Πανεπιστήμιο Πειραιώς – Τμήμα Πληροφορικής
Πρόγραμμα Μεταπτυχιακών Σπουδών
«Προηγμένα Συστήματα Πληροφορικής»

Μεταπτυχιακή Διατριβή

Τίτλος Διατριβής	**Μοντελοποίηση Χρηστών ηλεκτρονικού καταστήματος προϊόντων Πληροφορικής**
Ονοματεπώνυμο Φοιτητή	**Ρενιέρης Λουκάς του Ευαγγέλου**
Αριθμός Μητρώου	**ΜΠΣΠ/07031**
Κατεύθυνση	**Ευφυείς Τεχνολογίες Επικοινωνίας Ανθρώπου-Υπολογιστή**
Επιβλέποντες	**Μαρία Βίρβου, Καθηγήτρια**

Ημερομηνία Παράδοσης **Απρίλιος 2011**

Τριμελής Εξεταστική Επιτροπή

Βίρβου Μαρία Καθηγήτρια	Τσιχριντζής Γεώργιος Καθηγητής	Φούντας Ευάγγελος Καθηγητής

.....……..
Λουκάς Ε. Ρενιέρης
Πτυχιούχος Ο.Π.Α Μηχανικός Πληροφορικής

Περίληψη

Δεν αποτελεί είδηση το γεγονός, οτι πλέον το διαδίκτυο έχει εισβάλλει στην καθημερινότητα σημαντικής μερίδας του παγκόσμιου πληθυσμού, αποτελώντας αναπόσπαστο κομμάτι της. Την προβολή που παρέχει το μέσο, εκμεταλλεύτηκαν στο έπακρο κάθε λογής επιχειρήσεις, ώστε να εκθέσουν και να διακινήσουν τα προϊόντα τους σε αυτό, επεκτείνωντας έτσι την εμβέλεια τους σε παγκόσμιο επίπεδο και διεκδικώντας μεγαλύτερο μερίδιο από την πίτα της αγοράς, δίνοντας επιπλέον την ευκολία σε έναν πελάτη να κάνει αγορές από την άνεση του σπιτιού του. Στις αρχές του αυτό το εγχείρημα αντιμετωπίστηκε με επιφύλαξη και δυσπιστία από το αγοραστικό κοινό, όμως πλέον το ηλεκτρονικό εμπόριο αποτελεί αδιαμφισβήτητα έναν από τους δημοφιλέστερους τρόπους διακίνησης προϊόντων, με το φαινόμενο να παρουσιάζει διαρκώς αυξητικές τάσεις.

Στόχος της παρούσας διπλωματικής διατριβής, είναι η ανάδειξη και χρήση τεχνικών που μπορούν να οδηγήσουν σε βελτιστοποίηση της έκθεσης των προϊόντων στο κοινό, συμβάλλοντας αποτελεσματικά στην αύξηση του τζίρου ενός ηλεκτονικού καταστήματος καθώς και στην διευκόλυνση επιλογής προϊόντων από τον τελικό χρήστη. Σύμμαχός μας σε αυτή την προσπάθεια, θα είναι η θεωρία Μοντελοποίησης Χρηστών, η οποία θα βοηθήσει ώστε να διαχωρίσουμε τους χρήστες σε κατηγορίες ανάλογα με τις προτιμήσεις τους και τις ενέργειές τους, στοχεύοντας πλέον όχι σε ατομικό επίπεδο για την παροχή υπηρεσιών αλλά σε ομαδικό.

Για τον σκοπό αυτό δημιουργήθηκε ένα δοκιμαστικό ηλεκτρονικό κατάστημα προϊόντων πληροφορικής, μέσω του οποίου συλλέχθηκαν όλα τα απαραίτητα δεδομένα από τα οποία θα προσπαθήσουμε να εξηγήσουμε την σημασία της Μοντελοποίησης Χρηστών σε ένα σύγχρονο περιβάλλον ηλεκτρονικού εμπορίου.

Λέξεις Κλειδιά

Μοντελοποίηση Χρηστών, ιεραρχική συσταδοποίηση, δενδρόγραμμα, ομαδοποίηση, ηλεκτρονικό εμπόριο, ηλεκτρονικό κατάστημα, καλάθι αγοράς, προσαρμοστικό λογισμικό, φιλτράρισμα πληροφορίας, collaborative filtering, recommender systems, πίνακας απόστασης, πίνακας ομοιότητας, διάνυσμα γνωρισμάτων, ομοιογένεια δεδομένων, μηχανική μάθηση, αλγόριθμος κ-NN, κανονικοποίηση, μετασχηματισμός δεδομένων, PHP, MySql, Ajax, Wamp.

Abstract

It is common sense that more and more people nowadays use Internet in a daily basis and that it plays an important role in their life. Business all over the world try to take advantage of the fact that Internet has grown so much, by developing e-commerce sites so that they can expose their products at a more global level, trying to increase their profits and offering customers the ability to buy products with a simple click from the safety of their houses. At the start, e-commerce was faced by the customers with distrust, but today is undoubtedly one of the most popular types of commerce, attracting loyal fans day by day.

The target of this thesis is to elect the use of techniques that can contribute to the optimization of promoting products to the people, so that users can find more easily what they search for and for business to grow up. Our ally in this effort is the User Modeling theory through which we will try to cluster users in categories, depending on their moves and preferences, targeting this way to provide services to users as being a group, no individuals.

For this purpose a test-bed e-shop that sells technology products was created, so we could collect all the necessary data needed to interprete the effect of user modeling in a modern environment of e-commerce.

Keywords

User Modeling, hierarchical clustering, dendrogram, clustering, e-commerce, e-shop, shopping cart, adaptive software, information filtering, collaborative filtering, recommender systems, distance matrix, similarity matrix, features vector, data homogeneity, machine learning, k-NN algorithm, normalization, data transformation, PHP, MySql, Ajax, Wamp.

Ευχαριστίες

Σε αυτό το σημείο θα ήθελα να ευχαριστήσω όλο το διδακτικό και διοικητικό προσωπικό του μεταπτυχιακού προγράμματος «Προηγμένα Συστήματα Πληροφορικής» του Πανεπιστήμιου Πειραιά, για την προσπάθεια που κατέβαλλε, ώστε η φοίτησή μου σε αυτό να είναι όσο το δυνατόν καλύτερη και αποδοτικότερη.

Ιδιαίτερες ευχαριστίες, θα ήθελα να απευθύνω στην κα. Βίρβου, πρώτον για την τιμή που μου έκανε να μου εμπιστευτεί αυτή την εργασία και δεύτερον γιατί μέσω της παρακολούθησης των μαθημάτων της, ήρθα σε επαφή με έναν ιδιαιτέρως ενδιαφέροντα τομέα της Πληροφορικής και αποκόμισα αρκετές νέες και πάνω από όλα χρήσιμες γνώσεις.

Δεν θα μπορούσα να παραλείψω να ευχαριστήσω τον πολύ καλό μου φίλο και συνάδελφο Ηρακλή Κοντοδιό, για την άψογη συνεργασία που είχαμε τόσο για την διεκπεραίωση αυτής της εργασίας όσο και για την εν γένει συνεργασία μας εκτός πανεπιστημίου.

Τέλος, οφείλω ένα μεγάλο ευχαριστώ στους γονείς μου για την συμπαράσταση που μου παρείχαν, ώστε να μπορέσω να ανταπεξέλθω και να ολοκληρώσω ένα απαιτητικό πρόγραμμα σπουδών.

Πίνακας Περιεχομένων

1. ΕΙΣΑΓΩΓΗ

Στο κεφάλαιο αυτό επιχειρείται μια αρχική προσέγγιση του αντικειμένου που πραγματεύεται η διπλωματική εργασία, καθώς και τα βήματα που ακολουθήθηκαν προκειμένου να εκπονηθεί και τα οποία οδήγησαν στην δημιουργία ενός ολοκληρωμένου ηλεκτρονικού καταστήματος βασιζόμενου σε αρχές της θεωρίας της Μοντελοποίησης Χρηστών. Επιπρόσθετα ακολουθεί μία ανάλυση της δομής και των κεφαλαίων που απαρτίζουν την εργασία.

1.1 Αντικείμενο Διπλωματικής

Στο πλαίσιο της παγκοσμιοποίησης του εμπορίου μέσω του διαδικτύου και της εξαιρετικής άνθησης που παρουσιάζει το τελευταίο, ολοένα και μεγαλύτερος όγκος πληροφορίας γίνεται διαθέσιμος στον χρήστη-πελάτη.

Καθημερινά όλο και περισσότερες επιχειρήσεις επενδύουν στον μαγικό κόσμο του διαδικτύου, δημιουργώντας ηλεκτρονικά καταστήματα, εκθέτοντας έτσι τα προϊόντα ή τις υπηρεσίες τους σε ένα ευρύτερο κοινό. Για τον απλό χρήστη αυτό σημαίνει ότι από την άνεση του σπιτιού του και με μια πολύ απλή διαδικασία μπορεί να αποκτήσει έως και το πιο απίθανο αγαθό, ακόμα κι αν αυτό βρίσκεται στην άλλη άκρη της γης κι έχοντας να επιλέξει ανάμεσα σε μια τεράστια γκάμα. Για τον ιδιοκτήτη του ηλεκτρονικού καταστήματος είναι ένας δελεαστικός και σχετικά ανέξοδος τρόπος να διεκδικήσει ένα μεγαλύτερο κομμάτι από την πίτα της αγοράς και να διευρύνει το αγοραστικό του κοινό.

Το πρόβλημα που προκύπτει από την υπερβολική έκθεση του χρήστη – ειδικά του μη εξοικειομένου στην πληροφορική - σε αυτόν τον τεράστιο όγκο πληροφορίας, είναι ότι ο χρήστης μπορεί να επέλθει σε σύγχυση και τελικά να αποθαρρυνθεί από το να κάνει χρήση των νέων δυνατοτήτων που του προσφέρονται, καταφεύγοντας στους πατροπαράδοτους και ασφαλείς για τον ίδιο τρόπους αγοράς. Λύση σε αυτό το πρόβλημα μπορεί να αποτελέσει το ‘φιλτράρισμα’ της πληροφορίας που προσφέρεται στον χρήστη έτσι ώστε να προσεγγίσει σε κάποιον βαθμό τις ανάγκες του. Μία επιτυχημένη απόπειρα των παραπάνω, μπορεί να ενισχύσει τους δεσμούς εμπιστοσύνης μεταξύ καταστήματος – πελάτη και να καταστήσει τη σχέση τους μακροχρόνια.

Η παρούσα διπλωματική εργασία πραγματεύεται τεχνικές και μηχανισμούς βασισμένους στην θεωρία της Μοντελοποίησης Χρηστών, ώστε η πληροφορία που ανακτά ο χρήστης, να είναι όσο το δυνατόν πιο κοντινή στις επιθυμίες του. Για να αναδείξουμε τις ανωτέρω τεχνικές, δημιουργήθηκε ένα ολοκληρωμένο και λειτουργικό ηλεκτρονικό κατάστημα, το οποίο εξομειώνει σχεδόν όλες τις βασικές λειτουργίες ενός πραγματικού ηλεκτρονικού καταστήματος.

Μερικές από τις κυριότερες λειτουργίες που αναδεικνύονται μέσω της εφαρμογής, είναι η δημιουργία, συντήρηση και ανανέωση ενός ευέλικτου και δυναμικού μοντέλου χρηστών, που ως στόχο έχει την ένταξη των χρηστών σε έναν αριθμό ομάδων με κοινά χαρακτηριστικά, καθώς και η εξαγωγή συμπερασμάτων για τις σχηματιζόμενες ομάδες. Οι ομάδες είναι δυνατόν να σχηματιστούν μέσω τεσσάρων διαφορετικών μεθόδων ιεραρχικής συσταδοποίησης (hierarchical clustering). Η γνώση που προκύπτει από αυτές τις ομάδες συντελεί στο να έχουμε ένα προσαρμοστικό λογισμικό, του οποίου η διεπαφή προσαρμόζεται ανάλογα με την ομάδα στην οποία ανήκει ο χρήστης. Τέλος με την χρήση τεχνικών όπως ‘cross-sales’ το ηλεκτρονικό κατάστημα έχει τη δυνατότητα να έρθει ακόμα πιο κοντά στις προτιμήσεις του χρήστη.

1.2 Οργάνωση εργασίας

Η παρούσα διατριβή αποτελείται από επτά κεφάλαια. Στα κεφάλαια αυτά παρουσιάζεται αρχικά το θεωρητικό υπόβαθρο και ακολουθεί η εμβάθυνση όσον αφορά τον σχεδιασμό, την υλοποίηση και τη λειτουργία του ηλεκτρονικού καταστήματος.

Το παρών κεφάλαιο αποτελεί μία σύντομη περιγραφή του συστήματός μας και περιγράφει τη δομή του εγγράφου.

Στο δεύτερο κεφάλαιο εκθέτουμε τις υπάρχουσες (state of the art) τεχνολογίες που εφαρμόζονται σήμερα στον τομέα που εξετάζουμε, καθώς επίσης παραθέτουμε μερικές απόψεις για συγγράματα και δημοσιεύσεις που ασχολούνται με το προσκείμενο θέμα.

Στο τρίτο κεφάλαιο γίνεται η ανάλυση και ο σχεδιασμός του ηλεκτρονικου καταστήματος, οπού εξηγούνται μεταξύ άλλων η ανάλυση απαιτήσεων του συστήματος καθώς και παρατίθενται διάφορα διαγράμματα που δημιουργήθηκαν κατά την φάση αυτή.

Στο τέταρτο κεφάλαιο παρουσιάζεται το θεωρητικό υπόβαθρο πάνω στο οποίο στηρίζεται η υλοποίηση του συστήματος.

Στο πέμπτο κεφάλαιο μπορεί ένας μελλοντικός χρήστης του συστήματος να δει μια πλήρη περιγραφή του, και να πληροφορηθεί για όλες τις λειτουργίες του ηλεκτρονικού καταστήματος.

Στο έκτο κεφάλαιο αναλύεται διεξοδικά η κατασκευή του μοντέλου χρηστών το οποίο χρησιμοποιεί η εφαρμογή για να παράγει προτάσεις στους πελάτες και δίδονται λεπτομέρειες για τις μεθόδους ομαδοποίησης που χρησιμοποιήθηκαν.

Τέλος στο έβδομο κεφάλαιο επιχειρείται μία σύνοψη όσων αναπτύχθηκαν στα προηγούμενα κεφάλαια, καθώς και τα συμπεράσματα που εξήχθησαν από αυτήν την διπλωματική εργασία. Επίσης δίνονται μερικές ιδέες για το ποιες πρέπει να είναι οι μελλοντικές κινήσεις που θα καταστήσουν το σύστημα πιο ολοκληρωμένο και λειτουργικό.

2. ΑΝΑΣΚΟΠΗΣΗ ΠΕΔΙΟΥ

Πριν αναπτυχθεί το σύστημα το οποίο πραγματεύεται η παρούσα εργασία, είναι χρήσιμο να δοθούν μερικοί απαραίτητοι ορισμοί, καθώς να γίνει μια ανασκόπηση για τις τάσεις που υπάρχουν στην αγορά τα τελευταία χρόνια σε αυτόν τον τομέα της Πληροφορικής.

Επιγραμματικά θα περιγραφούν κάποια συστήματα που έχουν κερδίσει μεγάλο μερίδιο της αγοράς την δεκαετία που διανύουμε. Επίσης θα γίνει μία σύντομη ανάλυση και αναφορά σε μερικά συγγράματα και δημοσιεύσεις σχετικές με το αντικείμενο της εργασίας.

2.1 Γενικοί ορισμοί

Παρακάτω δίνονται δύο ορισμοί οι οποίοι έχουν άμεση σχέση με το αντικείμενο της παρούσας εργασίας.

2.1.1 Ηλεκτρονικό εμπόριο (E-Commerce)

Ως ηλεκτρονικό εμπόριο ορίζεται το εμπόριο που πραγματοποιείται με ηλεκτρονικά μέσα, αποτελεί δηλαδή μια ολοκληρωμένη συναλλαγή που πραγματοποιείται μέσω διαδικτύου - internet χωρίς να είναι απαραίτητη η φυσική παρουσία των συμβαλλομένων μερών (δηλαδή του πωλητή και του αγοραστή) (Π.Δ. 131/2003).

Ολοένα και αυξανόμενος είναι ο αριθμός των επιχειρήσεων που δραστηριοποιούνται στο ηλεκτρονικό εμπόριο, έτσι ώστε να μεγαλώσουν τον κύκλο εργασιών τους και ταυτόχρονα να προσφέρουν σύγχρονες υπηρεσίες εμπορίου στους πελάτες τους και σε περισσότερο προνομιακές τιμές.

2.1.2 Recommender Systems

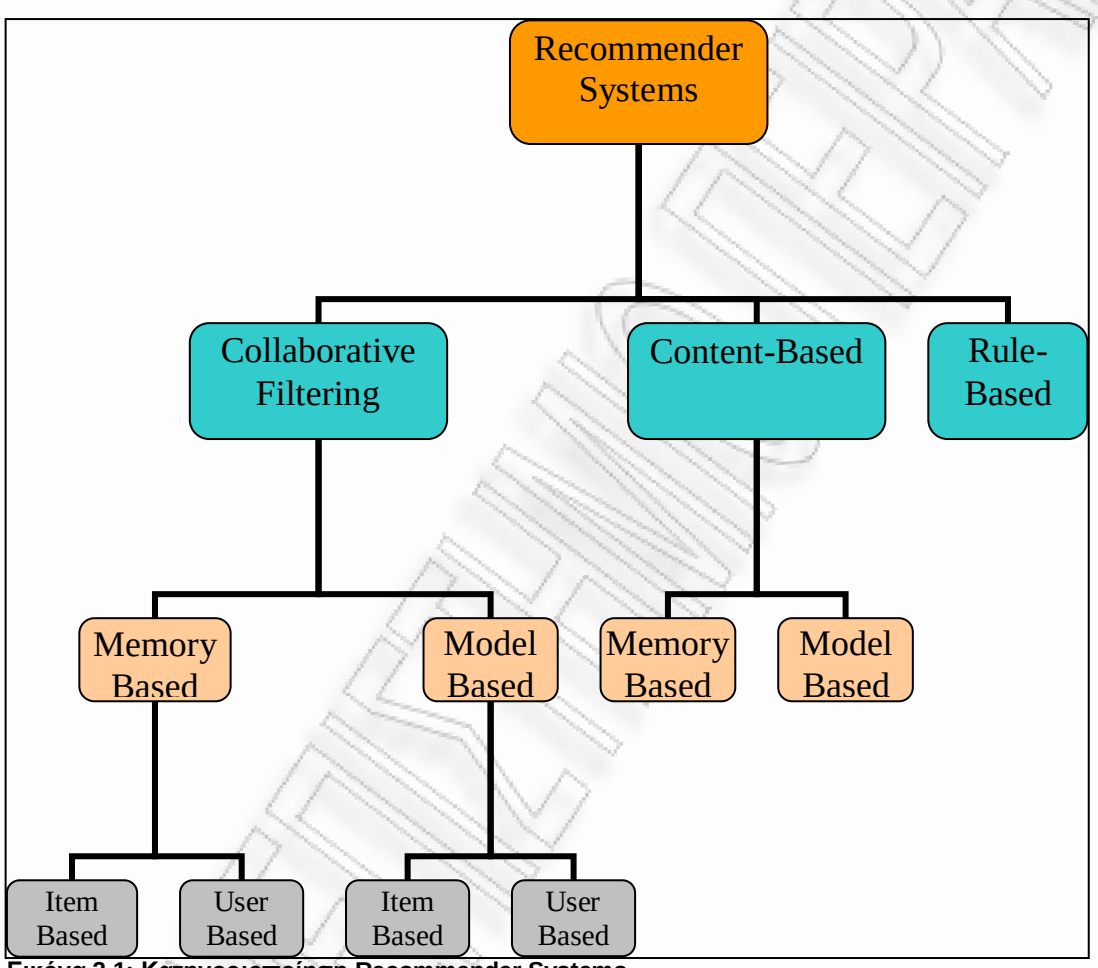
Το ενοποιημένο σύστημα Πληροφορικής το οποίο παράγει προτάσεις για περιεχόμενα, προϊόντα ή υπηρεσίες προς τους πιθανούς ή τους υπάρχοντες πελάτες μιας επιχείρησης τους οποίους πιθανολογείται ότι τους ενδιαφέρουν, περιγράφεται με τον όρο Recommender System.

Δηλαδή recommender system θεωρείται ένα σύστημα το οποίο λειτουργεί βάσει ενός συνόλου κανόνων που έχουν προκύψει από την συσσώρευση εμπειρίας, χωρίς απαραίτητα αυτό να βασίζεται σε όρους και τεχνικές τεχνητής νοημοσύνης όπως είναι π.χ. ο όρος μηχανική

μάθηση (machine learning) ή ευφυές σύστημα (intelligent system). Ήδη υπάρχουν αρκετά συστήματα στην αγορά που ενσωματώνουν ένα recommender system.

2.2 Κατηγοριοποίηση Recommender Systems

Σε αυτό το κεφάλαιο θα εξετάσουμε σε ποιες κατηγορίες διακρίνονται τα Recommender Systems με βάση διάφορες παραμέτρους. Όπως φαίνεται και στην εικόνα 2.1 τα Recommender Systems διαχωρίζονται σε τρείς κύριες κατηγορίες, τα Collaborative Filtering Systems, τα Content Based Systems και τα Rule Based Systems.



Εικόνα 2.1: Κατηγοριοποίηση Recommender Systems

2.2.1 Collaborative Filtering

Τα συστήματα Collaborative Filtering (CF) που ανήκουν στα Recommender Systems (RS) βασίζουν τις αποφάσεις που λαμβάνουν στις απόψεις των χρηστών. Π.χ. σε ένα ηλεκτρονικό κατάστημα το σύστημα καλείται να λάβει αποφάσεις για τις προτάσεις που πρέπει να κάνει στον χρήστη, βασιζόμενο στο πως οι χρήστες έχουν βαθμολογήσει τα προϊόντα. Η ιδέα είναι να μαντέψουμε πως ένας χρήστης θα βαθμολογούσε ένα προϊόν το οποίο δεν γνωρίζει, στηριζόμενοι στο πως άλλοι χρήστες το έχουν βαθμολογήσει. Με βάση αυτό το σκεπτικό μπορούμε π.χ. να υποθέσουμε ότι ο χρήστης θα ψήφιζε παρόμοια με τους ‘ομοιδεάτες’ του και ανόμοια με χρήστες με τους οποίους δεν έχουν κοινά χαρακτηριστικά.

Ο πιο διαδεδομένος αλγόριθμος CF, ο οποίος χρησιμοποιείται σε πολλά RS, είναι αυτός που βασίζεται στον αλγόριθμο k-nearest neighbor (kNN) και αρχικά προτάθηκε από τον Resnick. Ο αλγόριθμος αυτός στις μέρες μας καλείται user-user CF και είναι μια παραλλαγή του

αλγορίθμου kNN, ρυθμισμένου έτσι ώστε να έχει την καλύτερη δυνατή απόδοση. Μια άλλη παραλλαγή του αλγορίθμου kNN πρότεινε και ο Sarwar, ο οποίος βασίζεται στην ομοιότητα μεταξύ των αντικειμένων και όχι των χρηστών και ο οποίος είναι γνωστός ως αλγόριθμος item-item CF.

2.2.2 Model-Based εναντίον Memory-Based

Οι αλγόριθμοι πρόβλεψης σύμφωνα με τον Breese χωρίζονται σε δύο μεγάλες κατηγορίες: Τους αλγόριθμους που είναι memory-based και τους model-based. Οι memory-based αλγόριθμοι διατηρούν μια βάση δεδομένων με όλους τους χρήστες και τις προτιμήσεις τους και για κάθε πρόβλεψη που καλούνται να κάνουν, εκτελούν υπολογισμούς και αντλούν τα δεδομένα τους από ολόκληρη την βάση.

Αντίθετα οι model-based αλγόριθμοι δημιουργούν ένα περιγραφικό μοντέλο για τους χρήστες και τις προτιμήσεις τους και οι προβλέψεις τους στηρίζονται σε αυτό. Τα πλεονεκτήματα των memory-based αλγορίθμων είναι ότι είναι ευκολότεροι στην υλοποίηση και ευκολότεροι στο να δεχτούν νέα δεδομένα. Το μειονέκτημά τους είναι ότι με την αύξηση των χρηστών ή των προϊόντων μπορούν να γίνουν ασύμφοροι υπολογιστικά, δηλαδή η πολυπλοκότητά τους σε χώρο και χρόνο να αυξηθεί εκθετικά, σε αντίθεση με τους model-based αλγόριθμους οι οποίοι έχουν πολύ μικρότερες απαιτήσεις τόσο σε χώρο όσο και σε χρόνο, λόγω του γεγονότος ότι χρησιμοποιούν μόνο ένα υποσύνολο των δεδομένων εισόδου και ότι όλοι οι υπολογισμοί γίνονται offline (σε αντίθεση με τους model-based όπου οι υπολογισμοί πρέπει να γίνονται online). Το ότι χρησιμοποιούν μόνο ένα υποσύνολο των δεδομένων εισόδου μπορεί να είναι πλεονέκτημα γιατί έχουν σταθερή απόδοση ανεξάρτητα από τον συνολικό όγκο δεδομένων και άρα παράγουν προτάσεις πολύ γρήγορα, ωστόσο μπορεί να είναι και μειονέκτημα για την ακρίβεια των αποτελεσμάτων που παράγουν και κατά συνέπεια για οι προτάσεις που κάνουν στον χρήστη να μην είναι αρκετά ποιοτικές. Τέλος ένα μειονέκτημα των memory-based αλγορίθμων είναι η κακή αντιμετώπιση που έχουν σε καταστάσεις cold-start, δηλαδή σε περιπτώσεις νέων χρηστών όπου τα δεδομένα δεν επαρκούν για να προκύψουν ακριβείς προτάσεις για το νέο χρήστη. Οι καταστάσεις αυτές μπορούν να αντιμετωπιστούν με υβριδικά συστήματα τα οποία χρησιμοποιούν μια μίξη τεχνικών Collaborative Filtering και Content Based συστημάτων.

2.2.3 Αλγόριθμοι CF

Ένας κλασσικός αλγόριθμος Collaborative Filtering ακολουθεί την γενική μορφή:

1. Υπολόγισε τις σχέσεις ομοιότητας μεταξύ των χρηστών ή των αντικειμένων και αποθήκευσε τα αποτελέσματα σε ένα πίνακα n x m, όπου n το πλήθος των χρηστών και m το πλήθος των αντικειμένων.

2. Βρες τους ν πιο όμοιους χρήστες ή αντικείμενα και κάνε την αντιστοίχιση μεταξύ των ομοίων χρηστών για τα αντικείμενα που έδειξαν μεγαλύτερη προτίμηση.

3. Προέβλεψε τον βαθμό ενδιαφέροντος ενός χρήστη για ένα αντικείμενο.

4. Πρότεινε στον χρήστη τα top-N αντικείμενα σύμφωνα με τον βαθμό ενδιαφέροντος που υπολόγισες στο βήμα 3.

Κατά καιρούς έχουν προταθεί πολλοί αλγόριθμοι οι οποίοι προσπαθούν να δώσουν λύση στα προβλήματα που αντιμετωπίζουν οι αλγόριθμοι user-user και item-item, ωστόσο όλοι πάσχουν από κάποιες αδυναμίες. Τέτοιοι είναι αλγόριθμοι που βασίζονται σε Bayesian networks, clustering, AIN, SVD κ.α. Οι πιο διαδεδομένοι αλγόριθμοι πάντως στις μέρες μας και αυτοί που χρησιμοποιούνται στις περισσότερες εφαρμογές, είναι οι αλγόριθμοι user-user CF και item-item CF.

Αναμφίβολα, ο αλγόριθμος CF, που χρησιμοποιείται περισσότερο από οποιονδήποτε άλλον ειδικά σε ερευνητικό επίπεδο είναι ο kNN αλγόριθμος. Σε αυτόν, εάν υποθέσουμε ότι εξετάζουμε ένα σύστημα με n χρήστες και m αντικείμενα, οι προτιμήσεις των χρηστών εισάγονται σε έναν πίνακα n x m, στον οποίο η τιμή (i, j) συμολίζει την προτίμηση του χρήστη u_i για το αντικείμενο j ή είναι μηδενική εάν ο χρήστης δεν έχει εκφράσει την προτίμησή του για το

αντικείμενο αυτό. Παρακάτω θα εξετάσουμε μερικές σχέσεις υπολογισμού της ομοιότητας μεταξύ χρηστών και μεταξύ αντικειμένων, καθώς και σχέσεις για τις τιμές βάσει των οποίων γίνεται η πρόβλεψη για την πιθανότητα ένα αντικείμενο να αρέσει σε κάποιον χρήστη και κατά συνέπεια να του προτείνεται.

2.2.4 User-User αλγόριθμος

Ο user-user αλγόριθμος μπορεί να χωριστεί σε δύο στάδια. Στο πρώτο, υπολογίζονται οι ομοιότητες μεταξύ των χρηστών και το μοντέλο που προκύπτει αποθηκεύεται. Οι ομοιότητα μεταξύ των χρηστών μπορεί να μετρηθεί με πάρα πολλούς τρόπους, όμως ο πιο διαδεδομένος, είναι ο μηχανισμός GroupLens του Pearson correlation coefficient. Σύμφωνα με αυτόν η ομοιότητα μεταξύ δύο χρηστών u_i και u_j, υπολογίζεται από την σχέση:

$$W_{ij} = \frac{\sum_{k \in I}(R_{ik} - \overline{R_i})(R_{jk} - \overline{R_j})}{\sqrt{\sum_{k \in I}(R_{ik} - \overline{R_i})^2 \sum_{k \in I}(R_{jk} - \overline{R_j})^2}}$$

Όπου I είναι το σύνολο των αντικειμένων τα οποία έχουν βαθμολογηθεί και από τους δύο χρήστες, R_{ik} είναι η βαθμολογία του χρήστη u_i για το αντικείμενο k και $\overline{R_i}$ είναι η μέση τιμή όλων των βαθμολογιών του χρήστη u_i.

