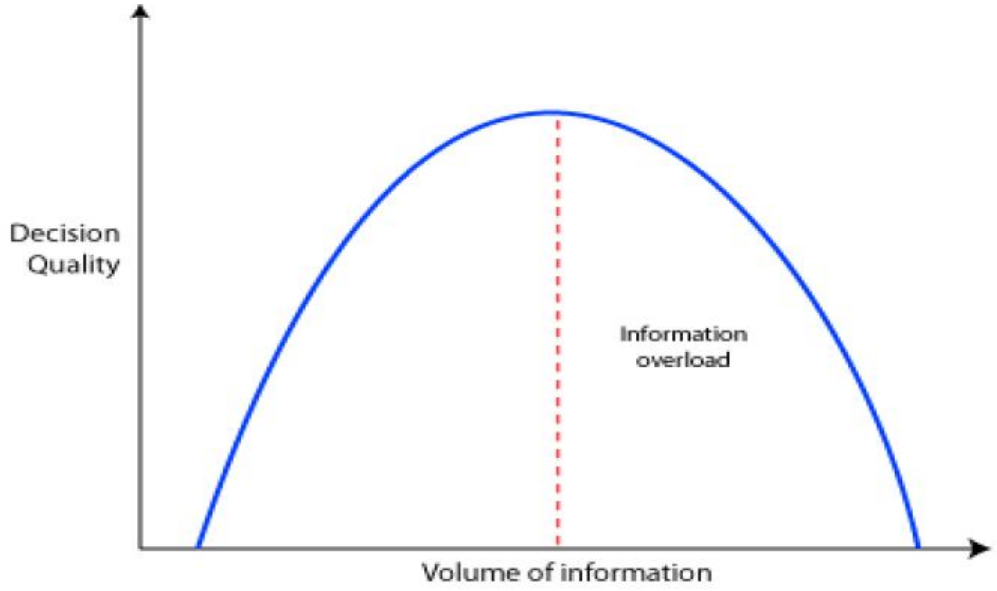
- Το σύστημα αποκτήσει εμπειρία για τον εκάστωτε χρήστη

Στα προβλήματα που αντιμετωπίζει η εφαρμογή τοποθετείται ότι δεν μπορεί αυτομάτως και εξ αρχής να κατηγοριοποιήσει τον χρήστη. Χρειάζεται να αποκτήσει εμπειρία μέσα από την χρήση της σελίδας από τον χρήστη. Πέραν αυτού όμως οι επιλογές του νέου χρήστη μπορεί να προκαλέσουν επιπλοκές στο σύστημα και να αποπροσανατολίσουν την κατεύθυνση του συστήματος.

3.4 Στόχος Συστήματος

Από την εφαρμογή της τεχνολογίας CF προσδοκάτε η δημιουργία μιας μηχανής αναζήτησης η οποία θα φιλτράρει τα δεδομένα και θα αποδίδει στον τελικό χρήστη προσωποποιούμενες προτάσεις. Δηλαδή στοχεύει από ένα τεράστιο όγκο πληροφοριών να είναι σε θέση να επιλέξει μόνο τις επιθυμητές από τον χρήστη. Το πρόβλημα που καλείται να λύση η εφαρμογή είναι υπαρκτό καθώς εκατομμύρια χρήστες επισκέπτονται συγκεκριμένα sites με χιλιάδες προϊόντα (Amazon), ή με εκατομμύρια αναρτήσεις (Google News). Το πρόβλημα γιγαντώνεται από την συνεχή ροή νέων πληροφοριών που εισάγονται, είτε από την πλευρά των χρηστών είτε από των προσφερόμενων υπηρεσιών, (προϊόντα, ειδήσεις, κλπ)

Πίνακας 8 Διαγραμματική απεικόνιση του φαινομένου Information Overload όπου λόγω του όγκου των πληροφοριών οι χρήστες χάνουν συχνά το στόχο τους

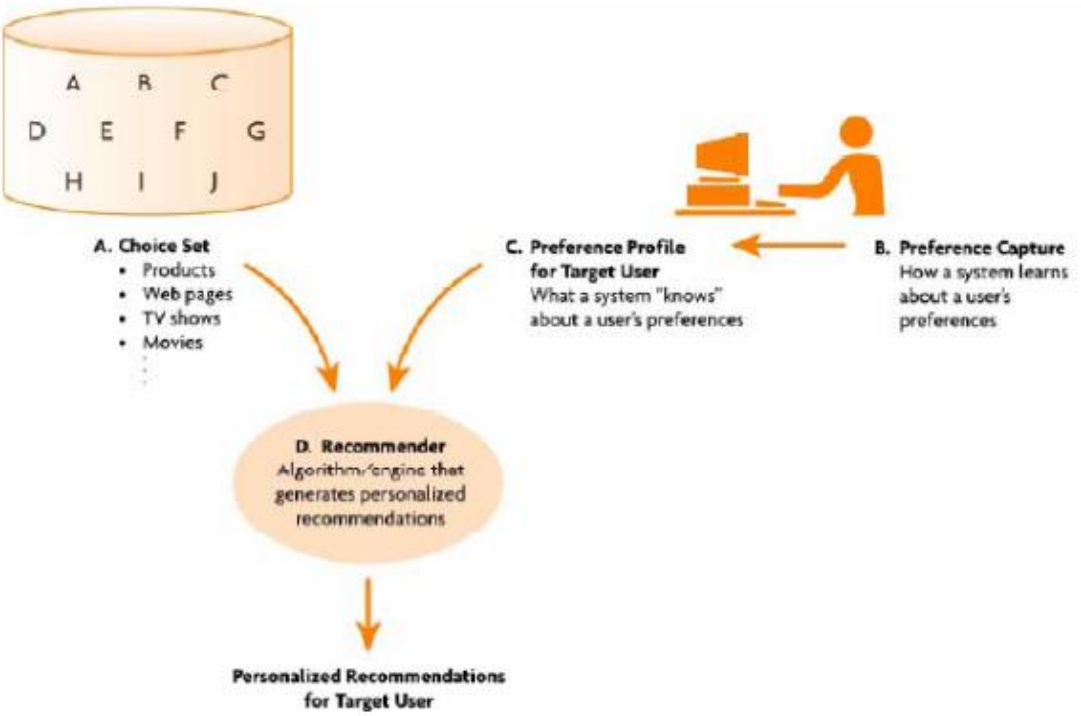


Στην εποχή που διανύουμε η οποία χαρακτηρίζεται από την έκρηξη των πληροφοριών η τεχνική το CF μπορεί να αποδειχθεί πολύ χρήσιμη, καθώς ο αριθμός των αντικειμένων σε μία μόνο κατηγορία (μουσική, ταινίες, βιβλία, ειδήσεις, ιστοσελίδες) έχει γίνει τόσο μεγάλος, ώστε ένα άτομο δεν δυνατό να τον προσπελάσει, προκειμένου να επιλέξει αυτά που τον ενδιαφέρουν. Αν η τεχνική στηριζόταν μόνο σε ένα σύστημα βαθμολόγησης το οποίο εντοπίζει το μέσο όρο για όλους τους χρήστες τότε θα αγνοούσε τις απαιτήσεις του συγκεκριμένου χρήστη, και θα ήταν ιδιαίτερα φτωχή σε περιπτώσεις όπου υπάρχει μεγάλη διακύμανση ενδιαφέροντος, όπως για παράδειγμα η πρόταση για συγκεκριμένο είδος μουσικής.

3.5 Γενική Δομή Συστήματος Collaborate Filtering

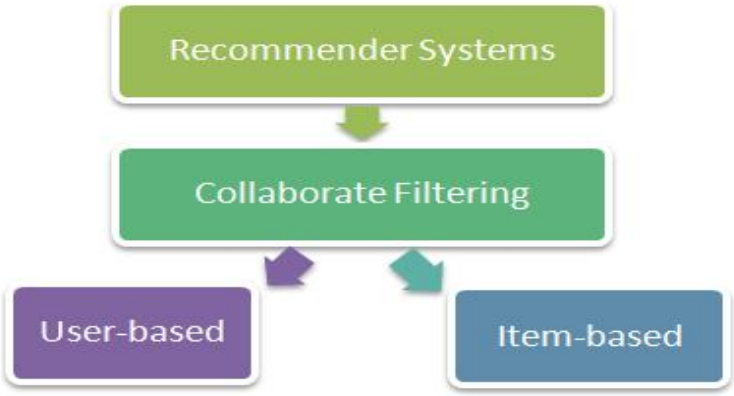
Οι τεχνικές CF χρησιμοποιούν βάσεις δεδομένων που προκύπτουν από τις επιλογές των χρηστών, ώστε να γίνουν προβλέψεις για προϊόντα που πιθανώς να χρειαστούν νέοι πελάτες με το ίδιο προφίλ. Σε ένα τυπικό σενάριο υπάρχει μια λίστα με χ χρήστες (χ1, χ2,...χν) και μια λίστα με ψ προϊόντα (ψ1, ψ2,...,ψν) και κάθε χρήστης (χν) έχει μια λίστα με προϊόντα (Ψχν), τα οποία ο χρήστης έχει βαθμολογήσει. Η

βαθμολόγηση μπορεί να είναι της κλίμακας από 1 έως 5 είτε μέσω ενδείξεων από τις οποίες ο χρήστης έχει να επιλέξει. Εντούτοις είναι πολύ πιθανό να μην υπάρχουν πάντα δεδομένα αξιολογήσεων αλλά αντίθετα να υπάρχουν διάφορα δυαδικά στοιχεία. (π.χ αν ένα αντικείμενο αγοράστηκε ή όχι.)



Εικόνα 11 Η δομή του συστήματος συνίσταται σε 4 στάδια

Στην συνέχεια η συνδυαστική σύγκριση της βαθμολόγησης του κάθε χρήστη για το κάθε προϊόν δίνει στο σύστημα την δυνατότητα να κάνει προβλέψεις για ένα καινούργιο χρήστη ο οποίος έχει δώσει στο σύστημα πληροφορίες σχετικά με τις προτιμήσεις του. Κατά μια έννοια το κλικ το κάθε χρήστη πάνω σε μια επιλογή (άρθρο, προϊόν, κλπ) αποτελεί μια ψήφο. Τα συστήματα Collaborate Filtering διακρίνονται σε Item-based και User-based.



Εικόνα 12 Διάκριση συστημάτων συνεργατικής διήθησης

3.6 Αλγόριθμοι

Στην τεχνολογία της συνεργατικής διήθησης δεδομένων «Collaborative Filtering» διακρίνονται δύο μοντέλα (όπως αναλύθηκαν παραπάνω) και εφαρμόζονται αντίστοιχα δύο τύποι αλγόριθμων, οι Memory-based και οι Model-based.

Στα CF χρησιμοποιούνται συνήθως δύο τρόποι προσέγγισης για την σύσταση την βασισμένη στην μνήμη «Memory-based» και βασισμένη στο μοντέλο «Model based». Στην «Memory-based» προσέγγιση το σύστημα επεξεργάζεται όλα τα αντικείμενα που έχει αξιολογήσει ο χρήστης για να μπορέσει να τον καταχωρίσει σε ομάδα κοινών ενδιαφερόντων. Το πρόβλημα είναι ότι τις περισσότερες φορές πρέπει να προσπελάσει μεγάλο όγκο δεδομένων. Αυτό έχει σαν αποτέλεσμα έχει να είναι αδύνατο μερικές φορές να πραγματοποιηθεί σύσταση σε πραγματικό χρόνο. Στην βασισμένη στο μοντέλο «Model-based» προσέγγιση δημιουργούμε ένα πρότυπο για τα δεδομένα που χρειαζόμαστε για την εκάστοτε σύσταση. Με άλλα λόγια, χρησιμοποιούμε μόνο τις αξιολογήσεις των χρηστών που μας χρειάζονται, όχι το σύνολο τους. Αυτή η προσέγγιση προσφέρει την ζητούμενη ταχύτητα και εξατομίκευση σε μια σύσταση.

Πλεονεκτήματα: Τα πλεονεκτήματα του φιλτραρίσματος συνεργασίας είναι αρκετά. Πρώτον, το σύστημα δεν χρειάζεται να έχει καμία γνώση πάνω στο αντικείμενο για το οποίο πραγματοποιείται η σύσταση. Δεύτερον, μέσα από την μεθοδο του φιλτραρίσματος μπορούν να πραγματοποιηθούν ασυνήθιστες συστάσεις αλλά, ταυτόχρονα, άκρως ουσιαστικές. Τρίτον, αν ακολουθήσουμε την «Model-based»

προσέγγιση θα έχουμε χαμηλές απαιτήσεις σε μνήμη και ελάχιστο χρόνο επεξεργασίας δεδομένων.

Μειονεκτήματα: Τα μειονεκτήματα είναι κυρίως δύο. Το πρώτο είναι ότι η ποσότητα των δεδομένων που επεξεργάζεται επηρεάζει την ποιότητα της σύστασης που λαμβάνει ο χρήστης. Το δεύτερο πρόβλημα που υπάρχει είναι το λεγόμενο «cold start problem». Δηλαδή, είναι δύσκολο να πραγματοποιηθεί μια σύσταση σε έναν νέο χρήστη που δεν έχει αξιολογήσει τίποτα ακόμα και για το νέο αντικείμενο είναι δύσκολο να συσταθεί αφού δεν έχει αξιολογηθεί από κανέναν χρήστη.
Όσον αφορά την μέθοδο «συνεργατική διήθηση με βάση τους χρήστες» αυτή αντιστοιχεί με τον αλγόριθμο Memory-based όπου γίνονται προβλέψεις μέσω της αξιολόγησης των χρηστών βασιζόμενοι σε στοιχεία προηγούμενων αξιολογήσεων που οι ίδιοι οι χρήστες έχουν κάνει. Η πρόβλεψη υπολογίζεται ως ένας μέσος όρος αξιολογήσεων του κάθε χρήστη αλλά και των χρηστών που το προφίλ τους έχει ομοιότητες. Για την υλοποίηση της πρόβλεψης δημιουργούνται πίνακες ομοιότητας ζευγών χρηστών οι όποιοι λειτουργούν παρασκηνιακά. Οι συγκριτικοί πίνακες λειτουργούν με διάδικο σύστημα με 1 ή 0 ανάλογα με την αξιολόγηση ή όχι ενός αντικειμένου από τον χρήστη. Παράλληλα έχει ορισθεί ένα κατώτατο όριο ομοιοτήτων για να αποφασιστεί αν υπάρχει τελικώς ομοιότητα μεταξύ των χρηστών του ζεύγους.

3.9.1 Αλγόριθμοι Βασισμένοι σε Μνήμη

Όσον αφορά τοτς αλγόριθμους που είναι βασισμένοι σε μνήμη θα πρέπει να αναφερθεί πως το σύστημα διατηρεί στην μνήμη όλες τις γνωστές βαθμολογίες/προτιμήσεις και τις χρησιμοποιεί για να βρει ομοιότητες ανάμεσα σε χρήστες ή αντικείμενα. Δύο χρήστες μοιάζουν όταν ενδιαφέρονται για παρόμοια πράγματα ενώ δύο αντικείμενα μοιάζουν όταν ένα σύνολο χρηστών τα αντιμετωπίζουν με παρόμοια αρέσκεια.
Στην περίπτωση που θέλουμε να βασιστούμε στις ομοιότητες χρηστών (user-based collaborative filtering), αναπαριστούμε τους χρήστες ως διανύσματα στο χώρο των αντικειμένων και υπολογίζουμε την ομοιότητα τους με βάση την απόσταση αυτών των διανυσμάτων. Όταν θέλουμε να εκτιμήσουμε την προτίμηση ενός χρήστη για ένα

άγνωστο αντικείμενο συγκεντρώνουμε τις προτιμήσεις των N κοντινότερων χρηστών που έχουν εκφράσει την προτίμηση τους για το αντικείμενο. Εκτιμούμε την προτίμηση του χρήστη εφαρμόζοντας μια συναθροιστική συνάρτηση, συνήθως τον σταθμισμένο μέσο όρο, πάνω στις τιμές που συγκεντρώσαμε.

Φορμαλιστικά, έστω s (u_i , u_j) η συνάρτηση που υπολογίζει την ομοιότητα ανάμεσα στους χρήστες u_i και u_j , $S \subseteq A$ το σύνολό των N όμοιων χρηστών, $r_{u,I}$ η προτίμηση του χρήστη u για το αντικείμενο I. Η γενική μορφή της συνάρτησης για τον υπολογισμό της εκτιμώμενης προτίμησης θα είναι $E(u,I) = aggr(S)$ και αν η συναθροιστική συνάρτηση είναινο ίναιν η συναθροιστική συ

$$E(u,I) = \frac{\sum_{u_o \in S} s(u,u_o) * r_{u_o,I}}{\sum_{u_o \in S} s(u,u_o)}$$

Αντίστοιχα όταν βασιζόμαστε στις ομοιότητες αντικειμένων (item-based collaborative filtering) αναπαριστούμε τα αντικείμενα ως διανύσματα στον χώρο των χρηστών και υπολογίζουμε τις αποστάσεις των διανυσμάτων. Όταν θέλουμε να εκτιμήσουμε την προτίμηση ενός χρήστη για ένα άγνωστο αντικείμενο βρίσκουμε τα N κοντινότερα αντικείμενα τα οποία έχει βαθμολογήσει ο χρήστης και υπολογίζουμε μέσω μιας συνάρτησης συνήθως με σταθμισμένο μέσο όρο, την τιμή βάσει των άλλων προτιμήσεων του. Φορμαλιστικά έστω s (Ii , $I\ j$) η ομοιότητα ανάμεσα στα αντικείμενα $I\ i$ και $I\ j$ και $SI \subseteq B$ το σύνολο των N όμοιων αντικειμένων. Η εκτίμηση της προτίμησης είναι $E(u,I)$ = $aggr$ (SI) και αν υποθέσουμε ότι χρησιμοποιούμε τον σταθμισμένο μέσο όρο

$$E(u,I) = \frac{\sum_{I_o \in SI} s(I,I_o) * r_{u,I_o}}{\sum_{I_o \in SI} s(I,I_o)}$$

3.9.2 Αλγόριθμοι βασισμένοι σε μοντέλο

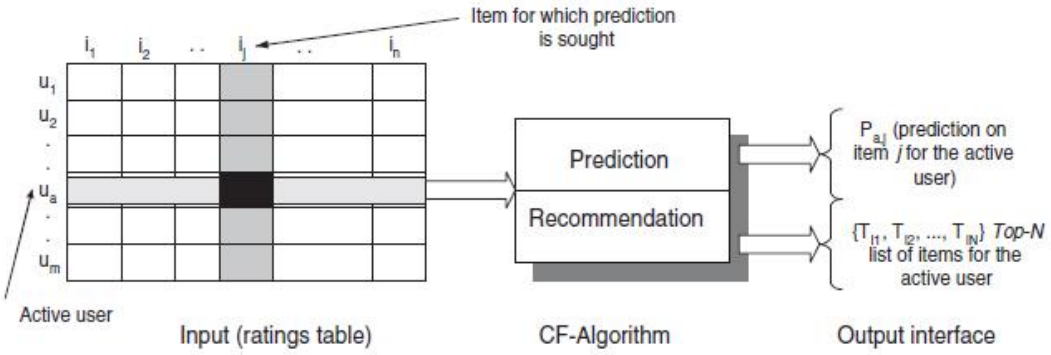
Τα περισσότερα συστήματα συνεργατικών τεχνικών βασίζονται σε αλγορίθμους μνήμης καθώς υλοποιούνται εύκολα. Παρουσιάζουν όμως πρόβλημα κλιμάκωσης και είναι ευαίσθητα στην εισαγωγή ψεύτικων προφίλ (profile injection). Σε αυτού του

τύπου τις επιθέσεις γίνεται προσπάθεια εισαγωγής ψεύτικων προφίλ χρηστών ώστε μέσω των προτιμήσεων να ευνοηθούν αντικείμενα που στην πραγματικότητα δεν θα ήταν το ίδιο δημοφιλή. Οι αλγόριθμοι που βασίζονται σε μοντέλο χρησιμοποιούν τα δεδομένα των προτιμήσεων/βαθμολογιών ως σύνολο εκπαίδευσης αλγόριθμων μηχανικής μάθησης για να παράξουν μοντέλα πρόβλεψης βαθμολογιών. Η επεξεργασία γίνεται σε μη πραγματικό χρόνο και έτσι μειώνεται το πρόβλημα κλιμάκωσης καθώς μπορεί να γίνει επεξεργασία μεγάλου αριθμό από προφίλ καθώς δεν απαιτείται απάντηση σε πραγματικό χρόνο. Η δυνατότητα της επεξεργασίας στο παρασκήνιο επιτρέπει την εκτέλεση πολύπλοκων αλγορίθμων αντιμετώπισης του θορύβου που προκαλούν οι επιθέσεις ψεύτικων προφίλ.

Τα μοντέλα πολλές φόρες τείνουν να έχουν λιγότερο ακριβή αποτελέσματα από τους αλγορίθμους μνήμης (O'Conner & Herlocker 1999). Για αυτό οι σχεδιαστές συστημάτων συστάσεων θα πρέπει να αναλογιστούν την σχέση κλιμάκωσης προς απόδοση πριν αποφασίσουν ποιες τεχνικές θα χρησιμοποιήσουν.

Δεδομένου ότι η πλειονότητα των χρηστών παρουσιάζουν πολλά ενδιαφέροντα βρίσκονται αντίστοιχα ταξινομημένοι σε διαφορικές και περισσότερες της μιας κατηγορίες χρηστών. Για να αντιμετωπιστεί αυτή η πολυπλοκότητα χρησιμοποιείται η μέθοδος ομαδοποίησης MinHash. Η μέθοδος MinHashing είναι μια μέθοδος ομαδοποίησης πιθανοτήτων. Αναθέτει ένα ζεύγος χρηστών της ίδιας κατηγορίας, με πιθανότητα ανάλογη της επικάλυψης των αντικειμένων που οι δύο χρήστες έχουν επιλέξει. Ο κάθε χρήστης αντιπροσωπεύεται από μια λίστα αντικειμένων, τα οποία αποτελούν το ιστορικό της αναζήτησης του. Η ομοιότητα μεταξύ των δύο χρηστών καθορίζεται από την επικάλυψη (κινά αντικείμενα) μεταξύ των λιστών του καθενός. Η ομοιότητα αποδίδεται από τον συντελεστή Jaccard που αναλύθηκε σε προηγούμενο κεφάλαιο. Ωστόσο η πρόβλεψη όταν πρόκειται για πραγματικό χρόνο δεν είναι πάντα εφικτή. Προκύπτουν για αυτό τεχνικές που περιορίζουν την αναζήτηση, όπως για παράδειγμα η δημιουργία ενός hash table, για να ανακαλύψουμε χρήστες που έχουν τουλάχιστον μια κοινή ψήφο- επιλογή.

Η βασική ιδέα του MinHashing είναι να υπολογίζεται μια ενιαία τιμή για κάθε χρήστη ώστε να μειώνεται ο όγκος των συσχετίσεων που μπορούν να προκύψουν. Επομένως μπορούμε να σκεφτούμε το MinHashing ως ένα αλγόριθμο που μπορεί να καταδείξει την πιθανότητα ομοιότητας, κάνοντας το με το να αντιστοιχεί σε μια θέση για κάθε χρήστη.



3.9.3 Σύγκριση Αλγόριθμων

Όπως προκύπτει από την έως τώρα ανάλυση ο παραδοσιακός αλγόριθμος collaborative filtering κάνει ελάχιστο ή καθόλου offline υπολογισμός, και ο online υπολογισμός εξαρτάται από τον αριθμό των πελατών και των αντικείμενων των καταλόγων. Ο αλγόριθμος είναι μη πρακτικός σε μεγάλα σύνολα στοιχείων, εκτός αν χρησιμοποιεί τη μείωση διαστατικότητας, δειγματοληψία, ή χωρισμός –εκ των οποίων όλα μειώνουν την ποιότητα σύστασης.

Τα βασισμένα στην αναζήτηση πρότυπα κατασκευάζουν τη λέξη κλειδί, την κατηγορία, και τις ενδείξεις των συντακτών offline, αλλά αποτυγχάνουν να παρέχουν συστάσεις με ενδιαφέροντες , επιλεγμένους τίτλους. Κάνουν επίσης κακές εκτιμήσεις για τους πελάτες με πολυάριθμες αγορές και εκτιμήσεις.

Το κλειδί για την εξελιξιμότητα και την απόδοση του ο αλγόριθμος item-to-item collaborative filtering είναι ότι δημιουργεί τους ακριβούς πίνακες για παρόμοια στοιχεία offline. Τα online συστατικά του αλγορίθμου -που ανατρέχει σε παρόμοια στοιχεία για τις αγορές και τις εκτιμήσεις του χρήστη- κάνουν εκτιμήσεις ανεξάρτητα από το μέγεθος καταλόγων ή το συνολικό αριθμό από τους πελάτες, εξαρτάται μόνο από το μέγεθος των τίτλων που ο χρήστης έχει αγοράσει ή έχει εκτιμήσει. Κατά συνέπεια, ο αλγόριθμος είναι γρήγορος ακόμη και για τα εξαιρετικά μεγάλα σύνολα στοιχείων. Επειδή ο αλγόριθμος συστήνει ιδιαίτερα υψηλά συσχετισμένα παρόμοια στοιχεία, η ποιότητα σύστασης είναι άριστη. Αντίθετα από τον παραδοσιακό αλγόριθμο collaborative filtering , ο αλγόριθμος αποδίδει επίσης καλά με περιορισμένα στοιχεία χρηστών, παράγοντας υψηλής ποιότητας συστάσεις με βάση μόνο δύο ή τρία στοιχεία.

Οι αλγόριθμοι σύστασης παρέχουν μια αποτελεσματική μορφή οροθετημένου μάρκετινγκ με τη δημιουργία μιας εξατομικευμένης εμπειρίας αγορών για κάθε πελάτη. Για τους μεγάλους λιανοπωλητές όπως το Amazon.com, ένας καλός αλγόριθμο σύστασης είναι εξελικτικός αναφορικά με μια πολύ μεγάλη βάση πελατών και έναν μεγάλο κατάλογο προϊόντων, απαιτεί μόνο κλάσματα δευτερολέπτου επεξεργασίας για να παραχθούν οι online συστάσεις, είναι σε θέση να αντιδράσει αμέσως στις αλλαγές στα στοιχεία ενός χρήστη, και κάνει αναγκαστικές συστάσεις για όλους τους χρήστε ανεξάρτητα από των αριθμό αγορών και εκτιμήσεων. Αντίθετα με άλλους αλγόριθμους, ο αλγόριθμος item-to-item collaborative filtering είναι ικανός να αντιμετωπίσει αυτήν την πρόκληση.

Στο μέλλον, αναμένουμε τη λιανική βιομηχανία να εφαρμόσει ευρύτερα τους αλγορίθμους σύστασης για οροθετημένο μάρκετινγκ, και online και offline. Ενώ οι επιχειρήσεις ηλεκτρονικού εμπορίου έχουν τα ευκολότερα οχήματα για την εξατομίκευση, τα αυξανόμενα ποσοστά μετατροπής της τεχνολογίας όπως αυτά συγκρίνονται με τις παραδοσιακές προσεγγίσεις ευρείας κλίμακας, θα καταστήσουν επίσης αναγκαίο για τους offline λιανοπωλητές τη χρήση αυτού στις ταχυδρομικές αποστολές, δελτία, και άλλες μορφές επικοινωνίας πελατών.