Στο δεύτερο στάδιο του αλγόριθμου γίνεται η πρόβλεψη για το αντικείμενο a και τον χρήστη u_i, ως εξής: Διαλέγουμε τους k κοντινότερους χρήστες οι οποίοι έχουν επίσης βαθμολογήσει το αντικείμενο a και υπολογίζοντας τις σταθμισμένες αποκλίσεις από τους μέσους όρους των χρηστών έχουμε την πρόβλεψη για τον χρήστη u_i. Ο τύπος είναι:

$$P_{ia} = \overline{R_i} + \frac{\sum_{u=1}^{k}(R_{ua} - \overline{R_u})W_{iu}}{\sum_{u=1}^{k} W_{iu}}$$

2.2.5 Item-item αλγόριθμος

Ο αντίστοιχος αλγόριθμος item-item εκτελεί ακριβώς τα ίδια βήματα με τον user-user. Δηλαδή υπολογίζει πρώτα τις ομοιότητες μεταξύ δύο αντικειμένων για όλα τα αντικείμενα. Ένας τρόπος υπολογισμός της ομοιότητας μεταξύ δύο αντικειμένων i και j είναι αυτός που προτείνεται στο [10]:

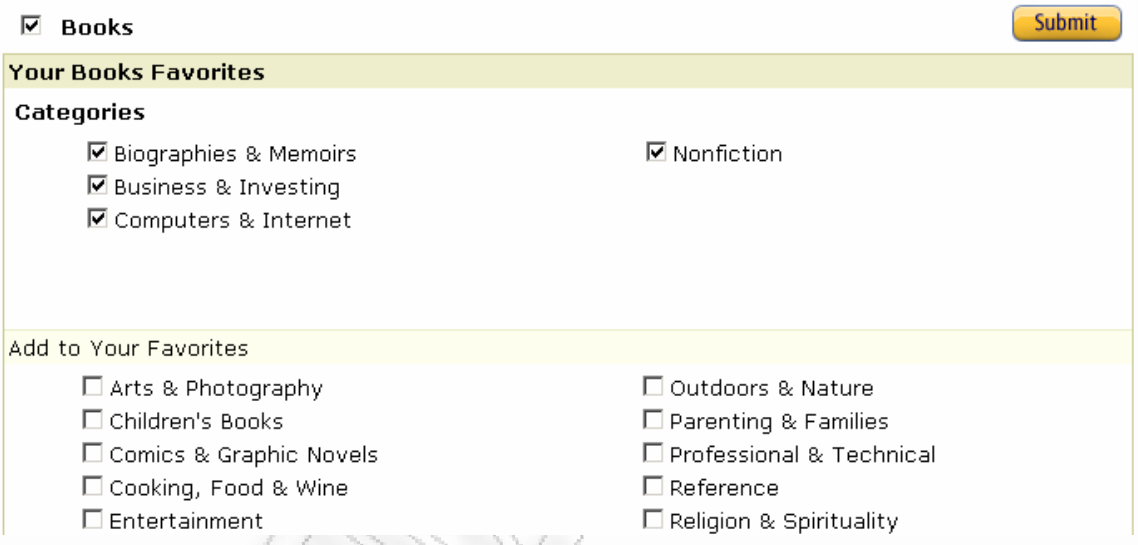
$$S_{i,j} = \frac{\sum_{u \in U}(R_{u,i} - \overline{R_u})(R_{u,j} - \overline{R_u})}{\sqrt{\sum_{u \in U}(R_{u,i} - \overline{R_u})^2 \sum_{u \in U}(R_{u,j} - \overline{R_u})^2}}$$

Ακολούθως η πρόβλεψη για το ζεύγος (χρήστης, αντικείμενο), (u,i) υπολογίζεται από τον τύπο:

$$P_{u,i} = \frac{\sum_{all\_similar\_items,N}(s_{i,N} * R_{u,N})}{\sum_{u \in U}|s_{i,N}|}$$

2.2.6 Content Based RS

Τα συστήματα περιεχομένου (Content based systems), όπως υποδεικνύει και το όνομά τους, βασίζονται στην ανάλυση του περιεχομένου προκειμένου να παράξουν προτάσεις στον χρήστη. Σε ένα τέτοιο σύστημα αναλύονται τα χαρακτηριστικά ενός αντικειμένου και στόχος είναι να υπολογιστεί η ομοιότητα και η συνάφειά του με τα άλλα αντικείμενα. Το πρόβλημα σε αυτά τα συστήματα είναι ότι επειδή δίνουν έμφαση στα αντικείμενα, πρέπει να βρεθούν τρόποι ώστε να γίνει η συσχέτιση τους με τους χρήστες. Έτσι οι χρήστες πρέπει να περιγράψουν με κάποιο τρόπο τις προτιμήσεις τους όσο καλύτερα γίνεται, ώστε να μπορεί να γίνει μια επιτυχημένη συσχέτιση. Συνήθως η συλλογή των προτιμήσεων των χρηστών γίνεται μέσω μιας διεπαφής, η οποία επιτρέπει στον χρήστη να καθορίσει τα ενδιαφέροντά του. Συχνά οι διεπαφές αυτές έχουν πολλά checkboxes που επιτρέπουν στον χρήστη την επιλογή διάφορων τιμών μιας ιδιότητας, π.χ την επιλογή κουζίνας ή αγαπημένο συγγραφέα/καλλιτέχνη, είδος μουσικής/βιβλίων (Εικ 2.2) κτλ. Είναι εμφανές ότι αν οι χρήστες αποτύχουν να αποδώσουν όσο πιο πιστά γίνεται, τις προτιμήσεις τους, τότε το σύστημα θα έχει μεγάλες πιθανότητες αποτυχίας στην πρόβλεψή του και πιθανότατα θα οδηγήσει σε ανεπιθύμητα αποτελέσματα. Για την αντιμετώπιση του προβλήματος έχουν προταθεί διάφορες μεθόδοι ‘profile indexing’, μεταξύ των οποίων οι [32] ‘inversion-based’ μέθοδοι και οι [31] ‘signature-based’.



Εικόνα 2.2: Παράδειγμα φόρμας συλλογής προτιμήσεων χρήστη στο Amazon.com όπου ένας χρήστης μπορεί να επιλέξει τις αγαπημένες του κατηγορίες βιβλίων.

Κάθε Content Based σύστημα, είναι προσανατολισμένο σε ένα συγκεκριμένο πεδίο και δεν μπορεί να είναι γενικού σκοπού. Η διαδικασία εκπαίδευσης του συστήματος γίνεται συνήθως με αλγόριθμους machine learning ή νευρωνικών δικτύων και επειδή είναι εξαιρετικά χρονοβόρα γίνεται offline. Τέλος το γεγονός ότι τα Content based συστήματα βασίζονται κυρίως στα αντικείμενα, τα καθιστά ως την ιδανική λύση για cold-start καταστάσεις, αφού το σύστημα παράγει προτάσεις ανεξάρτητα από το αν ένας χρήστης έχει εκδηλώσει προτιμήσεις.

2.2.7 Rule Based RS

Τα Rule Based συστήματα είναι τα λιγότερο δημοφιλή RS σε σχέση με τα προηγούμενα δύο. Συνήθως η χρήση τους είναι συμπληρωματική ενός CF ή ενός CB συστήματος. Ένα Rule based σύστημα αποτελείται από κανόνες για να παράγει προτάσεις βασισμένες στο ιστορικό των χρηστών. Π.χ αν ένας πελάτης ενός ηλεκτρονικού καταστήματος που πουλάει μουσική έχει αγοράσει ένα δίσκο από έναν συγκεκριμένο καλλιτέχνη, τότε μπορεί να υπάρχει ένας κανόνας τέτοιος ώστε αν ο καλλιτέχνης βγάλει νέο δίσκο να τον προτείνει στον πελάτη. Άλλος κανόνας μπορεί να είναι η πρόταση ενός βιβλίου ή ταινίας που είναι μέρος μιας σειράς (sequence) και της οποίας ο πελάτης έχει αγοράσει ένα ή περισσότερα προηγούμενα σκέλη.

Κυριότερο πλεονέκτημα των Rule based συστημάτων είναι η μηδενική πολυπλοκότητα που έχουν σε σύλληψη και υλοποίηση, καθώς αρκεί η θέσπιση κανόνων ωστέ να είναι δυνατή μια λογική συσχέτιση μεταξύ αντικειμένων. Επίσης πλεονέκτημα θεωρείται ότι μπορούν να δράσουν σαν συμπλήρωμα σε ένα CF σύστημα π.χ. για την αντιμετώπιση cold-start καταστάσεων. Το μεγαλύτερο μειονέκτημά τους είναι δεν μπορούν να προσφέρουν προσωποποιημένες προτάσεις στον ίδιο βαθμό με τα άλλα RS.

2.2.8 Υβριδικά Συστήματα

Υπάρχουν περιπτώσεις όπου ένα recommender system από τα παραπάνω δεν επαρκεί για να καλύψει όλες τις ανάγκες μιας επιχείρησης. Π.χ. ενώ ένα collaborative filtering σύστημα μπορεί να καλύπτει μεγάλο μέρος των αναγκών, ίσως κρίνεται απαραίτητη η προσθήκη τεχνικών content based ή rule based συστημάτων, ώστε για παράδειγμα να αντιμετωπιστεί το φαινόμενο cold-start. Σε πραγματικές εφαρμογές είναι πολύ συχνό το φαινόμενο ανάμιξης των διαφόρων συστημάτων, έτσι ώστε να προκύπτει ένα υβριδικό σύστημα πλήρως προσαρμοσμένο στις ανάγκες μιας επιχείρησης. Γενικά δεν υπάρχουν κανόνες που να καθορίζουν σαφώς πως θα χρησιμοποιηθούν ταυτόχρονα οι διάφορες τεχνικές. Είναι στην ευχέρια των αναλυτών-σχεδιαστών του συστήματος να εντοπίσουν τις απαιτήσεις του συστήματος και να χρησιμοποιήσουν όσα διαθέσιμα εργαλεία έχουν στην διαθεσή τους ώστε να φτάσουν στο επιθυμητό αποτέλεσμα.

2.3 Υπάρχουσα κατάσταση (state of the art)

Σε αυτό το σημείο θα παρουσιάσουμε μερικά διάσημα καταστήματα ηλεκτρονικού εμπορίου, τα οποία είτε στο πολύ πρόσφατο παρελθόν είτε τώρα, εφαρμόζουν διάφορες παραλλαγές της τεχνολογίας recommender systems. Για καθέ ένα από αυτά δίνεται μία σύντομη περιγραφή του τύπου της τεχνολογίας που χρησιμοποιεί. Λόγω της ταχύτητας με τις οποίες αλλάζουν αυτά τα συστήματα, ενδέχεται μερικά από αυτά να έχουν αλλάξει ή να έχουν καταργηθεί.

2.3.1 Amazon

Το Amazon.com είναι ίσως το δημοφιλέστερο ηλεκτρονικό κατάστημα στο διαδίκτυο. Στην αρχή της λειτουργίας του ο σκοπός του ήταν η πώληση βιβλίων, όμως με το πέρασμα του χρόνου γιγαντώθηκε και πλέον πουλάει σχεδόν οτιδήποτε. Δίνοντας έμφαση στον τομέα των πωλήσεων βιβλίων το Amazon™ χρησιμοποιεί μερικά από τα ακόλουθα recommender systems.

Πελάτες που αγόρασαν (Customers who bought): Όπως πολλά ηλεκτρονικά καταστήματα έτσι και το Amazon διαθέτει μια πληροφοριακή σελίδα για κάθε βιβλίο, δίνοντας ορισμένες σημαντικές πληροφορίες για αυτό. Το χαρακτηριστικό γνώρισμα ‘customers who bought’ βρίσκεται στην πληροφοριακή σελίδα κάθε βιβλίου και στην ουσία προσφέρει δύο ειδών προτάσεις στους πελάτες. Η πρώτη κατηγορία προτάσεων έχει να κάνει με βιβλία που αγοράστηκαν συχνότερα από πελάτες που αγόρασαν το βιβλίο το οποίο αγόρασε ο χρήστης. Η δεύτερη κατηγορία προτάσεων αφορά συγγραφείς των οποίων τα βιβλία αγοράστηκαν συχνότερα απο πελάτες οι οποίοι αγόρασαν βιβλία γραμμένα από τον συγγραφέα του συγκεκριμένου βιβλίου.

Eyes (μάτια): Το συγκεκριμένο χαρακτηριστικό επιτρέπει στους πελάτες να ειδοποιούνται μέσω e-mail για νέες αφίξεις προϊόντων στον διάθεση του Amazon. Οι πελάτες μπορούν να προσδιορίσουν το ενδιαφέρον τους με βάσει διάφορα κριτήρια, όπως π.χ. συγγραφέα, τίτλο, ημερομηνία έκδοσης και άλλα απλά ή σύνθετα κριτήρια. Δίνεται επίσης η δυνατότητα στους χρήστες να δηλώσουν τις προτιμήσεις τους κατευθείαν από οποιαδήποτε οθόνη αποτελεσμάτων αναζήτησης, δημιουργώντας έτσι μόνιμα κριτήρια ενδιαφέροντος.

Book Matcher: Το χαρακτηριστικό ‘Book Matcher’ δίνει τη δυνατότητα στους πελάτες να βαθμολογήσουν ένα βιβλίο το οποίο έχουν διαβάσει. Έτσι οι πελάτες μπορούν να αξιολογήσουν ένα βιβλίο σε μια κλίμακα πέντε σημείων, με χαμηλότερη βαθμολογία την ετικέτα ‘το μισώ’ και υψηλότερη το ‘το λατρεύω’. Οι χρήστες οι οποίο έχουν βαθμολογήσει έναν ικανό αριθμό βιβλίων

(συνθήκη ενεργοποίησης), μπορούν να απαιτήσουν να λάβουν προτάσεις για βιβλία τα οποία πιθανώς τους ενδιαφέρουν.

Σχόλια πελατών (Customer comments): Η υπηρεσία αυτή προσφέρει στους πελάτες την δυνατότητα να λάβουν προτάσεις για ένα βιβλίο, βασισμένες στις απόψεις άλλων χρηστών για το βιβλίο αυτό. Στην πληροφοριακή σελίδα κάθε βιβλίου υπάρχει μια λίστα με βαθμολογήσεις άλλων χρηστών καθώς και τις απόψεις τους για το υπό αγορά βιβλίο.

2.3.2 eBay

FeedBack Profile: Η γνωστή πλατφόρμα αγοραπωλησιών eBay™ διαθέτει την λειτουργία 'Feedback profile' η οποία επιτρέπει σε κάθε χρήστη του eBay.com είτε αυτός είναι αγοραστής είτε πωλητής να αξιολογήσει άλλους χρήστες με τους οποίους έχει συνεργαστεί, συμβάλλοντας με αυτόν τον τρόπο στην δημιουργία μιας πιο ολοκληρωμένης εικόνας του προφίλ του χρήστη. Αυτού του είδους η αξιολόγηση ενός χρήστη αποτελείται από μια βαθμολόγηση ικανοποίησης με επιλογή ενός εκ των τριών διαθέσιμων επιλογών (ικανοποιημένος, ουδέτερος, μη ικανοποιημένος) και συνοδεύεται από ένα συγκεκριμένο σχόλιο για τον άλλο πελάτη. Το χαρακτηριστικό αυτό αποσκοπεί σε ένα recommender system για τους επίδοξους αγοραστές οι οποίοι θα μπορούν να δουν τα προφίλ των πωλητών. Το προφίλ αυτό αποτελείται από ένα πίνακα αριθμών με βαθμολογήσεις για διάφορες χρονικές περιόδους (τελευταία εβδομάδα, τελευταίο μήνα, τελευταίο εξάμηνο), καθώς και μια συνολική αποτίμηση για κάθε χρήστη (π.χ. 394 θετικές ψήφοι από 685 μοναδικούς πελάτες).

2.3.3 CDNOW

My CDNOW: Η πλατφόρμα πώλησης μουσικής CDNOW™ επιτρέπει στους πελάτες της τη δημιουργία μια βιβλιοθήκης, η οποία βασίζεται σε αγαπημένους τους δίσκους και καλλιτέχνες. Οι πελάτες καταθέτουν τις προτιμήσεις τους για το ποιοι δίσκοι τους αρέσουν και για το ποιοι είναι οι αγαπημένοι τους καλλιτέχνες. Μετά από μία αγορά ενός δίσκου, ο δίσκος αυτός προστίθεται αυτόματα στην βιβλιοθήκη σε μια κατηγορία με την ετικέτα 'own it'. Αν και η λίστα των αγορασμένων είναι μια σχετικά ισχυρή ένδειξη για τις θετικές προτιμήσεις του χρήστη, ωστόσο οι χρήστες έχουν την δυνατότητα να διασαφηνίσουν τις προτιμήσεις τους βάζοντας δύο ετικέτες 'own it and like it' ή 'own it but dislike it'. Όταν ένας χρήστης ζητήσει προτάσεις, το σύστημα 'προβλέπει' έξι άλμπουμ τα οποία ενδέχεται να αρέσουν στον χρήστη, βασισμένο στις αγορές που έχει κάνει. Σε αυτή την λίστα προτάσεων, ο χρήστης έχει την δυνατότητα να αξιολογήσει κάθε δίσκο που περιέχει επιλέγοντας ένα από τα τρία σχόλια 'own-it', move to wish list' ή 'not for me'. Αν ο χρήστης επιλέξει μια από τις παραπάνω επιλογές, τότε η λίστα ανανεώνεται αυτόματα με βάσει τα χαρακτηριστικά της επιλογής αυτής.

Album advisor: Το χαρακτηριστικό αυτό παρέχει δύο λειτουργίες. Στην πρώτη (single album mode) όταν ένας χρήστης μεταβεί στην πληροφοριακή σελίδα ενός δίσκου, το σύστημα προτείνει δέκα δίσκους σχετικούς με αυτόν που ενδιαφέρεται ο χρήστης. Στην δεύτερη λειτουργία (multiple artist mode) οι χρήστες καλούνται να εισάγουν μέχρι τρεις καλλιτέχνες που τους αρέσουν και το σύστημα προτείνει δέκα δίσκους οι οποίοι σχετίζονται με τους καλλιτέχνες αυτούς.

2.3.4 Levis

Style Finder: Η δυνατότητα Style Finder επιτρέπει στους πελάτες να λαμβάνουν προτάσεις σε άρθρα που αφορούν τα προϊόντα της εταιρίας.Οι πελάτες συμπληρώνουν μια φόρμα με στοιχεία που αφορούν το φύλο τους, την ηλικία τους κτλ. Έπειτα τους δίνονται τρεις κατηγορίες 'μουσική', 'στυλ' και 'διασκέδαση'και καλούνται να βαθμολογήσουν τουλάχιστον τέσσερις υποκατηγορίες από την κάθε κατηγορία. Η βαθμολόγηση γίνεται σε μία κλίμακα επτά σημείων με χαμηλότερη βαθμολογία το 'leave it' και υψηλότερη το 'love it'. Αφού οι χρήστες ολοκληρώσουν την παραπάνω διαδικασία μπορούν να επιλέξουν τη λειτουργία 'get recommendations'. Το σύστημα επιστρέφει έξι μικρογραφίες προϊόντων που είναι οι προτάσεις

ρουχισμού για τους χρήστες βασισμένες στο προφίλ που έχτισε ο χρήστης με την παραπάνω διαδικασία. Φυσικά οι χρήστες μπορούν να αξιολογήσουν καθεμία από αυτές τις προτάσεις και αναλόγως το σύστημα μπορεί να αλλάξει από μία έως και τις έξι προτάσεις του.

2.3.5 Moviefinder

Match Maker: Η λειτουργία Match maker επιτρέπει στους πελάτες του Moviefinder.com να εντοπίσουν ταινίες οι οποίες έχουν παρόμοια 'διάθεση, θέμα, κατηγορία ή ηθοποιούς' με μία ταινία που οι ίδιοι θα επιλέξουν. Με την χρήση της υπηρεσίας οι πελάτες λαμβάνουν μια λίστα με προτάσεις για ταινίες που ταιριάζουν στα κριτήρια που προσδιόρησαν παραπάνω, καθώς και για ταινίες στις οποίες σκηνοθέτης ή ηθοποιοί συμμετείχαν στην παραπάνω ταινία.

We Predict: Η υπηρεσία αυτή προτείνει ταινίες στους πελάτες με βάση ταινίες που έχουν δει και βαθμολογήσει στο παρελθόν. Οι πελάτες μπορούν να βαθμολογήσουν μία ταινία την οποία έχουν δει σε μία κλίμακα πέντε σημείων, από το A έως το F. Η βαθμολογίες αυτές χρησιμοποιούνται με δύο τρόπους. Καθώς οι χρήστες περιηγούνται στο site, κάθε ταινία (που δεν έχει βαθμολογηθεί από τον χρήστη) έχει μια ετικέτα με μία πρόβλεψη για τον συγκεκριμένο χρήστη (ή go see it ή forget it). Επίσης όταν οι χρήστες κάνουν μια αναζήτηση για ταινίες τα αποτελέσματα μπορούν να είναι ταξινομημένα με βάση την πρόβλεψη που έχει κάνει το σύστημα για τον εκάστοτε χρήστη.

2.3.6 Vogoo PHP Lib

Η 'Vogoo PHP Lib' είναι μια βιβλιοθήκη ελεύθερη για χρήση (open source), η οποία προσφέρει ένα API στον προγραμματιστή της PHP, ώστε να εφαρμόσει τεχνικές Collaborative Filtering σε οποιοδήποτε website. Μερικά από τα κύρια χαρακτηριστικά που προσφέρει η βιβλιοθήκη είναι:

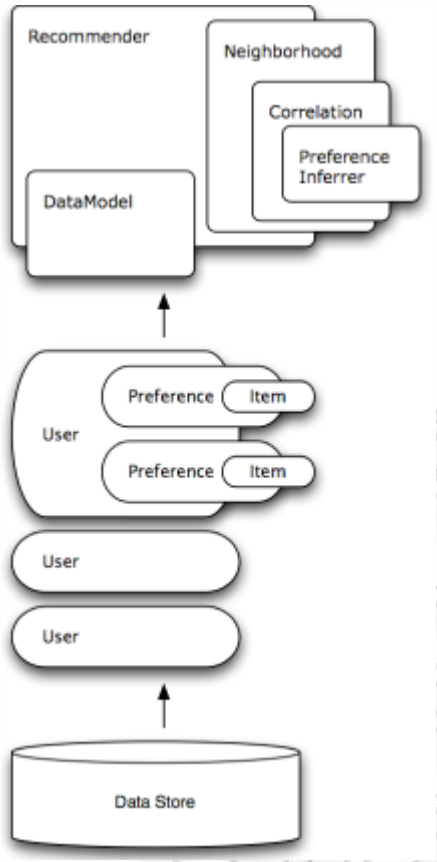
- Μηχανισμό βασισμένο σε CF που προβλέπει την ψήφο του πελάτη.
- Αυτόματοι υπολογισμοί πρόθεσης ψήφου με βάση τα click και τις αγορές.
- Προτάσεις βασισμένες σε χρήστες και όχι σε αντικείμενα.
- Μηχανισμό πραγματικού χρόνου για item-based CF με το χαρακτηριστικό 'σε αυτούς που άρεσε το x αντικείμενο άρεσε και το y'.
- Στοχευμένες διαφημίσεις και μηνύματα σε πραγματικό χρόνο.
- Χαρακτηριστικά γνωρίσματα και υπηρεσίες CF για χρήστες που δεν είναι μέλη.
- Δυνατότητα στα μέλη να προσδιορίζουν σε πιο ποσοστό είναι ευχαριστημένοι από τα αντικείμενα που τους προτείνονται.
- Γρήγορο υπολογισμό ομοιοτήτων μεταξύ μελών.
- Δυνατότητα περιορισμού των αντικειμένων που επιστρέφουν οι item-based συναρτήσεις.
- Χειρισμός πολλαπλών κατηγοριών αντικειμένων.

2.3.7 Cofi

Η Cofi είναι μία βιβλιοθήκη γραμμένη σε Java, ελεύθερη για χρήση η οποία προσφέρει δυνατότητες για Collaborative Filtering. Εμπνευστής αυτής της βιβλιοθήκης είναι ο Daniel Lemire, ο οποίος υποστηρίζει ότι δημιούργησε αυτή τη βιβλιοθήκη για να βοηθήσει τους προγραμματιστές ώστε να μην χρειάζεται να διαβάσουν τόση βιβλιογραφία για να ενσωματώσουν έναν αλγόριθμο CF σε ένα site. Έπίσης ο Lemire πρότεινε τρία νέα σχήματα CF [38], το Slope One Scheme, το Weighted Slope One Scheme και το BI-POLAR Slope One Scheme.

2.3.8 Mahout (Taste 2)

Το 'Taste 2' πρόκειται για την εξέλιξη της open source Java βιβλιοθήκης 'Taste' η οποία είναι πλέον ενσωματωμένη στο σύστημα Mahout του Apache. Η βιβλιοθήκη αυτή είναι μια μηχανή collaborative filtering η οποία δέχεται ως είσοδο τις προτιμήσεις των χρηστών για κάποια προϊόντα και επιστρέφει τις προβλέψεις για τις προτιμήσεις τους, για άλλα προϊόντα. Προσφέρει ένα πλούσιο σύνολο από εξαρτήματα, από τα οποία μπορεί να κατασκευαστεί ένα πλήρως παραμετροποιημένο και προσαρμοσμένο στις ανάγκες μας, recommender system. Επίσης εκτός από τη Java, το Mahout μπορεί να χρησιμοποιηθεί και μέσω web services. Μερικές από τις αφηρημένες κλάσεις που προσφέρει το Mahout είναι η DataModel, UserSimilarity, ItemSimilarity, UserNeighborhood και η Recommender μερικές από τις οποίες μπορούμε να διακρίνουμε και στην απεικόνιση της αρχιτεκτονικής του συστήματος (Εικ 2.3). Τέλος το Mahout, υποστηρίζει memory-based, item-based και slope-one recommenders, ενώ προς το παρόν δεν υποστηρίζει model recommenders.



Εικόνα 2.3: Διάγραμμα που απεικονίζει τις σχέσεις μεταξύ διαφόρων υποσυστημάτων του Mahout για έναν user-based recommender

2.3.9 Netflix

Το Netflix είναι ένα ηλεκτρονικό κατάστημα ενοικίασης ταινιών. Το Netflix χρησιμοποιεί ένα από τα πιο επιτυχημένα collaborative filtering systems, ώστε να προτείνει στους χρήστες του ταινίες. Η επιτυχία του οφείλεται στο οτι το σύστημα δεν είναι προσανατολισμένο στο να προτείνει μόνο νέες ταινίες που είναι στη μόδα αλλά και παλαιότερες, έτσι ώστε να εκμεταλλεύεται όλο το φάσμα της τεράστιας βιβλιοθήκης ταινιών που διαθέτει. Σύμφωνα με την εταιρία, τα μέλη εμπιστεύονται τις προτάσεις που τους κάνει το σύστημα σε ποσοστό κοντά στο 60%.

Το σύστημα που παράγει αυτές τις προτάσεις ονομάζεται CineMatch. Πρόκειται για μια βάση δεδομένων η οποία χρησιμοποιεί πληροφορίες από τρεις πηγές, για να διαπιστώσει ποιες ταινίες ενδέχεται να αρέσουν στους χρήστες, οι οποίες είναι οι εξής:

- Οι ταινίες, οι οποίες είναι οργανωμένες σε ομάδες.
- Οι αξιολογήσεις (ratings) των χρηστών, οι νοικιασμένες ταινίες τρέχουσες και παλιές.
- Οι μέσοι όροι βαθμολογιών όλων των χρηστών

Το σύστημα έχει τη δυνατότητα να κάνει χιλιάδες προτάσεις το δευτερόλεπτο ή δισεκατομμύρια τη μέρα. Ανοίγοντας έναν λογαριασμό στο Netflix, ο χρήστης εισάγει μερικές προτιμήσεις. Κάθε ταινία που νοικιάζει έχει δικαίωμα να την βαθμολογήσει από 1 μέχρι 5 αστέρια. Όσο περισσότερες ταινίες βαθμολογήσει ο χρήστης τόσο ακριβέστερες είναι οι προτάσεις που του γίνονται. Κατά μέσο όρο οι χρήστες το Netflix βαθμολογούν περίπου 200 ταινίες οι οποίες συγκρίνονται με ένα σύνολο βαθμολογήσεων γύρω στα 2 δισεκατομμύρια. Σύμφωνα πάντα με το Netflix, οι προβλέψεις που παράγει το σύστημα έχουν ακρίβεια μισού αστεριού κατά 75% των φορών και το 50% των χρηστών οι οποίοι ενοικιάζουν προτεινόμενες ταινίες τις βαθμολογούν με 5 αστέρια. Η στατιστική μέθοδος που χρησιμοποιείται είναι multivariate regression και το σύστημα ενημερώνεται πολύ συχνά με τα νέα δεδομένα.

Παρόλο που το σύστημα κρίνεται αρκετά επιτυχημένο η Netflix τον Οκτώβριο του 2006 προκύρηξε ένα διαγωνισμό πενταετούς διάρκειας, τον λεγόμενο 'The Netflix prize' με σκοπό να βελτιωθεί ακόμα περισσότερο η ακρίβεια του συστήματος. Για τον σκοπό αυτό έχει δώσει ένα dataset 100 εκατομμυρίων εγγραφών που περιέχουν βαθμολογήσεις χρηστών για μια περίοδο 7 χρόνων και προσκαλεί οποιόνδήποτε ενδιαφερόμενο να τα χρησιμοποιήσει ώστε να ανακαλύψει έναν πιο ακριβή αλγόριθμο. Στόχος είναι η βελτίωση της ακρίβειας του αλγορίθμου κατά 10%, με έπαθλο 1 εκατομμύριο δολλάρια. Άποψη της εταιρίας ήταν ότι τα έξοδα του επάθλου θα υπερκαλυφθούν από την αύξηση των εσόδων η οποία θα υπάρξει σε ενδεχόμενη αύξηση της ακρίβειας του αλγορίθμου. Τελικά στις 21 Σεπτεμβρίου 2009 το έπαθλο απονεμήθηκε σε μία ομάδα με όνομα 'BellKor's Pragmatic Chaos'.

2.4 Ερευνητικό έργο

Εκτός από τα παραπάνω εμπορικά συστήματα, έχουν αναπτυχθεί και εξεταστεί πολλά ερευνητικά συστήματα τα οποία ασχολούνται με το αντικείμενο του φιλτραρίσματος πληροφοριών.

Ένα δοκιμαστικό σύστημα (test-bed) το οποίο δημιουργήθηκε με σκοπό την ομαδοποίηση χρηστών ενός ηλεκτρονικού video club είναι το σύστημα Vision.com [1]. Το σύστημα αυτό στάθηκε η αφορμή ώστε να προσανατολιστεί η παρούσα εργασία σε αυτόν τον τομέα έρευνας. Στόχος του συστήματος, είναι η σύγκριση τεσσάρων αλγορίθμων ομαδοποίησης (hierarchica, fuzzy c-means, spectral and artificial immune network (AIN)-based clustering).

Για τη συλλογή των δεδομένων χρησιμοποιήθηκαν τα προφίλ 150 χρηστών, δηλαδή 150 διανύσματα τα οποία τροφοδότησαν τους αλγόριθμους ομαδοποίησης. Το σύστημα παρέχει στον χρήστη την δυνατότητα να επιλέξει ταινίες από τέσσερις κατηγορίες. Οι κινήσεις τις οποίες κάνει ο χρήστης κατά την περιήγησή του στο κατάστημα (π.χ. επιλογή ταινίας, επιλογή κατηγορίας), καταγράφονται στην βάση δεδομένων, ώστε να είναι δυνατή η μελλοντική επεξεργασία τους.