3.7 Περιορισμοί Εφαρμογής

Οι αλγόριθμοι που χρησιμοποιούνται για συμβουλευτικές υπηρεσίες στο ηλεκτρονικό εμπόριο συχνά δρουν σε ένα ανταγωνιστικό περιβάλλον, ειδικά για μεγάλα ηλεκτρονικά καταστήματα, όπως το eBay και το Amazon. Συχνά ένα συμβουλευτικό σύστημα προωθεί γρήγορα και ακριβή αποτελέσματα προτάσεις προσελκύουν το ενδιαφέρον των πελατών και προσφέρουν κέρδη στις εταιρείες. Για τα συστήματα CF η απόδοση έγκυρων αποτελεσμάτων εξαρτάται από το πόσο καλά οργανωμένα εισάγονται οι πληροφορίες το όποιο είναι άλλωστε και το χαρακτηριστικό της λειτουργίας τους – η εισαγωγή δεδομένων

Υπάρχουν αρκετοί περιορισμοί στις βάσεις μνήμης των τεχνικών CF. Χαρακτηριστικά αναφέρεται πως είναι αναξιόπιστες όταν ο χρήστης δεν έχει συμπληρώσει όλα τα απαιτούμενα πεδία με αποτέλεσμα να μην μπορεί να υπάρξει ταύτιση των δεδομένων ώστε να γίνει κάποια πρόβλεψη.

Πίνακας 9 Παράδειγμα έλλειψης απαιτούμενων πληροφοριών για την λήψη πρόβλεψης

(a)

Alice: (like) Shrek, Snow White, (dislike) Superman
Bob: (like) Snow White, Superman, (dislike) spiderman
Chris: (like) spiderman, (dislike) Snow white
Tony: (like) Shrek, (dislike) Spiderman

(b)

	Shrek	Snow White	Spider-man	Super-man
Alice	Like	Like		Dislike
Bob		Like	Dislike	Like
Chris		Dislike	Like	
Tony	Like		Dislike	?

Αραιή Αναπαράσταση Δεδομένων –Data Sparsity : Στην πραγματικότητα πολλά διαφημιστικά προωθητικά συστήματα χρησιμοποιούνται για να βελτιστοποιήσουν την διαχείριση μεγάλων ποσοτήτων προϊόντων. Οι πίνακες στοιχείων των χρηστών χρησιμοποιούνται για συνδυαστικό φιλτράρισμα το οποίο στις περιπτώσεις που οι χρήστες δεν συμπληρώνουν ολοκληρωμένα αμφισβητείται η αξιοπιστία του.

Η ελλιπής συμπλήρωση πινάκων στοιχείων από τους χρήστες είναι σύνηθες φαινόμενο. Παρατηρείται κυρίως όταν οι πληροφορίες ζητούνται να εισαχθούν στην αρχή όταν ακόμα η σχέση μεταξύ του χρήστη και του προγράμματος είναι κρύα.

Επίσης τα νέα προϊόντα δεν μπορούν να εκτιμηθούν μέχρι κάποιος χρήστης τα βαθμολογήσει και συνήθως οι για αυτό το λόγο στα πρώτα στάδια το σύστημα δεν μπορεί να κάνει σωστές προβλέψεις λόγω έλλειψης ιστορικού του προϊόντος.

Κάλυψη – Coverage : Μπορεί να οριστεί ως για ποια προϊόντα μπορεί το σύστημα να κάνει προβλέψεις στους υποψήφιους πελάτες. Η μειωμένη κάλυψη παρουσιάζεται όταν οι χρήστες βαθμολογούν μόνο ένα μικρό μέρος των προϊόντων, με αποτέλεσμα οι αλγόριθμοι να μην μπορούν να καταλήξουν σε ασφαλή συμπεράσματα.

Μεταβλητότητα Γειτονικά Δεδομένων – Neighbor Transitivity : Το πρόβλημα παρουσιάζεται από την ύπαρξη βάσεων δεδομένων με λίγες πληροφορίες που προκαλείται όταν οι χρήστες δεν έχουν εισάγει όλα τα δεδομένα που απαιτούνται. Ως αποτέλεσμα το σύστημα μπορεί να ταυτίσει μόνο μερικές από τις απαντήσεις με αυτές των άλλων χρηστών και δεν μπορεί να καταλήξει σε συμπεράσματα. Αυτό

επηρεάζει κυρίως συστήματα που συγκρίνουν τα δείγματα σε ζευγάρια – και ενώ δηλαδή έχει δημιουργηθεί ένα τέτοια ζευγάρι δύο πελατών – η έλλειψη στοιχείων για το ένα δεν μπορεί να τελεσφορήσει αποτελέσματα για το άλλο.

Όπως διαπιστώνεται η βασική περιοριστική παράμετρος στην επίτευξη πρόβλεψης είναι ο ανθρώπινος παράγοντας και συγκεκριμένα η έλλειψη πνεύματος συνεργασίας με το σύστημα.

Οι ευφυείς βοηθοί στους ιστοχώρους, ιδιαίτερα σε ιστοχώρους ηλεκτρονικού εμπορίου, γίνονται όλο και περισσότερο κοινοί. Ένα βασικό μειονέκτημα είναι ότι, πολύ συχνά αυτοί οι οδηγοί απόφασης αποτελούνται από έναν τεράστιο κατάλογο ερωτήσεων, οι οποίες πρέπει να απαντηθούν προκειμένου να βρεθεί το καταλληλότερο στοιχείο. Αυτό έχει ως αποτέλεσμα να δημιουργούνται διάλογοι μεγάλης διάρκειας, οι οποίοι οδηγούν στην μείωση του ενδιαφέροντος του χρήστη καθώς δεν τον ενθαρρύνουν στην χρήση αυτού του είδους των ιστοχώρων. Επομένως, στόχος είναι η ύπαρξη ενός ευφυή βοηθού, ο οποίος θα υποβάλλει τον *ελάχιστο αριθμό* ερωτήσεων στην κατάσταση διαλόγου, ελαχιστοποιώντας ταυτόχρονα και τον αριθμό των κύκλων αίτημα-απάντησης. Η μείωση του μήκους του διαλόγου μπορεί να επιτευχθεί μέσω της παραγωγής *επικεντρωμένων* ***(focused)*** ερωτήσεων στον χρήστη. (Aha D., 1995)

3.8 Αντιμετώπιση Προβλημάτων

Για την άμβλυνση των προβλημάτων που δημιουργούνται από την έλλειψη πληροφοριών έχουν προταθεί αρκετές τεχνικές αντιμετώπισης. Τεχνικές μείωσης της διάστασης του φαινομένου όπως η SVD Singular Value Decomposition η οποία δεν λαμβάνει υπόψη τα ερωτηματολόγια που παρουσιάζουν ελλείψεις, ώστε να μειώσουν τα φαινόμενα αναντιστοιχίας. Η μέθοδος της SVD στηρίζεται στο ότι είναι προτιμότερο να μην μπορούν να γίνουν προβλέψεις για μεγαλύτερο χρονικό διάστημα λόγω έλλειψης ιστορικού από το να δίνονται εσφαλμένες προτάσεις στους καταναλωτές. Το πρόβλημα της μεθόδου είναι ότι καθώς απορρίπτει τα ελλιπή βιογραφικά, υπάρχει ο κίνδυνος να χαθούν και σημαντικές πληροφορίες.

Μια άλλη τεχνική η (LSI) Latent Semantic Indexing (LSI) χρησιμοποιεί την αρχή ανάκτησης πληροφοριών. Η μέθοδος αξιοποιεί όλα τα ερωτηματολόγια που συμπληρώθηκαν ανεξαρτήτου της πυκνότητας των δεδομένων και από τις ομοιότητες

που εντοπίζει μεταξύ του καταναλωτικού προφίλ των χρηστών προσπαθεί να κάνει προβλέψεις. (Χ. Χριστάκου, 2012)

4. ΠΑΡΑΔΕΙΓΜΑΤΑ ΕΦΑΡΜΟΓΩΝ ΕΞΟΡΥΞΗΣ ΓΝΩΣΗΣ ΣΤΟΝ ΠΑΓΚΟΣΜΙΟ ΙΣΤΟ

Ο γενικός σκοπός του τομέα του web usage mining είναι να συνδυάσει πληροφορίες που αφορούν τα πρότυπα των περιηγήσεων των χρηστών, που αποτελούν σημαντική πηγή εξαγωγής συμπερασμάτων για τη βελτίωση των σελίδων αλλά και την ομαδοποίηση των χρηστών. Τα αποτελέσματα που παράγονται από την ανάλυση των web log αρχείων μπορούν να χρησιμοποιηθούν για πολλούς σκοπούς όπως στην προσωποποίηση του περιεχομένου των σελίδων, στη βελτίωση της περιήγησης των χρηστών με προανάκληση (prefetching) και επαναποθήκευση (caching) δεδομένων, στη βελτίωση σχεδίασης των σελίδων και τέλος σε σελίδες όπως ηλεκτρονικού εμπορίου στην ικανοποίηση του πελάτη. Όλα αυτά αναλύονται στις ακόλουθες παραγράφους.

4.1 Προσωποποίηση Περιεχομένου

Οι τεχνικές που υπάρχουν μπορούν να χρησιμοποιηθούν για να παρέχουν προσωποποιημένα δεδομένα στου χρήστες δίνοντας τους την αίσθηση πως υπάρχει ειδικό περιεχόμενο σε σελίδες ειδικά για αυτούς και τα ενδιαφέροντά τους. Για παράδειγμα είναι δυνατόν να γίνει πρόβλεψη της συμπεριφοράς ενός χρήστη σε πραγματικό χρόνο εάν συγκριθούν τα τρέχοντα πρότυπα περιήγησης με τα τυπικά πρότυπα που έχουν παλαιότερα εξαχθεί από τα web log αρχεία. Σε αυτή την περιοχή τα συστήματα (recommendation systems) που προτείνουν σε χρήστες συνδέσμους με περιεχόμενο που πιθανό τους ενδιαφέρει [42][43][44][45] είναι πολύ κοινά χρησιμοποιούμενα. Οι προσωποποιημένοι χάρτες [46] είναι ένα παράδειγμα συστήματος που προτείνει συνδέσμους. Το πιο γνωστό πλέον σύστημα που προτείνει συνδέσμους είναι το google. Εκεί κάθε υπηρεσία που προσφέρεται έχει ειδική περιοχή που προτείνει συνδέσμους στο χρήστη με χρήση αρχείων cookies. Ακόμα και οι αναζητήσεις στη μηχανή αναζήτησης της google είναι προσωποποιημένες για κάθε υπολογιστή. Λαμβάνει υπ’όψιν το ιστορικό αναζητήσεων και σε μελλοντικές αναζητήσεις δίνει αντίστοιχα αποτελέσματα που είναι πιο ενδιαφέροντα για κάθε χρήστη. Ένα άλλο παράδειγμα είναι τα ηλεκτρονικά καταστήματα που ανάλογα με

τις προηγούμενες αγορές ενός χρήστη του προτείνει στην κεντρική σελίδα προσφορές της κατηγορίας προϊόντων που προτιμάει.

4.2 Προανάκληση και Επαναποθήκευση Δεδομένων

Τα αποτελέσματα που προέρχονται από την εφαρμογή τεχνικών web usage mining μπορούν να χρησιμοποιηθούν για τη βελτίωση της απόδοσης του διακομιστή και γενικά των εφαρμογών διαδικτύου. Τυπικά οι τεχνικές web usage mining μπορούν να χρησιμοποιηθούν για τη δημιουργία κατάλληλων στρατηγικών προανάκλησης και επαναποθήκευσης δεδομένων για τη μείωση του χρόνου απόκρισης των διακομιστών

4.3 Υποστήριξη στο Σχεδιασμό Σελίδων

Η ευχρηστία (usability) είναι ένα από τα σημαντικότερα ζητήματα στο σχεδιασμό και την υλοποίηση των σελίδων. Με τις υπάρχουσες τεχνικές μπορεί να δοθούν κατευθύνσεις για τη βελτίωση του σχεδιασμού των εφαρμογών διαδικτύου. Οι προσαρμοστικές σελίδες εκφράζουν ένα επιπλέον βήμα στην βελτίωση της ευχρηστίας τους. Σε αυτή την περίπτωση το περιεχόμενο και η δομή μπορεί δυναμικά να προσαρμόζεται και να αναδιοργανώνεται ανάλογα με τη συμπεριφορά των χρηστών.

4.4 Ηλεκτρονικό Εμπόριο

Η εξόρυξη γνώσης σε εμπορικές σελίδες είναι πολύ σημαντική στη βελτίωση των παρεχόμενων υπηρεσιών, την ικανοποίηση του κάθε πελάτη αλλά και στην αύξηση των κερδών της επιχείρησης. Η διαχείριση των σχέσεων πελάτη και επιχείρησης (customer relationship management) μπορεί να βοηθηθεί από τη χρήση τεχνικών web usage mining. Με αυτό το σκεπτικό δίνεται έμφαση στα εξής:

- Προσέλκυση πελατών
- Διατήρηση πελατών
- Ανταλλαγή πωλήσεων
- Ενεργή παρουσία πελατών

4.5 Μελέτη Συγκεκριμένων Περιπτώσεων

Θεωρείται ως δεδομένο πως το ηλεκτρονικό εμπόριο αποτελεί το μέλλον και την εξέλιξη του παραδοσιακού εμπορίου. Για να μπορέσει μια εταιρεία να γίνει ανταγωνιστική στο ηλεκτρονικό εμπόριο θα πρέπει να εφαρμόσει μεθόδους προσέλκυσης του πελάτη και να μπορέσει να προβλέψει τις ανάγκες του. Για αυτό τον λόγο έχουν αναπτυχθεί λογάριθμοι προσδιορισμού των αναγκών του καταναλωτή και αντίστοιχα συστήματα συστάσεων. Η υιοθέτηση μηχανισμών προσέλκυσης νέων πελατών θα πρέπει να γίνεται παράλληλα με εφαρμογή συστημάτων διαχείρισης της εταιρείας, (συστήματα ERP που αναπτύχθηκαν στο 2^{ο} κεφάλαιο). Σε αντίθετη περίπτωση, όταν δηλαδή η εταιρεία αναπτύσσεται μονοδιάστατα, χωρίς την βελτίωση της οργάνωσης της το αποτέλεσμα είναι να εκτίθεται δημοσίως. (Σιωμίκος Γ., 2002), (Φούκη Ι, 2013) Για την κατανόηση της σημαντικότητας αυτής της διττής ανάπτυξης μιας εταιρείας παρουσιάζονται και αναλύονται κάποια αρκετά διαδεδομένα παραδείγματα ιστοσελίδων. Συγκεκριμένα παρουσιάζονται οι εξής ιστότοποι (Η. Marmanis, 2008)

- Εφημερίδα Guardian
- Cinematch - Ενοικιάσεις Ταινιών
- Twitter – Αλγόριθμος (Follower of Follower)
- eBay
- Amazon
- YouTube

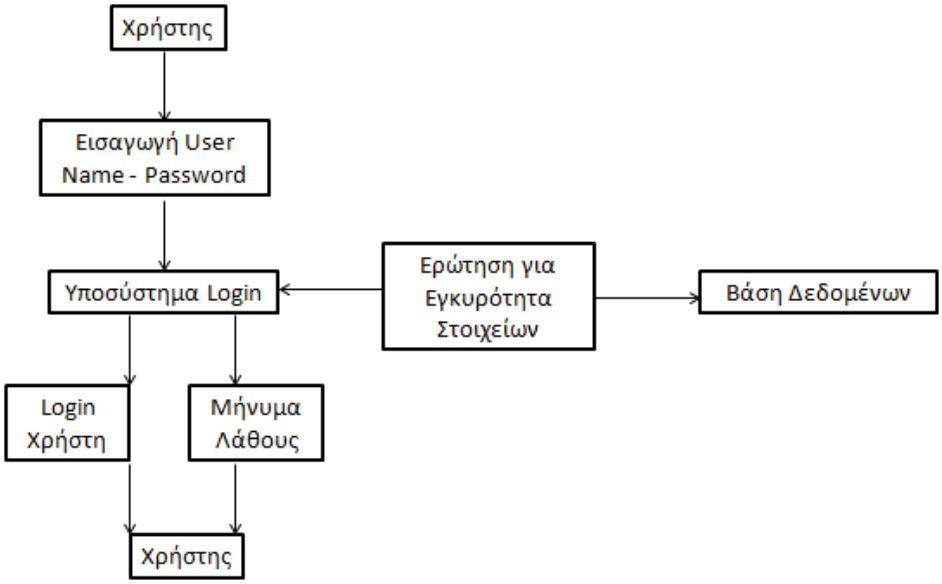
Στόχος της ανάλυσης είναι η τεκμηρίωση της άποψης ότι η επιτυχία των ιστότοπων σε θέματα σύστασης οφείλεται στην συνδυασμένη χρήση αλγόριθμων φιλτραρίσματος και του συστήματος διαχείρισης των βάσεων δεδομένων τους.

4.6 Εφημερίδα Guardian

Η ανάπτυξη ειδησεογραφικών ιστοσελίδων προέβαλε ως επιτακτική ανάγκη, στην εύρεση τρόπων που θα προσαρμόζονταν στις ανάγκες, τις ιδιαιτερότητες κάθε

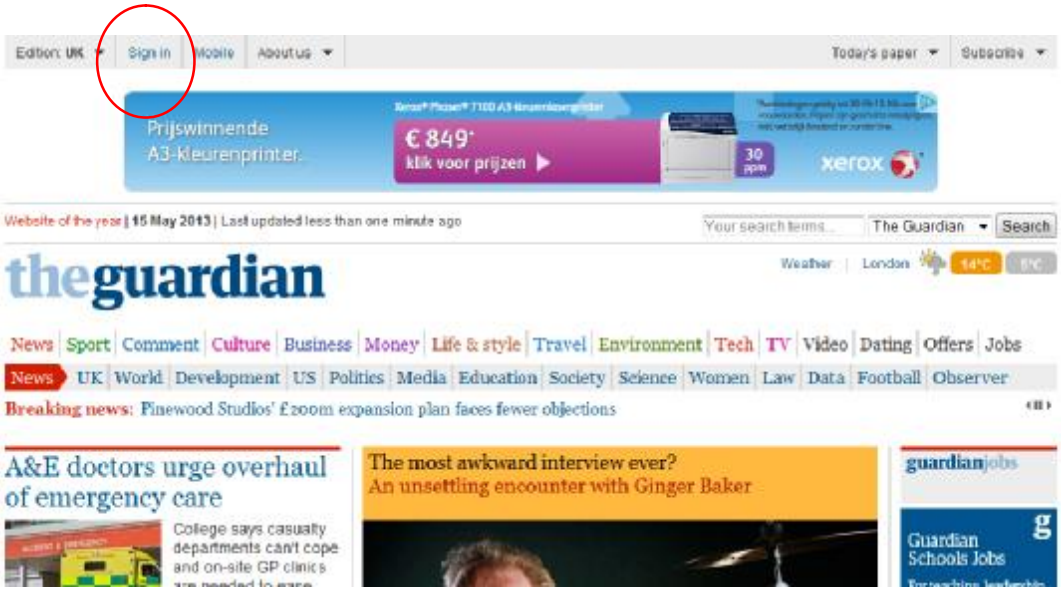
ατόμου, ώστε να του εξασφαλίσουν το επιθυμητό, δηλαδή την αδιάλειπτη και συνεχή εύκολη πρόσβαση στην πληροφορία. Μια από τις πρώτες σε επισκεψιμότητα σελίδες σε ευρωπαϊκό επίπεδο είναι η εφημερίδα Guardian.

Στην αρχική της σελίδα (www.guardian.co.uk) η εφημερίδα έχει πάνω δεξιά την επιλογή εγγραφής στο σύστημα. Η εγγραφή μπορεί να γίνει μέσα σε ελάχιστα δευτερόλεπτα μέσω του λογαριασμού του Facebook που διαθέτει ο χρήσης. Όταν ο χρήστης δώσει τα αναγνωριστικά του στοιχεία (username και password), εξετάζεται – μετά από επικοινωνία με τη βάση δεδομένων – αν αντιστοιχούν σε εγγεγραμμένο χρήστη. Αν ναι, εισάγεται στο σύστημα αρχίζει την πλοήγησή του στην πύλη. Επίσης προσφέρεται και η εναλλακτική, αν δεν είναι γραμμένος χρήστης, να συμπληρώσει τη φόρμα εγγραφής και να καταχωρηθεί στο σύστημα. Το διάγραμμα ροής δεδομένων της εφαρμογής φαίνεται στην επόμενη εικόνα



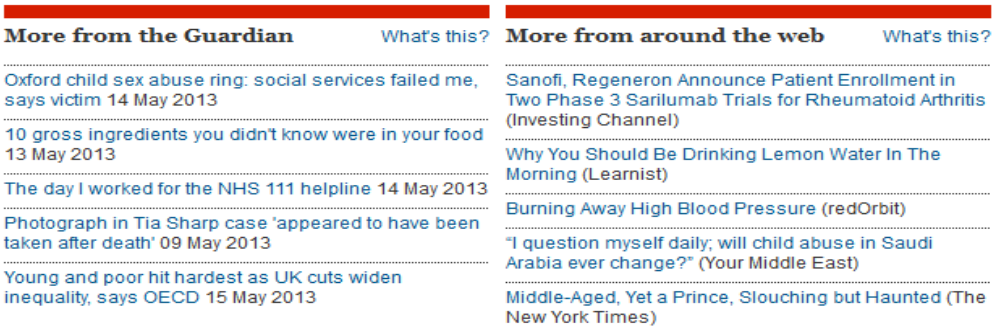
Εικόνα 13 Εφαρμογή εισόδου επισκέπτη

Μετά την εγγραφή του χρήστη στο σύστημα, κάθε άρθρο που διάβαζει ο χρήστης καταγράφεται και αποθηκεύεται σε μια βάση δεδομένων με το ιστορικό των τελευταίων τριάντα ημερών. Ως αποτέλεσμα της διαδικασίας αυτής, κάθε φορά που ο χρήστης διαβάζει ένα άρθρο στο τέλος της σελίδας του προτείνονται άρθρα που ομοιότητες στο περιεχόμενο με αυτά που έχει διαβάσει στο παρελθόν.



Εικόνα 14 Απόσπασμα αρχικής σελίδας

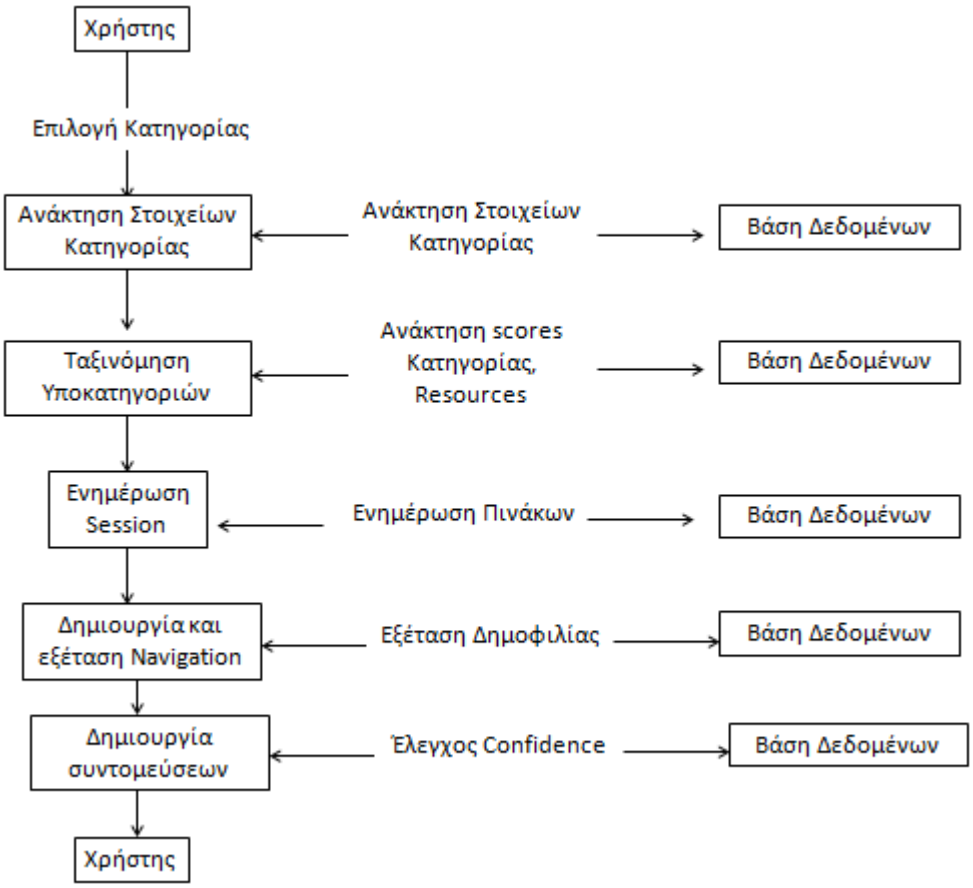
Ως στόχος του συστήματος είναι να καταφέρει να προτείνει άρθρα στον χρήστη τα οποία θα ήθελε πολύ αλλά θα ήταν δύσκολο να τα βρεί κάτω από άλλες συνθήκες.
Η ιστοσελίδα χρησιμοποιεί επίσης και μια μέθοδο μη προσωποποιημένης σύστασης που βασίζεται στα πιο δημοφιλή άρθρα που έχουν δημοσιευτεί. Η δημοτικότητα ενός άρθρου φαίνεται από το πόσες ενέργειες έχουν γίνει από τους χρήστες πάνω στο άρθρο. Ενέργεια μπορεί να θεωρηθεί η ανάγνωση ενός άρθρου, το ποστάρισμα του σε κάποιο blog ή το να γίνει αποστολή του άρθρου με e-mail. Η μέθοδος αυτή έχει στόχο να προσελκύσει νέους χρήστες που δεν έχουν κάνει εγγραφή ακόμα



Εικόνα 15 Οι σχετικές προτάσεις της εφημερίδας, βάσει του προηγούμενου άρθρου που αναγνώστηκε

Η ιστοσελίδα χρησιμοποιεί ως πηγή σύστασης τα κειμενικά χαρακτηριστικά καθώς και λέξεις κλειδιά όπως: θέμα και συγγραφέας. Με βάση ένα σύνολο

χαρακτηριστικών γνωρισμάτων στοιχείων, το σύστημα προσπαθεί να δημιουργήσει ένα μοντέλο για κάθε χρήστη που του επιτρέπει να ταξινομήσει τα άγνωστα στοιχεία σε ενδιαφέροντα μη ενδιαφέροντα για αυτόν. Η σύσταση στηρίζεται στο φιλτράρισμα περιεχομένου και η εφαρμογή του αλγόριθμου είναι απλή καθώς δεν χρειάζονται σύνθετες συσχετίσεις, (π.χ. την γνώμη των άλλων αναγνωστών). Ο αλγόριθμός παρουσιάζει την ακόλουθη μορφή.



Εικόνα 16 Εφαρμογή πλοήγησης

4.7 Cinematch - Ενοικίασεις Ταινιών

Η εταιρεία Netflix, που δραστηριοποιείται στον χώρο ενοικιάσεως ταινιών[3] μέσω διαδικτύου θέλοντας να προσφέρει όσο το δυνατόν καλύτερες υπηρεσίες στους συνδρομητές, σχεδίασαν ένα σύστημα συστάσεων (αλγόριθμος) με την βοήθεια του οποίου θα προτείνονταν ταινίες στους συνδρομητές με βάσει το προφίλ τους. Ο χρησιμοποιούμενος αλγόριθμος είναι ένα σύστημα συνεργατικού φιλτραρίσματος (Collaborative Filtering System) με την ονομασία. Το CineMatch είναι μια βάση δεδομένων που χρησιμοποιεί πληροφορίες από τρείς πηγές για να αποφασίσει ποιες ταινίες θα προτείνει στον χρήστη. (Μ. Κωνσταντίνου, 2012)

- Αρχικά οι ίδιες οι ταινίες, οι οποίες οργανώνονται σε ομάδες κοινών ταινιών, βάσει του περιεχομένου τους. Με τον όρο «περιεχόμενο ταινίας» καλείται ένα σύνολο στοιχείων που μπορούν να καθορίσουν λεκτικά τις παραμέτρους μιας ταινίας. Μια σημαντική παράμετρος είναι το είδος μιας ταινίας, ή καλύτερα τα είδη στα οποία ανήκει. Λαμβάνονται υπόψη ακόμα οι συντελεστές της ταινίας, δηλαδή ο σκηνοθέτης, ο σεναριογράφος και οι ηθοποιοί. Τέλος, εξετάζεται και η περίληψη της ταινίας. Προφανώς, από το κείμενο της περίληψης(μαζί με την προσθήκη του τίτλου) αφαιρούνται οι συνηθισμένες και χωρίς σημασία λέξεις (όπως σύνδεσμοι, αντωνυμίες κ.α.) και υπολογίζεται η συχνότητα των λέξεων για να καθοριστούν αυτές που είναι σημαντικές και χαρακτηρίζουν την ταινία.