Από τα δεδομένα που συλλέγονται υπολογίζεται για κάθε χρήστη ξεχωριστά, ο βαθμός ενδιαφέροντός του για μια κατηγορία ταινιών ή μία μεμονωμένη ταινία. Άλλα χαρακτηριστικά που παίζουν ρόλο στον υπολογισμό, είναι η τιμή της ταινίας, οι ηθοποιοί που συμμετέχουν και ο σκηνοθέτης. Ένα παράδειγμα υπολογισμού του βαθμού ενδιαφέροντος ενός χρήστη για μια ταινία την οποία έχει αγοράσει είναι το ακόλουθο:

$$\text{InterestOnBoughtMovie} = \frac{VisitsOnBoughtMovie}{VisitsOnAllBoughtMovies}$$

Το σύστημα κάνει διαχωρισμό των δεδομένων σε τρία μέρη. Το ένα περιέχει στατιστικά στοιχεία των επισκέψεων κάθε χρήστη σε μία ταινία. Το δεύτερο περιέχει αντίστοιχα στοιχεία για τις κινήσεις του χρήστη που αφορούν το καλάθι αγοράς και τέλος το τρίτο αφορά στοιχεία των ταινιών τις οποίες έχει αγοράσει ο χρήστης. Τα δεδομένα και των τριών αυτών κατηγοριών μεταφράζονται σε διανύσματα τα οποία έχουν 80 χαρακτηριστικά γνωρίσματα έκαστο, στα

οποία περιλαμβάνονται οι κατηγορίες, οι τιμές των ταινιών ηθοποιοί και σκηνοθέτες. Τα διανύσματα αυτά τροφοδοτούν τους αλγόριθμους ομαδοποίησης και έτσι γίνεται η σύγκριση μεταξύ των αποτελεσμάτων για τις προκύπτουσες ομάδες.

Η εξέλιξη της έρευνας συνεχίστηκε [5] με την δημιουργία στερεοτύπων η οποία βασίστηκε στα δεδομένα τα οποία συλλέχθηκαν μέσω της δοκιμαστικής πλατφόρμας Vision.com. Σε αυτήν τη δημοσίευση [5] οι συγγραφείς περιγράφουν τη δημιουργία διπλών στερεοτύπων με την χρήση ενός Immune Network System (INS). Το INS εφαρμόστηκε για τα δεδομένα των 150 χρηστών και κατηγοριοποίησε τα ενδιαφέροντα των χρηστών καθώς και τις ταινίες. Η διπλή αυτή κατηγοριοποίηση έγινε με ιεραρχικό τρόπο και σαν αποτέλεσμα δημιουργήθηκαν πολλαπλά επίπεδα στερεοτύπων. Τα στερεότυπα αυτά χρησιμοποιήθηκαν στην εφαρμογή ωστέ να μπορεί δυναμικά να εξαχθεί ένα συμπέρασμα για τα ενδιαφέροντα του χρήστη, βασισμένα μόνο σε μικρό αριθμό κινήσεών του.

Μια παρόμοια προσπάθεια επιχειρείται στο [11], όπου ένα Artificial Immune System (AIN) εφαρμόστηκε σε περίπου 1000 χρήστες των οποίων είχαν συλλεχθεί οι προτιμήσεις πλοήγησης στο διαδίκτυο (Favorites ή Bookmarks). Ιδιαίτερη έμφαση σε αυτή την έρευνα δόθηκε στην κατασκευή ενός μέτρου ομοιότητας των προφίλ των χρηστών, το οποίο θα μπορούσε να οδηγήσει στην ορθότερη τροφοδότηση του AIN.

3. ΑΝΑΛΥΣΗ ΚΑΙ ΣΧΕΔΙΑΣΜΟΣ ΣΥΣΤΗΜΑΤΟΣ

Απο τις σημαντικότερες φάσεις στον κύκλο ζωής ενός λογισμικού είναι η φάση της ανάλυσης και της σχεδίασής του. Στην ουσία πρόκειται για τα θεμέλια, πάνω στα οποία θα χτιστεί σταδιακά ένα λογισμικό άρα η συμβολή αυτής της φάσης παίζει σημαίνοντα ρόλο στην επιτυχή δημιουργία του και στον βαθμό στον οποίο θα καλύπτει τους χρήστες για τους οποίους προορίζεται.

3.1 Διαδικασία ανάπτυξης λογισμικού

Η διαδικασία ανάπτυξης λογισμικού χωρίζεται σε τρεις κύριες φάσεις οι οποίες κατά σειρά είναι:

- Συλλογή Απαιτήσεων
- Σχεδιασμός
- Υλοποίηση

Κατά την συλλογή των απαιτήσεων ενός συστήματος ορίζονται οι δραστηριότητες, οι κίνδυνοι και ένα δοκιμαστικό σχέδιο του συστήματος. Σε αυτή τη φάση δηλαδή ορίζεται εν συντομία ΤΙ πρέπει να κάνει το σύστημα. Πιο αναλυτικά στο στάδιο αυτό εστιάζουμε στο να παρέχουμε μία μηχανική περιγραφή των αντικειμένων, των λειτουργιών και των καταστάσεων του συστήματος λογισμικού.

Ο σχεδιασμός του συστήματος καθορίζει το ΠΩΣ το σύστημα θα υλοποιήσει τις λειτουργίες οι οποίες περιγράφονται κατά την συλλογή των απαιτήσεων και γενικά ορίζει την δομή που θα έχει το σύστημα χωρίς ωστόσο να μπαίνει σε λεπτομέρειες υλοποίησης. Δηλαδή η σχεδίαση λογισμικού εστιάζει στον τρόπο με τον οποίο οι απαιτήσεις του λογισμικού, μπορούν να υλοποιηθούν.

Τέλος στη φάση της υλοποίησης παράγεται ο πηγαίος κώδικας και η τεκμηρίωση του συστήματος καθώς επίσης γίνονται και δοκιμαστικοί έλεγχοι του λογισμικού που έχει παραχθεί ώστε αυτό να επικυρωθεί.

3.1.1 Λειτουργικές Απαιτήσεις

1. Εγγραφή νέου χρήστη (Sign-In)

Περιγραφή: Ο χρήστης ο οποίος θέλει να γίνει μέλος του ηλεκτρονικού καταστήματος, εισάγει τα απαραίτητα στοιχεία που απαιτούνται για να καταχωρηθεί στο σύστημα. Σε περίπτωση που όλα τα στοιχεία που θα εισάγει είναι σωστα, το σύστημα καταχωρεί τον χρήστη σαν νέο πελάτη

και δημιουργεί ένα μοναδικό κωδικό για αυτόν. Αν το σύστημα εντοπίσει λάθος στην καταχώρηση των στοιχείων, τότε εμφανίζει το κατάλληλο μήνυμα λάθους.

Είσοδος: Στοιχεία χρήστη

Επεξεργασία: Το σύστημα ελέγχει εάν υπάρχει χρήστης με το ίδιο user-id, το οποίο επέλεξε ο νέος χρήστης. Εάν δεν υπάρχει και όλα τα υπόλοιπα στοιχεία είναι αποδεκτά, το σύστημα καταχωρεί τον πελάτη στην βάση δεδομένων και ενημερώνει τον χρήστη για την επιτυχή ολοκλήρωση της διαδικασίας. Στην αντίθετη περίπτωση, το σύστημα εμφανίζει κατάλληλο μήνυμα λάθους στον χρήστη.

Έξοδος: Νέα εγγραφή στον πίνακα πελατών ή μήνυμα λάθους.

2. Εισαγωγή χρήστη στο κατάστημα (Log-In)

Περιγραφή: Ο χρήστης ο οποίος επιθυμεί να κάνει log-in στο ηλεκτρονικό κατάστημα πρέπει να εισάγει το αναγνωριστικό χρήστη (user-id) και το συνθηματικό (password). Το σύστημα προχωράει στην ταυτοποίηση των στοιχείων αυτών. Αν ο χρήστης έδωσε σωστά στοιχεία τότε συνδέεται στο σύστημα, αλλιώς εμφανίζει μήνυμα λάθους.

Είσοδος: Αναγνωριστικό χρήστη (user-id), Συνθηματικό χρήστη (password)

Επεξεργασία: Το σύστημα αναζητά στην βάση πελατών το αναγνωριστικό χρήστη που έδωσε ο χρήστης και ακολούθως κάνει την ταυτοποίηση του συνθηματικού. Αν η διαδικασία είναι επιτυχημένη τότε το σύστημα καλωσορίζει τον χρήστη και πλέον αυτός έχει πρόσβαση σε όλες τις υπηρεσίες του καταστήματος που είναι διαθέσιμες για τους εγγεγραμμένους χρήστες. Σε διαφορετική περίπτωση το σύστημα εμφανίζει στον χρήστη ένα μήνυμα λάθους και τον οδηγεί στο να επαναεισάγει τα ζητούμενα στοιχεία.

Έξοδος: Καταγραφή στο session state της εφαρμογής ότι ο χρήστης είναι πλέον συνδεδεμένος ή μήνυμα λάθους.

3. Επεξεργασία στοιχείων χρήστη

Περιγραφή: Ο χρήστης ο οποίος επιθυμεί να αλλάξει τα στοιχεία σύνδεσης (password) ή τα προσωπικά του στοιχεία (όνομα, επώνυμο, διεύθυνση, e-mail κτλ) πρέπει προηγουμένως να έχει κάνει επιτυχή είσοδο στο σύστημα (log-in).

Είσοδος: Αναγνωριστικό χρήστη (user-id), Συνθηματικό χρήστη (password), προαιρετικά νέα προσωπικά στοιχεία [όνομα, επώνυμο, διεύθυνση, e-mail].

Επεξεργασία: Το σύστημα αναζητά στην βάση πελατών το αναγνωριστικό χρήστη που έδωσε ο χρήστης και ακολούθως κάνει την ταυτοποίηση του συνθηματικού. Αν η διαδικασία είναι επιτυχημένη τότε το σύστημα εμφανίζει στον χρήστη την φόρμα συμπλήρωσης των προσωπικών του στοιχείων. Αφού ο χρήστης συμπληρώσει τα επιθυμητά στοιχεία το σύστημα ελέγχει την ορθότητα τους και αν δεν υπάρχει πρόβλημα εμφανίζει μήνυμα επιτυχούς αλλαγής των στοιχείων και στέλνει ενημερωτικό e-mail στον χρήστη. Αν τα στοιχεία έχουν κάποιο λάθος τότε εμφανίζει το ανάλογο μήνυμα λάθους και προτρέπει τον χρήστη να το διορθώσει.

Έξοδος: Ανανέωση του πίνακα πελατών στην βάση δεδομένων σε περίπτωση επιτυχημένης ολοκλήρωσης της διαδικασίας και αποστολή ενημερωτικού e-mail ή εμφάνιση μηνύματος λάθους και επανάληψη της διαδικασίας.

4.Προσθήκη προϊόντος στο καλάθι αγοράς

Περιγραφή: Ο χρήστης επιλέγει να προσθέσει ένα προϊόν στο καλάθι αγοράς.

Είσοδος: Αναγνωριστικό προϊόντος

Επεξεργασία: Το σύστημα ψάχνει αν το id του προϊόντος ώστε να διαπιστωθεί αν το προϊόν υπάρχει ήδη στο καλάθι αγοράς. Αν δεν υπάρχει ήδη, τότε ελέγχει την διαθεσιμότητα του προϊόντος και αν το προϊόν είναι διαθέσιμο ενημερώνει τον πίνακα basket στη βάση, αλλιώς ενημερώνει τον χρήστη. Αν το προϊόν έχει ήδη προστεθεί στο καλάθι, τότε ενημερώνει τους μετρητές για το προϊόν αυξάνοντας τον κατά 1.

Έξοδος: Ενημέρωση της βάσης και ανανέωση στην οθόνη του καλαθιού αγοράς ώστε να απεικονίζει τα σωστά προϊόντα ή ενημέρωση του χρήστη με μήνυμα ότι το προϊόν δεν είναι διαθέσιμο.

5. Αφαίρεση προϊόντος από το καλάθι αγοράς

Περιγραφή: Ο χρήστης επιλέγει να αφαιρέσει ένα προϊόν από το καλάθι αγοράς.

Είσοδος: Αναγνωριστικό προϊόντος (product_id), ποσότητα (quantity)

Επεξεργασία: Το σύστημα αναζητά το αναγνωριστικό προϊόντος στον πίνακα basket της βάσης. Αν η ποσότητα του προϊόντος είναι 1 τότε αφαιρεί το προϊόν από το καλάθι αγοράς και απο τον πίνακα basket. Σε περίπτωση που η ποσότητα του προϊόντος είναι μεγαλύτερη του ενός, μειώνεται κατά την ποσότητα που απαίτησε ο χρήστης και ενημερώνεται κατάλληλα ο πίνακας basket.

Έξοδος: Ενημέρωση της βάσης με τις νέες τιμές για το καλάθι αγοράς και ανανέωση της οθόνης ώστε το καλάθι αγοράς να απεικονίζει τα νέα δεδομένα.

6. Ολοκλήρωση Αγοράς προϊόντος (-ων) (Check-out)

Περιγραφή: Ο χρήστης επιλέγει να ολοκληρώσει τις αγορές του πατώντας τον σύνδεσμο Check-Out.

Είσοδος: Αναγνωριστικό χρήστη (user_id), συνθηματικό χρήστη (password)

Επεξεργασία: Για να ολοκληρώσει ο χρήστης τις αγορές του απαραίτητη προϋπόθεση είναι να έχει κάνει εισαγωγή στο σύστημα. Το σύστημα ελέγχει αν ο χρήστης έχει συνδεθεί και αν ναι τον ανακατευθύνει στην σελίδα οπού καλείται να επιβεβαιώσει την αγορά, αλλιώς τον παραπέμπει στη σελίδα σύνδεσης για να κάνει log-in.

Έξοδος: Το σύστημα ενημερώνει τους μετρητές των προϊόντων στην βάση δεδομένων και ενημερώνει τον χρήστη ότι η συναλλαγή ολοκληρώθηκε επιτυχώς με μήνυμα στην οθόνη και e-mail επιβεβαίωσης με τα στοιχεία της παραγγελίας. Σε περίπτωση που δεν είναι συνδεδεμένος του εμφανίζει το κατάλληλο μήνυμα που τον προτρέπει να συνδεθεί.

7. Επιλογή προϊόντος

Περιγραφή: Ο χρήστης κατά την περίηγησή του στο site επιλέγει ένα αντικείμενο κάνοντας κλικ επάνω του.

Είσοδος: Κλικ στον σύνδεσμο ή την εικόνα ενός προϊόντος (product_id).

Επεξεργασία: Το σύστημα ανακτά το προϊόν που αντιστοιχεί στον σωστό κωδικό και εμφανίζει στον χρήστη λεπτομέρειες για το προϊόν, όπως τα χαρακτηριστικά του, μία σύντομη ή αναλυτική περιγραφή και εικόνες του προϊόντος (με δυνατότητα μεγέθυνσης). Τέλος ενημερώνονται οι μετρητές στην βάση δεδομένων που αντιστοιχούν στο προϊόν για τον συγκεκριμένο χρήστη.

Έξοδος: Ενημέρωση μετρητών επισκεψιμότητας προϊόντος στην βάση δεδομένων και εμφάνιση στην οθόνη των χαρακτηριστικών του προϊόντος.

8. Ψήφιση προϊόντος

Περιγραφή: Ο χρήστης επιθυμεί να αξιολογήσει ένα προϊόν σε μία κλίμακα 5 σημείων.

Είσοδος: product_id, user_id, rating

Επεξεργασία: Το σύστημα ελέγχει αν ο συγκεκριμένος χρήστης έχει ξαναβαθμολογήσει το προϊόν. Αν όχι τότε ενημερώνει τους μετρητές ψήφισης του προϊόντος και εμφανίζει στον χρήστη την μέση βαθμολογία του προϊόντος.

Έξοδος: Εμφάνιση της μέσης βαθμολογίας του προϊόντος και της βαθμολογίας που έδωσε ο χρήστης καθώς και ενημέρωση των μετρητών ψήφισης στην βάση δεδομένων.

9. Ψήφιση site

Περιγραφή: Ο χρήστης αξιολογεί το site σε μία κλίμακα 5 σημείων.

Είσοδος: user_id, rating

Επεξεργασία: Το σύστημα ελέγχει αν ο χρήστης είναι συνδεδεμένος και αν δεν είναι τον παραπέμπει να συνδεθεί. Αφού ψηφίσει ενημερώνει την οθόνη με τον μέσο όρο βαθμολόγησης του site από όλους τους χρήστες και με την δική του βαθμολογία, όπως επίσης και τους μετρητές ψήφισης του site στην βάση δεδομένων.

Έξοδος: Εμφάνιση στην οθόνη της μέσης βαθμολογίας του site και της βαθμολογίας που έδωσε ο χρήστης καθώς και ενημέρωση των μετρητών στην βάση.

10. Αναζήτηση προϊόντος

Περιγραφή: Ο χρήστης ψάχνει για κάποιο προϊόν συγκεκριμένα χαρακτηριστικά.

Είσοδος: περιγραφή, [cluster_id], [κατηγορία]

Επεξεργασία: Το σύστημα ψάχνει την βάση να δει αν ο χρήστης ανήκει σε κάποιο cluster. Το σύστημα αναζητεί προϊόντα με βάση την περιγραφή και πιθανώς την κατηγορία που δόθηκε από τον χρήστη. Αν ο χρήστης ανήκει σε κάποιο cluster τότε τα προϊόντα που επιστρέφονται στον χρήστη ταξινομούνται με βάση της προτίμησης της ομάδας στην οποία ανήκει, αλλιώς επιστρέφονται ταξινομημένα με αλφαβητική σειρά. Προϊόντα που έχει ήδη αγοράσει ο χρήστης δεν παραλείπονται από τα αποτελέσματα της αναζήτησης.

Έξοδος: Λίστα προϊόντων που δημιουργείται βάσει των κριτηρίων αναζήτησης.

3.1.2 Μη Λειτουργικές Απαιτήσεις

1. Απαιτήσεις σχεδίασης συστήματος

Η σχεδίαση της εφαρμογής (δημιουργία UML διαγραμμάτων) θα γίνει με χρήση της εφαρμογής Rational Rose.

2. Βάση Δεδομένων

Η εφαρμογή να αντλεί και να αποθηκεύει τα δεδομένα που απαιτούνται, από μια βάση δεδομένων MySql.

3. Απαιτήσεις Υλοποίησης εφαρμογής

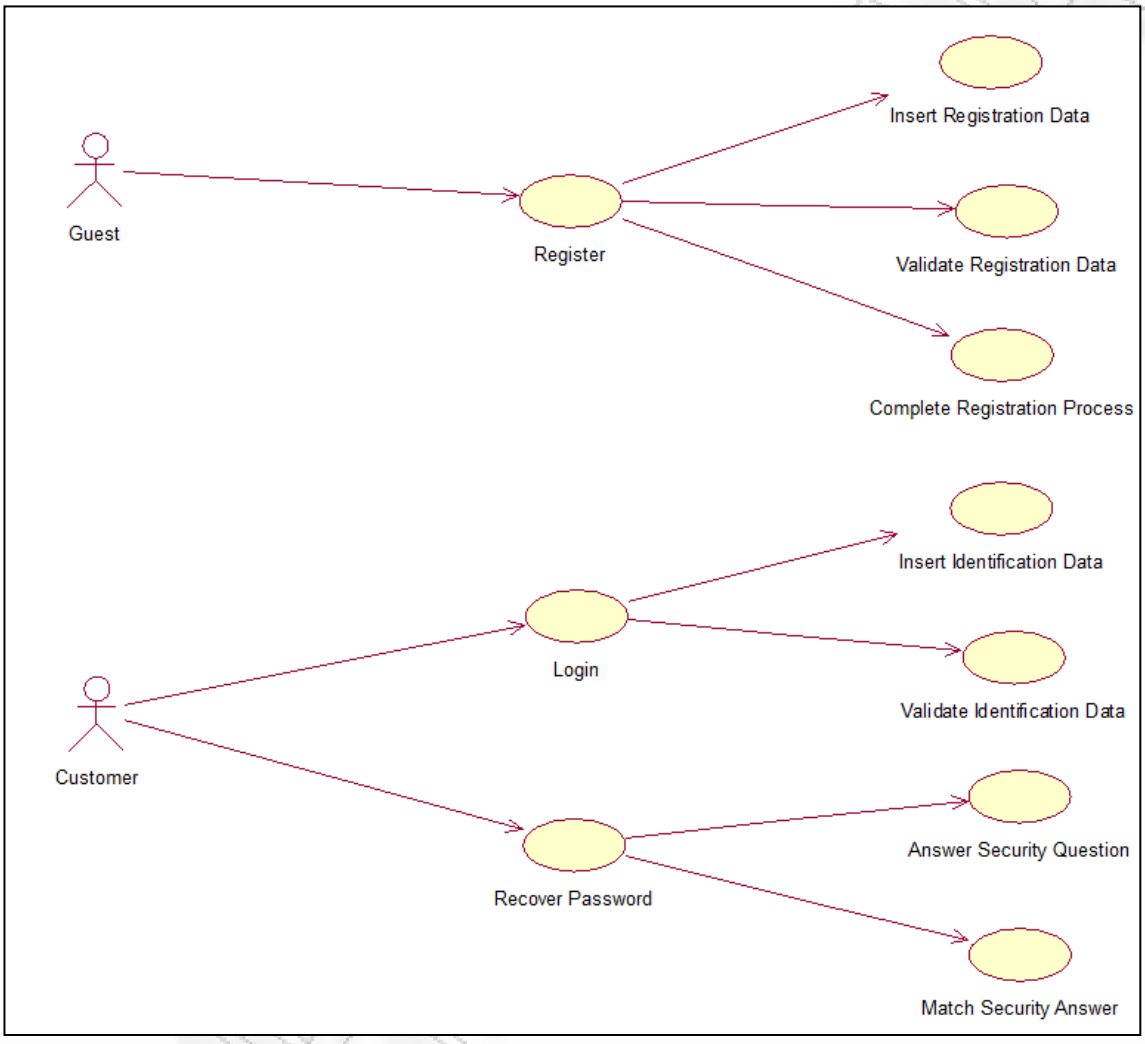
Η κύρια γλώσσα προγραμματισμού που θα χρησιμοποιηθεί για την υλοποίηση της εφαρμογής να είναι η server-side scripting language PHP.

3.2 Διαγράμματα περιπτώσεων χρήσης (Use case diagrams)

Το κυριότερο ζήτημα στην επιτυχημένη κατασκευή ενός πληροφοριακού συστήματος είναι η κατασκευή του σωστού συστήματος, δηλαδή ενός συστήματος που να ικανοποιεί τις απαιτήσεις των χρηστών του. Δεν είναι λίγα τα πληροφοριακά συστήματα που αποτυγχάνουν να ικανοποιήσουν τον βασικότατο αυτόν στόχο, και αυτό οφείλεται στην αποτυχία του συστήματος να ικανοποιήσει τις προδιαγραφές του χρήστη. Τα μοντέλα περιπτώσεων χρήσης δίνουν έμφαση στην λειτουργικότητα ενός συστήματος, όπως αυτή μπορεί να γίνει αντιληπτή από τους χρήστες του. Μία περίπτωση χρήσης απεικονίζει στην ουσία μία συναλλαγή (αλληλεπίδραση) ενός χρήστη με το σύστημα. Ένα σενάριο αποτελείται από μια ακολουθία βημάτων τα οποία περιγράφουν την αλληλεπίδραση του χρήστη με το σύστημα. Στην ουσία ένα διάγραμμα περίπτωσης χρήσης περιγράφει μια σειρά σεναρίων από τα οποία ένα μπορεί να έχει θετική έκβαση για τον χρήστη και τα υπόλοιπα προσπαθούν να περιγράψουν προβληματικές καταστάσεις στις οποίες μπορεί να βρεθεί ο χρήστης είτε λόγω αστοχίας του συστήματος είτε λόγω αβλεψίας του χρήστη. Σε κάθε περίπτωση η λειτουργία ενός συστήματος θα πρέπει να είναι απρόσκοπτη και σε αυτό βοηθάνε τα διαγράμματα περιπτώσεων χρήσης ώστε να

προβλέπουμε όλες τις πιθανές κι απίθανες ενέργειες του χρήστη και να μπορούμε να τις διαχειριστούμε.

3.2.1 Διάγραμμα χρήσης Σύνδεσης και Εγγραφής χρήστη



Εικ. 3.1: Use Case 'Customer Register & Log-In'

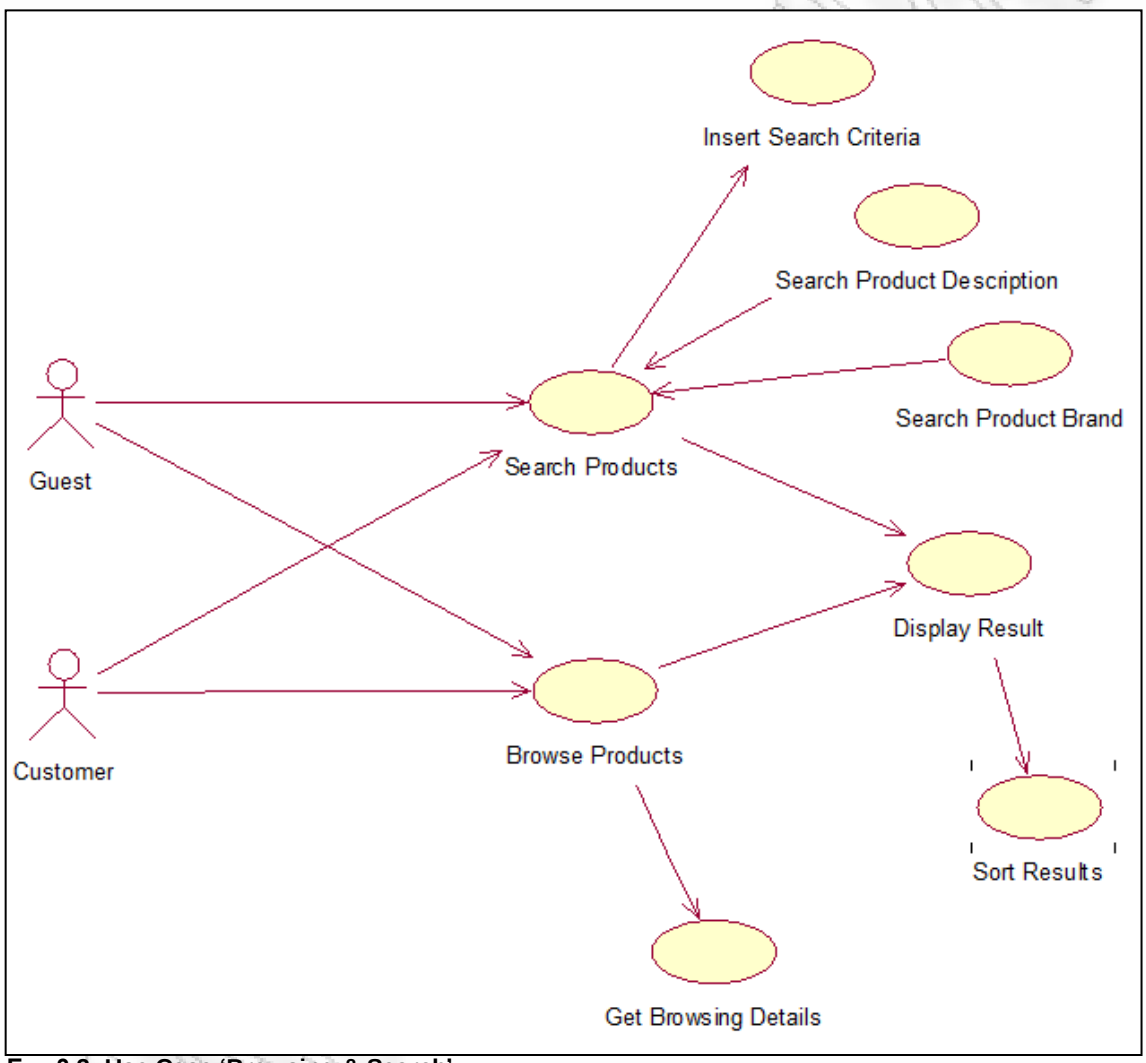
Όπως φαίνεται στο παραπάνω διάγραμμα χρήσης (Εικ. 3.1), περιγράφονται τρεις κύριες λειτουργίες του συστήματος. Η μία είναι η διαδικασία Register της οποία ενεργοποιός είναι ο Guest. Ως Guest ορίζεται ο επισκέπτης του ηλεκτρονικού καταστήματος, ο οποίος όμως δεν έχει γίνει μέλος του. Για να γίνει μέλος πρέπει να εγγραφεί στο σύστημα. Την εγγραφή στο σύστημα απεικονίζει η διαδικασία Register. Η εγγραφή στο ηλεκτρονικό κατάστημα αποτελείται απο τρεις διαδικασίες. Πρώτη είναι η διαδικασία εισαγωγής δεδομένων (Insert Registration Data) κατά την οποία ο χρήστης καλείται να συμπληρώσει τα προσωπικά του στοιχεία. Αφού ολοκληρωθεί η συμπλήρωση των στοιχείων ακολουθεί η διαδικασία επαλήθευσης των στοιχείων (Validate Registration Data) ή οποία αν είναι επιτυχημένη ακολουθείται από την διαδικασία Ολοκλήρωσης της εγγραφής (Complete Registration Process), η οποία εισάγει τον χρήστη στο σύστημα ως νέο πιστοποιημένο πελάτη (Customer).

Η διαδικασία Login έχει ως ενεργοποιό τον Customer, ο οποίος συμβολίζει ένα εγγεγραμμένο μέλος του συστήματος. Σύμφωνα με αυτή την διαδικασία ο πελάτης καλείται να

εισάγει τα πιστοποιητικά εισόδου στο σύστημα (διαδικασία Insert Identification Data) και αφού το σύστημα τα εγκρίνει (διαδικασία Validate Identification Data), ο χρήστης είναι πλέον αναγνωρισμένος από το σύστημα.

Τέλος περιγράφεται η διαδικασία κατά την οποία ένας χρήστης ζητά την ανάκτηση του συνθηματικού εισόδου και καλείται να εισάγει την απάντηση στην μυστική ερώτηση (διαδικασία Answer Security Question) την οποία το σύστημα εξετάζει αν είναι σωστή ώστε να αποστείλει τα ανακτημένα στοιχεία στον χρήστη (διαδικασία Match Security Answer).

3.2.2 Διάγραμμα χρήσης Περιήγησης και Αναζήτησης

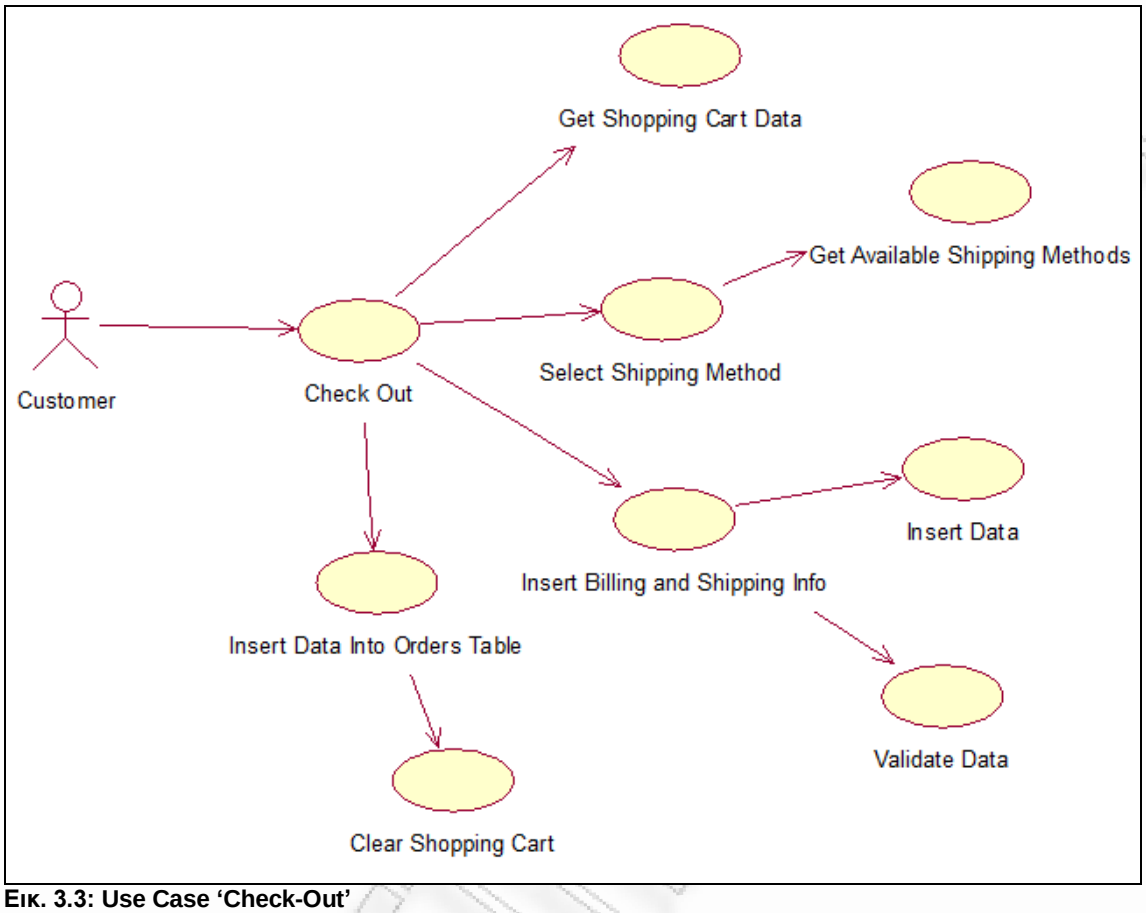


Εικ. 3.2: Use Case 'Browsing & Search'

Στο παραπάνω διάγραμμα (Εικ. 3.2) περιγράφονται οι λειτουργίες περιήγησης (Browse Products) και αναζήτησης (Search Products). Κύριοι ενεργοποιοί είναι και πάλι ο Guest και ο Customer οι οποίοι συμμετείχαν και στο πρώτο διάγραμμα (Εικ. 3.1). Στην διαδικασία αναζήτησης οι δύο ενεργοποιοί εισάγουν τα κριτήρια αναζήτησης και το σύστημα αναζητά τα κριτήρια αυτά στην περιγραφή ή με βάση την καρηγορία του προϊόντος και επιστρέφει τα αποτελέσματα κατάλληλα ταξινομημένα στην οθόνη.

Στην διαδικασία Browse Products οι δύο ενεργοποιοί ζητούν να δουν λεπτομέρειες για ένα προϊόν επιλέγοντάς το και το σύστημα επιστρέφει τις πληροφορίες για το προϊόν στην οθόνη του χρήστη.

3.2.3 Διάγραμμα χρήσης Ολοκλήρωσης αγοράς (Check-Out)

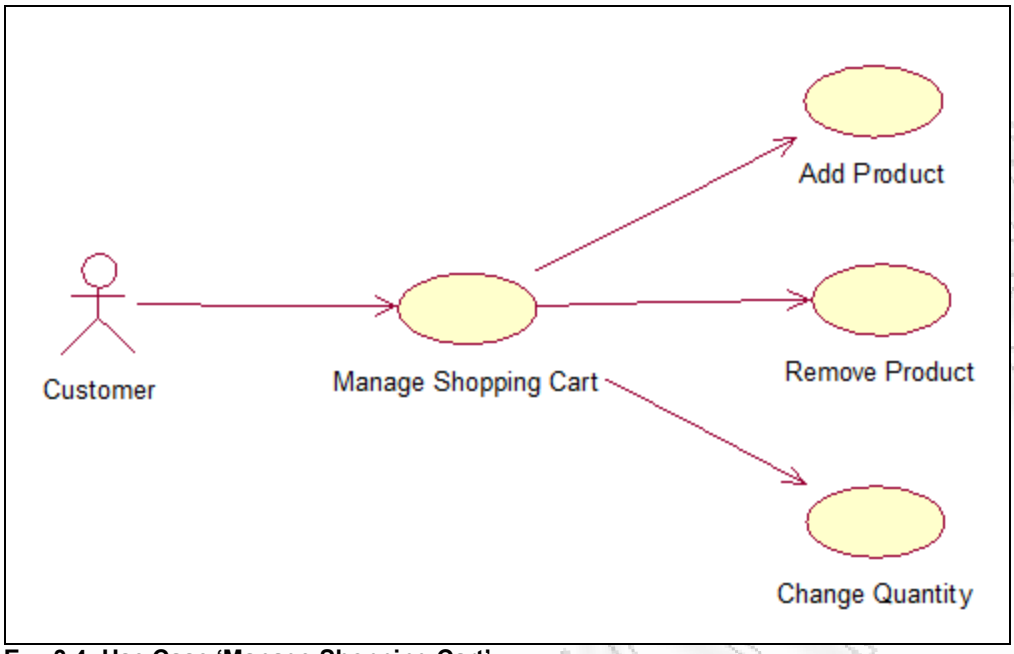


Εικ. 3.3: Use Case 'Check-Out'

Το διάγραμμα χρήσης Check-Out (Εικ. 3.3) περιγράφει πώς ένας εγγεγραμμένος χρήστης του ηλεκτρονικού καταστήματος, μπορεί να ολοκληρώσει μία αγορά. Κατά την διαδικασία αυτή, ο χρήστης καλείται να επιλέξει την μέθοδο αποστολής των προϊόντων (Select Shipping Method) και να εισάγει την μέθοδο πληρωμής (Insert Billing and Shipping Info). Το σύστημα αξιολογεί τα στοιχεία αυτά ως προς την ορθότητά τους και εφόσον είναι σωστά ανακτά τα δεδομένα των προϊόντων που υπάρχουν στο καλάθι αγοράς (Get Shopping Cart Data). Αν όλα πάνε καλά όλα τα δεδομένα της παραγγελίας αποθηκεύονται στον πίνακα που διατηρεί το ιστορικό των αγορών για όλους τους πελάτες (orders).

3.2.4 Διάγραμμα χρήσης διαχείρησης καλαθιού αγοράς

Το διάγραμμα χρήσης για την διαχείρηση του καλαθιού αγοράς (Εικ. 3.4) περιγράφει τις επιλογές που έχει ένας χρήστης σε σχέση με τα περιεχόμενα του καλαθιού. Ο χρήστης έχει την δυνατότητα να προσθέσει ένα ή περισσότερα προϊόντα στο καλάθι αγοράς (Add Product) επιλέγοντας το κουμπί Add to Cart που υπάρχει δίπλα σε κάθε προϊόν. Επίσης ο χρήστης μπορεί να αφαιρέσει εάν επιθυμεί ένα προϊόν από το καλάθι αγοράς (Remove Product). Τέλος μπαίνοντας στη σελίδα διαχείρησης του καλαθιού αγοράς, ο χρήστης μπορεί να αυξομειώσει όσο επιθυμεί τις ποσότητες των προϊόντων που υπάρχουν στο καλάθι.



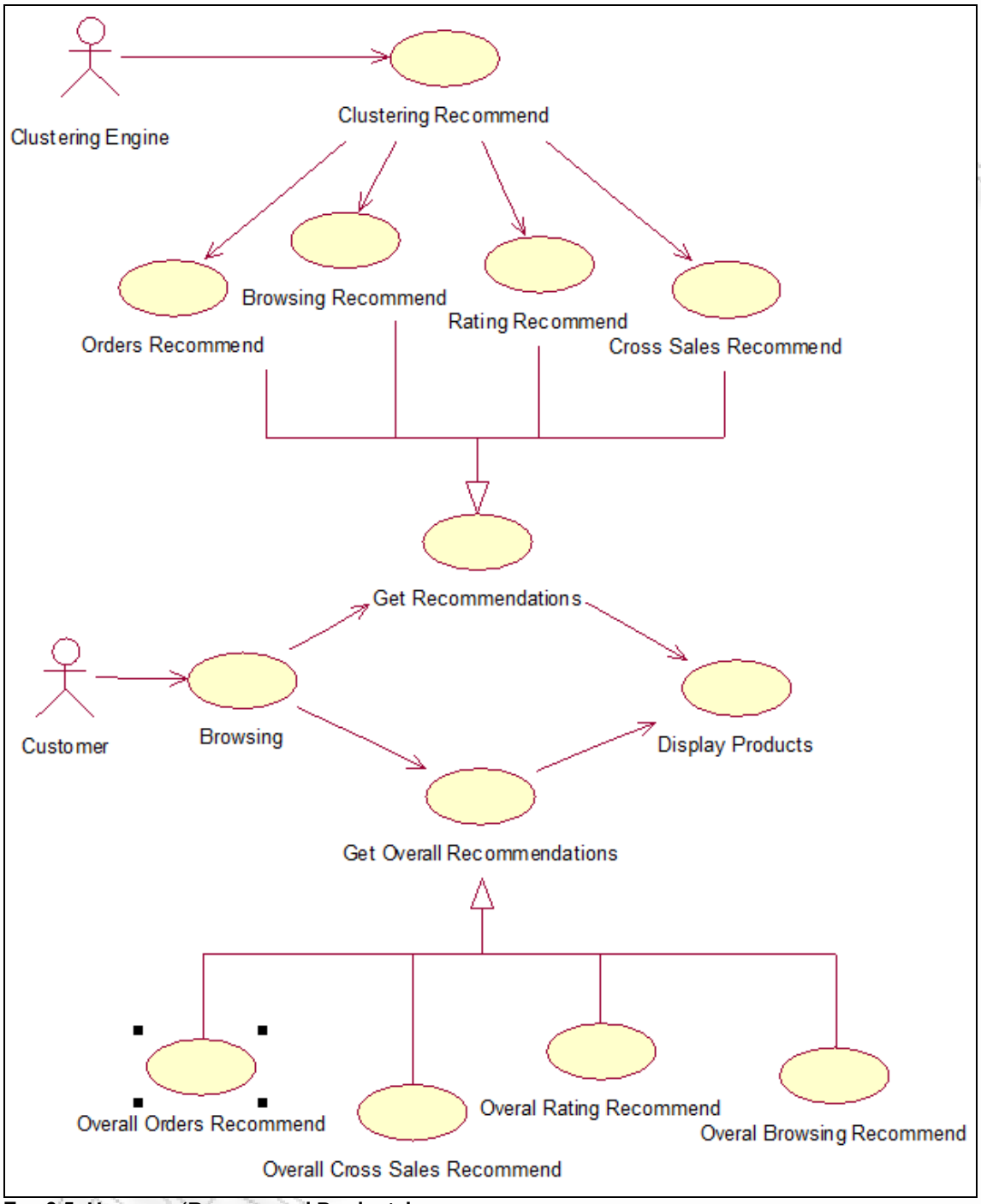
Εικ. 3.4: Use Case 'Manage Shopping Cart'

3.2.5 Διάγραμμα χρήσης διαχείρησης προτάσεων

Στο διάγραμμα χρήσης διαχείρησης προτάσεων (Εικ. 3.5), εξηγείται η διαδικασία με την οποία το σύστημα προτάσεων (Recommender System), προσαρμόζει το περιβάλλον περιήγησης του χρήστη διαμορφώνοντας την διεπαφή του καταστήματος ανάλογα με τις προτάσεις απο εξάγονται από τον μηχανισμό ομαδοποίησης.

Όταν ένας χρήστης (ο οποίος έχει κάνει login), περιηγείται στο ηλεκτρονικό κατάστημα και επιλέγει κάποιο προϊόν, τότε το σύστημα συμπληρώνει (Display Products) κάποιες επιλεγμένες θέσεις στην οθόνη (placeholders) με άλλα προϊόντα που πιθανώς να ενδιαφέρουν τον χρήστη στο πλαίσιο της στοχευμένης προώθησης των προϊόντων. Ο ενεργοποιός Clustering Engine είναι ο μηχανισμός που διαθέτει το ηλεκτρονικό κατάστημα, ο οποίος είναι υπεύθυνος για την ομαδοποίηση των εγγεγραμμένων χρηστών. Η διαδικασία Clustering Recommends είναι αυτή η οποία εκμεταλλεύεται τα στοιχεία της ομαδοποίησης κάνοντας εξόρυξη δεδομένων (data mining). Τα στοιχεία αυτά χωρίζονται σε τέσσερις κατηγορίες, ανάλογα με την προέλευσή τους. Η πρώτη κατηγορία αφορά δεδομένα που εξάγονται με βάση τις πληροφορίες που αφορούν αγορές τις οποίες έχουν πραγματοποιήσει οι χρήστες της ομάδας. Η δεύτερη κατηγορία προέρχεται από τα δεδομένα της ομάδας που αφορούν στην επισκεψιμότητα των πελατών σε διάφορα προϊόντα. Δεδομένα που αφορούν την βαθμολόγηση σε προϊόντα που έχει κάνει η ομάδα καθώς και βαθμολόγηση του καταστήματος, συντελούν την τρίτη κατηγορία προτάσεων. Τέλος η τέταρτη κατηγορία αφορά προϊόντα τα οποία έχουν αγοράσει οι χρήστες της ομάδας, που έχουν αγοράσει το προϊόν το οποίο κοιτάζει ο χρήστης (Cross-Sales).

Τέλος, για κάθε κατηγορία από αυτές που αναλύθηκαν παραπάνω, εξάγονται δεδομένα προτάσεων, συνολικά για όλους τους χρήστες και όχι μόνο για τους χρήστες που ανήκουν στην ομάδα του χρήστη (διαδικασία Get Overall Recommendations).



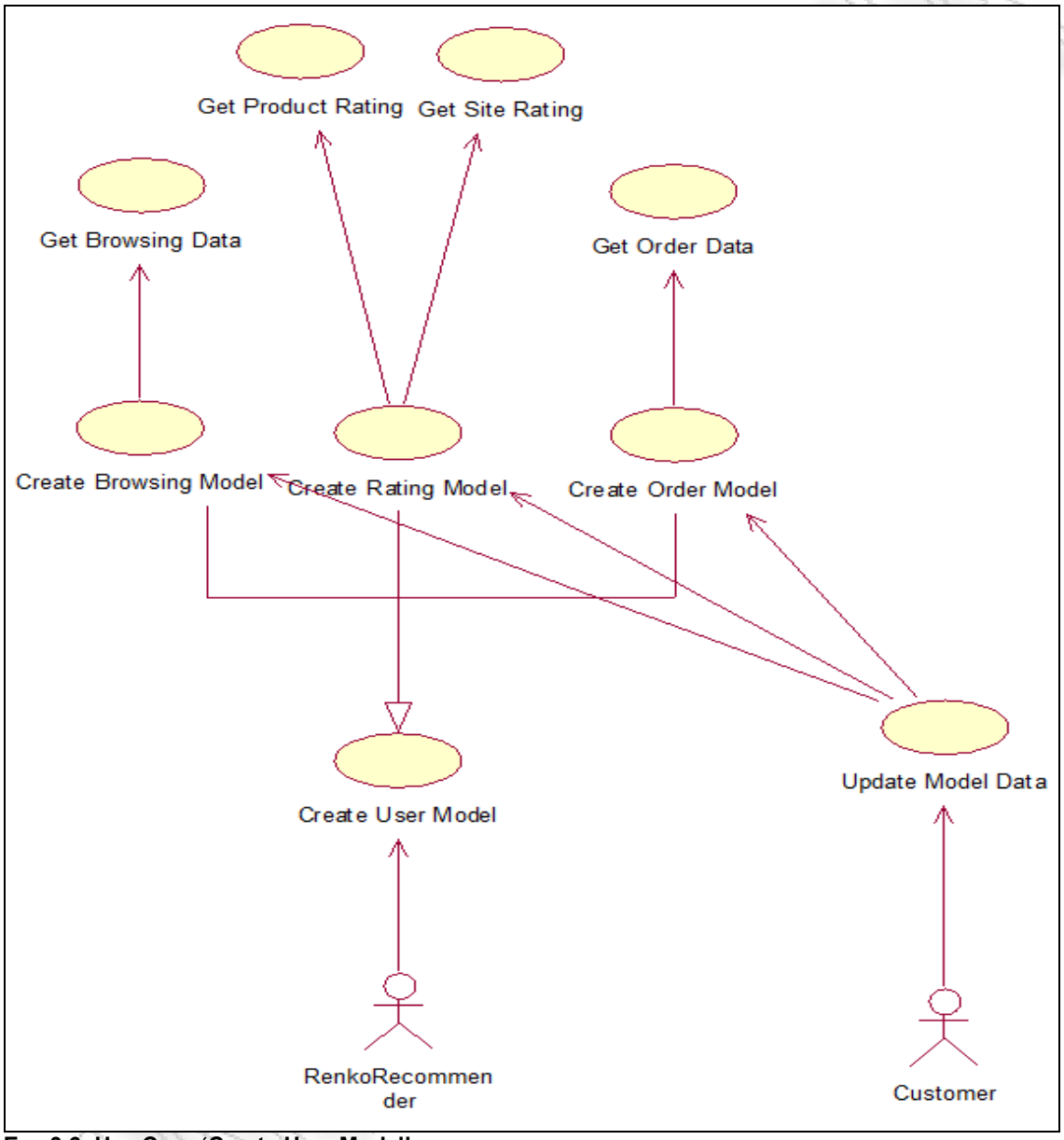
Εικ. 3.5: Use case 'Recommend Products'

3.2.6 Διάγραμμα χρήσης Δημιουργίας μοντέλου χρηστών

Το διάγραμμα χρήσης δημιουργίας μοντέλου χρηστών (Εικ. 3.6) περιγράφει μία από τις κυριότερες λειτουργίες του συστήματος που περιγράφουμε. Περιγράφει την δημιουργία του μοντέλου χρηστών βάσει των οποίων το σύστημα παράγει προτάσεις για κάθε χρήστη.

Στο σύστημα αυτό παρατηρούμε ότι συμμετέχει ένας νέος ενεργοποιός, ο RenkoRecommender. Έτσι έχουμε ονομάσει το σύστημα το οποίο είναι υπεύθυνο για την παραγωγή προτάσεων. Παρατηρούμε από το διάγραμμα ότι το κύριο μοντέλο χρήστη είναι στην

ουσία σύνθεση τριών μικρότερων μοντέλων (Browsing model, Rating model, Order model) τα οποία παράγονται από την επεξεργασία των δεδομένων αυτών των κατηγοριών που αφορούν σε κάθε ομάδα χρηστών. Σε κάθε ένα από αυτά τα μοντέλα συμμετέχει και ο χρήστης με τις έμμεσες πληροφορίες που παράγει με τις κινήσεις και την εν γένει συμπεριφορά του μέσα στο ηλεκτρονικό κατάστημα.



Εικ. 3.6: Use Case 'Create User Model'

Επίσης είναι εμφανές από το διάγραμμα ότι το μοντέλο που προκύπτει από την βαθμολογία των χρηστών είναι σύνθεση δεδομένων από δεδομένα που αφορούν την ψήφιση προϊόντων και από δεδομένα που αφορούν στην βαθμολόγηση του site. Τέλος όλα τα παραπανω δεδομένα συνεισφέρουν στην συναρμολόγηση ενός συνολικού μοντέλου χρήστη μέσω του οποίου αποφασίζονται όλες οι προτάσεις που γίνονται στους χρήστες.

4. ΘΕΩΡΗΤΙΚΟ ΥΠΟΒΑΘΡΟ

Στο κεφάλαιο αυτό θα παρατεθούν σημαντικά κομμάτια της θεωρίας μοντελοποίησης χρηστών καθώς και τεχνικές ιεραρχικές ομαδοποίησης και μέτρησης ομοιότητας. Η αφομοίωση όλων τα προαναφερθέντων στοιχείων είναι απαραίτητο συστατικό για την διεκπεραίωση της εργασίας.

4.1 Μοντελοποίηση Χρηστών

Ένας απλός χρήστης του διαδικτύου στις μέρες μας, βομβαρδίζεται στην κυριολεξία από έναν τεράστιο όγκο πληροφορίας, γεγονός το οποίο αρκετές φορές μπορεί να αποθαρρύνει ή να απομακρύνει τους νέους ειδικά χρήστες, από την ενασχόλησή τους με αυτό. Η απάντηση της επιστήμης της Πληροφορικής στο πρόβλημα αυτό είναι ένας σχετικά καινούριος και ιδιαίτερα αναπτυσσόμενος κλάδος, ο οποίος δεν είναι άλλος από την Μοντελοποίηση Χρηστών. Ο κλάδος αυτός στηρίζεται σε μεγάλο βαθμό σε τεχνικές Τεχνητής Νοημοσύνης, με σκοπό την δημιουργία ευφυών συστημάτων, τα οποία μέσω της παρατήρησης της συμπεριφοράς των χρηστών έχουν την δυνατότητα να προσαρμόζονται στις πραγματικές ανάγκες των χρηστών. Ανεξάρτητως όμως από την βοήθεια που παρέχει στους χρήστες, η Μοντελοποίηση Χρηστών είναι και ένα δυνατό εργαλείο στα χέρια των εταιριών που δραστηριοποιούνται στο διαδίκτυο, με το οποίο μπορούν να προωθήσουν αποτελεσματικότερα τα προϊόντα ή τις υπηρεσίες που εμπορεύονται. Είναι φανερό ότι μια επιτυχημένη εφαρμογή της θεωρίας της Μοντελοποίησης Χρηστών, βοηθά τόσο τον χρήστη να κάνει καλύτερες επιλογές, όσο και τις εταιρίες που την εφαρμόζουν, να επεκτείνουν το αγοραστικό τους κοινό και να ενισχύσουν τους δεσμούς εμπιστοσύνης με αυτό.

4.2 Κατηγοριοποίηση Μοντέλου Χρήστη

Υπάρχουν κάποιοι βασικοί παράγοντες βάσει των οποίων μπορεί να χαρακτηριστεί ένα μοντέλο χρήστη. Η διττή φύση των χαρακτηριστικών αυτών δεν εμποδίζει τον σχεδιαστή του συστήματος να κάνει χρήση συνδυασμών που απαιτούνται ώστε να έχει το επιθυμητό αποτέλεσμα. Δηλαδή όχι μόνο τα χαρακτηριστικά αυτά δεν είναι αμοιβαίως αποκλειόμενα, αντίθετως αποτελεί συνήθης πρακτική η ταυτόχρονη χρησιμοποίησή τους. Είναι στρατηγικής σημασίας στην σχεδίαση και άρα στην αποδοτικότητα ενός μοντέλου χρήστη, η σωστή επιλογή (χωρίς να αποκλείονται συνδυασμοί τους) των ακόλουθων χαρακτηριστικών:

4.2.1 Βαθμός Εξειδίκευσης

Ο βαθμός εξειδίκευσης πραγματεύεται το εάν σε κάθε χρήστη αναλογεί ένα μοντέλο ή το μοντέλο αντιπροσωπεύει μια τάξη χρηστών. Παράδειγμα μοντέλου προσαρμοσμένου σε μεμονωμένο χρήστη αποτελεί το σύστημα RESCUER, το οποίο ελέγχει χρήστες του λειτουργικού συστήματος UNIX για τυχόν συντακτικά ή λογικά λάθη. Στο ηλεκτρονικό μας κατάστημα κατασκευάστηκε ένα έμμεσο μοντέλο που αφορά ομάδες χρηστών κι όχι κάθε χρήστη μεμονωμένα. Η λογική στην οποία επιλέχτηκε το μοντέλο να αφορά ομάδα χρηστών, βασίζεται στο γεγονός ότι για την διοίκηση ενός ηλεκτρονικού καταστήματος θα είναι ευκολότερο να χαράξει ευκολότερα την στρατηγική πωλήσεων ή στοχευμένης διαφήμισης για περιορισμένο αριθμό αντιπροσωπευτικών ομάδων, παρά για κάθε χρήστη ξεχωριστά. Επίσης σε περίπτωση που ο αριθμός των χρηστών μεγαλώσει, ο αριθμός των ομάδων μπορεί να παραμείνει όπως έχει προαποφασιστεί, συμβάλλοντας έτσι στην μείωση της πολυπλοκότητας του συστήματος.

4.2.2 Τροποποιησιμότητα

Ένα μοντέλο διαχωρίζεται όσον αφορά την τροποποιησιμότητά του σε στατικό ή δυναμικό. Είναι προφανές ότι στατικό είναι ένα μοντέλο το οποίο καθορίζεται a priori από τον διαχειριστή του συστήματος και δεν αλλάζει καθόλη την διάρκεια της ζωής του λογισμικού. Στατικά μοντέλα μπορεί να θεωρηθούν χρήσιμα σε εφαρμογές που είναι γνωστές κι ευδιάκριτες οι ομάδες χρηστών και έχουν συγκεκριμένα χαρακτηριστικά. Αντίθετα με την δυσκαμψία ενός στατικού

μοντέλου, τα δυναμικά μοντέλα στηρίζονται στην παρακολούθηση των ενεργειών των χρηστών και εξελίσσονται βάσει αυτών.

Οι απαιτήσεις των χρηστών ενός ηλεκτρονικού καταστήματος πληροφορικής είναι φύσει δυναμικές, λόγω των ραγδαίων εξελίξεων και αλλαγών σε καθημερινό πλέον επίπεδο, οπότε το μοντέλο χρήστη που χρησιμοποιήθηκε είναι δυναμικό.

4.2.3 Τρόπος απόκτησης

Στην περίπτωση που το μοντέλο του χρήστη καθορίζεται από τον σχεδιαστή του συστήματος ή οι χρήστες συμμετέχουν οικιοθελώς στον σχηματισμό του, τότε ο τρόπος απόκτησης του μοντέλου είναι άμεσος. Τα μέσα απόκτησης μπορεί να είναι έντυπα ερωτηματολόγια, συνεντεύξεις από τους χρήστες ή χρήση του προγράμματος για το οποίο προορίζεται το μοντέλο. Ο άμεσος τρόπος απόκτησης ονομάζεται και ενεργητική απόκτηση του μοντέλου επειδή ο χρήστης συμμετέχει ενεργά σε αυτήν. Αντίθετα, έμμεση ή παθητική απόκτηση του μοντέλου χρήστη, ονομάζεται αυτή στην οποία συνάγονται αυτόματα από το σύστημα μέσω της παρακολούθησης της συμπεριφοράς των χρηστών κατά την διάρκεια της χρήσης του προγράμματος.

Ο τρόπος απόκτησης των δεδομένων για το μοντέλο που χρησιμοποιήθηκε στην παρούσα εργασία ήταν έμμεσος. Οι κινήσεις κάθε εγγεγραμμένου χρήστη καταγράφονται όσον αφορά κάποια χαρακτηριστικά γνωρίσματα του μοντέλου, χωρίς ο χρήστης να το γνωρίζει. Η διαφάνεια αυτού του τρόπου περισυλλογής των δεδομένων, μας εξασφάλισε και την κατά όσο το δυνατό, μεγαλύτερη αμεροληψία των χρηστών που συμμετείχαν στην διαδικασία. Ο τρόπος απόκτησης των δεδομένων θα αναλυθεί εκτενέστερα σε μεταγενέστερο κεφάλαιο.

4.2.4 Χρονική έκταση

Τα μοντέλα χρηστών ως προς την χρονική τους έκταση διακρίνονται σε βραχυπρόθεσμα και μακρυπρόθεσμα. Βραχυπρόθεσμο ονομάζεται το μοντέλο, όταν το σύστημα εξάγει συμπεράσματα για τη συμπεριφορά του χρήστη κατά τη διάρκεια της τρεχουσας συνεδρείας του χρήστη με το λογισμικό. Αντίθετα μακροπρόθεσμο είναι το μοντέλο χρήστη το οποίο καταγράφει μόνιμα τη συμπεριφορά και τα χαρακτηριστικά του χρήστη και τα αποθηκεύει ώστε να είναι πάντα ενημερωμένο. Το μοντέλο χρήστη στην παρούσα εργασία έχει μακροπρόθεσμο χαρακτήρα.

4.3 Επιδράσεις Μοντελοποίησης σε ένα ηλεκτρονικό κατάστημα

Όπως προαναφέρθηκε, ο ρόλος ενός μοντέλου χρήστη σε μία σύγχρονη εφαρμογή ηλεκτρονικού εμπορίου, είναι κατα κύριο λόγο να βοηθήσει το χρήστη στην σταχυολόγηση της πληροφορίας μέσα από έναν τεράστιο όγκο δεδομένων. Με απλά λόγια να διευκολύνει τον χρήστη στην τελική του επιλογή. Σε περίπτωση που το μοντέλο εφαρμοστεί με επιτυχία και ο χρήστης μείνει ικανοποιημένος τότε είναι βέβαιο ότι αυτό θα έχει θετικό αντίκτυπο και άρα θα οδηγήσει σε αύξηση των πωλήσεων που είναι κι ο αντικειμενικός στόχος.

Στον αντίποδα ένα κακοσχεδιασμένο μοντέλο είναι πιθανό, στην καλύτερη περίπτωση να μην έχει τα αναμενόμενα θετικά αποτελέσματα και στην χειρότερη περίπτωση να επιδράσει αρνητικά στην λειτουργία του καταστήματος απωθόντας τους πελάτες. Παρακάτω παραθέτουμε πιο αναλυτικά τις θετικές και τις αρνητικές επιδράσεις που μπορεί να έχει σε μια εφαρμοηγή το μοντέλο χρήστη.