- Δεύτερον οι βαθμολογίες που έχει δώσει ο χρήστης σε προηγούμενες ταινίες που έχει δει. Η βαθμολογία μπορεί να είναι από ένα έως πέντε αστέρια (πέντε είναι η καλλίτερη). Ταινίες άξιες σύστασης θεωρούνται αυτές που λαμβάνουν βαθμολογία 4-5, ενώ απορριπτέες αυτές που λαμβάνουν 1-3.Αυτό είναι απαραίτητο προκειμένου να γίνουν γνωστές οι προτιμήσεις του χρήστη και να κατασκευαστεί το προφίλ.

- Τέλος οι συνδυασμένες βαθμολογίες όλων των χρηστών της υπηρεσίας. (Χ. Χριστάκου, 2012)

[3] Η εταιρεία παρέχει τη μεγαλύτερη συλλογή ταινιών, όμως αυτή είναι διαθέσιμη μόνο στην Αμερική και κυρίως γίνεται μέσω της ταχυδρομικής υπηρεσίες, ενώ μόνο ένας μικρός αριθμός τίτλων είναι διαθέσιμος για προβολή μέσω διαδικτύου.

Εικόνα 17 Αποτελέσματα Recommender System στο Netflix

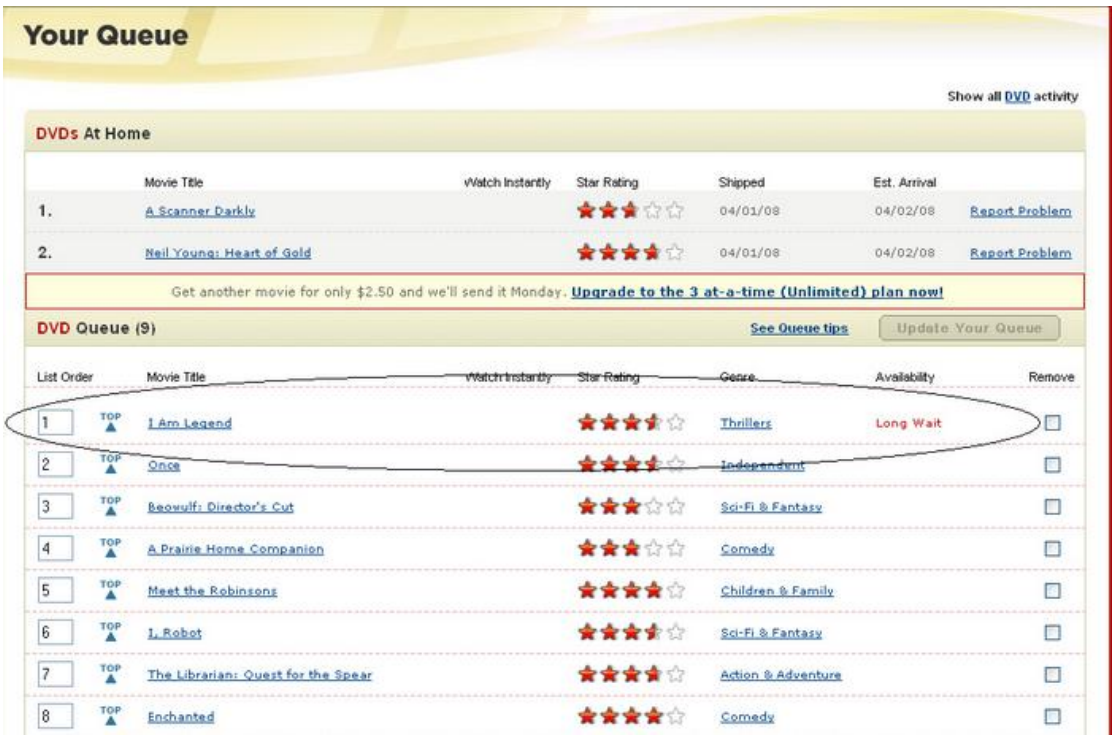
Στην συνέχεια συνδυάζει αυτές τις πληροφορίες μεταξύ τους με σκοπό να βρεί παρόμοιες ταινίες με τις ταινίες που του άρεσαν στο παρελθόν. Το Cinematch, εβδομαδιαία, επεξεργάζεται τις παλιότερες αξιολογήσεις ταινιών από τους συνδρομητές με σκοπό να καταχωρίσει τις ταινίες σε λίστες ταινιών με ομοιότητες. Για να βρεθεί η ομοιότητα ανάμεσα στις ταινίες, χρησιμοποιείται ο συντελεστής γραμμικής συσχέτισης. (Φ. Μιχελινάκης, 2011)

Η συσχέτιση μετρά το βαθμό συνάφειας αλληλεπίδρασης ανάμεσα σε δύο ή περισσότερες ταινίες. Πρακτικά, σημαίνει ότι από την τιμή ενός δείκτη (συντελεστή συσχέτισης) κατανοούμε πόσο έντονη ή χαλαρή είναι η συσχέτιση δύο ταινιών. Η διαδικασία συσχέτισης παρουσιάζεται στις κατηγορικές μεταβλητές του Netflix με μια παραλλαγή του συντελεστή Pearson.

Το στατιστικό τεστ X^2 (Pearson Chi-Square test) είναι το πιο δημοφιλές μη παραμετρικό τεστ.

Καθώς έχουμε στη διάθεση μας ένα δείγμα ποιοτικών δεδομένων οργανωμένο σε ονομαστικές κατηγορίες, στοχεύουμε στην χρήση των δεδομένων αυτών ώστε να προσδιοριστεί η αναλογία του πληθυσμού που ανήκει στην κάθε κατηγορία. Για την επίτευξη αυτού του στόχου διατυπώνεται μια μηδενική υπόθεση, που είτε δηλώνει ότι δεν υπάρχει κάποια συγκεκριμένη προτίμηση στις διαθέσιμες ονομαστικές κατηγορίες *(no-preference null-hypothesis)*, είτε δηλώνει ότι τα ποσοστά που προτιμώνται από τα υποκείμενα δε διαφέρουν από τα ποσοστά άλλων πληθυσμών οι οποίοι αποτελούν το σημείο αναφοράς *(nodifference from a comparison population)*.

Και στις δύο περιπτώσεις, αυτό που προσδιορίζει η μηδενική υπόθεση είναι ο αναμενόμενος αριθμός (expected frequency – fe) των υποκειμένων που ανήκει σε κάθε ονομαστική κατηγορία. Ο έλεγχος υποθέσεων που ακολουθεί αξιολογεί αυτή τη μηδενική υπόθεση, συγκρίνοντας τον αριθμό των υποκειμένων που αναμένεται σε κάθε ονομαστική κατηγορία με τον αριθμό των υποκειμένων που παρατηρείται ότι ανήκει σε κάθε ονομαστική κατηγορία (observed frequency - fo), με βάση τις μετρήσεις του δείγματος.



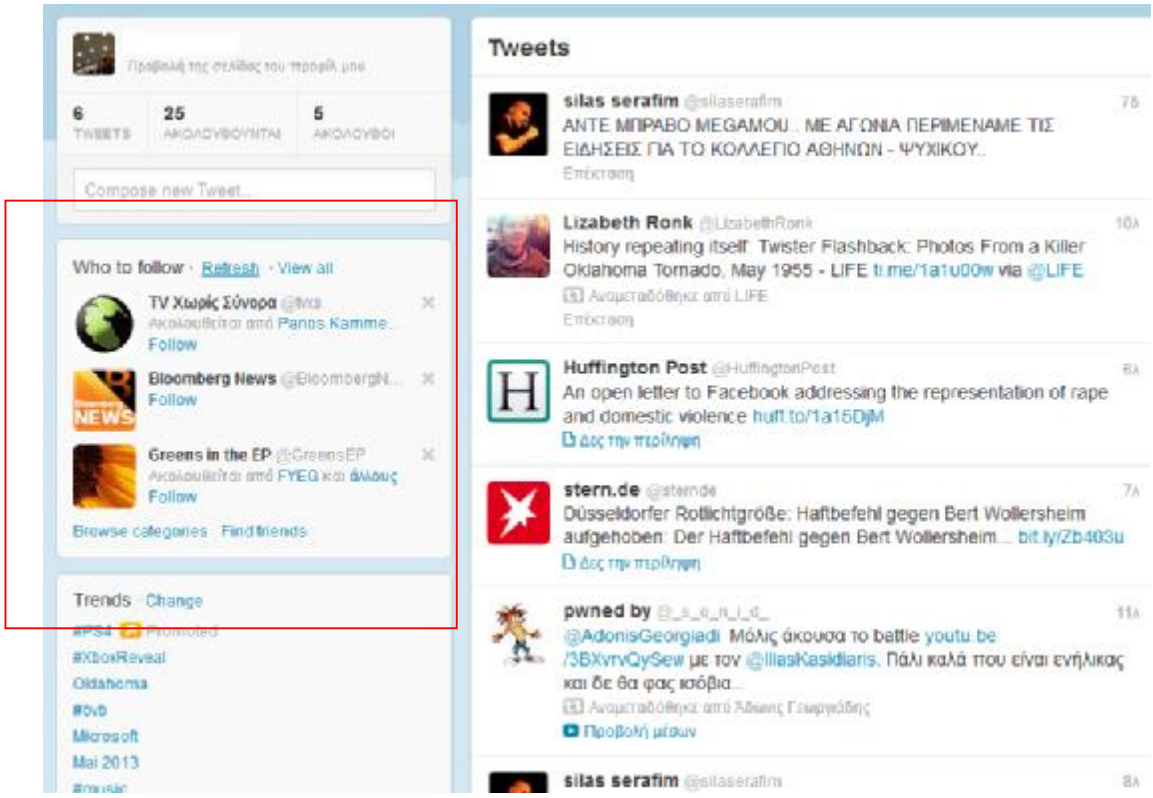
Εικόνα 18 Κατάταξη των ταινιών βάσει της λίστας αναμονής που υπάρχει για αυτές. Για κάθε μια ταίνία δίνεται η συνολική αξιολόγηση που έχει λάβει από το κοινό

Το σύστημα έχει τρία ευδιάκριτα στρώματα, το γραφικό περιβάλλον διαπροσωπείας χρήστη (GUI), το κύριο στρώμα εφαρμογής που εφαρμόζει τις στρατηγικές σύστασης (π.χ., μηχανή ψηφοφορίας) και το στρώμα αποθήκευσης πληροφοριών που αποτελείται από τη βάση δεδομένων ταινιών και τη βάση δεδομένων των προτιμήσεων των χρηστών.

4.8 Twitter – Follower of Follower

Το Twitter είναι μια από τις πιο αξιοσημείωτες micro-blogging υπηρεσίες, που απασχολεί ένα μοντέλο κοινωνικού δικτύου που ονομάζεται «following», στο οποίο κάθε χρήστης μπορεί να διαλέξει εκείνον που αυτός θέλει να «ακολουθεί» (follow), δηλαδή από τον οποίο μπορεί να λαμβάνει tweets χωρίς να απαιτείται ο τελευταίος να παρέχει άδεια πρώτα.

Το Micro-blogging είναι ένα είδος επικοινωνίας που εμφανίζεται ολοένα και περισσότερο στο προσκήνιο τα τελευταία χρόνια. Επιτρέπει στους χρήστες να δημοσιεύουν σύντομα μηνύματα ενημέρωσης, τα οποία μπορεί να υποβάλλονται σε πολλά διαφορετικά κανάλια, συμπεριλαμβανομένου του Web και της υπηρεσίας έκδοσης μηνυμάτων. Μία από τις περισσότερο διακεκριμένες υπηρεσίες του microblogging είναι το Twitter. Αναφερόμαστε στους χρήστες του twitter ως «Twitterers», και τα μικρά μηνύματα ενημέρωσης που δημοσιεύονται από τους χρήστες «tweets».



Εικόνα 19 Απόσπασμα της αρχικής σελίδας του Twitter. Στην δεξιά στήλη υπάρχει το παράθυρο σύστασης ποιόν να ακολουθήσει ο χρήστης

Αυτό επιτρέπει στους twitterers να δημοσιεύουν tweets (με ένα όριο 140 χαρακτήρων).Το Twitter επίσης παρέχει τη λειτουργικότητα της κοινωνικής δικτύωσης. Αντίθετα προς τις υπηρεσίες άλλων κοινωνικών δικτύων που απαιτούν οι χρήστες να στέλνουν προσκλήσεις φιλίας προς τους άλλους χρήστες για να τους κάνουν φίλους, το Twitter περιλαμβάνει ένα μοντέλο κοινωνικής δικτύωσης καλούμενο «ακόλουθος» (following),στο οποίο κάθε twitterer επιτρέπεται να διαλέγει ποιόν θέλει να ακολουθεί χωρίς απαίτηση κάποιας άδειας. Αντιστρόφως, αυτός μπορεί επίσης να ακολουθείται από άλλους χωρίς τη χορήγηση άδειας πρώτα.

Το Twitter έγινε πολύ δημοφιλές από την πρώτη μέρα που εμφανίστηκε. Ενδεικτικά, το πρώτο εξάμηνο του 2010 ήταν καταγεγραμμένοι στο Twitter πάνω από 100 εκατομμύρια χρήστες, οι οποίοι συνέτασσαν πάνω από 65 εκατομμύρια tweets την ημέρα. Αυτό έχει τραβήξει το αυξανόμενο ενδιαφέρον της ερευνητικής κοινότητας. Έχουν γίνει εργασίες για τη μελέτη των τοπολογικών και γεωγραφικών ιδιοτήτων του κοινωνικού δικτύου που σχηματίστηκε από τους twitterers και από τους followers τους. Επιπροσθέτως, έχουν πραγματοποιηθεί έρευνες για τον προσδιορισμό της ταυτότητας των twitterers που επηρεάζουν («influential»).