4.3.1 Θετικές επιδράσεις

Οι τρεις κυριότεροι τρόποι με τους οποίους μπορεί να επιδράσει θετικά το μοντέλο χρήστη σε ένα ηλεκτρονικό κατάστημα είναι οι ακόλουθοι:

- Μετατροπή επισκεπτών σε αγοραστές (Browsers Into Buyers)

- Διασταύρωση πωλήσεων (Cross-Sales)
- Βελτίωση σχέσης καταστήματος πελάτη

Η μετατροπή των εν δυνάμει πελατών σε αγοραστές είναι η διαδικασία κατά την οποία ένας χρήστης του ηλεκτρονικού καταστήματος περιηγείται στο κατάστημα χωρίς να προβαίνει σε αγορά. Ένα μοντέλο χρήστη το οποίο έχει τη δυνατότητα να καταγράφει τις κινήσεις του χρήστη (π.χ. ποιες κατηγορίες επισκέπτεται συχνότερα) είναι σε θέση αναλύοντάς τες, να σχηματίσει γνώμη για τις προτιμήσεις του και να αναλόγως να του προτείνει προϊόντα που ενδεχομένως τον ενδιαφέρουν. Εάν ο χρήστης δελεαστεί απο τις προτάσεις που του κάνει το σύστημα και οδηγηθεί σε αγορά αυτό σημαίνει ότι ο στόχος της μετατροπής 'Browser Into Buyer' επετεύχθει.

Η διασταύρωση των πωλήσεων είναι η τεχνική κατά την οποία προτείνονται στον χρήστη προϊόντα σχετικά με αυτά που έχει ήδη αγοράσει ή πρόκειται να αγοράσει. Επιπρόσθετα στον χρήστη που αγοράζει ένα προϊόν ή βρίσκεται ένα βήμα πριν ολοκληρώσει την αγορά (check-out) προτείνονται προϊόντα τα οποία έχουν αγοράσει άλλοι χρήστες μαζί με το συγκεκριμένο προϊόν. Είναι προφανές ότι αν οι προτάσεις τις οποίες θα κάνει το σύστημα στον πελάτη είναι επιτυχημένες, θα οδηγήσουν στην αύξηση του μέσου όρου των πωλήσεων. Αν λοιπόν αυτές οι προτάσεις προκύπτουν από εξαγωγή συμπερασμάτων από το μοντέλο χρηστών, είναι πολυ πιθανό αυτές να προσεγγίσουν με μεγαλύτερη ακρίβεια τις προτιμήσεις του χρήστη.

Η βελτίωση της σχέσης καταστήματος – πελάτη είναι μια επίσης πολυ σημαντική θετική επίδραση που μπορεί να έχει η εφαρμογή ενός μοντέλου χρηστών. Σε ένα άκρως ανταγωνιστικό περιβάλλον, όπου ο πελάτης με μερικά μόνο κλικ μπορεί να βρεθεί σε ένα ανταγωνιστικό site, είναι στρατηγικής σημασίας η επένδυση στο χτίσιμο μιας σχέσης εμπιστοσύνης με τον πελάτη. Όσο πιο ικανοποιημένος μένει ένας πελάτης από τις υπηρεσίες που του παρέχονται, τόσο δυσκολότερα θα προτιμήσει να αφήσει το δικό μας κατάστημα για ένα ανταγωνιστικό. Αν ένας πελάτης λοιπόν χρησιμοποιεί σε μεγάλο βαθμό το σύστημα και τον ικανοποιούν οι προτάσεις που του κάνει, τόσο πιο 'πιστός' θα γίνεται, με αποτέλεσμα είτε να το χρησιμοποιεί συχνότερα ή ακόμα να προτείνει και στο περιβάλλον του να το επισκεφθεί. Το πόσο σημαντική είναι η επένδυση στη βελτίωση της σχέσης καταστήματός – πελάτη, φαίνεται και από το γεγονός ότι ακόμα κι αν κάποιος ανταγωνιστής χτίσει ένα σύστημα παρόμοιο με το δικό μας, είναι τέτοια η φύση της ψυχολογίας του πελάτη που πολυ δύσκολα θα πειστεί να αλλάξει περιβάλλον. Εξάλλου, για να φτάσει ο ανταγωνιστής στο επίπεδο γνώσης που εμείς ήδη κατέχουμε για ένα πελάτη (και άρα να ικανοποιούν τις ανάγκες του), θα χρειαστεί σημαντικός χρόνος και υπομονή από τον πελάτη, παράγοντες οι οποίοι καθιστούν ακόμη δυσκολότερη την μετάβαση του πελάτη σε ένα ανταγωνιστικό κατάστημα.

4.3.2 Αρνητικές επιδράσεις

Η ποιότητα των προτάσεων που κάνει ένα σύστημα σε έναν πελάτη, έχει βαρύνουσα επίδραση στην μελλοντική αγοραστική του συμπεριφορά. Για αυτό είναι μείζονος σημασίας η προσοχή στην σχεδίαση του μοντέλου χρήστη, ώστε να μην παρουσιαστούν αρνητικές επιδράσεις από την χρήση του, δύο από τις οποίες είναι οι παρακάτω:

- Υπερβολική παρεμβατικότητα
- False negative and false positive error

Μία σημαντική επίπτωση που ενδέχεται να έχει σε ένα σύστημα, ένα κακοσχεδιασμένο μοντέλο χρήστη, είναι ο υψηλός βαθμός παρεμβατικότητας του συστήματος κατά την αλληλεπίδραση του με τον χρήστη. Αυτό μπορεί να στερήσει από τον χρήστη την πρωτοβουλία κινήσεων και να του περιορίζει σημαντικά το βαθμό ελευθερίας, καθιστώντας συχνά δυσάρεστη την εμπειρία της αλληλεπίδρασης του χρήστη με το λογισμικό. Για παράδειγμα πολλοί χρήστες ενοχλούνται με την χρήση πολλών αναδυόμενων παραθύρων με διαφημιστικό περιεχόμενο, η οποία τους εμποδίζει να συνεχίζουν απρόσκοπτα την περιήγησή τους στο κατάστημα. Άλλοι χρήστες μπορεί να αισθανθούν ενοχλημένοι όταν βομβαρδίζονται από διαφημιστικά e-mail (τα οποία είναι και πιθανό να μαρκαριστούν σαν spam από πολλές εταιρίες) που τους αποστέλλει

το ηλεκτρονικό κατάστημα ανά πολύ τακτά χρονικά διαστήματα. Ακόμα συχνές αλλαγές στην διεπαφή ενός ηλεκτρονικού καταστήματος, αποτελούν αρνητικό παράγοντα, καθώς οι χρήστες δεν μπορούν να εξοικιωθούν με την λειτουργία του και ενδεχομένως να δυσκολεύονται να βρουν κάτι το οποίο ψάχνουν. Αυτοί και αρκετοί ακόμα παράγοντες μεγαλύτερης ή μικρότερης βαρύτητας, μπορούν να αποτελέσουν τροχοπέδη στην εύρυθμη αλληλεπίδραση ενός συστήματος με τους χρήστες και να έχουν ως συνέπεια τον εκνευρισμό των τελευταίων και την αναζήτησή τους για καλύτερες λύσεις.

Εξίσου σοβαρές αρνητικές επιπτώσεις στην εμπειρία ενός χρήστη με ένα ηλεκτρονικό κατάστημα μπορεί να έχουν οι λανθασμένες εντυπώσεις που έχει το σύστημα για αυτόν, με συνέπεια να κάνει δύο τύπους λαθών όσον αφορά τις προτάσεις του.

Ο πρώτος τύπος λάθους είναι το λεγόμενο 'false negative error'. Αυτός αναφέρεται στο λάθος που κάνει το σύστημα (ή καλύτερα την αβλεψία) να μην προτείνει στον χρήστη ένα προϊόν, αν και ο χρήστης θα έδειχνε ιδιαίτερο ενδιαφέρον για αυτό. Η συνέπεια αύτης της λανθασμένης προσέγγισης, είναι η μη βελτίωση της εμπειρίας του χρήστη και η μηδενική συνεισφορά της μοντελοποίησης στην αύξηση των πωλήσεων του καταστήματος, αφού δεν δίνεται στον χρήστη μια πρόταση η οποία με πολλές πιθανότητες θα οδηγούσε σε αγορά.

Ο δεύτερος τύπος λάθους, ο οποίος και έχει σαφώς δυσμενέστερες συνέπειες, είναι το 'false positive error', στην αποφυγή του οποίου πρέπει να δίνεται ιδιαίτερη βαρύτητα. Αύτος ο τύπος λάθους, αφορά σε προϊόντα τα οποία το σύστημα προτείνει στο χρήστη ο οποίος προβαίνει στην αγορά τους και κατόπιν ανακαλύπτει ότι αυτό το προϊόν είτε δεν το χρειαζόταν είτε δεν καλύπτει τις ανάγκες του. Ο χρήστης αισθάνεται ότι το σύστημα τον 'παρέσυρε' σε αυτή την αγορά και μπορεί να νιώθει προδομένος από την επιλογή του, γεγονός το οποίο μειώνει την αξιοπιστία του συστήματος και τείνει να οδηγήσει έναν χρήστη στην αναζήτηση εναλλακτικών λύσεων.

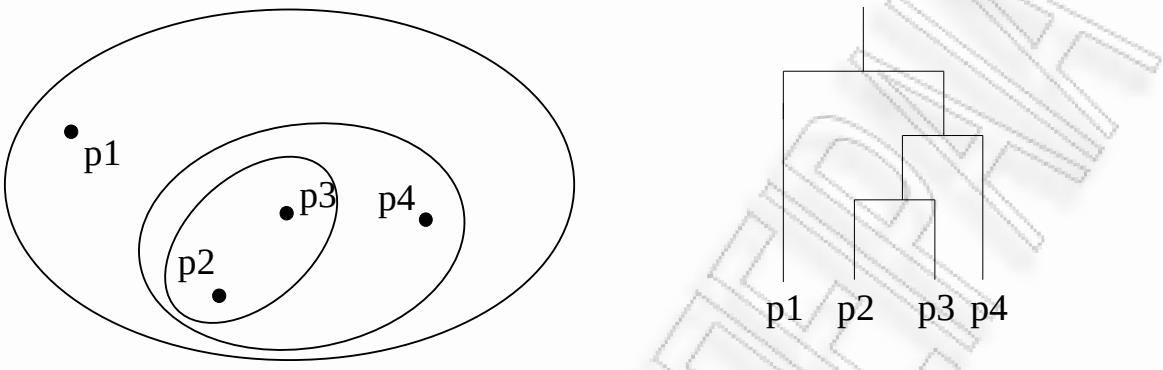
4.4 Ιεραρχική Ομαδοποίηση (Συσταδοποίηση)

Η ομαδοποίηση συχνά καλείται και μάθηση χωρίς επίβλεψη (unsupervised learning) γιατί το σύστημα εκπαιδεύεται από τα χαρακτηριστικά των αντικειμένων (χρηστών) και τα κατηγοριοποιεί αυτόματα σε ομάδες (συστάδες). Δηλαδή δεν υπάρχει εξωτερική επέμβαση για το ποια αντικείμενα θα τοποθετηθούν σε ποια ομαδα. Αυτό που προσδιορίζουμε είναι ο κανόνας βάσει του οποίου θα προκύψουν οι ομάδες ο οποίος εξαρτάται από την απόσταση (ομοιότητα) μεταξύ των αντικειμένων ή των ομάδων. Αντίθετα, στην μάθηση με επίβλεψη (supervised learning or classification, όπως π.χ. LDA Decision tree, Naïve Bayes) γνωρίζουμε εκ των προτέρων την ομαδοποίηση των αντικειμένων και η μηχανή πρέπει να βρει τους καλύτερους κανόνες ομαδοποίησης.
Ως ομάδα ορίζεται μια συλλογή από αντικείμενα τα οποία παρουσιάζουν ομοιότητες μεταξύ τους (βρίσκονται κοντά) και έχουν διαφορές (απέχουν) από αντικείμενα που ανήκουν σε άλλες ομάδες. Η ομοιότητα των αντικειμένων υπολογίζεται βάσει των τιμών των εκάστοτε χαρακτηριστικών που περιγράφουν τα αντικείμενα. Δηλαδή ένα αντικείμενο a είναι ένα διάνυσμα n τιμών : $x = (x_1, x_2, \ldots x_n)$, όπου x_i είναι η τιμή του i-οστού χαρακτηριστικού γνωρίσματος (feature) του αντικειμένου και n η διάσταση του αντικειμένου. Συνήθως το κάθε αντικείμενο εκφράζεται με ένα διάνυσμα των χαρακτηριστικών του και επομένως η ανομοιότητα μεταξύ δύο αντικειμένων μπορεί να προσδιοριστεί με ένα μέτρο απόστασης μεταξύ των αντίστοιχων διανύσματων.

Η ανάλυση ενός συνόλου αντικειμένων σε ομάδες είναι η διαδικασία που αποσκοπεί στο διαχωρισμό του συνόλου αυτού σε υποσύνολα, τέτοια ώστε να υπάρχει ομοιογένεια μέσα σε κάθε υποσύνολο και ανομοιογένεια μεταξύ των στοιχείων που ανήκουν σε ξεχωριστά υποσύνολα. Στην ιεραρχική ομαδοποίηση, οι ομάδες που προκύπτουν από αυτή τη διαδικασία, είναι οργανωμένες ιεραρχικά, έτσι ώστε σε κάθε στάδιο της ιεραρχίας, τα στοιχεία μιας ομάδας να είναι πιο όμοια μεταξύ τους από αυτά που ανήκουν σε άλλη ομάδα. Για να ολοκληρωθεί ο αλγόριθμος δεν χρειάζεται να υποθέσουμε συγκεκριμένο αριθμό ομάδων. Οποιοσδήποτε αριθμός ομάδων μπορεί να επιτευχθεί 'κόβοντας' το δενδρόγραμμα στο κατάλληλο επίπεδο.

Σημαντικό ρόλο στην ομαδοποίηση διαδραματίζει η απόσταση (ή ομοιότητα) μεταξύ των αντικειμένων, βάσει των οποίων δημιουργούνται οι ομάδες. Δύο είναι οι κύριες μεθόδοι

ιεραρχικής ομαδοποίησης: η Συσσωρευτική μέθοδος (Agglomerative Hierarchical Clustering) και η Διαιρετική μέθοδος (Divisive Hierarchical Clustering). Παρακάτω θα αναλυθεί η Συσσωρευτική μέθοδος και οι παραλλαγές της οι οποίες χρησιμοποιήθηκαν για την δημιουργία των ομάδων χρηστών του ηλεκτρονικού καταστήματος.



Εικόνα 4.1: Ιεραρχική Ομαδοποίηση και αναπαράστασή της με δενδρόγραμμα (dendrogram)

4.5 Συσσωρευτική μέθοδος (Agglomerative Hierarchical Clustering)

Η συσσωρευτική μέθοδος είναι μια από κάτω προς τα πάνω (bottom up) προσέγγιση Ιεραρχικής ομαδοποίησης, κατά την οποία οι ομάδες συντίθενται βήμα βήμα ξεκινώντας από τόσες ομάδες όσα και τα στοιχεία που πρόκειται να ομαδοποιηθούν και καταλήγοντας σε μια υπερομάδα η οποία περιέχει όλες τις άλλες ομάδες. Βασικότερη αδυναμία της συσσωρευτικής μεθόδου, (όπως και όλων των ιεραρχικών μεθόδων) είναι ότι απο τη στιγμή που πραγματοποιηθεί ένα βήμα συγχώνευσης αυτό δεν είναι δυνατό να αναιρεθεί. Το θετικό είναι ότι αφού δεν εξετάζεται το σύνολο των πιθανών επιλογών, το υπολογιστικό κόστος μειώνεται σημαντικά.

Τα αποτελέσματα της ιεραρχικής ομαδοποίησης είναι δυνατόν να αναπαρασταθούν διαγραμματικά με ένα ανεστραμμένο δένδρο το οποίο ονομάζεται δενδρόγραμμα (βλ. Εικόνα 3.1). Τα φύλλα του δένδρου απαρτίζονται από τα στοιχεία που θα ομαδοποιηθούν, ενώ οι παράλληλες γραμμές προς τον οριζόντιο άξονα παριστάνουν τις ομάδες.

4.6 Αλγόριθμος Ιεραρχικής Ομαδοποίησης (Συσσωρευτική Μέθοδος)

Η συσσωρευτική μέθοδος είναι η πιο δημοφιλής μέθοδος ιεραρχικής ομαδοποίησης την οποία χρησιμοποιούν και τα περισσότερα στατιστικά πακέτα. Ο αλγόριθμος είναι πολύ απλός και αποτελείται από τέσσερα βήματα:

Βήμα 1:

Ορίζουμε *N* ομάδες, τόσα όσα κι ο αριθμός των στοιχείων που θα ομαδοποιηθούν. Κάθε ομάδα αποτελείται από ένα και μόνο στοιχείο. Επίσης ορίζουμε έναν πίνακα απόστασης (ή ομοιότητας) *N x N*, ο οποίος περιέχει τις αποστάσεις ανά δύο των N στοιχείων. Αν ως d(i,j) ορίζεται η απόσταση μεταξύ των στοιχείων i και j, τότε ισχύουν τα εξής:

- d(i , j) = d(j , i) (1)
- d(i, i) = 0 (2)

Είναι προφανές λόγω των σχέσεων (1) και (2) ότι ο πίνακας απόστασης (Εικ. 3.2) είναι συμμετρικός και όλα τα στοιχεία της διαγωνιου του είναι μηδενικά. Για τον λόγο αυτό στον αλγόριθμό μπορούμε να κάνουμε χρήση μόνο του άνω τριγωνικού πίνακα ή του κάτω τριγωνικού.

Βήμα 2:

Αναζητούμε στον πίνακα αποστάσης το ζεύγος των ομάδων με την μικρότερη απόσταση μεταξύ τους. Αν υποθέσουμε ότι οι ομάδες με τη μικρότερη απόσταση είναι οι U και V, τότε σημειώνουμε την απόσταση d_{UV} και συνεχίζουμε στο επόμενο βήμα.

Βήμα 3:

Συγχωνεύουμε τις ομάδες που επισυμάνθηκαν στο βήμα 2 σε μια νέα ομάδα, έστω T. Ανανεώνουμε τον πίνακα απόστασης διαγράφοντας τις γραμμές και τις στήλες που αντιστοιχούν στις ομάδες U και V και προσθέτουμε μια νέα γραμμή και στήλη στον πίνακα που αντιστοιχεί στη νεοσύστατη ομάδα T. Κατόπιν συμπληρώνουμε τη νέα γραμμή και στήλη με τις αποστάσεις των υπολοίπων ομάδων απο την ομάδα T.

Έστω d_{UT} η ελάχιστη απόσταση μεταξύ όλων των στοιχείων της ομάδας U με όλα τα στοιχεία της ομάδας T. Τότε η νέα απόσταση της ομάδας d_{UT} δίνεται από τη σχέση:

$$d_{UT} = \min(dUi, dTj) \ \forall \ U_i \in U, \ \forall \ j \in T_j$$

Βήμα 4:

Επαναλαμβάνουμε τα βήματα 2 και 3 (N-1) φορές μέχρι να απομείνει μόνο μία ομάδα η οποία θα περιέχει και τα N στοιχεία. Κάθε φορά που φτάνουμε σε αυτό το βήμα καταγράφουμε την ομάδα που δημιουργήθηκε καθώς και την τιμή d_{UV} από το δεύτερο βήμα η οποία είναι το επίπεδο (απόσταση) της ομάδας.

Περιληπτικά με χρήση ψευδοκώδικα ο αλγόριθμος έχει ως εξής:

1: Υπολογισμός του Πίνακα Απόστασης
2: Έστω κάθε σημείο αποτελεί και μια ομάδα
3: ***Repeat***
4: Συγχώνευση των δύο κοντινότερων ομάδων
5: Ενημέρωση του Πίνακα Απόστασης
6: ***Until*** *να μείνει μία μόνο ομάδα*

Dist	A	B	C	D	E	F
A	0.00	0.71	5.66	3.61	4.24	3.20
B	0.71	0.00	4.95	2.92	3.54	2.50
C	5.66	4.95	0.00	2.24	1.41	2.50
D	3.61	2.92	2.24	0.00	1.00	0.50
E	4.24	3.54	1.41	1.00	0.00	1.12
F	3.20	2.50	2.50	0.50	1.12	0.00

Εικόνα 3.2: Πίνακας απόστασης στον οποίο διακρίνονται οι άνω και κάτω τριγωνικοί πίνακες καθώς και με κόκκινο η απόσταση βάσει του οποίου επιλέγονται οι ομάδες στο βήμα 2

Ο παραπάνω αλγόριθμός έχει πολλές παραλλαγές, οι οποίες οφείλονται στις πολλές διαφορετικές μεθόδους που υπάρχουν για τον υπολογισμό της απόστασης μεταξύ δύο ομάδων. Από τις μεθόδους αυτές θα εξετάσουμε τέσσερις από τις πλέον δημοφιλείς οι οποίες έχουν υλοποιηθεί και εφαρμοστεί στην διαδικασία ομαδοποίησης των χρηστών του ηλεκτρονικού μας καταστήματος.

4.6.1 Μέθοδος του ελαχίστου (Min ή Single Linkage)

Η μέθοδος single linkage, γνωστή και ως ομαδοποίηση κοντινότερου γείτονα (nearest-neighbor clustering), έχει πολλές επιθυμητές θεωρητικές ιδιότητες, ωστόσο στην πράξη φαίνεται να μην

έχει το ίδιο επιθυμητα αποτελέσματα με τις άλλες μεθόδους. Στην μέθοδο αυτή, ως απόσταση μεταξύ δύο ομάδων θεωρούμε την απόσταση με την μικρότερη τιμή (από ολους τους πιθανούς συνδυασμούς στοιχείων μεταξύ των δύο ομάδων) και δίνεται από τον τύπο:

$$D_{KL} = \min_{\substack{i \in C_K \\ j \in C_L}} d(x_i, x_j)$$

Το βασικό μειονέκτημα της μεθόδου του ελαχίστου είναι η λεγόμενη επίπτωση αλυσίδας (chaining effect), η οποία οφείλεται στον τοπικό χαρακτήρα του κριτηρίου. Έτσι η μέθοδος είναι ευαίσθητη σε δεδομένα θορύβου τα οποία μπορεί να οδηγήσουν στην δημιουργία μεγάλων ομάδων. Η μέθοδος ευνοεί την εύρεση μη ελλειπτικών ομάδων (non-elliptical shape clusters). Με την χρήση κατάλληλων δομών δεδομένων, η μέθοδος μπορεί να ολοκληρωθεί σε πολυωνυμικό χρόνο $O(n^2)$, όπου n είναι το πλήθος των αντικειμένων.

4.6.2 Μέθοδος πλήρους συνδεσιμότητας (Max ή Complete Linkage)

Στη μέθοδο complete linkage η απόσταση μεταξύ δύο ομάδων είναι εκείνη με την μεγαλύτερη τιμή από όλες τις πιθανές αποστάσεις μεταξύ όλων των στοιχείων των ομάδων αυτών. Η απόσταση υπολογίζεται από τον ακόλουθο τύπο:

$$D_{KL} = \max_{\substack{i \in C_K \\ j \in C_L}} d(x_i, x_j)$$

Ενώ η μέθοδος πλήρους συνδεσιμότητας δεν πάσχει από το *chaining effect* το οποίο εμφανίζεται στη μέθοδο single linkage, εντούτοις είναι περισσότερο ευαίσθητο σε ακραίες τιμές (outliers). Ανεξάρτητα από αυτό, έχει παρατηρηθεί ότι η συγκεκριμένη μέθοδος παράγει καλύτερες και πιο συμπυκνωμένες ομάδες από τις αντίστοιχες της μεθόδου single linkage. Στην χειρότερη περίπτωση ο αλγόριθμος έχει πολυπλοκότητα $O(n^2 \log n)$, όπου n το πλήθος των αντικειμένων.

4.6.3 Μέθοδος Μέσου όρου ομάδας (Average Linkage)

Η μέθοδος average linkage, αποτελεί την μέση οδό μεταξύ των δύο παραπάνω μεθόδων αφού προσπαθεί να συμβιβάσει την ευαισθησία της complete linkage στις ακραίες τιμές και την τάση της single linkage να σχηματίζει μεγάλες μη σφαιρικές ομάδες που δεν αντανακλούν ακριβώς την πραγματικότητα. Ως απόσταση μεταξύ των δύο ομάδων θεωρούμε την μέση απόσταση τους η οποία δίνεται από τον τύπο:

$$D_{KL} = \frac{1}{n_K n_L} \sum_{i \in C_K} \sum_{j \in C_L} d(x_i, x_j)$$

Η μέθοδος average linkage έχει την τάση να δημιουργεί ομάδες με μικρές αποκλίσεις, ωστόσο λόγω του ότι στον υπολογισμό της απόστασης λαμβάνουν μέρος όλα τα σημεία μιας ομάδας, έχει την τάση να επηρεάζεται λιγότερο από τις δύο προηγούμενες ομάδες από τα ακραία σημεία. Η πολυπλοκότητα του αλγορίθμου είναι $O(n^2 \log n)$, όπου n το πλήθος των αντικειμένων.

4.6.4 Μέθοδος Ward

Στη μέθοδο Ward η απόσταση μεταξύ δύο ομάδων ορίζεται ως η αύξηση στο άθροισμα των τετραγωνικών αποκλίσεων (αποστάσεων) κάθε στοιχείου των ομάδων, που προκύπτει από την ένωση των ομάδων αυτών. Στα θετικά της μεθόδου ward καταγράφεται το γεγονός ότι έχει την

τάση να δημιουργεί ομάδες με την μικρότερη δυνατή διακύμανση μεταξύ των μελών της ομάδας, όπως και το ότι δημιουργεί ομάδες με παρόμοιο αριθμό μελών (ισοπληθείς).

$$D_{KL} = \frac{\left\| \bar{x}_K - \bar{x}_L \right\|^2}{\left(\frac{1}{n_K} + \frac{1}{n_L} \right)}$$

Η μέθοδος Ward μπορεί να χρησιμοποιηθεί και στην επιλογή της αρχικής ομάδας (seed) για τον αλγόριθμο k-means.

4.7 Πίνακας απόστασης (ή ομοιότητας)

Σημαντικό ρόλο, όπως είδαμε παραπάνω, παίζει στους αλγόριθμους ιεραρχικής ομαδοποίησης που αναλύουμε, παίζει η απόσταση μεταξύ αντικειμένων ή ομάδων. Όλοι οι αλγόριθμοι δέχονται ως είσοδο, έναν πίνακα απόστασης (ή ομοιότητας), ο οποίος είναι ένας συγκεντρωτικός πίνακας (aggregated matrix), των επιμέρους πίνακων απόστασης που υπολογίζονται από τα χαρακτηριστικά στοιχεία των αντικειμένων και συγκεκριμένα για το ηλεκτρονικό μας κατάστημα, των χρηστών.

Συνήθως το πρόβλημα στον υπολογισμό του τελικού πίνακα ο οποίος θα τροφοδοτήσει τον εκάστοτε αλγόριθμο, είναι η ανομοιογένεια των χαρακτηριστικών γνωρισμάτων (multivariate data), τα οποία πριν προστεθούν μεταξύ τους πρέπει να ομογενοποιηθούν.

4.7.1 Ορισμός απόστασης - ομοιότητας

Η απόσταση εκφράζει στην ουσία την ανομοιότητα μεταξύ δύο αντικειμένων και έχει τις εξής ιδιότητες:

- $d_{ij} \geq 0$, η απόσταση μεταξύ δύο σημείων είναι πάντα θετική ή 0
- $d_{ii} = 0$, (ανακλαστική ιδιότητα)
- $d_{ij} = d_{ji}$ (συμμετρική ιδιότητα)
- $d_{ij} \leq d_{ik} + d_{jk}$ (τριγωνική ανισότητα)

Ομοιότητα ονομάζεται το μέγεθος που αντανακλά τη δύναμη της σχέσης μεταξύ δύο αντικειμένων. Η ομοιότητα είναι ένα μέγεθος το οποίο είναι δύσκολο να αποτιμηθεί και η δυσκολία αυτή έγγειται στο γεγονός, πως τα χαρακτηριστικά ενός αντικειμένου μπορεί να μην είναι μετρήσιμα π.χ. με συνεχείς μεταβλητές. Για παράδειγμα έννοιες όπως χρώμα, σχήμα κτλ, είναι χαρακτηριστικά τα οποία πρέπει να υποστούν κάποια επεξεργασία προκειμένου να είναι μετρήσιμα και έτσι να μπορούμε να υπολογίσουμε την απόσταση μεταξύ τους.

Έστω ότι με s_{ij} συμβολίζεται η ομοιότητα και με δ_{ij} η ανομοιότητα μεταξύ δύο αντικειμένων.

Εάν $s_{ij} \in [0,1]$, τότε ισχύει η σχέση: $s_{ij} = 1 - \delta_{ij}$ με $\delta_{ij} \in [0,1]$. Επομένως αφού η ανομοιότητα είναι σχετικά εύκολο να υπολογιστεί, από την παραπάνω σχέση είμαστε σε θέση να υπολογίσουμε και την ομοιότητα μεταξύ των γνωρισμάτων δύο αντικειμένων.

Γενικά αν είναι εφικτή η μέτρηση της ομοιότητας ή ανομοιότητας μεταξύ δύο αντικειμένων τότε είμαστε σε θέση:

- Να διαχωρίσουμε (ποιοτικά) τα αντικείμενα μεταξύ τους
- Να τα ομαδοποιήσουμε (π.χ. με hierarchical clustering ή k-means clustering)
- Αφού ομαδοποιήσουμε τα δεδομένα να δώσουμε ένα χαρακτηρισμό στην ομάδα.
- Να εντάξουμε ένα νέο αντικείμενο σε μια από τις υπάρχουσες ομάδες
- Να προβλέψουμε τη συμπεριφορά ενός νέου αντικειμένου

- Να κάνουμε data mining.
- Να ανακαλύψουμε κάποια δομή μέσα στο σύνολο των δεδομένων.
- Και τέλος να πάρουμε αποφάσεις και να κάνουμε σχεδιασμό βασιζόμενοι στην δομή των δεδομένων και τις προβλέψεις που μπορούμε να εξάγουμε από αυτα.

4.7.2 Διαδικασία υπολογισμού πίνακα απόστασης

Για να υπολογίσουμε τον πίνακα απόστασης ενός συνόλου αντικειμένων τα οποία αποτελούνται από πολυμεταβλητά δεδομένα ακολουθούμε τα εξής βήματα:

Για κάθε γνώρισμα:

1) Μετατροπή των δεδομένων εισόδου σε συντεταγμένες βάσει της κλίμακας μέτρησης
2) Υπολογισμός πίνακα απόστασης του γνωρίσματος βασιζόμενοι στις συντεταγμένες του πρώτου βήματος
3) Κανονικοποίηση του πίνακα στο εύρος τιμών [0,1]
4) Προσθέτουμε τον πίνακα απόστασης του γνωρίσματος σε έναν συγκεντρωτικό πίνακα απόστασης.

Αν όλα τα δεδομένα είναι ισοβαρή, δηλαδή έχουν την ίδια βαρύτητα, τότε στο τέλος της διαδικασίας διαιρούμε όλα τα στοιχεία του πίνακα απόστασης με το πλήθος των γνωρισμάτων έτσι ώστε και ο συγκεντρωτικός πίνακας να είναι κανονικοποιημένος, δηλαδή όλες οι τιμές του να ανήκουν στο διάστημα [0,1].

4.7.3 Πολυμεταβλητά χαρακτηριστικά γνωρίσματα

Όπως αναφέρθηκε παραπάνω, για να υπολογιστεί η απόσταση μεταξύ δύο αντικειμένων, τα οποία έχουν ανομοιογενή χαρακτηριστικά γνωρίσματα, θα πρέπει αυτά τα δεδομένα να αναπαρασταθούν σε μετρήσιμα μεγέθη. Επίσης ο πίνακας απόστασης ενδέχεται να επηρεαστεί από δεδομένα τα οποία θα έχουν διαφορετικές μονάδες μέτρησης κι έτσι η ομαδοποίηση των αντικειμένων να μην είναι ακριβής. Λύση στο παραπάνω πρόβλημα μπορεί να δώσει η κανονικοποίηση των δεδομένων. Τα χαρακτηριστικά γνωρίσματα ενός αντικειμένου διακρίνονται με βάση το είδος των τιμών που παίρνουν σε:

1) Ποσοτικά (Quantitative features)
    i) Συνεχείς τιμές (continuous values π.χ. βάρος, ύψος)
    ii) Διακριτές τιμές (discrete values π.χ. πλήθος αντικειμένων)
    iii) Τιμές Διαστημάτων (interval values, π.χ. διάρκεια γεγονότος)
2) Ποιοτικά (Qualitative features)
    i) Ονομαστικά (nominal or unordered (π.χ. χρώμα)
    ii) Κατηγορικά (ordinal, π.χ. ζεστό ή κρύο, ranking (1 to 5)
    iii) Δυαδικά (binary, π.χ. Ναι ή Όχι, Συμφωνώ ή Διαφωνώ κτλ)

4.7.4 Μεθόδοι μέτρησης απόστασης για Συνεχείς Μεταβλητές

Όλες οι τιμές που μετρούν συνεχή μεγέθη όπως χρόνος, βάρος, τιμή κτλ, μπορούν να αναπαρασταθούν ως σημεία στον χώρο και κατά συνέπεια η απόστασή τους μπορεί να μετρηθεί με τις γνωστές μεθόδους μέτρησης απόστασης. Μερικές από τις πλέον γνωστές μεθόδους μέτρησης απόστασης συνεχών μεταβλητών είναι οι ακόλουθες.

- Ευκλείδια απόσταση, η οποία δίνεται από τη σχέση $d_{ij} = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2}$

- City block (Manhattan) distance η οποία υπολογίζεται με τη σχέση $d_{ij} = \sum_{k=1}^{n} \left| x_{ik} - x_{jk} \right|$
- Chebyshev distance που δίνεται από τη σχέση $d_{ij} = \max_{k} \left| x_{ik} - x_{jk} \right|$
- Απόσταση Minkowski τάξης λ, δίνεται από τον τύπο $d_{ij} = \sqrt[l]{\sum_{k=1}^{n} (x_{ik} - x_{jk})^{l}}$. Η απόσταση Minkowski είναι η γενικότερος τύπος απόστασης αφού για λ=1 προκύπτει ο τύπος City block distance, για λ=2 προκύπτει ο τύπος της Ευκλείδιας απόστασης ενώ για $\lambda=\infty$ έχουμε τον τύπο Chebyshev distance.

Άλλες λιγότερο δημοφιλείς μεθόδοι είναι οι:

- Canberra distance
- Bray Curtis distance (Sorensen distance)
- Angular separation
- Correlation coefficient

4.7.5 Μεθόδοι μέτρησης ομοιότητας για Δυαδικές Μεταβλητές

Συχνά ερχόμαστε αντιμέτωποι με μεταβλητές οι οποίες παίρνουν μόνο δύο τιμές. Παράδειγμα τέτοιων μεταβλητών είναι αυτές που απαντούν σε ερωτήματα της φύσεως Ναι ή Όχι, True or False, Επιτυχία ή Αποτυχία, Θετικό ή Αρνητικό, 0 ή 1 και ονομάζονται δυαδικές μεταβλητές (binary).

Η ομοιότητα ή ανομοιότητα (απόσταση) αυτών των μεταβλητών μπορούν να μετρηθούν με όρους συχνότητας εμφάνισης των τιμών αυτών. Ας παραθέσουμε ένα παράδειγμα ενός αντικειμένου που έχει ως γνωρίσματα τρεις δυαδικές μεταβλητές και να υπολογίσουμε την ανομοιότητα δύο στιγμιοτύπων του αντικειμένου αυτού.

Έστω λοιπόν το αντικείμενο ‘φρούτο’ το οποίο αποτελείται από τις ιδιότητες [Γλυκό, Σφαιρικό, Τραγανό]. Στον πίνακα που ακολουθεί φαίνονται δύο στιγμιότηπα της τάξης αυτής.

Αντικείμενο	Γλυκό	Σφαιρικό	Τραγανό
i = Μήλο	Ναι	Ναι	Ναι
j = Μπανάνα	Ναι	Όχι	Όχι

Πίνακας 3.1. Παράδειγμα αντικειμένων με γνωρίσματα δυαδικες μεταβλητές

Από τον παραπάνω πίνακα προκύπτουν τα εξής διανύσματα για τα δύο αντικείμενα:

Μήλο = [1,1,1]

Μπανάνα = [1,0,0]

Όπου η τιμή ‘Ναι’ ισούται με 1 και η τιμή ‘Όχι’ με 0.

Έστω:

p: ο αριθμός των μεταβλητών με τιμή 1 ταυτόχρονα και για τα δύο αντικείμενα.

q: ο αριθμός των μεταβλητών με τιμή 1 για το αντικείμενο i και τιμή 0 για το αντικείμενο j.

r: ο αριθμός των μεταβλητών με τιμή 0 για το αντικείμενο i και τιμή 1 για το αντικείμενο j.

s: ο αριθμός των μεταβλητών με τιμή 0 ταυτόχρονα και για τα δύο αντικείμενα.

t: το συνολικό πλήθος, δηλαδή t = p + q + r + s.

Οι βασικοί τύποι για μέτρηση της απόστασης για αντικείμενα με δυαδικές μεταβλητές είναι:

- Simple Matching Distance: $d_{ij} = \frac{q+r}{t}$

- Jaccard's Distance: $d_{ij} = \frac{q+r}{p+q+r}$
- Hamming Distance: $d_{ij} = q + r$

Για τα παραπάνω δεδομένα λοιπόν έχουμε: p=1, q=2, r=0, s=0 οπότε t=3. Άρα με βάση τον τύπο του Hamming έχουμε $d_{ij} = q + r = 2$

Επίσης ο B. S. Everit (1978) πρότεινε άλλους δέκα τρόπους μέτρησης αυτού του τύπου των μεταβλητών:

1. $d_{ij} = \frac{p+s}{t}$ Simple Matching Coefficient
2. $d_{ij} = \frac{p}{t}$
3. $d_{ij} = \frac{p}{p+q+r}$ Jaccard's Coefficient
4. $d_{ij} = \frac{2p}{2p+q+r}$
5. $d_{ij} = \frac{2(p+s)}{2(p+s)+q+r}$
6. $d_{ij} = \frac{p+s}{p+s+2(q+r)}$
7. $d_{ij} = \frac{p}{p+s+2(q+r)}$
8. $d_{ij} = \frac{p}{p+2(q+r)}$
9. $d_{ij} = \frac{p}{q+r}$
10. $d_{ij} = \frac{p+s}{q+r}$

Η μέθοδος που χρησιμοποιήσαμε για την μέτρηση της απόστασης δυαδικών μεταβλητών στην εφαρμογή μας είναι η μέθοδος του Hamming η οποία αναφέρθηκε παραπάνω.

Σύμφωνα με την μέθοδο αυτή αν σχηματίζουμε μία λέξη (word) από ψηφία 0 και 1 που έχουν το ίδιο μήκος, στην οποία κάθε ψηφίο συμβολίζει ένα δυαδικό γνώρισμα, τότε μπορούμε να μετρήσουμε σε πόσες θέσεις τα ψηφία διαφέρουν μεταξύ των δύο λέξεων.

4.7.6 Μεθόδοι μέτρησης απόστασης για κατηγορικές (ordinal) μεταβλητές

Σε αυτή την κατηγορία μπορούν να ενταχθούν και οι ονομαστικές μεταβλητές. Η διαφορά μεταξύ κατηγορικών (ordinal) και ονομαστικών (nominal) μεταβλητών είναι ότι οι ονομαστικές εκφράζουν καλύτερα μεγέθη επιλογής χωρίς σειρά, ενώ οι κατηγορικές δίνουν μεγαλύτερη έμφαση στη σειρά.

Παράδειγματα που περιγράφονται από αυτού του τύπου τις μεταβλητές είναι τα παρακάτω:

- Rating: -2=Strongly Disagree, -1=Disagree, 0=Neutral, 1=Agree, 2=Strongly Agree
- Priority: 1=Καλύτερος, μεγαλύτερος αριθμός = χαμηλότερη σημασία
- Ordering: Ακολουθία τιμών με βάση τις ετικέτες (π.χ. Ταξίαρχος, Συντ/ρχης...)

Οι κυριότερες μεθόδοι υπολογισμού της απόστασης αυτών των μεταβλητών είναι οι εξής:

- Normalize Rank Transformation
- Spearman Distance
- Footrule Distance
- Kendall Distance
- Cayley Distance
- Hamming Distance
- Ulam Distance
- Chebychev Distance

Στην παρούσα εργασία για την μέτρηση της απόστασης χρησιμοποιήθηκε η πρώτη μέθοδος, δηλαδή η Normalize Rank Transformation. Οι κατηγορικές μεταβλητές μπορούν να μετασχηματιστούν σε συνεχείς (ποσοτικές) μέσω κανονικοποίησης. Μετά την κανονικοποίησή τους μπορούν να λογίζονται ως συνεχείς και κατά συνέπεια η απόσταση τους μπορεί να μετρηθεί με τους τρόπους που υπολογίζονται οι συνεχείς. Αναλυτικά τα βήματα που ακολουθούμε είναι τα παρακάτω:

- Μετατρέπουμε την μεταβλητή σε Rank (r=1, R= max value)
- Κανονικοποιούμε το rank σε μία τιμή μεταξύ [0,1] με τον τύπο x = $\frac{r-1}{R-1}$

Παράδειγμα:

Ας πούμε ότι έχουμε μία μεταβλητή τύπου Rating που παίρνει τιμές [-2, -1, 0, 1, 2].

Έχουμε αρχική σειρά: [-2, -1, 0, 1, 2]. Την μετατρέπουμε σε Rank μεταβλητή αρχίζοντας να μετράμε από το 1, δηλαδή έχουμε την ακολουθία: [1, 2, 3, 4, 5]. Από την τελευταία ακολουθία προκύπτει ότι r=1 και R=5. Άρα η κανονικοποιημένη ακολουθία είναι [0, ¼, ½, ¾, 1]. Έτσι για να υπολογίσουμε την απόσταση μεταξύ δύο κατηγορικών μεταβλητών δεν έχουμε παρά να χρησιμοποιήσουμε τις μετρικές που υπάρχουν για τις συνεχείς μεταβλητές όπως π.χ. είναι η ευκλείδια απόσταση.

5. ΠΕΡΙΓΡΑΦΗ ΕΦΑΡΜΟΓΗΣ

Στο κομμάτι της υλοποίησης της συγκεκριμένης διατριβής, σαν βασικός στόχος ήταν η δημιουργία ενός ηλεκτρονικού καταστήματος με χρήση ενός αλγορίθμου Clustering, ο οποίος ομαδοποιεί τους χρήστες ανάλογα με τις κινήσεις τους μέσα στον ιστότοπο. Για την υλοποίηση αυτή χρησιμοποιήθηκαν οι γλώσσες προγραμματισμού php, html, xml, javascript οι οποίες είναι από τις πιο διαδεδομένες γλώσσες προγραμματισμού web εφαρμογών με ανοιχτό κώδικα. Επίσης χρησιμοποιήθηκε και η MySql για την δημιουργία μιας βάσης δεδομένων που θα υποστηρίζει το ηλεκτρονικό μας κατάστημα.

Στο συγκεκριμένο κεφάλαιο περιγράφονται οι τεχνικές προγραμματισμού που εφαρμόστηκαν όπως η ajax, προγραμματιστικά εργαλεία που χρησιμοποιήθηκαν για την υλοποίηση της συγκεκριμένης εφαρμογής καθώς και εργαλεία που βοήθησαν στην εξαγωγή στατιστικών και άλλων συμπερασμάτων όπως το matlab.

5.1 Πλατφόρμα και Προγραμματιστικά Εργαλεία

Σε αυτή την ενότητα περιγράφονται τα εργαλεία και ο λόγος για τον οποίο επιλέχθηκαν, για την υλοποίηση του ηλεκτρονικού καταστήματος.

5.1.1 Επιλογή Γλώσσας Προγραμματισμού

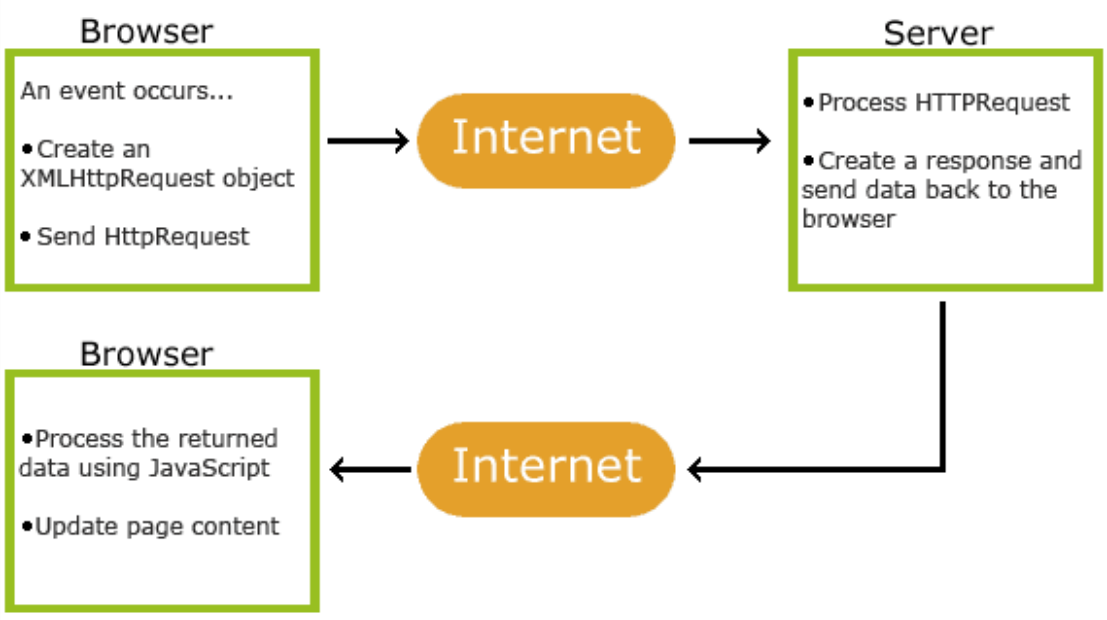
Οι γλώσσες προγραμματισμού που επιλέχθηκαν είναι ανοιχτού κώδικα για την διευκόλυνσή μας στην εύρεση εργαλείων και πλατφορμών από το διαδίκτυο. Παρακάτω θα περιγράψουμε μία μία τις γλώσσες και κάποιες σύγχρονες μεθόδους υλοποίησης που χρήζουν αναφοράς.

1. Το μεγαλύτερο κομμάτι της εφαρμογής μας υλοποιήθηκε με την PHP 5 η οποία είναι μία γλώσσα προγραμματισμού για τη δημιουργία δυναμικών ιστοσελίδων όπως είναι το ηλεκτρονικό μας κατάστημα. Μια σελίδα PHP για να εμφανιστεί χρειάζεται έναν server τύπου apache (στον οποίο θα αναφερθούμε αργότερα) για την επεξεργασία των εντολών και από έναν φυλλομετρητή ο οποίος απεικονίζει τα αποτελέσματα του apache server. Τα βασικότερα πλεονεκτήματα της συγκεκριμένης γλώσσας είναι ο κύριος λόγος για τον οποίο χρησιμοποιήθηκε και είναι τα ακόλουθα:
    a. Επειδή η PHP είναι μια γλώσσα που εκτελείται από την πλευρά του server, ο χρήστης δεν χρειάζεται να έχει εγκατεστημένη καμία ειδική μηχανή αναζήτησης ή plug-ins για να δεί το αποτέλεσμα του κώδικα.
    b. Τρέχει σε όλα τα λειτουργικά συστήματα Mac,Windows,Linux κ.α., καθώς και όλους τους φυλλομετρητές όπως είναι ο internet explorer,Mozilla firefox, saphari, opera, google chrome κ.α.
    c. Επειδή είναι script γλώσσα δεν χρειάζεται κάποιον compiler παρά μόνο ένα notepap στο οποίο γράφεις τον κώδικα και στο τέλος το αποθηκεύεις με την κατάληξη .php .
    d. Είναι ανοιχτού κώδικα με αποτέλεσμα να υπάρχουν πολλά παραδείγματα και manual από την κοινότητα της php, τα οποία διατίθενται δωρεάν.
2. Η HTML σε συνδυασμό με τα CSS(Cascading Style Sheets) είναι η γλώσσες χαρακτηρισμού με τις οποίες δώσαμε σχήμα και χρώμα στο ηλεκτρονικό μας κατάστημα. Χρησιμοποιόντας την html δημιουργήσαμε τη φόρμα εγγραφής στην ιστοσελίδα, τις φόρμες ψηφοφορίας και γενικά της εμφάνισης της ιστοσελίδας.
3. Η javascript είναι άλλη μια γλώσσα προγραμματισμού script η οποία όμως σε αντίθεση με την php τρέχει στον φυλλομετρητή και όχι στον server. Αυτό σημαίνει ότι η επεξεργασία του κώδικα Javascript και η παραγωγή του τελικού περιεχομένου HTML δεν πραγματοποιείται στον server, αλλά στο πρόγραμμα περιήγησης των επισκεπτών. Αυτή η διαφορά έχει και πλεονεκτήματα και μειονεκτήματα για καθεμιά από τις δύο γλώσσες. Συγκεκριμένα, η Javascript δεν έχει καμία απαίτηση από πλευράς δυνατοτήτων του server για να εκτελεστεί (επεξεργαστική ισχύ, συμβατό λογισμικό διακομιστή), αλλά βασίζεται στις δυνατότητες του φυλλομετρητή των επισκεπτών. Επίσης μπορεί να ενσωματωθεί σε στατικές σελίδες HTML. Παρόλα αυτά, οι δυνατότητές της είναι σημαντικά μικρότερες από αυτές της PHP και δεν παρέχει συνδεσιμότητα με βάσεις δεδομένων.
4. Για την διαχείριση των δεδομένων τις ιστοσελίδας (πελάτες, προϊόντα, αγορές) και πολλά που θα αναφέρουμε σε άλλη παράγραφο, χρησιμοποιήσαμε άλλη μια γλώσσα ανοιχτού κώδικα, την MySql. Η MySql είναι μία γλώσσα υπολογιστών στις βάσεις δεδομένων, που σχεδιάστηκε για τη διαχείριση δεδομένων, σε ένα σύστημα διαχείρισης σχεσιακών βάσεων δεδομένων (Relational Database Management System, RDBMS) και η οποία, αρχικά, βασίστηκε στη σχεσιακή άλγεβρα. Η γλώσσα περιλαμβάνει δυνατότητες ανάκτησης και ενημέρωσης δεδομένων, δημιουργίας και τροποποίησης σχημάτων και σχεσιακών πινάκων, αλλά και ελέγχου πρόσβασης στα δεδομένα.
5. Η XML (αγγλ. αρκτ. από το Extensible Markup Language) είναι μία γλώσσα σήμανσης, που περιέχει ένα σύνολο κανόνων για την ηλεκτρονική κωδικοποίηση κειμένων. Ορίζεται, κυρίως, στην προδιαγραφή XML 1.0 (XML 1.0 Specification), που δημιούργησε ο διεθνής οργανισμός προτύπων W3C (World Wide Web Consortium), αλλά και σε διάφορες άλλες σχετικές προδιαγραφές ανοιχτών προτύπων.

    Η XML σχεδιάστηκε δίνοντας έμφαση στην απλότητα, τη γενικότητα και τη χρησιμότητα στο Διαδίκτυο. Είναι μία μορφοποίηση δεδομένων κειμένου, με ισχυρή υποστήριξη Unicode για όλες τις γλώσσες του κόσμου. Αν και η σχεδίαση της XML εστιάζει στα κείμενα, χρησιμοποιείται ευρέως για την αναπαράσταση αυθαίρετων δομών δεδομένων, που προκύπτουν για παράδειγμα στις υπηρεσίες ιστού.

6. Τέλος θα αναφέρουμε και μία μέθοδο προγραμματισμού επονομαζόμενη ως Ajax η οποία είναι πολύ διαδεδομένη στο διαδίκτυο. Η Ajax(Asynchronous Javascript and XML, είναι ένα σύνολο τεχνικών για την δημιουργία άκρως διαδραστικών ιστοσελίδων και εφαρμογών στο ίντερνετ. Με την χρήση της Ajax τα δεδομένα μπορούν να μεταφερθούν ασύγχρονα μέσω του αντικειμένου XMLHttpRequest και η σελιδα να συνεχιζει την φόρτωση της χωρίς να περιμένει απάντηση από τον server. Η Ajax είναι ένας από τους πρόσφατους και σημαντικότερους τρόπους βελτίωσης της online εμπειρίας των χρηστών και δημιουργίας νέας και καινοτομικής λειτουργικότητας web. Επιτρέποντας σε συγκεκριμένα τμήματα μιας ιστοσελίδας να προβάλλονται χωρίς ανανέωση ολόκληρης της σελίδας, βελτιώνει σημαντικά την εμπειρία των εφαρμογών web. Επιτρέπει επίσης στους προγραμματιστές web να δημιουργούν διαισθητικές και καινοτομικές διαδραστικές διαδικασίες

   Η τεχνολογία Ajax, αν και αρκετά νέα, έχει αρχίσει και χρησιμοποιείται από όλο και περισσότερες δικτυακές εφαρμογές κερδίζοντας συνεχώς έδαφος. Πρώτη η Google, η πασίγνωστη μηχανή αναζήτησης, δημιούργησε αρκετή εντύπωση χρησιμοποιώντας την Ajax στην δοκιμαστική υπηρεσία Google Suggest. Έπειτα ακολούθησαν και πάρα πολλές άλλες υπηρεσίες με σημαντικότερες τις Google Maps, Gmail, Youtube, Facebook κ.α. που χρησιμοποιούν σε μεγάλο βαθμό την τεχνολογία Ajax. Επίσης ο νέος ιστότοπος Microsoft Live (http://www.live.com) με υπηρεσίες αντίστοιχες της Google, καθώς και η τελευταία έκδοση της σελίδας της Yahoo (http://www.yahoo.com), χρησιμοποιούν εκτενώς Ajax.
   Το Google Spreadsheets είναι μια νέα εφαρμογή από τα Google Labs και μπορεί να βρεθεί στη διεύθυνση http://spreadsheets.google.com. Πρόκειται για μια web εφαρμογή λογιστικών φύλλων με κυριότερο χαρακτηριστικό της ότι είναι φτιαγμένη εξ ολοκλήρου με τη χρησιμοποίηση της Ajax, γεγονός που την καθιστά άκρως ανταγωνιστική. Η τεχνολογία Ajax περιλαμβάνει τα εξής:
   • Παρουσίαση χρησιμοποιώντας XHTML και CSS, με βάση τα standards
   • Δυναμική παρουσίαση μέσω του Document Object Model
   • Ανταλλαγή και διαχείριση πληροφοριών χρησιμοποιώντας XML
   • Ασύγχρονη μεταφορά δεδομένων μέσω του XMLHttpRequest
   • Και την Javascript που τα δένει όλα μαζί



Εικόνα 5.1: Πως δουλεύει η AJAX (Πηγή: http://www.gtsamis.gr/?p=1057)

5.1.2 Επιλογή Λειτουργικού Συστήματος

Έχοντας να δουλέψουμε με τις παραπάνω γλώσσες είχαμε την επιλογή να αναπτύξουμε τον κώδικά μας σε οποιοδήποτε περιβάλλον και λειτουργικό σύστημα θέλαμε. Επιλέχθηκαν τα

Windows XP της εταιρίας Microsoft λόγω της συμβατότητας που προσφέρουν, αφού αποτελούν μακράν την πιο διαδεδομένη πλατφόρμα, αλλά και λόγω της πληθώρας των προγραμμάτων και εργαλείων για ανάπτυξη κώδικα που διατίθενται σε αυτά.

Εδώ να αναφέρουμε ότι ο παραγόμενος κώδικας είναι πολύ εύκολο να χρησιμοποιηθεί σε όλα τα λειτουργικά συστήματα, αρκεί βέβαια να είναι εγκατεστημένος ένας apache server,η php, και η mysql τα οποία εγκαθίστανται πολύ εύκολα και χωρίς ειδικές γνώσεις με την πλατφόρμα wamp.

5.1.3 Επιλογή Εργαλείου Ανάπτυξης Κώδικα και σχεδιασμού βάσης δεδομένων

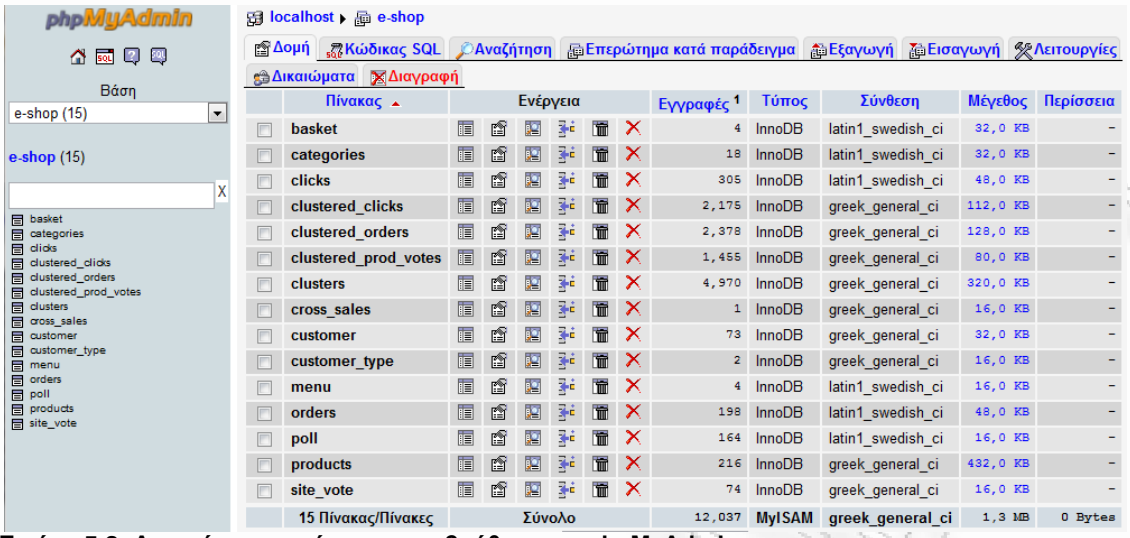
Για την ανάπτυξη του κώδικα χρησιμοποιήθηκε το εργαλείο notepad++ στο οποίο μπορούν να συνταχθούν όλες οι παραπάνω γλώσσες που αναφέρθηκαν. Για να λειτουργήσει όμως ο κώδικας χρειάζεται εγκατάσταση μίας ομάδας server για την μετάφραση του κώδικα και την δημιουργία της βάσης δεδομένων. Για την εγκατάσταση αυτή χρησιμοποιήσαμε τον wampserver. Ο WampServer είναι ένα web Windows περιβάλλον ανάπτυξης ανοικτού κώδικα, το οποίο μας επιτρέπει να δημιουργήσουμε εφαρμογές web με Apache, PHP και βάση δεδομένων MySQL. Επίσης, για την καλύτερη διαχείριση της βάσης δεδομένων έχει ενσωματωμένο το phpmyadmin .

1. Ο Apache HTTP γνωστός και απλά σαν Apache είναι ένας εξυπηρετητής του παγκόσμιου ιστού (web). Όποτε ένας χρήστης επισκέπτεται ένα ιστότοπο το πρόγραμμα πλοήγησης (browser) επικοινωνεί με έναν διακομιστή (server) μέσω του πρωτοκόλλου HTTP, ο οποίος παράγει τις ιστοσελίδες και τις αποστέλλει στο πρόγραμμα πλοήγησης. Ο Apache είναι ένας από τους δημοφιλέστερους, εν μέρει γιατί λειτουργεί σε διάφορες πλατφόρμες όπως τα Windows, το Linux, το Unixκαι το Mac OS X. Συντηρείται τώρα από μια κοινότητα ανοικτού κώδικα με επιτήρηση από το Ίδρυμα Λογισμικού Apache (Apache Software Foundation).

Η MySQL είναι ένα σύστημα διαχείρισης σχεσιακών βάσεων δεδομένων (RDBMS) το οποίο μετρά περισσότερες από 11 εκατομμύρια εγκαταστάσεις. Έλαβε το όνομά του από την κόρη του Μόντυ Βιντένιους, την Μάι. Το πρόγραμμα τρέχει έναν εξυπηρετητή (server) παρέχοντας πρόσβαση πολλών χρηστών σε ένα σύνολο βάσεων δεδομένων.

Το phpMyAdmin είναι ένα δωρεάν εργαλείο, λογισμικό γραμμένο σε PHP που προορίζεται για την διαχείρηση της MySQL μέσω του World Wide Web. Το phpMyAdmin υποστηρίζει ένα ευρύ φάσμα δράσεων με MySQL. Οι πιο συχνά χρησιμοποιούμενες λειτουργίες βρίσκονται ή υποστηρίζονται από το περιβάλλον εργασίας χρήστη (βάσεις δεδομένων διαχείρισης, πίνακες, πεδία, σχέσεις, ευρετήρια, οι χρήστες, δικαιώματα, κ.λπ.), ενώ παρέχεται και η δυνατότητα άμεσης εκτέλεσης μιας δήλωσης SQL.

Ακολουθεί σχηματική αναπαράσταση της βάσης δεδομένων του ηλεκτρονικού καταστήματος που χτίστηκε με τη βοήθεια του phpmyadmin .



Εικόνα 5.2: Απεικόνιση πινάκων με τη βοήθεια του phpMyAdmin.

5.2 Περιγραφή λειτουργιών ηλεκτρονικού καταστήματος

Σε αυτή τη παράγραφο θα περιγράψουμε τις λειτουργίες του ηλεκτρονικού καταστήματος με απεικόνιση τμημάτων κώδικα και εικόνων.