Το Twitter χρησιμοποιεί ένα αλγόριθμο που βασίζεται στην υπόθεση ότι «εάν κάποιος που ακολουθεί ο χρήστης Α, ακολουθεί με την σειρά του κάποιον χρήστη Β πιθανώς ο χρήστης Α να θέλει να ακολουθήσει τον Β» Αυτή η μέθοδος είναι αρκετά αποτελεσματική καθώς, είναι πολύ πιθανόν, χρήστες που έχουν κοινούς ακόλουθους να θέλουν να ακολουθηθούν και μεταξύ τους.

η “following” σχέση να είναι ένας ισχυρός δείκτης της ομοιότητας μεταξύ των χρηστών. Με άλλα λόγια, ένας twitterer ακολουθεί ένα φίλο επειδή ενδιαφέρεται για τα θέματα που ο φίλος δημοσιεύει στα tweets και ο φίλος ακολουθεί πίσω επειδή βρίσκει ότι μοιράζονται θέμα όμοιου ενδιαφέροντος. Αυτό το φαινόμενο ονομάζεται «ομοφιλία», και έχει παρατηρηθεί σε πολλά κοινωνικά δίκτυα, πράγμα το οποίο είναι πολύ σημαντικό.

Στο περιβάλλον του Twitter η ομοφιλία υποδηλώνει ότι ένας twitterer ακολουθεί ένα φίλο επειδή ενδιαφέρεται για τα θέματα που ο φίλος δημοσιεύει, και ο φίλος ακολουθεί πάλι επειδή βρίσκει ότι μοιράζονται όμοια θεματικά ενδιαφέροντα.

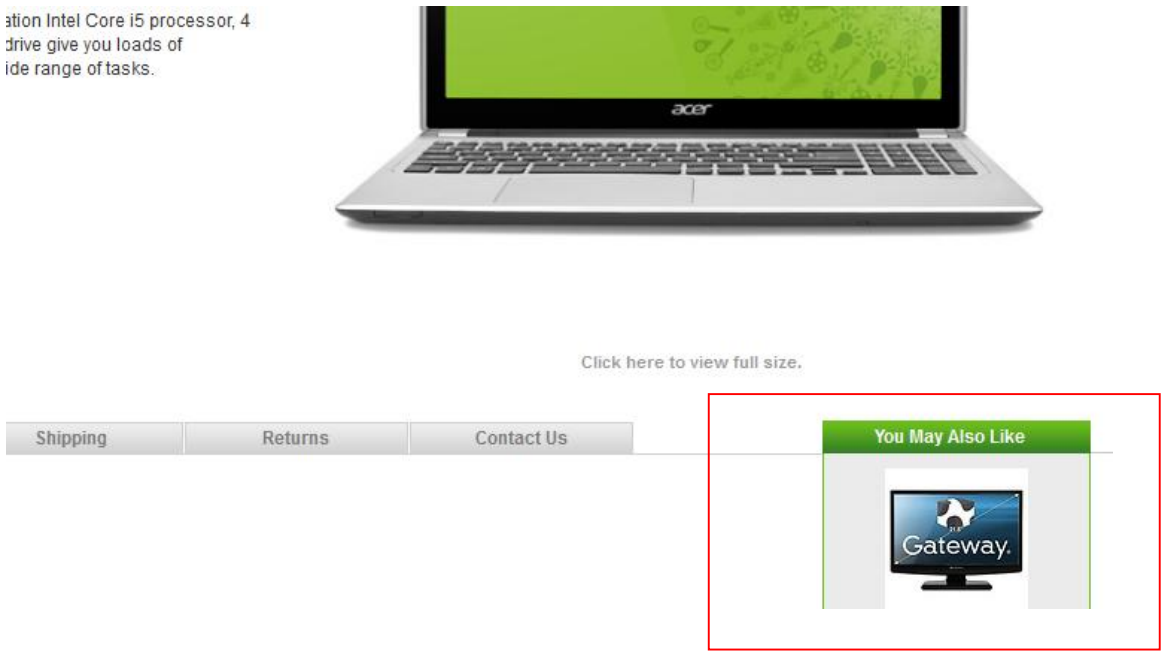
Η παρουσία της ομοφιλίας υποδηλώνει ότι υπάρχουν twitterers οι οποίοι με σοβαρότητα διαλέγουν φίλους να ακολουθήσουν. Αυτό είναι βασικό, διότι η αναγνώριση των twitterers οι οποίοι επηρεάζουν βασιζόμενοι στις «following»

σχέσεις, δεν θα είχε κανένα νόημα εάν κανένας twitterer δεν είναι διαλέγει με σοβαρότητα ποιους να κάνει follow.

4.9 eBay

Το eBay είναι ένα κατάστημα ηλεκτρονικού εμπορίου που εκτελεί την μεσολάβηση πώλησης διάφορων προϊόντων. Η ιστοσελίδα χρησιμοποιεί ως πηγή σύστασης τα χαρακτηριστικά του αντικειμένου. Όλα τα προϊόντα που εμπορεύεται η σελίδα διαθέτουν λεπτομερή χαρακτηριστικά (λέξεις κλειδιά) όπως: ονομασία, κατασκευαστή, χρονολογία κατασκευής.

Στην εφαρμογή «ίσως να σου αρέσει αυτό» οι συστάσεις βασίζονται στην αγορά ενός προϊόντος από τον χρήστη. Από την στιγμή που ο χρήστης θα πραγματοποιήσει την αγορά, το σύστημα θα του κάνει προτάσεις που να σχετίζονται με τα προϊόντα που έχει αγοράσει στο παρελθόν. Η σύσταση στηρίζεται στην συσχέτιση με βάση το περιεχόμενο αντλώντας πληροφορίες από τα προϊόντα για τα οποία έχει ενδιαφερθεί ο καταναλωτής χωρίς να τον συσχετίζει με τους υπόλοιπους.



Τα συστήματα διαχείρισης θεματικών καταλόγων στο διαδίκτυο, εφαρμόζουν διαδικασίες εξόρυξης γνώσης από τις πλοηγήσεις των χρηστών με σκοπό τη δημιουργία ομάδων χρηστών με κοινά ενδιαφέροντα και κατ’επέκταση την εξατομίκευση των καταλόγων σε επίπεδο ομάδων χρηστών.

Το περιεχόμενο του καταλόγου, η ιεραρχία του, καθώς και το σύνολο των πλοηγήσεων των χρηστών είναι αποθηκευμένο με τη μορφή μίας σχεσιακής βάσης δεδομένων.

Ένας θεματικός κατάλογος μπορεί να παρασταθεί με τη μορφή ενός κατευθυνόμενου γράφου, οπού οι κόμβοι αποτελούν τις κατηγορίες και οι ακμές τις σχέσεις κατηγορία-υποκατηγορία.

Στα συστήματα διαχείρισης θεμταικών καταλόγων ενσωματώνεται η λειτουργία αναζήτησης σε κατηγορίες οι οποίες αποτελούν απογόνους της εκάστοτε κατηγορίας που διαλέγει ο χρήστης να πραγματοποιήσει την αναζήτησή του. Συνεπώς θα πρέπει να υπάρχει ένας αποδοτικός τρόπος να συγκεντρώνεται το σύνολο των υποκατηγοριών μίας κατηγορίας.

η λύση προσφέρει η χρήση ενός Interval Labeling Scheme. Σύμφωνα με αυτό,αρχικά εντοπίζεται το spanning tree του γράφου. Σε κάθε κόμβο του spanning tree (δηλαδή κάθε θεματική ενότητα) δίνεται ένα ζεύγος τιμών [a,b], όπου το b είναι ο αριθμός που δίνεται στον κόμβο κατά την postorder διάσχιση του δέντρου και a το ελάχιστο b που εμφανίζεται στους απογόνους του κόμβου. Στη συνέχεια, η διαδικασία επεκτείνεται σε όλο το γράφο ως εξής: Για κάθε ζευγάρι κόμβων p,q, που

αντιστοιχεί σε σχέση κατηγορία - υποκατηγορία, ο κόμβος p κληρονομεί το ζεύγος τιμών

[a,b] του q, εκτός αν το διάστημα $[a_p,b_p]$ περιλαμβάνει το $[a_q,b_q]$. Ικανή και αναγκαία συνθήκη για να είναι ένας κόμβος q απόγονος ενός κόμβου p, είναι κάποιο από τα διαστήματα [a,b] του κόμβου p να περιλαμβάνει τον postorder αριθμό του q.

Η συνεισφορά της συγκεκριμένης τεχνικής στη λειτουργία της αναζήτησης είναι πολύ σημαντική, καθώς επιτρέπει τη γρήγορη και αποδοτική εύρεση του συνόλου των υποκατηγοριών μίας κατηγορίας (στην οποία εκτελείται αναζήτηση), με σκοπό να περιορίζεται σε αυτό η αναζήτηση κατηγοριών και resources.

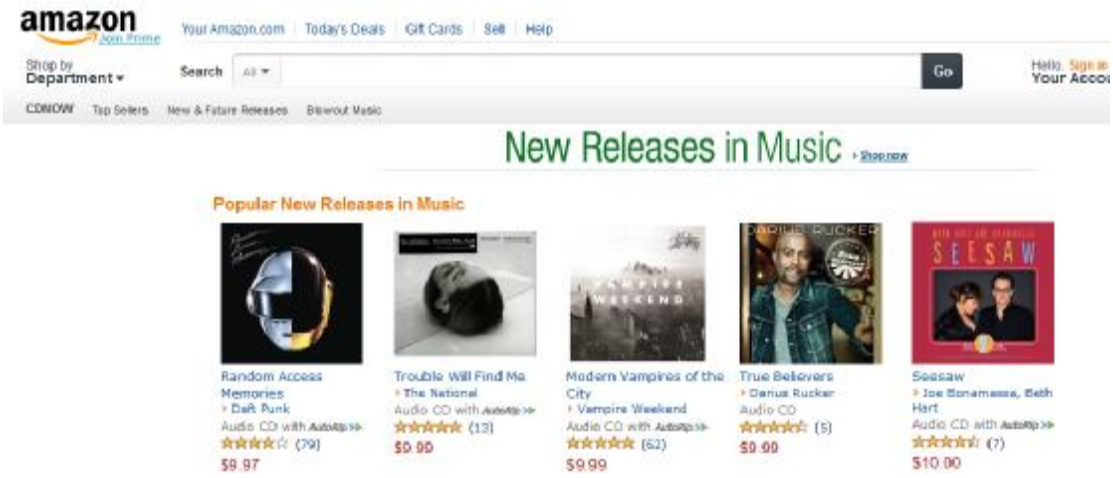
4.10 Amazon

Στο ηλεκτρονικό εμπόριο χρησιμοποιείται κυρίως η ιστορική προσέγγιση, σύμφωνα με την οποία τα συστήματα διατηρούν έναν κατάλογο των αγορασμένων προϊόντων και των αντίστοιχων εκτιμήσεων των χρηστών, ως παραμέτρους χρήστη (προφίλ). Αυτό συμβαίνει στο δημοφιλέστερο σύστημα recommender στο ηλεκτρονικό εμπόριο, το Amazon.com Το Amazon είναι ένα κατάστημα ηλεκτρονικού εμπορίου που αναλαμβάνει την πώληση διάφορων προϊόντων, όλων των κατηγοριών εμπορίου. Η ιστοσελίδα χρησιμοποιεί διάφορες μεθόδους σύστασης που στηρίζονται σε τρεις πηγές για την τελική πρόταση προς τον χρήστη.

- Η πρώτη πηγή είναι η συμπεριφορά του χρήστη, δηλαδή ποια αντικείμενα έχει αγοράσει στο παρελθόν ή πρόσθεσε στο καλάθι αγορών ή πως αξιολόγησε το αντικείμενο μετά από την αγορά του (Rating-commend).

- Η δεύτερη πηγή είναι τα χαρακτηριστικά του αντικειμένου. Όλα τα προϊόντα που εμπορεύεται η σελίδα διαθέτουν λεπτομερή χαρακτηριστικά (λέξεις κλειδιά) όπως: ονομασία, κατασκευαστή, χρονολογία κατασκευής.

- Η τρίτη πηγή που λαμβάνεται υπόψη για την τελική σύσταση είναι οι ομοιότητες ανάμεσα στους χρήστες.



Οι συστάσεις βασίζονται στην αγορά ενός προϊόντος από τον χρήστη. Από την στιγμή που ο χρήστης θα πραγματοποιήσει την αγορά, το σύστημα θα του κάνει προτάσεις που θα εμφανιστούν σε δύο διαφορετικές λίστες. Στην πρώτη λίστα το σύστημα

βρίσκει άλλους χρήστες που έχουν αγοράσει το ίδιο προιόν με τον χρήστη. Στην συνέχεια προτείνει στον χρήστη προϊόντα που δεν έχει αγοράσει αλλά έχουν αγοράσει οι άλλοι χρήστες που έχουν αγοράσει το ίδιο προϊόν. Ουσιαστικά, αυτές οι προτάσεις στηρίζονται στο ότι χρήστες με κοινές αγορές θα έχουν και κοινά ενδιαφέροντα.