5.2.1 Εγγραφή στο ηλεκτρονικό κατάστημα

Κατά την εγγραφή του χρήστη ζητούνται τα παρακάτω στοιχεία: ονοματεπώνυμο, διεύθυνση, φύλλο, user name, e-mail και κωδικός. Δεν έχει προστεθεί κάποιο ερωτηματολόγιο διότι η ομαδοποίηση που εφαρμόζεται είναι βάσει των κινήσεων τους οι οποίες είναι οι εξής:

- Αγορά(orders) προϊόντων
- Επίσκεψη προϊόντων(clicks)
- Βαθμολογία προϊόντων(product vote)
- Ενδιαφέρον εμφάνισης περισσότερων χαρακτηριστικών σε κάθε προϊόν.(advanced user)
- Βαθμολογία του ηλεκτρονικού καταστήματος

Εικόνα 5.3. Φόρμα εγγραφής στο ηλεκτρονικό κατάστημα

Κάνοντας αριστερό κλικ στο κουμπί Register γίνονται έλεγχοι ορθότητας του e-mail, password και συμπλήρωσης όλων των πεδίων. Αν δεν παρουσιαστεί κάποιο σφάλμα τότε εκτελείται ο παρακάτω κώδικας για την εισαγωγή του πελάτη στη βάση δεδομένων.

```
$query="INSERT INTO customer (username, password, email, gender, fname, lname,
street_name, street_number)
        VALUES ( '$_POST[username]' , '$_POST[password]', '$_POST[email]',
'$_POST[gender]', '$_POST[fname]', '$_POST[lname]', '$_POST[street_name]',
'$_POST[street_number]')";
 $ret = mysql_query($query);
```

5.2.2 Προσθήκη προϊόντος στο καλάθι αγορών και ολοκλήρωση αγορών

Για την προσθήκη ενός προϊόντος στο καλάθι αγορών (Εικ. 4.4) χρειάζεται απλά να πατηθεί ο σύνδεσμος Buy ή όπου υπάρχει το εικονίδιο με σχήμα ενός καροτσιού super market.

	Product Name	Item Price	Count	Total Price	Del
	Core 2 Quad 9300 Tra	223.00 €	1 ADD	223 €	X
	X1550 256MB TVO-DVI	26.00 €	2 ADD	52 €	X
	DFI Lanparty DK P45	159.00 €	1 ADD	159 €	X
	CoolerMaster Storm Sniper w/Window Case	154.00 €	1 ADD	154 €	X
Return to shopping		Confirm Order			
Total Price:		588 €			

Εικόνα 5.4. Παράδειγμα καλαθιού αγορών ενός πελάτη

Στο καλάθι αγορών ο χρήστης έχει τις εξής δυνατότητες:

1. Επιστροφή στην σελίδα του τελευταίου προϊόντος που επέλεξε για το καλάθι αγορών πατώντας τον σύνδεσμο Return to shopping.
2. Επιλογή αριθμού τεμαχίων ανά προϊόν πληκτρολογώντας έναν αριθμό στη στήλη Count και πατώντας το κουμπί ADD βλέποντας αυτομάτως την ανανέωση των τιμών.
3. Αφαίρεση οποιουδήποτε προϊόντος από το καλάθι αγορών πατώντας το X στη στήλη Del
4. Και τέλος μπορεί να ολοκληρώσει την παραγγελία του πατώντας τον σύνδεσμο Confirm Order.

Αν ο χρήστης επιλέξει την ολοκλήρωση της αγοράς γίνεται προσθήκη των προϊόντων στον πίνακα αγορών (orders) στην βάση δεδομένων και γίνεται διαγραφή των προϊόντων από το καλάθι αγορών. Παρακάτω παρουσιάζεται ο αντίστοιχος κώδικας που εκτελείται.

```
$query1="SELECT * From basket where customer_id = $user_id";
$ret1 = mysql_query($query1);
$num_of_rows = mysql_num_rows($ret1);

while( $row1 = mysql_fetch_array($ret1)){
$product_id = $row1['product_id'];
$count = $row1['count'];
```

```
$query = "INSERT INTO orders (cust_id, product_id,count )
        VALUES ('$user_id','$product_id','$count')";
$ret = mysql_query($query);
}
mysql_query("DELETE FROM basket WHERE customer_id= $user_id");
```

5.2.3 Προεπισκόπηση προϊόντων

Όταν ο χρήστης θελήσει να δει τις λεπτομέρειες ενός προϊόντος (δηλαδή κάνει κλικ επάνω στο σύνδεσμο του προϊόντος), αυξάνονται οι μετρητές επισκεψιμότητας του προϊόντος για τον συγκεκριμένο πελάτη στη βάση δεδομένων (πίνακας clicks).

Με αυτό τον τρόπο μπορούν να ελεγχτούν οι προτιμήσεις ενός πελάτη και να εξαχθούν χρήσιμες πληροφορίες από τη μοντελοποίηση που εφαρμόζεται στο ηλεκτρονικό κατάστημα. Ο κώδικας που εκτελείται είναι ο παρακάτω:

```
$query = "SELECT * FROM clicks WHERE customer_id=$user and product_id=$product_id
$ret = mysql_query($query);
$num = mysql_num_rows($ret);
if($num==0)
{
//Εάν δεν έχει ξανά επισκεφτεί το προιον γίνεται εισαγωγή καινούριας εγγραφής
$query = "INSERT INTO clicks (customer_id, cat_id, product_id, clicks_number) VALUES
($user,$cat_id,$product_id, 1)";
$ret = mysql_query($query);
}else{
//Εάν έχει ξανά επισκεφτεί το προιον γίνεται ενημέρωση της υπάρχουσας εγγραφής
προσθέτοντας ένα click στο πεδίο count του πίνακα clicks.
$row = mysql_fetch_array($ret);
$clicks = $row['clicks_number'];
$clicks++;
$query = "UPDATE clicks set clicks_number=$clicks where customer_id=$user and
product_id=$product_id";
$ret = mysql_query($query);
}
```

Εικόνα 5.5: Μεγέθυνση προϊόντος για καλύτερη προεπισκόπηση

Με το που εισέλθει ο χρήστης στην πληροφοριακή σελίδα έχει την δυνατότητα άμεσα να πληροφορηθεί για τα βασικά χαρακτηριστικά του προϊόντος καθώς και την τιμή του. Επίσης αν θέλει να το αγοράσει μπορεί να το προσθέσει στο καλάθι αγοράς πατώντας τον σύνδεσμο Buy now. Επιπρόσθετα μία χρήσιμη λειτουργία είναι η λειτουργία της μεγέθυνσης του προϊόντος (Εικ. 4.5) που προσφέρει μια καλύτερη προεπισκόπησή του στον χρήστη. Τέλος ο χρήστης αν δεν καλύπτεται από την βασική περιγραφή του προϊόντος μπορεί να δει μια αναλυτική περιγραφή του προϊόντος πατώντας το σύνδεσμο Click to see more characteristics (Εικ. 4.6) και αν θέλει να τα αποκρύψει μπορεί να το κάνει πατώντας τον σύνδεσμο Hide Characteristics. Έαν ένας χρήστης έχει πατήσει περισσότερες από 15 φορές τον σύνδεσμο αυτόν τότε η αναλυτική περιγραφή θα είναι ανοιχτή σε κάθε προϊόν από προεπιλογή.

Εικόνα 5.6: Πληροφοριακή σελίδα προϊόντος. Με κόκκινη υπογράμμιση φαίνεται ο σύνδεσμος που εμφανίζει περισσότερες λεπτομέρειες για το προϊόν.

5.2.4 Βαθμολόγηση προϊόντων

Η διαδικασία ψήφισης ενός προϊόντος διαδραματίζει σημαντικό ρόλο στην μετέπειτα ομαδοποίηση και μοντελοποίηση των χρηστών. Η σημαντικότητά της έγκειται στο γεγονός πως για μια διαδικασία που είναι προαιρετική, ο χρήστης αφιερώνει αυθόρμητα λίγο από τον χρόνο του για να βαθμολογήσει ένα προϊόν θέλωντας να αποτυπώσει με αυτόν τον τρόπο τα συναισθήματά του για το προϊόν αυτό.

Έτσι στη σελίδα κάθε προϊόντος δίνεται η δυνατότητα στο χρήστη να βαθμολογήσει το προϊόν μόνο μία φορά σε μία κλίμακα 5 σημείων (κακό, μέτριο, καλό, πολύ καλό, ιδανικό), και αμέσως μετά εμφανίζεται το αποτέλεσμα της ψηφοφορίας όλων των χρηστών (Εικ. 4.7).

Ενδεικτικά παρατίθεται ένα μικρό κομμάτι κώδικα που εκτελείται κατά την βαθμολόγιση του προιόντος.

```
INSERT INTO poll (product_id, customer_id, vote_rating, ip_address)
VALUES ('$product_id','$user_id','$customers_vote','$customers_ip')
```

Εικόνα 5.7. Παράδειγμα φόρμας και αποτελέσματος βαθμολογίας προϊόντος.

5.2.5 Βαθμολόγηση του ηλεκτρονικού καταστήματος

Οποιοσδήποτε εγγεγραμμένος χρήστης του ηλεκτρονικού καταστήματος διαθέτει το δικαίωμα να το βαθμολογήσει μία φορά για κάθε νέα έκδοση του (Εικ 4.8). Έτσι λαμβάνεται η γνώμη των χρηστών και επιτυγχάνεται η συνεχής βελτίωση της ιστοσελίδας. Επίσης η βαθμολογία που δίδεται από κάθε χρήστη, συμπεριλαμβάνεται στη μοντελοποίηση χρηστών για την αναπροσαρμογή της ιστοσελίδας για κάθε ομάδα χρηστών.

Εικόνα 5.8. Παράδειγμα φόρμας και αποτελέσματος βαθμολογίας ιστοσελίδας.

Ενδεικτικά παρατίθεται ένα μικρό κομμάτι κώδικα που εκτελείται κατά την βαθμολόγηση του ηλεκτρονικού καταστήματος.

```
INSERT INTO site_vote (customer_id, site_rate)
VALUES ('$user_id','$customers_vote')
```

5.2.6 Αναζήτηση προϊόντος

Για την πιο γρήγορη εύρεση ενός προϊόντος προστέθηκε ένα παράθυρο αναζήτησης όπως συμβαίνει στις περισσότερες ηλεκτρονικές σελίδες. Ως εκ τούτου δεν θα μπορούσε να λείπει ένα τέτοιο εργαλείο από το ηλεκτρονικό μας κατάστημα. Κατά την αναζήτηση ενός προϊόντος ή ομάδας προϊόντων, εκτελείτε ένα ερώτημα στη βάση δεδομένων και συγκεκριμένα στα πεδία που αφορούν τον πίνακα προιόντων. Για ταχύτερη αναζήτηση σε μερικά πεδία προστέθηκαν indexes. Ωστόσο αποφεύχθηκε η εκτεταμένη χρήση τους γιατί κάτι τέτοιο θα επιβάρυνε την λειτουργία της βάσης συνολικά. Η αναζήτηση ενός προϊόντος μπορεί να γίνει με τρεις τρόπους:

1. Αναζήτηση προϊόντος σε συγκεκριμένη κατηγορία.

Σε αυτήν την περίπτωση ο χρήστης επιλέγει κατηγορία στο παράθυρο της αναζήτησης και μετά προχωράει στην αναζήτηση. Μπορεί επίσης να πληκτρολογήσει μια λέξη προς αναζήτηση και να περιορίσει τα αποτελέσματα επιλέγοντας και κατηγορία (Εικ 4.9).Παρακάτω ακολουθεί η απεικόνιση του παραθύρου αναζήτησης.

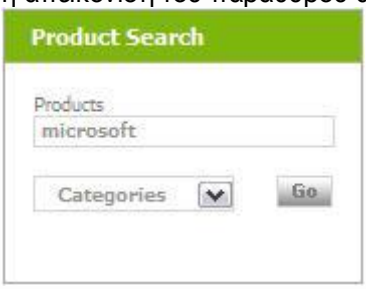
Εικόνα 5.9. Παράδειγμα φόρμας αναζήτησης με επιλογή κατηγορίας.

Παρακάτω ακολουθεί ενδεικτικά ένα μικρό κομμάτι κώδικα για την εκτέλεση της αναζήτησης.

*SELECT * FROM products WHERE cat_id = $cat_id and (text LIKE '%$word%' OR marka LIKE '%$word%')*

2. Αναζήτηση προϊόντος σε όλες τις κατηγορίες

Η αναζήτηση προϊόντος σε όλες τις κατηγορίες είναι η πιο απλή. Το μόνο που έχει να κάνει ο χρήστης είναι να πληκτρολόγηση μια περιγραφή ενός προϊόντος. Παρακάτω ακολουθεί η απεικόνιση του παραθύρου αναζήτησης.



Εικόνα 5.10. Παράδειγμα φόρμας αναζήτησης χωρίς επιλογή κατηγορίας.

Παρακάτω ακολουθεί ενδεικτικά ένα μικρό κομμάτι κώδικα για την εκτέλεση της αναζήτησης.

*SELECT * FROM products WHERE product_name LIKE '%$word%' OR text LIKE '%$word%' OR marka LIKE '%$word%'*

3. Αναζήτηση μέσω της βοηθητικής λειτουργίας AutoSuggest

Σε αυτή τη περίπτωση έχει χρησιμοποιηθεί η τεχνική AJAX κατά την οποία, καθώς ο χρήστης πληκτρολογεί ένα γράμμα στέλνεται ασύγχρονα (Postback) ένα ερώτημα στη βάση δεδομένων και επιστρέφονται άμεσα τα αποτελέσματα του ερωτήματος δίχως να γίνει ανανέωση της ιστοσελίδας. Κάθε φορά που αλλάζει το μήκος της λέξης το περιεχόμενο της λέξης που πληκτρολογεί ο χρήστης επαναλαμβάνεται η διαδικασία. Αυτή η τεχνική χρησιμοποιήθηκε για να υπάρχει ένα πιο σύγχρονο και φιλικό περιβάλλον προς τους χρήστες – πελάτες του ηλεκτρονικού καταστήματος. Παρακάτω ακολουθεί η απεικόνιση του παραθύρου αναζήτησης.

Εικόνα 5.11 Παράδειγμα φόρμας αναζήτησης με AJAX auto suggestion.

Στην παραπάνω εικόνα (Εικ 4.11), φαίνεται πως αλλάζει δυναμικά το παράθυρο με τα αποτελέσματα της αναζήτησης, καθώς αλλάζει το περιεχόμενο της λέξης προς αναζήτηση. Για την υλοποίηση αυτής της τεχνικής έγινε χρήση πολλών γλωσσών προγραμματισμού όπως είναι η XML, JavaScript, MySQL, PHP και HTML.

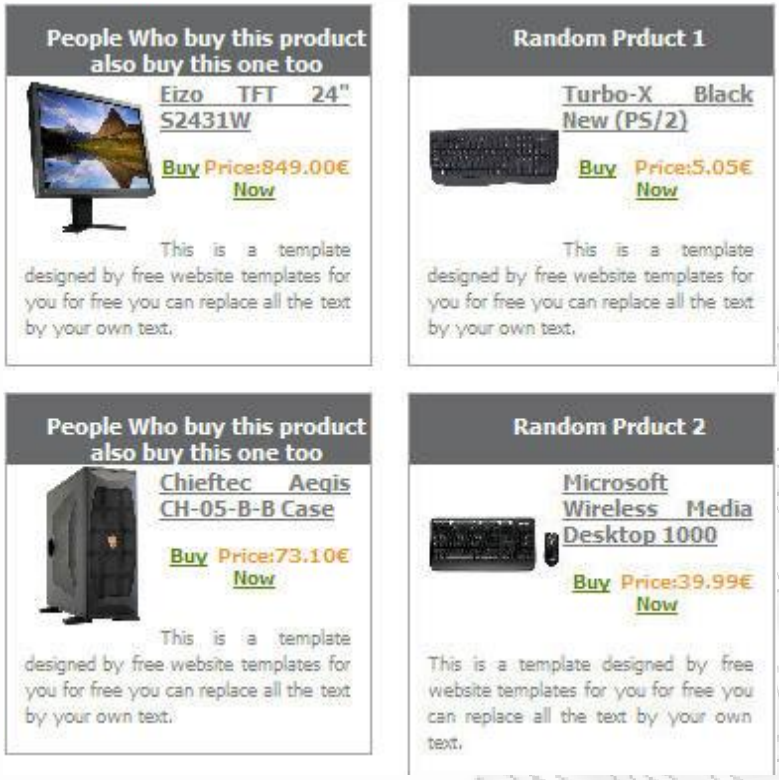
5.3 Διασταύρωση πωλήσεων (Cross-Sales)

Στη σελίδα ενός προϊόντος (πχ. product_page.php?product=65&cat_id=3) εμφανίζονται στην δεξιά στήλη, όπως απεικονίζεται παρακάτω, δύο προϊόντα τα οποία έχουν τις περισσότερες συνδυαζόμενες αγορές με το συγκεκριμένο προϊόν. Για τον υπολογισμό και εμφάνιση αυτών των δύο προϊόντων αρχικά ελέγχονται οι αγορές των χρηστών που βρίσκονται στην ίδια ομάδα (cluster) με τον χρήστη που έχει κάνει log in. Σε περίπτωση που το προϊόν της σελίδας δεν είναι αρκετά αρεστό από την ομάδα (cluster) του χρήστη, και δεν υπάρχουν συνδυαζόμενες αγορές, τότε ελέγχονται οι αγορές όλων των εγγεγραμμένων χρηστών του ηλεκτρονικού καταστήματος ανεξαρτήτου ομάδας(cluster). Αν πάλι δεν βρεθούν αποτελέσματα συνδυαζόμενων αγορών αυτού του προϊόντος , εμφανίζονται τυχαία προϊόντα της κατηγορίας που ανήκει το προϊόν.

Η μέθοδος αυτή ονομάζεται cross sales και εφαρμόζεται στα περισσότερα ηλεκτρονικά καταστήματα. Στη περίπτωση της διπλωματικής μας η προσπάθεια που έγινε είναι η διασταύρωση πωλήσεων να μην γίνεται για όλους τους χρήστες, αλλά μόνο σε αυτούς που προέκυψαν από την μοντελοποίηση (ανήκουν στην ίδια ομάδα. Το γενικότερο πνεύμα της τεχνικής του cross sales είναι ότι όποιος έχει αγοράσει αυτό το προϊόν, αγόρασε και ένα άλλο προϊόν. Με αυτόν τον τρόπο παρακινείται ο χρήστης να επεκτείνει την παραγγελία του βάζοντας στο καλάθι αγορών του περισσότερα προϊόντα κάποια από τα οποία πιθανώς να μην γνώριζε καν την ύπαρξή τους. Παρακάτω παρατίθεται ένα απόσπασμα κώδικα που χρησιμοποιείται για την εφαρμογή του cross-sales.

1. Μαζεύουμε όλες τις αγορές που εμπεριέχουν αυτό το προϊόν εξαιρώντας τις αγορές που έχει κάνει ο χρήστης που επισκέπτεται τη σελιδα.
   *SELECT * FROM orders*
   WHERE product_id= $product_id AND cust_id != $user_id
2. Συγκεντρώνονται όλες οι συνδυασμένες αγορές των υπόλοιπων χρηστών.
   *SELECT * FROM orders*
   WHERE cust_id = $cust_id and product_id != $product_id and Date ='$date'
3. Εισάγεται στον πίνακα cross_sales ο κωδικός προιόντος,αριθμός τεμαχίων.
   INSERT INTO cross_sales (product_id, count)
   VALUES ($product_cross_id,$count_cross)
4. Υπολογίζεται το προιόν με τις περισσότερες αγορές και εμφανίζονται με φθίνουσα σειρά ώστε να διαλέξουμε τα δύο πρώτα για την εμφάνιση στην ιστοσελίδα.
   *SELECT *,SUM(count) As A FROM cross_sales GROUP BY product_id*
   ORDER BY A DESC

Παρακάτω ακολουθεί απεικόνιση των προϊόντων που προκύπτουν από τη μέθοδο cross-sales(Εικ. 5.11, αριστερό μέρος) και η απεικόνιση των τυχαίων προϊόντων σε περίπτωση μη εύρεσης συνδυασμένης αγοράς (Εικ 5.11, δεξί μέρος).



Εικόνα 5.11. Παράδειγμα Cross-Sales (αριστερά επιτυχημένο cross-sales, δεξιά τυχαία προιόντα)

5.4 Εμφάνιση προϊόντων χωρίς την χρήση της μοντελοποίησης

Κάθε υπέρ-κατηγορία(menu) όπως είναι τα computers, laptop, peripherals, games έχουν τη δική τους σελίδα που εμφανίζονται τα προϊόντα των υποκατηγοριών τους. Σε αυτή τη σελίδα εκτός από τα τυχαία προϊόντα της κάθε υποκατηγορίας εμφανίζονται και άλλα δυο τα όποια είναι τα πιο δημοφιλή αυτής της υπέρ-κατηγορίας.

5.4.1 Δημοφιλέστερο προϊόν βάσει αγορών

Στη δεξιά κολώνα των υπέρ-κατηγοριών , το πρώτο προϊόν που εμφανίζεται, είναι το πιο δημοφιλές προϊόν αυτού του μενού βάσει των αγορών. Εμφανίζοντας αυτό το προϊόν στο χρήστη-πελάτη πετυχαίνουμε την προσέλκυση του και την παρότρυνση του, για αγορά ενός από τα πιο δημοφιλή προϊόντα του καταστήματος για τα οποία πιο πριν δεν γνώριζε την απήχηση τους στον υπόλοιπο κόσμο που επισκέπτεται το συγκεκριμένο ηλεκτρονικό κατάστημα. Παρακάτω ακολουθεί ένα μικρό κομμάτι κώδικα που εκτελείται για την εύρεση του προϊόντος με τις περισσότερες αγορές αυτού του menu από την βάση δεδομένων.

```
SELECT * , sum( count )
FROM orders AS a
JOIN products AS b ON a.product_id = b.product_id
WHERE menu_id = $menu_id
GROUP BY a.product_id
ORDER BY menu_id, sum( count ) DESC";
```

5.4.2 Δημοφιλέστερο προϊόν βάσει επισκεψιμότητας

Στη δεξιά κολώνα των υπέρ-κατηγοριών , το δεύτερο προϊόν που εμφανίζεται, είναι το πιο δημοφιλές προϊόν αυτού του μενού βάσει των επισκέψεων που δέχεται από τους εγγεγραμμένους χρήστες. Η επίσκεψη ενός προϊόντος μπορεί να μην έχει την ίδια βαρύτητα με την αγορά του, παρ' όλα αυτά είναι ένας σημαντικός δείκτης για την σημαντικότητα του. Εμφανίζοντας αυτό το προϊόν στο χρήστη-πελάτη πετυχαίνουμε την προσέλκυση και την παρότρυνση του, για αγορά ενός από τα πιο δημοφιλή προϊόντα του καταστήματος για τα οποία πιο πριν δεν γνώριζε την απήχηση τους στον υπόλοιπο κόσμο που επισκέπτεται το συγκεκριμένο ηλεκτρονικό κατάστημα. Παρακάτω ακολουθεί ένα μικρό κομμάτι κώδικα που εκτελείται για την εύρεση του προϊόντος με τις περισσότερες αγορές αυτού του menu από την βάση δεδομένων. Παρακάτω ακολουθεί ένα μικρό κομμάτι κώδικα που εκτελείται για την εύρεση του προϊόντος με τα περισσότερα clicks αυτού του menu από την βάση δεδομένων.

```
SELECT * , sum( clicks_number )
FROM clicks AS a
JOIN products AS b ON a.product_id = b.product_id
WHERE menu_id = $menu_id
GROUP BY a.product_id
ORDER BY sum( clicks_number ) DESC";
```

Εικόνα 5.12. Παράδειγμα προϊόντων most ordered(αριστερά)-most visited (δεξιά) για το menu laptop

5.4.3 Προϊόν με τη μεγαλύτερη βαθμολογία κατά μέσο όρο (Top Rating)

Στο δεξί πάνω άκρο της πρώτης σελίδας απεικονίζεται το προϊόν με τη μεγαλύτερη βαθμολογία κατά μέσο όρο, που έδωσαν μόνο οι εγγεγραμμένοι χρήστες. Αυτό το προϊόν κρίνεται σημαντικό και άξιο εμφάνισης διότι η βαθμολογία του προέρχεται είτε επειδή ο πελάτης το αγόρασε και έμεινε ικανοποιημένος οπότε και το βαθμολόγησε, είτε επειδή προσέλκυσε τον χρήστη να το βαθμολογήσει βάσει χαρακτηριστικών και τιμής ενδεικτικό της πρόθεσής του να το αγοράσεις. Ο κώδικας για την εμφάνιση του προϊόντος η και πραγματική του απεικόνιση φαίνονται στην εικόνα 5.13. Με την παρακάτω εντολή sql ζητείται ο καλύτερος μέσος όρος με τους περισσότερους ψήφους.

```
SELECT avg( vote_rating ) , count( vote_rating ) , product_id
FROM  poll`
GROUP BY product_id
ORDER BY avg( vote_rating ) DESC , count( vote_rating ) DESC
```

Έχοντας το product_id εκτελούμε την παρακάτω sql για την προσκόμιση όλων των στοιχείων του προϊόντος για την εμφάνιση του.

```
SELECT * FROM products WHERE product_id = 11
```

AVG	Count	Procuct_id
5.0000	4	11
5.0000	2	73
5.0000	2	29
5.0000	2	40
5.0000	2	88
5.0000	1	168
5.0000	1	93
5.0000	1	154
5.0000	1	87
5.0000	1	156
5.0000	1	6

Εικόνα 5.13. Παράδειγμα προϊόντος συγκέντρωσης μεγαλύτερης βαθμολογίας (αριστερά η απεικόνιση στην ιστοσελίδα, δεξιά ένα κομμάτι του πίνακα με τη καλύτερη βαθμολογία.

5.4.4 Προϊόν με τις περισσότερες αγορές όλων των χρηστών(Best Seller)

Στο κέντρο της δεξιάς στήλης της αρχικής σελίδας απεικονίζεται το προϊόν το οποίο έχει αγοραστεί περισσότερο από το σύνολο των χρηστών. Αυτό το προϊόν έχει ιδιαίτερη βαρύτητα διότι αντιπροσωπεύει τις πραγματικές επιθυμίες της πλειοψηφίας του αγοραστικού μας κοινού. Ο κώδικας για την εμφάνιση του προϊόντος και η πραγματική του απεικόνιση στην ιστοσελίδα παρατίθενται παρακάτω.

Product_id	Count
20	9
113	6
78	6
72	6
69	6
51	5
77	5
92	5
63	4

Εικόνα 5.14. Παράδειγμα προϊόντος με τις περισσότερες αγορές (αριστερά η απεικόνιση στην ιστοσελίδα, δεξιά ένα κομμάτι του πίνακα με τον αριθμό πωλήσεων και product_id)

Στον παρακάτω κώδικα sql ζητείται το άθροισμα των αγορών ανά προϊόν σε φθίνουσα σειρά.

```
SELECT product_id Product_id, sum( count ) Count
FROM orders
GROUP BY product_id
ORDER BY sum( count ) DESC
```

Έχοντας τον κωδικό του best seller προϊόντος εκτελούμε την παρακάτω sql για την προσκόμιση όλων των στοιχείων του προϊόντος για την εμφάνιση του.

```
SELECT * FROM products
WHERE product_id = 20
```

5.4.5 Προϊόν με τη μεγαλύτερη επισκεψιμότητα (most visited)

Στο κάτω μέρος της δεξιά στήλης απεικονίζεται το προϊόν με τη μεγαλύτερη επισκεψιμότητα βάσει των click που καταγράφονται στην βάση δεδομένων κάθε φορά που ένας χρήστης επισκέπτεται το προϊόν αυτό. Αυτή η κατηγορία διαφήμισης μπορεί να μην έχει την ίδια βαρύτητα με την βαθμολογία και την αγορά ενός προϊόντος , αλλά δείχνει το ενδιαφέρον (έστω και μικρότερο) των πελατών και οφείλουμε να το λάβουμε υπόψιν μας. Ο κώδικας για την εμφάνιση του προϊόντος και η πραγματική του απεικόνιση στην ιστοσελίδα παρατίθενται παρακάτω. Στον παρακάτω κώδικα sql ζητείται το άθροισμα των επισκέψεων ανά προϊόν σε φθίνουσα σειρά.

```
SELECT product_id Product, cat_id Category, sum( clicks_number ) Sum_Clicks
FROM clicks
GROUP BY product_id
ORDER BY sum( clicks_number ) DESC
```

Έχοντας τον κωδικό του most visited προϊόντος εκτελούμε την παρακάτω εντολή sql για την προσκόμιση όλων των στοιχείων του προϊόντος για την εμφάνιση του.

```
SELECT * FROM products
WHERE product_id = 52
```

Product	Category	Sum_Clicks
52	1	75
78	11	65
69	10	60
7	2	50
4	1	48
11	2	41
53	1	38
20	10	34
58	2	33
29	11	33

Εικόνα 5.15. Παράδειγμα προϊόντος με τις περισσότερες επισκέψεις (αριστερά η απεικόνιση στην ιστοσελίδα, δεξιά ένα κομμάτι του πίνακα με τον αριθμό επισκέψεων και product_id,cat_id)

5.4.6 Η νεότερη άφιξη προϊόντος στο ηλεκτρονικό κατάστημα (Fresh)

Στο πάνω άκρο της κεντρική στήλης έχει τοποθετηθεί το πιο πρόσφατο προϊόν του καταστήματος. Εμφανίζοντας κάθε μέρα τις καινούριες αφίξεις πετυχαίνουμε την εφαρμογή διαφήμισης όπως συμβαίνει με τα φυλλάδια στα σούπερ μάρκετ ή όπως συμβαίνει στα περισσότερα e-shop. Εμφανίζεται το πιο καινούριο προϊόν για την προσέλκυση με σκοπό την επίσκεψη του και αγορά του. Στον παρακάτω κώδικα sql ζητείται το προϊόν με την πιο πρόσφατη ημερομηνία καταχώρησης.

```
SELECT * FROM products
ORDER BY DateAdded DESC
```

Εικόνα 5.16. Παράδειγμα προϊόντος με την πιο πρόσφατη ημερομηνία καταχώρησης (αριστερά όπως απεικονίζεται στο e-shop και δεξιά το αποτέλεσμα αναζήτησης στη βάση δεδομένων)

5.5 Εμφάνιση προϊόντων με βάση την μοντελοποίηση χρηστών

Σε περίοπτη θέση (κεντρικά στην αρχική σελίδα) εμφανίζονται τα προϊόντα που βασίζονται στην μοντελοποίηση χρηστών. Τα προϊόντα αυτά προσαρμόζονται δυναμικά σύμφωνα με τις προτιμήσεις του εκάστοτε χρήστη. Αν με την πάροδο του καιρού ο χρήστης αλλάξει συνήθειες τότε η μοντελοποίηση τον κατατάσσει σε άλλη ομάδα (cluster) με αποτέλεσμα να αλλάζουν και τα προϊόντα εμφάνισης. Τα προϊόντα αυτά έχουν ιδιαίτερη βαρύτητα διότι κατά μεγάλο ποσοστό ανήκουν στις αγαπημένες κατηγορίες προϊόντων του εγγεγραμμένου χρήστη. Η τοποθέτηση των προϊόντον αυτών σε κεντρικό σημείο της σελίδας είναι επίσης κομβικής σημασίας αφού προκαλεί το άμεσο ενδιαφέρον του πελάτη. Παρακάτω θα περιγραφεί ο τρόπος με τον οποίο διαλέγονται προς εμφάνιση τα προϊόντα αυτά.