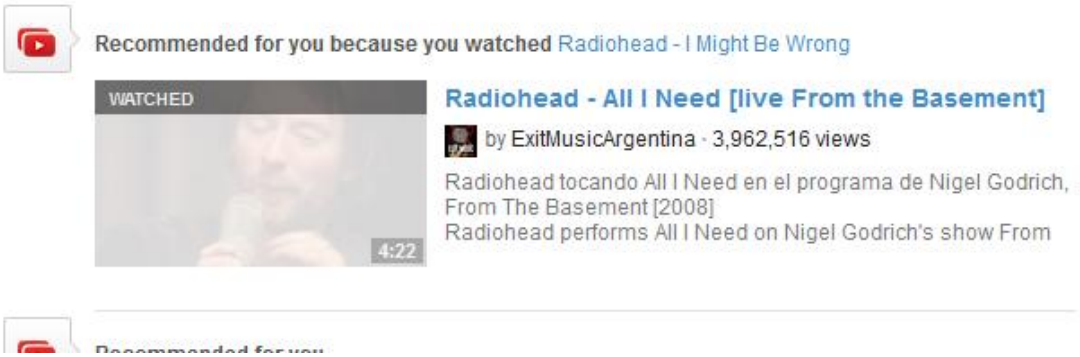
Εικόνα 20 Το συνεργατικό φιλτράρισμα με βάση το προϊόν εφαρμόζεται ευρέως στο Amazon.com

Εικόνα 21 Παράδειγμα εξήγησης σύστασης που χρησιμοποιεί συνεργατικό φιλτράρισμα με βάση το προϊόν

Μια από τις πιο κοινές μεθόδους για να καθορίσει την ομοιότητα είναι ο υπολογισμός γωνίας συνημίτονου. Το σύστημα σύστασης Amazon.com χρησιμοποιεί το μέτρο συνημιτόνου για να αποφασιστεί η ομοιότητα μεταξύ κάθε δύο στοιχείων που αγοράζονται από κάθε πελάτη και για να καθιερωθεί η μήτρα στοιχείων, η οποία περιέχει τις σχέσεις στοιχείο προς στοιχείο.

4.11 YouTube

Πρόσφατα, κάποια από τα δημοφιλέστερα με την μεγαλύτερη συμμετοχή Κοινωνικά Μέσα έχουν προσθέσει στις υπηρεσίες που προσφέρουν και μηχανές συστημάτων προτάσεων, όπως το YouTube που πλέον στην αρχική του σελίδα συμπεριλαμβάνει προτάσεις για video βάσει όσων έχει δει προηγουμένως ο χρήστης και όσων έχει δηλώσει ως αγαπημένα. Η εφαρμογή αυτή είχε ως αποτέλεσμα να αυξηθεί ο αριθμός των χρηστών που επισκέπτονται την αρχική σελίδα, η συχνότητα των επισκέψεων στο YouTube και ο αριθμός των πελατών που εγγράφονται στην συγκεκριμένη πλατφόρμα.



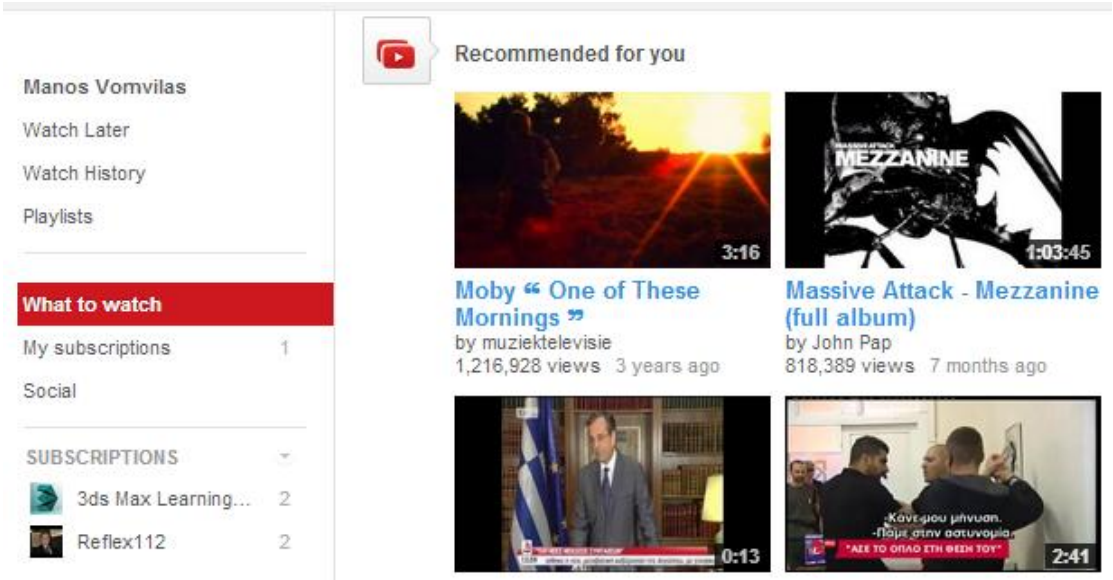
Εικόνα 22 Παράδειγμα εξήγησης σύστασης που χρησιμοποιεί συνεργατικό φιλτράρισμα με βάση το προϊόν

Ορισμένα από τα καλύτερα παραδείγματα συστημάτων προτάσεων όπως αυτό του YouTube εφαρμόζουν ευρέως συνεργατικό φιλτράρισμα με βάση την μνήμη για την εξαγωγή προβλέψεων. Ο βασικός λόγος που συμβαίνει αυτό είναι επειδή το Συνεργατικό Φιλτράρισμα στηρίζεται στους χρήστες, σε παρελθοντικές προτιμήσεις τους και σε κοινά ενδιαφέροντα με άλλους χρήστες και όχι στο περιεχόμενο των προϊόντων, το οποίο αλλάζει δυναμικά και πολύ συχνά δεν είναι αρκετά σαφές ώστε

να οδηγήσει σε ακριβείς προβλέψεις. Επιπλέον, η επιλογή memory-based συστημάτων οφείλεται μεταξύ άλλων στο ότι είναι εύκολα και απλά τόσο στην δημιουργία όσο και στην εφαρμογή τους, καθώς και στο ότι είναι σταθερά και δεν επηρεάζονται σε σημαντικό βαθμό από το συνεχώς μεταβαλλόμενο dataset (προσθήκη νέων χρηστών, προϊόντων ή βαθμολογιών).

Η εφαρμογή αυτή αποτελεί τμήμα της εφαρμογής πλοήγησης και αναλαμβάνει τη δημιουργία και εμφάνιση στο χρήστη συντομεύσεων, που προκύπτουν μετά από σύγκριση της τρέχουσας πλοήγησής του με δημοφιλή αποθηκευμένα τμήματα πλοηγήσεων.

Όσο ο χρήστης πλοηγείται στην πύλη, το σύστημα εξετάζει ένα παράθυρο ν-τελευταιών κατηγοριών που έχει επισκεφθεί (navigation window). Εάν οι κατηγορίες αυτές είναι δημοφιλείς - δηλαδή έχουν ανιχνευθεί δημοφιλή subpatterns μήκους 1 που ταυτίζονται με αυτές-, τότε ελέγχεται κατά πόσο αυτή η ακολουθία κατηγοριών εντοπίζεται σε δημοφιλή subpatterns μήκους ν+1 που είναι αποθηκευμένα στη βάση. Εάν εντοπιστούν τέτοια subpatterns, τότε δημιουργούνται υποψήφιες συντομεύσεις. Αν το confidence της μετάβασης προς την υποψήφια συντόμευση είναι μεγαλύτερο από ένα κάτω όριο, τότε η συντόμευση αυτή εμφανίζεται στο χρήστη. Όμοια και για την περίπτωση των δημοφιλών L-subpatterns, δηλαδή των ακολουθιών κατηγοριών στις οποίες οι χρήστες έχουν επιλέξει τουλάχιστον ένα resource. Το διάγραμμα ροής δεδομένων της εφαρμογής online personalization φαίνεται στην ακόλουθη εικόνα.



Επιχείρηση	Είδος Φιλτραρίσματος	Τομέας Ανάπτυξης	Σύστημα ERP	Χαρακτηρισμός
Εφημερίδα Guardian	Διήθηση με βάση το περιεχόμενο	Social Media	Αρχειοθέτηση άρθρων και αναγνωστών	Λειτουργεί έως σήμερα, χωρίς ιδιαίτερη καινοτομία
Cinematch	Συνδιαστική διήθηση με βάση το προϊόν και το πελάτη	e-commerce	Αρχειοθέτηση Ταινιών, Χρηστών, Συσχέτιση Χρηστών	Καινοτομεί στο χώρο, αλλά χρήζει βελτίωσης
Twitter	Συνδιαστική διήθηση με βάση τον πελάτη	Social Network	Συσχέτιση Χρηστών	Καινοτομεί και αναπτύσσεται
eBay	Διήθηση με βάση το περιεχόμενο	e-commerce	Αρχειοθέτηση προϊόντων και πελατών	Λειτουργεί έως σήμερα
Amazon	Συνδιαστική διήθηση με βάση το προϊόν	e-commerce	Αρχειοθέτηση προϊόντων και πελατών και Συσχέτιση Πελατών	Καινοτομεί και αναπτύσσεται
YouTube	Συνδιαστική διήθηση με βάση το προϊόν	Social Media	Αρχειοθέτηση Δεδομένων , Χρηστών και συσχέτιση δεδομένων	Καινοτομεί και αναπτύσσεται

ΣΥΜΠΕΡΑΣΜΑΤΑ ΜΕΛΕΤΗΣ

Είναι γεγονός η καταλυτική εξέλιξη της τεχνολογίας και συγκεκριμένα του διαδικτύου, η οποία έχει επιφέρει το φαινόμενο της υπερπληροφόρησης. Πολλές φορές, οι χρήστες του διαδικτύου δεν έχουν τη δυνατότητα να παρακολουθήσουν τις εξελίξεις σχετικά με τα αντικείμενα που τους ενδιαφέρουν, καθώς και να οργανώσουν και διαχειριστούν τον τεράστιο όγκο των πληροφοριών. Είναι αναγκαίο να αναφερθεί ότι ο σχεδιασμός των συστημάτων συστάσεων αποσκοπούσε στη διήθηση και την οργάνωση του συνόλου της πληροφορίας, καθώς και στην πιο εύκολη εύρεση χρήσιμης πληροφορίας για τους χρήστες. Στη σύγχρονη εποχή, πολλές γνωστές εφαρμογές υποστηρίζουν τους αλγόριθμους συστάσεων, με χαρακτηριστικό παράδειγμα το youtube.com ή την amazon.com. Η βασική τους θέση και ο κύριος ρόλος τους είναι να προτείνουν στους χρήστες «προϊόντα», αντίστοιχα με τα ενδιαφέροντά τους. Σαφέστατα, υπάρχουν ποικίλοι και διάφοροι αλγόριθμοι συστάσεων. Οι αλγόριθμοι βάσει περιεχομένου (content based) και οι αλγόριθμοι βάσει συνεργατικής διήθησης (user/item based) είναι οι βασικές κατηγορίες στις οποίες διακρίνονται. Όπως είναι φυσικό, η καθεμία από αυτές παρουσιάζει πολυάριθμα πλεονεκτήματα και μειονεκτήματα και θεωρείται κατάλληλη για συγκεκριμένες περιστάσεις.

Οι υπεύθυνοι παροχής ηλεκτρικών υπηρεσιών και οι χρήστες των υπηρεσιών αυτών εξυπηρετούνται από τα Συστήματα Προτάσεων. Είναι πολλά τα οφέλη που υπάρχουν από την αξιοποίηση ενός ανάλογου συστήματος, με κυριότερο την αύξηση των πωλήσεων των προϊόντων τους σε σχέση με τα άτομα που παρέχουν υπηρεσίες και προϊόντα ηλεκτρονικά μέσω του διαδικτύου. Αυτό οφείλεται στο γεγονός ότι οι προτάσεις του συστήματος είναι αντίστοιχες με τις ανάγκες και τα ενδιαφέροντα του χρήστη. Το κύριο χαρακτηριστικό των συστημάτων προτάσεων αποσκοπεί στην εξυπηρέτηση του χρήστη, αφού οι ενδιαφέρουσες συστάσεις συμβάλλουν στη βελτίωση της εμπειρίας του και στην αύξηση της εμπιστοσύνης του στο σύστημα. Συνάμα, ένα Recommender System θεωρείται πετυχημένο, αφού έχει τη δυνατότητα να προτείνει στον χρήστη προϊόντα που θα δυσκολευόταν να βρει από μόνος του, διότι δεν ανήκουν στην λίστα με τα πιο δημοφιλή προϊόντα.

Σε περιπτώσεις κατά τις οποίες ο αριθμός των χρηστών ξεπερνά τον αριθμό των προϊόντων - οι περιπτώσεις αυτές είναι οι πιο δημοφιλείς και συχνές στις διαδικτυακές εφαρμογές συστημάτων προτάσεων - τότε τα συστήματα προτάσεων με βάση τα προϊόντα πιθανόν να έχουν καλύτερες επιδόσεις. Η διαπίστωση αυτή επικεντρώνεται στην πρόβλεψη ότι στο άμεσο μέλλον το Amazon μελλοντικά θα έχει ικανοποιητικότερα αποτελέσματα συστάσεων συγκριτικά με το ebay. Αυτό συμβαίνει επειδή το Amazon εφαρμόζει αλγόριθμο συνεργατικής διήθησης με βάση το προϊόν, ενώ το ebay εφαρμόζει απλή συσχέτιση σύμφωνα με το περιεχόμενο.

Στην περίπτωση κατά την οποία το φιλτράρισμα επιτυγχάνεται με βάση τον χρήστη, οι προβλέψεις παρουσιάζουν μεγαλύτερη ακρίβεια στην περίπτωση που τα αποθηκευμένα προϊόντα είναι κατά πολύ περισσότερα από τους χρήστες.

Η εξαγωγή συστάσεων στα Κοινωνικά Μέσα αποτελεί μια μελλοντική τάση στα Συστήματα Προτάσεων. Κυριαρχεί η άποψη ότι η χρήση δεδομένων που προέρχονται αυτόματα από τα κοινωνικά μέσα, καθώς και η εξαγωγή προτάσεων που βασίζονται σε κοινωνικά χαρακτηριστικά αποτελούν το μέλλον στο συγκεκριμένο επιστημονικό πεδίο και σημειώνουν ήδη ιδιαίτερη προσοχή.

Βιβλιογραφία

Adam F. and O'Doherty P. , Lessons from enterprise resource-planning implementations in Ireland: towards smaller and shorter ERP project [Βιβλίο]. - [s.l.] : Journal of Information Technology V.15, 2000.

Addicted A Κατασκευή Mobile Site για Smart phones [Ηλεκτρονικό] // http://www.addicted.gr/mobile-sites/. - 2012.

Aha D. Bankert R., A Comparative Evaluation of Sequential Feature Selection Algorithms [Βιβλίο]. - [s.l.] : in proceedings of AI & Statistics Workshop, 1995.

Altec Software A Τεχνικά Χαρακτηριστικά Altec Software Atlantis ERP [Βιβλίο]. - Αθήνα : [s.n.], 2008.

Auriol E. Manago M. , Althoff S. Wess S., Dittrich S. , Intergrating Induction and Case Based Reasoning [Βιβλίο]. - France : Springer, 1994.

Burke R. B Hybrid Recommender Systems. Survey and Experiments. User Modeling and User-Adapted Interaction [Βιβλίο]. - 2002.

Deloitte Digital D The Dawn of Mobile Influence [Βιβλίο]. - USA : Delloite Digital, 2012.

Dien D. D E-business development for competitive advantagesQ a case study [Βιβλίο]. - [s.l.] : Information and Management, 2002.

Ellsworth J. & Ellsworth W. E Επιχειρηματικές Εφαρμογές με το Internet [Βιβλίο]. - Αθήνα : Εκδόσεις Γκιούρδας, 1997.

Elmasri & Navathe E Θεμελιώδεις Αρχές Συστημάτων Βάσεων Δεδομένων [Βιβλίο]. - Αθήνα : εκδ. Δίαυλος, 2007.

Forrester F Mobile Commerce Forecast 2011 to 2012 Forrester Report [Βιβλίο]. - USA : [s.n.], 2012.

H. Marmanis D. Babenko , Algorithms of the Intelligent Web, in Artificial Intelligence [Βιβλίο]. - 2008.

Hausman A. H A multi-method investigation og consumer motivations in impulse buying behavior [Βιβλίο]. - [s.l.] : Journal of Consumer Marketing, 2000. - Τόμ. 17.

info.magento.com Magento [Ηλεκτρονικό]. - http://info.magento.com/rs/magentocommerce/images/Magento_Mobile_Datasheet.pdf, 2011.

Malaga M [Βιβλίο]. - 2005.

Manber U. Patel A. , Robinson J. , Experience with Personalization on Yahoo! [Βιβλίο]. - [s.l.] : Communications of the ACM, 2000.

Mobasher B. Cooley R., Srivastava J. Automatic Personalization Based on Web Usage MINING [Βιβλίο]. - [s.l.] : Communications of the ACM, 2000.

Montainer M. Beatriz Lopez, Josef De la Rosa , A taxonomy of Recommender agents on the internet [Βιβλίο]. - [s.l.] : Artificial Intelligence Review, 2003.

Silberschatz & Korth & Sudarshan S Συστήματα Βάσεων Δεδομένων, η Πλήρης Θεωρία των Βάσεων Δεδομένων [Βιβλίο]. - Αθήνα : εκδ. Μ. Γκιούρδας, 2011.

Sven R. Beck M. Freitag B. , Generating Recommendation Dialogues from Product Models [Βιβλίο]. - Passau : University of Passau - Germany, 2003.

Synergic Software S Κατασκευή ιστοσελίδων για smart phones και tablets [Ηλεκτρονικό] // http://www.synergic.gr/blog. - 2012.

T. Wailgum T EPR Definition and Solutions [Ηλεκτρονικό] // www.cio.com. - 18 3 2013.

Turban E., J. Lee, D. King, and H.M. Chung , Electronic commerce: a managerial perspective [Βιβλίο]. - NJ : Prentice Hall Upper Saddle River, 2000.

www.espa.gr [Ηλεκτρονικό]. - 2013.

www.magentocommerce.com [Ηλεκτρονικό]. - 2013.

Z. I. Magabe Z Open Access Technology [Βιβλίο]. - Stockholm : Royal Institute of Technology, 2006.

Α.Σ.Αδριανοπούλου Β. Ασίκη, Ε. Βασιλειάδη, Ι. Μίνη, Γ. Παναγιωτοπούλου, Ι. Παπακυριακοπούλου , Τα Πληροφοριακά Συστήματα Enterprise Resource Planning (ERP) στην Ελληνική Επιχείρηση [Βιβλίο]. - Αθήνα : [s.n.], 2000.

Αθανασάκης Εμ. Α Διείσδυση και Ανάπτυξη του Ηλεκτρονικού Εμπορίου στις Ελληνικές Επιχειρήσεις [Βιβλίο]. - Κρήτη : Τεχνολογικό Εκπαιδευτικό Ίδρυμα Κρήτης, 2012.

Β. Γκιντσιούδης Β Σχεδίαση και Ανάπτυξη Εξατομικευμένης Εφαρμογής Κινητού Ηλεκτρονικού Εμπορίου [Βιβλίο]. - Θεσσαλονίκη : Πανεπιστήμιο Μακεδονίας - Μεταπτυχιακή Εργασία, 2008.

Δ. Παρούτσας Δ Αναζήτηση Πληροφοριών στον Παγκόσμιο Ιστό [Ηλεκτρονικό] // Η Εκπαίδευση στο Δημοτικό Σχολείο. - http://paroutsas.jmc.gr/search.htm, 2013.

Διάλεξη 10η : Κινητό Εμπόριο Δ Ψηφιακό Περιεχόμενο και Ηλεκτρονικό Εμπόριο (Δ' Εξάμηνο) [Βιβλίο]. - Πανεπιστήμιο Αιγαίου : Σχολή Κοινωνικών Επιστημών , 2011.

Ζώτος Ζ Εξατομικευμένη αναζήτηση πληροφορίας με χρήση σημασιολογικών δικτύων [Βιβλίο]. - Πάτρα : Πολυτεχνική Σχολή Πατρών, 2007.

Κάπρος Κ Εξατομικευμένη αναζήτηση πληροφορίας στο διαδίκτυο [Βιβλίο]. - Πάτρα : Πολυτεχνική Σχολή Πατρών , 2006.

Καρακατσούλης Δ. Κ Υλοποιήση Ηλεκτρονικού Καταστήματος YorBooks [Βιβλίο]. - Πάτρα : Πανεπιστήμιο Πατρών Πολυτεχνική Σχολή - Τμήμα Μηχανικών Ηλεκτρονικών Υπολογιστών και Πληροφορικής, 2011.

Κατσίκας Σ. & Μήτρου Λ. Κ Ασφάλεια Πληροφοριακών & Επικοινωνιακών Συστημάτων στο Χώρο του Ηλεκτρονικού Επιχειρείν [Βιβλίο]. - Αθήνα : Ομάδα Εργασίας Β1 του e-businessforum, 2002. - Τόμ. διαθέσιμο στον διαδικτυακό τόπο "www.ebusinessforum.gr".

Κατσουλάκος Γ. Κ Νέα Οικονομία, Διαδίκτυο και Ηλεκτρονικό Εμπόριο [Βιβλίο]. - Αθήνα : Εκδόσεις Κέρκυρα, 2001.

Κολλάρας Κ Ο Σημασιολογικός ιστός [Βιβλίο]. - Πάτρα : Πανεπιστήμιο Πατρών, 2007.

Κουρής Κ Εφαρμογή Τεχνικών Data Mining σε Συστήματα Ηλεκτρονικού Εμπορίου [Βιβλίο]. - Πάτρα : Πολυτεχνική Σχολή Πατρών, 2006.

Λ. Στάμκου Λ Πληροφοριακά Συστήματα στο Λιανικό Εμπόριο [Βιβλίο]. - Θεσσαλονίκη : Μεταπτυχιακό Τμήμα Εφαρμοσμένης Επιχειρηματικής Πληροφορίας, 2011.

Μ. Κωνσταντίνου Μ Στστήματα Συστάσεων σε Ηλεκτρονικά καταστήματα Λιανικής [Βιβλίο]. - Λάρισα : Τμήμα Τεχνολογίας Πληροφορικής και Τηλεπικοινωνιών, 2012.

Μ. Ρήγκου Μ Personalization, Τεχνολογίες & Υπηρεσίες [Βιβλίο]. - Πάτρα : Πανεπιστήμιο Πατρών, 2012.

Μανωλόπουλος & Παπαδόπουλος Μ Συστήματα Βάσεων Δεδομένων – Θεωρία και Πρακτική Εφαρμογή [Βιβλίο]. - Αθήνα : [s.n.], 2009.

Μαρκέλλου Μ Τεχνικές και συστήματα διαχείρισης γνώσης στο διαδίκτυο [Βιβλίο]. - Πάτρα : Πολυτεχνική Σχολή Πατρών, 2005.

Μεττούρης Μ Υλοποίηση εφαρμογής εξόρυξης δεδομένων σε αποτελέσματα εντοπισμού της θέσης κινητού χρήστη και αξιοποίηση της πληροφορίας σε M-commerce εφαρμογές [Βιβλίο]. - Πάτρα : Πανεπιστήμιο Πατρών, 2008.

Ξένος & Χριστοδουλάκης Ξ Βάσεις Δεδομένων [Βιβλίο]. - Πάτρα : ΕΑΠ, 2000.

Οικονόμου & Αργυρόπουλος Ο [Βιβλίο]. - 1995.

Π. Παναγιωτόπουλος Π Εφαρμογή Πολυκριτήριας Μεθοδολογίας AHP για την Επιλογή ERP [Βιβλίο]. - Αθήνα : Εθνικό Μετσόβιο Πολυτεχνείο, 2007.

Παπανικολάου Π Συλλογή, αξιοποίηση και επεξεργασία πληροφοριών που παρέχουν τα κοινωνικά δίκτυα για υποστήριξη εγαρμογών που τρέχουν σε περιβάλλοντα κοινωνικών δικτύων (Facebook) [Βιβλίο]. - Πάτρα : Πανεπιστήμιο Πατρών, 2011.

Πασχόπουλος Α. Π Ηλεκτρονικό Εμπόριο [Βιβλίο]. - Αθήνα : Εκδόσεις Κλειδάριθμος, 2007.

Πασχόπουλος Α. Σκάλτσας Π. , Η λεκτρονικό Εμπόριο [Βιβλίο]. - Αθήνα : Κλειδάριθμος 2η Έκδοση, 2001.

Πλέγας Π Αλγόριθμοι και τεχνικές εξατομικευμένης αναζήτησης σε διαδικτυακά περιβάλλοντα με χρήση υποκείμενων σημασιολογιών [Βιβλίο]. - Πάτρα : Πολυτεχνική Σχολή Πατρών , 2013.

Πλέγας Π Τεχνικές εξατομικευμένης αναζήτησης στον παγίσμιο ιστό [Βιβλίο]. - Πάτρα : Πανεπιστήμιο Πατρών , 2008.

Σάββας Ι. & Μαυρέλλης Ν. Σ Ελληνικά ERP & Εμπορικές - Λογιστικές Εφαρμογές Financial RAM [Βιβλίο]. - Αθήνα : [s.n.], 2005.

Σιωμίκος Γ. Σ Συμπεριφορά Καταναλωτή και Στρατηγικό Μάρκετινγκ [Βιβλίο]. - Αθήνα : Εκδόσεις Σταμούλη, 2002.

Σωχωράκη Ε. Ραφαηλία Β. , Ηλεκτρονικό Εμπόριο μέσω Κινητών Τηλεφώνων, Μελέτη για την Ασφάλεια των Συναλλαγων [Βιβλίο]. - Ηράκλειο : Τεχνολογικό Εκπαιδευτικό Ίδρυμα Κρήτης, 2009.

Ταράτσα Τ Εξόρυξη γνώσης σε κοινωνικά δικτύα [Βιβλίο]. - Αθήνα : Πανεπιστήμιο Πειραιά, 2011.

Τζιάστα Α. Τ Ηλεκτρονικό Εμπόριο και Ηλεκτρονικό Μάρκετινγκ [Βιβλίο]. - Αθήνα : Χαροκόπειο Πανεπιστήμιο - Τμήμα Οικιακής Οικονομίας και Οικολογίας, 2011.

Τσιράκης Τ Αλγόριθμοι και τεχνικές εξόρυξης δεδομένων από ροές δεδομένων στον παγκόσμιο ιστό [Βιβλίο]. - Πάτρα : Πολυτεχνική Σχολή Πατρών , 2006.

Τσόπογλου Σ. Τ Συγκριτική Ανάλυση και Μελέτη ERP Συστημάτων [Βιβλίο]. - Μακεδονία : Μεταπτυχιακό Πρόγραμμα Εφαρμοσμένης Πληροφορικής - Πανεπιστήμιο Μακεδονίας, 2013.

Φ. Μιχελινάκης Φ Μετάδοση σε Πραγματικό Χρόνο Ροών Πολυμέσων Πάνω απο Δίκτυα Ομοτίμων Κόμβων [Βιβλίο]. - Αθήνα : Εθνικό Μετσόβειο Πολυτεχνείο, 2011.

Φαλιάγκα Φ Εξόρυξη γνώσης στον παγκόσμιο ιστό και εφαρμογές σε συτήματα συστάσεων [Βιβλίο]. - Πάτρα : Πολυτεχνική Σχολή Πατρών , 2012.

Φούκη Ι Φ Έρευνα Αγοράς για Κινητές Συσκευές [Βιβλίο]. - Λάρισα : Σχολή Τεχνολογικών Εφαρμογών - Τμήμα Τεχνολογίας Πληροφορικής και Τηλεπικοινωνιών, 2013.

Χ. Χριστάκου Χ Εισηγητικά Συστήματα Βασισμένα σε Μοντελοποιήση Προτιμήσεων Χρήστη και Μεθόδους Διήθησης της Πληροφορίας [Βιβλίο]. - Αθήνα : Εθνικό Μετσόβιο Πολυτεχνείο, 2012.

Χρυσοχού Χ. Χ Ανάλυση και Σχεδιασμός Συστημάτων ERP. Υλοποιήση και Ανάπτυξη σε Εταιρεία Επεξεργασίας Ύδατος [Βιβλίο]. - Αθήνα : Πανεπιστήμιο Πειραιά - Μεταπτυχιακό Πρόγραμμα Σπουδών Οργάνωση και Διοίκηση Βιομηχανικών Συστημάτων, 2008.