5.5.1 Προϊόν με τις περισσότερες αγορές χρηστών που ανήκουν στην ίδια ομάδα

Το προϊόν αυτό προέρχεται από την κατηγορία με τις περισσότερες αγορές των εγγεγραμμένων χρηστών που ανήκουν στην ίδια ομάδα “cluster”. Κάθε φορά που ανανεώνεται η σελίδα αυτή, εμφανίζεται ένα τυχαίο προϊόν από την κατηγορία αυτή εκτός των προϊόντων που έχουν είδη αποκτηθεί από το συνδεδεμένο χρήστη. Με αυτό τον τρόπο ο χρήστης λαμβάνει διαφημίσεις για προϊόντα που κατά μεγάλο ποσοστό τον ενδιαφέρουν και ενημερώνεται άμεσα για αυτά. Επίσης δεν επαναλαμβάνονται οι διαφημίσεις των είδη αποκτηθέντων προϊόντων με σκοπό την αποφυγή της αποστροφής του χρήστη, αντικρίζοντας συνεχώς προϊόντα που πλέον δεν του προκαλούν κάποιο ενδιαφέρον. Βέβαια σε περίπτωση που επιθυμεί να αποκτήσει ξανά κάποιο προϊόν μπορεί απλά να το επισκεφτεί και να το προσθέσει στο καλάθι αγορών του.

Σε περίπτωση που ο χρήστης δεν έχει καταταχτεί σε κάποια ομάδα “cluster” τότε εμφανίζεται ένα τυχαίο προϊόν ανεξαρτήτου κατηγορίας. Το ίδιο ισχύει και για έναν μη εγγεγραμμένο χρήστη για τον οποίο δεν υπάρχουν στοιχεία στη βάση και δεν μπορεί να πάρει μέρος στη μοντελοποίηση.

Στον παρακάτω κώδικα sql ζητείται η κατηγορία με τις περισσότερες αγορές, από τον πίνακα αγορών των user του ίδιου cluster (clustered_order). Ο πίνακας clustered_orders δημιουργείται κατά την εκτέλεση του αλγορίθμου clustering, ο οποίος δεν εκτελείται κατά την είσοδο του χρήστη για την αποφυγή καθυστερήσεων φόρτωσης.

```
SELECT cat_id , sum( count ) AS TOTAL
FROM `clustered_orders`
WHERE customer_id = $user
GROUP BY cat_id
ORDER BY `TOTAL` DESC
```

Η κατηγορία αυτή αποθηκεύεται σε μία session μεταβλητή που διαρκεί μέχρι ο χρήστης να αποσυνδεθεί από τη σελίδα.

```
$_SESSION['best_cat_id'];
```

Εικόνα 5.16. Παράδειγμα προϊόντος από την best seller κατηγορία των χρηστών του ίδιου cluster.

Τέλος, γίνεται η αναζήτηση στον πίνακα προϊόντων με παράμετρο τη παραπάνω session μεταβλητή και επιστρέφονται όλα τα στοιχεία του προϊόντος.

*SELECT * FROM products*

WHERE cat_id=$_SESSION['best_cat_id'];

5.5.2 Προϊόν της κατηγορίας με την μεγαλύτερη επισκεψιμότητα χρηστών που ανήκουν στην ίδια ομάδα

Το προϊόν αυτό προέρχεται από την κατηγορία με την μεγαλύτερη επισκεψιμότητα των εγγεγραμμένων χρηστών που ανήκουν στην ίδια ομάδα “cluster”. Κάθε φορά που ανανεώνεται η σελίδα, εμφανίζεται ένα τυχαίο προϊόν από την κατηγορία αυτή εκτός των προϊόντων που έχουν είδη αποκτηθεί από το συνδεδεμένο χρήστη. Με αυτό τον τρόπο ο χρήστης λαμβάνει διαφημίσεις για προϊόντα που κατά μεγάλο ποσοστό τον ενδιαφέρουν και ενημερώνεται άμεσα για αυτά. Επίσης δεν επαναλαμβάνονται οι διαφημίσεις των είδη αποκτηθέντων προϊόντων με σκοπό την αποφυγή της αποστροφής του χρήστη, αντικρίζοντας συνεχώς προϊόντα που πλέον δεν του προκαλούν κάποιο ενδιαφέρον. Βέβαια σε περίπτωση που επιθυμεί να αποκτήσει ξανά κάποιο προϊόν μπορεί απλά να το επισκεφτεί και να το προσθέσει στο καλάθι αγορών του.

Σε περίπτωση που ο χρήστης δεν έχει καταταχτεί σε κάποια ομάδα “cluster” ή δεν είναι εγγεγραμμένος, εμφανίζεται ένα τυχαίο προϊόν ανεξαρτήτου κατηγορίας

Εικόνα 5.17 Παράδειγμα προϊόντος από την best visited κατηγορία των χρηστών του ίδιου cluster.

Στον παρακάτω κώδικα sql ζητείται η κατηγορία με τα περισσότερα clicks, από τον πίνακα ‘click’ των χρηστών της ίδιας ομάδας (clustered_clicks). Ο πίνακας clustered_clicks δημιουργείται κατά την εκτέλεση του αλγορίθμου clustering.

```
SELECT cat_id, sum( count ) AS TOTAL
FROM `clustered_clicks`
WHERE customer_id = $user
GROUP BY cat_id
ORDER BY `TOTAL` DESC
```

Η κατηγορία αυτή αποθηκεύεται σε μία session μεταβλητή που διαρκεί μέχρι ο χρήστης να αποσυνδεθεί από τη σελίδα.

```
$_SESSION['best_click_cat_id'];
```

Τέλος, γίνεται η αναζήτηση στον πίνακα προϊόντων με παράμετρο τη παραπάνω session μεταβλητή και επιστρέφονται όλα τα στοιχεία του προϊόντος.

```
SELECT * FROM products
WHERE cat_id=$_SESSION['best_click_cat_id'];
```

5.5.3 Προϊόν με την καλύτερη κατά μέσο όρο βαθμολογία των χρηστών της ίδιας ομάδας

Το προϊόν αυτό προέρχεται από την κατηγορία με την μεγαλύτερη κατά μέσω όρο βαθμολογία των εγγεγραμμένων χρηστών που ανήκουν στην ίδια ομάδα “cluster”. Κάθε φορά που ανανεώνεται η σελίδα, εμφανίζεται το ίδιο προϊόν από την κατηγορία αυτή εκτός των προϊόντων που έχουν είδη αποκτηθεί από το συνδεδεμένο χρήστη. Με αυτό τον τρόπο ο χρήστης λαμβάνει διαφημίσεις για προϊόντα που κατά μεγάλο ποσοστό τον ενδιαφέρουν και ενημερώνεται άμεσα για αυτά. Επίσης δεν επαναλαμβάνονται οι διαφημίσεις των είδη αποκτηθέντων προϊόντων με σκοπό την αποφυγή της αποστροφής του χρήστη, αντικρίζοντας συνεχώς προϊόντα που πλέον δεν του προκαλούν κάποιο ενδιαφέρον. Βέβαια σε περίπτωση που επιθυμεί να αποκτήσει ξανά κάποιο προϊόν μπορεί απλά να το επισκεφτεί και να το προσθέσει στο καλάθι αγορών του.

Σε περίπτωση που ο χρήστης δεν έχει καταταχτεί σε κάποια ομάδα “cluster” τότε εμφανίζεται ένα τυχαίο προϊόν ανεξαρτήτου κατηγορίας. Το ίδιο ισχύει και για έναν μη εγγεγραμμένο χρήστη για τον οποίο δεν υπάρχουν στοιχεία στη βάση και δεν μπορεί να πάρει μέρος στη μοντελοποίηση. Ο κώδικας για την εμφάνιση του προϊόντος η και πραγματική του απεικόνιση στην ιστοσελίδα παρατίθενται παρακάτω.

Εικόνα 5.18 Παράδειγμα προϊόντος με την μεγαλύτερη μέσο όρο βαθμολογία που έδωσαν οι χρήστες του ίδιου cluster.

Στον παρακάτω κώδικα sql ζητείται το προιον με τη μεγαλυτερη βαθμολογια κατά μέσω ορο, από τον πίνακα poll των user του ίδιου cluster (clustered_votes). Ο πίνακας clustered_votes δημιουργείται κατά την εκτέλεση του αλγορίθμου clustering.

```
SELECT * , avg( vote ) , count( vote )
FROM `clustered_prod_votes`
```

WHERE customer_id =$user
GROUP BY product_id
ORDER BY avg(vote) DESC , count(vote) DESC

Το προϊόν αυτό αποθηκεύεται σε μία session μεταβλητή που διαρκεί μέχρι ο χρήστης να αποσυνδεθεί από τη σελίδα.

$_SESSION['clustered_best_illected']

Τέλος, γίνεται η αναζήτηση στον πίνακα προϊόντων με παράμετρο τη παραπάνω session μεταβλητή και επιστρέφονται όλα τα στοιχεία του προϊόντος.

*SELECT * FROM products*
WHERE product_id = $_SESSION['clustered_best_illected']

5.5.4 Το πιο πρόσφατο προϊόν από την κατηγορία με τις περισσότερες αγορές των χρηστών της ίδιας ομάδας

Το προϊόν αυτό είναι το πιο νεοεισαχθέν της κατηγορίας με τις περισσότερες πωλήσεις των χρηστών που ανήκουν στο cluster του εκάστοτε συνδεδεμένου χρήστη. Αξίζει να σημειωθεί ότι εάν έχει ήδη αποκτηθεί αυτό το προϊόν, τότε εμφανίζεται το αμέσως επόμενο ημερομηνιακά από την ίδια κατηγορία. Με αυτό τον τρόπο ο χρήστης λαμβάνει διαφημίσεις για προϊόντα που κατά μεγάλο ποσοστό τον ενδιαφέρουν και ενημερώνεται για τις νέες κυκλοφορίες άμεσα. Επίσης δεν επαναλαμβάνονται οι διαφημίσεις των είδη αποκτηθέντων προϊόντων με σκοπό την αποφυγή της αποστροφής του χρήστη, αντικρίζοντας συνεχώς προϊόντα που πλέον δεν του προκαλούν κάποιο ενδιαφέρον. Βέβαια σε περίπτωση που επιθυμεί να αποκτήσει ξανά κάποιο προϊόν μπορεί απλά να το επισκεφτεί και να το προσθέσει στο καλάθι αγορών του.

Σε περίπτωση που ο χρήστης δεν έχει καταταχτεί σε κάποια ομάδα “cluster” τότε εμφανίζεται ένα τυχαίο προϊόν ανεξαρτήτου κατηγορίας. Το ίδιο ισχύει και για έναν μη εγγεγραμμένο χρήστη για τον οποίο δεν υπάρχουν στοιχεία στη βάση και δεν μπορεί να πάρει μέρος στη μοντελοποίηση.

Ο κώδικας για την εμφάνιση του προϊόντος η και πραγματική του απεικόνιση στην ιστοσελίδα παρατίθενται παρακάτω.

Εικόνα 5.19. Παράδειγμα νεοεισαχθέντος προϊόντος από την best seller κατηγορία των χρηστών του ίδιου cluster.

Στον παρακάτω κώδικα sql ζητείται το νεότερο προϊόν βάσει ημερομηνίας καταχώρισης από την best seller κατηγορία των χρηστών του ίδιου cluster. Ουσιαστικά αναζητά την κατηγορία με τις περισσότερες αγορές στον πίνακα (clustered_order). Ο πίνακας clustered_orders δημιουργείται κατά την εκτέλεση του αλγορίθμου clustering, ο οποίος δεν εκτελείται κατά την είσοδο του χρήστη για την αποφυγή καθυστερήσεων φόρτωσης.

SELECT cat_id , sum(count) AS TOTAL

```
FROM `clustered_orders`
WHERE customer_id = $user
GROUP BY cat_id
ORDER BY `TOTAL` DESC
```

Η κατηγορια αυτη αποθηκεύεται σε μία session μεταβλητή που διαρκεί μέχρι ο χρήστης να αποσυνδεθεί από τη σελίδα.

```
$_SESSION['best_cat_id'];
```

Τέλος, γίνεται η αναζήτηση στον πίνακα προϊόντων με παράμετρο τη παραπάνω session μεταβλητή και επιστρέφονται όλα τα στοιχεία του προϊόντος.

```
SELECT * FROM products
WHERE cat_id = $_SESSION['best_cat_id']
ORDER BY DateAdded DESC
```

6. ΛΕΠΤΟΜΕΡΕΙΕΣ ΥΛΟΠΟΙΗΣΗΣ

Σε αυτό το κεφάλαιο θα αναλυθούν λεπτομερώς όλα τα βήματα που ακολουθήθηκαν προκειμένου να διαχωριστούν οι χρήστες του ηλεκτρονικού καταστήματος σε ομάδες. Χρησιμοποιήθηκαν 72 διαφορετικοί χρήστες προκειμένου να συλλεχθούν τα απαραίτητα δεδομένα. Για τον σκοπό αυτό όλες οι εκδόσεις του site έχουν ανεβεί στην διεύθυνση http://www.renko-webhosting.gr/e-shop, όπου πλέον βρίσκεται η τελική έκδοσή του.

6.1 Επιλογή χαρακτηριστικών γνωρισμάτων

Το πρώτο στάδιο και ίσως το σημαντικότερο στην διαδικασία της ομαδοποίησης των χρηστών είναι η επιλογή των χαρακτηριστικών γνωρισμάτων (Features) που διαθέτουν οι χρήστες. Η ποιότητα των αποτελεσμάτων της ομαδοποίησης εξαρτάται άμεσα από την ποιότητα των χαρακτηριστικών γνωρισμάτων, συνεπώς απαιτείται να δοθεί ιδιαίτερη βαρύτητα στην επιλογή τους.

Όπως αναφέρθηκε επιγραμματικά και στο προηγούμενο κεφάλαιο, τα δεδομένα αυτά συλλέγονται μέσω των κινήσεων που κάνει ένας εγγεγραμμένος χρήστης στο ηλεκτρονικό κατάστημα. Το σύνολο των χαρακτηριστικών γνωρισμάτων αφού αναπαρασταθεί με τα κατάλληλα αριθμητικά δεδομένα θα σχηματίσει ένα διάνυσμα για κάθε χρήστη και το σύνολο των διανυσμάτων θα τροφοδοτήσει τον αλγόριθμο ομαδοποίησης. Στις επόμενες ενότητες αναλύονται τα χαρακτηριστικά γνωρίσματα που επιλέχθηκαν.

6.1.1 Μέσος όρος αξίας αγορών

Αντιπροσωπευτικός της αγοραστικής δύναμης ενός πελάτη είναι ο μέσος όρος της αξίας των αγορών πού έχει πραγματοποιήσει. Όσες περισσότερες αγορές πραγματοποιήσει ένας πελάτης τόσο καλύτερες ενδείξεις θα έχουμε για το εύρος τιμών για το οποίο κινείται. Έτσι μεταξύ δύο χρηστών ένα καλό μέτρο σύγκρισης είναι τα επίπεδα μέσου κόστους των αγορών τους. Για κάθε χρήστη i λοιπόν υπολογίζουμε τον μέσο όρο της αξίας των αγορών του από τον τύπο:

$$\text{MeanValue}[i] = \frac{TotalMoneySpent}{TotalItemsBought}$$

Τα στοιχεία που χρειαζόμαστε για τον υπολογισμό της παραπάνω τιμής θα τα αντλήσουμε από την βάση δεδομένων με τα ακόλουθα ερωτήματα SQL από τον πίνακα Orders (Εικ 6.1):

order_id	cust_id	product_id	count	Date
104	30	215	1	2010-09-23 21:32:49
105	30	76	1	2010-09-23 21:32:49
106	31	116	1	2010-09-23 23:07:12
107	31	121	1	2010-09-23 23:07:12

Εικ 6.1. Στιγμιότυπο του πίνακα Orders όπου φαίνονται οι αγορές δύο πελατών

$query = "SELECT product_id, count FROM orders

WHERE cust_id = $customer_id

ORDER BY cust_id";

Με το παραπάνω ερώτημα αντλούμε όλους τους κωδικούς προϊόντων (product_id) και τις ποσότητες (count) για τον χρήστη με id cust_id. Ακολούθως πολλαπλασιάζουμε την ποσότητα κάθε προϊόντος με την τιμή αγοράς και προσθέτουμε στον συνολικό μετρητή. Η ονοματολογία των μεταβλητών είναι επεξηγηματική ώστε να μην απαιτείται περαιτέρω ανάλυση.

$Product_Price = $row['price'];

*$Order_Price = $Product_Price * $Product_Count;*

$Total_Money_Spend = $Total_Money_Spend + $Order_Price;

$nProducts = $nProducts + $Product_Count;

Τέλος για να υπολογίσουμε το meanValue για κάθε πελάτη εκτελούμε την πράξη

$MeanValue[$i] = $Total_Money_Spend / $nProducts;

Τα αποτελέσματα της προηγούμενης διαδικασίας αποθηκεύονται σε ένα πίνακα 1 x n (π.χ. Πίν. 6.1), όπου n το πλήθος των πελατών τέτοιος ώστε Array[i] να συμβολίζει τον μέσο όρο αξίας αγορών του πελάτη με customer_id = i.

56	205	98	329	176

Πίνακας 6.1. Παράδειγμα πίνακα με τους μέσους όρους αξίας αγορών για 5 πελάτες

6.1.2 Σύνολο αγορών προϊόντων ανά κατηγορία

Σκοπός μας μέσω αυτού του γνωρίσματος είναι να υπολογίσουμε τον βαθμό ενδιαφέροντος ενός χρήστη για κάθε μία από τις 18 κατηγορίες προϊόντων που διαθέτει το ηλεκτρονικό κατάστημα. Στην ουσία δηλαδή πρόκειται για ένα σύνολο 18 γνωρισμάτων κάθε ένα από τα οποία συμβολίζει τον βαθμό ενδιαφέροντος του χρήστη για την συγκεκριμένη κατηγορία. Άρα αυτό που έχουμε να κάνουμε για κάθε πελάτη είναι να αθροίσουμε το σύνολο των προϊόντων που έχει αγοράσει από κάθε κατηγορία. Δηλαδή με ψευδοκώδικα

∀ *πελάτη U*

∀ *κατηγορία I*

∀ *αγορά*

Έστω id το αναγνωριστικό του προϊόντος

If προϊόν με product_id=id ανήκει στην κατηγορία I

Πρόσθεσε 1 στο βαθμό ενδιαφέροντος του πελάτη U για την κατηγορία I

EndIf

Loop

Loop

Loop

Έτσι για κάθε κατηγορία έχουμε ένα πίνακα 1 x n, όπου n το πλήθος των πελατών όπου κάθε τιμή έστω Array[i] αναπαριστά τον βαθμό ενδιαφέροντος του χρήστη i για την κατηγορία αυτή.

Λόγω ότι ο κώδικας της διαδικασίας αυτής είναι ιδιαίτερα μακροσκελής παραθέτουμε μόνο ένα ερώτημα SQL, το οποίο φέρνει τις συνολικές αγορές ενός πελάτη για μία κατηγορία.

```
Query = "SELECT cust_id , sum( count ) as SUM
FROM `orders` AS a
JOIN products AS b ON a.product_id = b.product_id
WHERE b.cat_id = $cat_id
GROUP BY a.cust_id
```

6.1.3 Σύνολο επισκέψεων προϊόντων ανά κατηγορία

Όπως και το προηγούμενο γνώρισμα, έτσι και αυτό στην ουσία αποτελείται απο 18 γνωρίσματα, ένα για κάθε κατηγορία προϊόντος. Σκοπός κάθε ενός από τα γνωρίσματα αυτά είναι ο υπολογισμός του βαθμού ενδιαφέροντος ενός χρήστη για την αντίστοιχη κατηγορία, από την επισκεψιμότητα που έχει αυτή (παρ. 5.2.3). Άρα αυτό που έχουμε να κάνουμε για κάθε πελάτη είναι να αθροίσουμε το σύνολο των φορών που έχει κλικάρει τα προϊόντα κάθε κατηγορίας. Δηλαδή με ψευδοκώδικα

```
∀ πελάτη U
        ∀ κατηγορία I
                ∀ επίσκεψη
                        Έστω id το αναγνωριστικό του προϊόντος
                        If προϊόν με product_id=id ανήκει στην κατηγορία I
                                Πρόσθεσε 1 στο βαθμό ενδιαφέροντος του πελάτη U για την κατηγορία I
                        EndIf
                Loop
        Loop
Loop
```

Έτσι για κάθε κατηγορία έχουμε ένα πίνακα 1 x n, όπου n το πλήθος των πελάτων και η τιμή Array[i] εκφράζει τον βαθμό ενδιαφέροντος του χρήστη i για την κατηγορία αυτή. Ενδεικτικά ο κώδικας SQL μέσω του οποίου αντλούμε τα συνολικά click από τον πίνακα Clikcs (Εικ. 6.2) ενός πελάτη για κάθε κατηγορία:

```
query = "SELECT customer_id , sum( clicks_number ) as SUM
FROM `clicks` AS a
JOIN products AS b ON a.product_id = b.product_id
WHERE b.cat_id = $cat_id
GROUP BY a.customer_id
```

click_id	customer_id	cat_id	product_id	clicks_number
165	15	3	13	1
166	15	21	159	1
167	18	1	54	1
168	18	21	163	1

Εικ 6.2. Στιγμιότυπο του πίνακα Cliks όπου φαίνονται οι επισκέψεις δύο πελατών

6.1.4 Βαθμολόγηση καταστήματος

Η βαθμολόγηση που δίνει ένας πελάτης στο ηλεκτρονικό κατάστημα, είναι ενδεικτική του βαθμού ικανοποίησής του από τις υπηρεσίες που του παρέχονται. Έτσι μπορούμε να πούμε ότι δύο χρήστες οι οποίοι αξιολογούν το site με τον ίδιο βαθμό έχουν ένα κοινό χαρακτηριστικό το οποίο μπορεί να συμβάλλει στην ομαδοποίησή τους. Κάθε χρήστης έχει την ευχέρεια να βαθμολογήσει το κατάστημα σε μία κλίμακα 5 σημείων με την χειρότερη βαθμολογία να αντιστοιχεί στον αριθμό 1 και την υψηλότερη στον βαθμό 5. Επομένως είναι πολύ εύκολος ο υπολογισμός του βαθμού ενδιαφέροντος ενός χρήστη, αρκεί να αντλήσουμε τα δεδομένα για κάθε πελάτη από τον πίνακα site_vote. Έτσι για αυτό το γνώρισμα προκύπτει ένας πίνακας 1x n, όπου n το πλήθος των πελατών έτσι ώστε το στοιχείο Rating[i] να απεικονίζει την βαθμολογία με την οποία ψήφισε το κατάστημα ο πελάτης i. Η εντολή SQL με την οποία προσκομίζουμε τα δεδομένα ψήφισης από την βάση δεδομένων είναι:

query = "SELECT site_rate FROM site_vote

WHERE customer_id = $customer_id";

vote_id	customer_id	site_rate
1	3	4
2	13	3
3	2	3
4	15	4

Εικόνα 6.3. Στιγμιότυπο του πίνακα site_vote, όπου διακρίνονται οι ψήφοι 4 πελατών

6.1.5 Βαθμολόγηση προϊόντων κατηγορίας

Κάθε εγγεγραμμένος χρήστης μπορεί να ψηφίσει οποιοδήποτε προϊόν επιθυμεί. Προφανώς δεν μπορούμε να χρίσουμε την βαθμολογία κάθε προϊόντος ως χαρακτηριστικό γνώρισμα λόγω του υπερβολικά μεγάλου αριθμού προϊόντων που διαθέτει ένα ηλεκτρονικό κατάστημα. Αν γινόταν αυτό θα καταλήγαμε σε ένα τεράστιο διάνυσματικό προφίλ για κάθε χρήστη και είναι πιθανόν ο αλγόριθμος ομαδοποίησης να μην ολοκληρωνόταν σε ρεαλιστικό χρόνο. Έτσι αποφασίστηκε όπως και παραπάνω η δημιουργία 18 χαρακτηριστικών γνωρισμάτων, ένα για κάθε κατηγορία προϊόντων. Η τιμή καθενός από αυτά θα είναι ο μέσος όρος της βαθμολογίας που έχει δώσει ο χρήστης για τα προϊόντα της κατηγορίας δηλαδή θα δίνεται από τον τύπο:

$\text{AvgR}_{i,j} = \frac{\sum_{1}^{n} R_k}{n}$, όπου $\text{AvgR}_{i,j}$ είναι ο μέσος όρος βαθμολογίας της κατηγορίας i για τον χρήστη j, R_k η βαθμολογία του προϊόντος k και n ο συνολικός αριθμός προϊόντων της κατηγορίας i τα οποία έχει βαθμολογήσει ο χρήστης. Π.χ αν υποθέσουμε οτι ο πελάτης με customer_id = 2 (Εικ 6.4) έχει ψηφίσει 3 προϊόντα της κατηγορίας 1 (CPU) με βαθμολογίες 2, 4 και 3 τότε ο μέσος όρος βαθμολογίας του χρήστη για την κατηγορία 1 είναι

$\text{Avg}_{1,2} = \frac{2+4+3}{3} = 3$. Οπότε για κάθε κατηγορία έχουμε ένα πίνακα 1 x n, όπου n το πλήθος των πελατών και η τιμή Array[i] εκφράζει τον βαθμό ενδιαφέροντος του χρήστη i για την κατηγορία αυτή, με βάση τον μέσο όρο βαθμολόγησης των προϊόντων.

product_id	customer_id ▲	vote_rating
1	2	2
2	2	4
4	2	3

Εικόνα 6.4. Στιγμιότυπο του πίνακα poll για τον πελάτη με id = 2

6.1.6 Προχωρημένοι χρήστες

Ένας χρήστης χαρακτηρίζεται ως προχωρημένος από το σύστημα εάν έχει εμφανίσει περισσότερες από 15 φορές τα αναλυτικά χαρακτηριστικά προϊόντων. Αυτή είναι μία σύμβαση η οποία έγινε, ώστε να αξιολογήσουμε την χρησιμότητα ενός τέτοιου χαρακτηριστικού. Προφανώς η ιδιότητα 'advanced user' είναι μια Boolean μεταβλητή η οποία μπορεί να πάρει την τιμή 0 αν ο χρήστης δεν είναι προχωρημένος ή 1 αν είναι. Κατά συνέπεια έχουμε ένα πίνακα 1 x n, όπου n ο αριθμός των χρηστών και η τιμή κάθε θέσης του πίνακα είναι 0 ή 1.

6.2 Επεξεργασία και μετασχηματισμός δεδομένων

Έχοντας συλλέξει τα δεδομένα από την προηγούμενη ενότητα πρέπει να υπολογίσουμε τους πίνακες ομοιότητας για καθένα από τα χαρακτηριστικά γνωρίσματα και έπειτα να τα ομογενοποιήσουμε ώστε να αθροίσουμε όλους τους πίνακες σε ένα συγκεντρωτικό πίνακα ομοιότητας. Όλα τα δεδομένα είναι αποθηκευμένα σε πίνακες στη μνήμη μεγέθους 1 x n, όπου n το πλήθος των πελατών. Κάθε πίνακας απόστασης που θα προκύψει για τα χαρακτηριστικά γνωρίσματα θα έχει μέγεθος n x n και κάθε στοιχείο Array[i][j] θα ισούται με την απόσταση (ανομοιότητα) μεταξύ των χρηστών i και j.

6.2.1 Υπολογισμός πίνακα ομοιότητας για τον μέσο όρο αξίας αγορών

Για να υπολογίσουμε την ομοιότητα δύο χρηστών με βάση τον μέσο όρο αξίας αγορών πρέπει να χρησιμοποιήσουμε μία μέθοδο μέτρησης απόστασης για συνεχείς μεταβλητές. Τέτοιες μεθόδοι περιγράφηκαν διεξοδικά στο κεφάλαιο 4.7.4. Η μέθοδος που χρησιμοποιούμε στη συγκεκριμένη περίπτωση είναι η City-Block distance. Είναι στην διακριτική ευχέρια του διαχειριστή του συστήματος να επιλέξει κάποια άλλη μέθοδο. Στην παρούσα διατριβή έχουν υλοποιηθεί η City Block distance και η ευκλείδια απόσταση για την μέτρηση απόστασης συνεχών μεταβλητών.

Όπως περιγράψαμε στο κεφάλαιο 6.1.1, έχουμε στη διάθεσή μας ένα πίνακα 1 x n, ο οποίος περιέχει τον μέσο όρο της αξίας των αγορών κάθε πελάτη. Σε αυτό το σημείο πρέπει να τονιστεί ότι όπως όλοι οι πίνακες, έτσι κι αυτός, είναι ταξινομημένοι σε αύξουσα σειρά με βάση το αναγνωριστικό του πελάτη (customer_id), για να υπάρχει μία συνέπεια. Χρησιμοποιούμε λοιπόν την μέθοδο City Block distance, η οποία δέχεται ως είσοδο ένα πίνακα 1 x n και επιστρέφει ένα πίνακα απόστασης n x n, ο οποίος περιέχει όλες τις αποστάσεις μεταξύ όλων των χρηστών ανά δύο. Η function CityBlockDistance($arr) έχει ως εξής:

```
function CityBlockDistance($arr)
{
        $d;
        $count;
        $count = count($arr);
                for ($i=0; $i<$count; $i++){
                        for ($j=0; $j<$count; $j++){
                                $d[$i][$j] = abs($arr[$i]-$arr[$j]);
                        }
                }
        return $d;
}
```

Ο πίνακας που επιστρέφει η παραπάνω συνάρτηση πρέπει να κανονικοποιηθεί ώστε να μπορεί να προστεθεί με τους υπόλοιπους πίνακες ομοιότητας σε έναν συγκεντρωτικό πίνακα. Για τον λόγο αυτό έχει δημιουργηθεί η μέθοδος NormalizeArray η οποία δέχεται ως όρισμα (by reference) ένα πίνακα n x n και επιστρέφει τον πίνακα αυτόν κανονικοποιημένο δηλαδή μέσα στο πεδίο τιμών [0, 1]. Παρακάτω παρατίθεται η function NormalizeArray.

```
function NormalizeArray(&$arr)
{
	$count;
	$count = count($arr);
	$maxVal = -10000000;
	// Vres to megalytero stoixeio toy pinaka
	for($i=0;$i<$count;$i++){
		for($j=0;$j<$count;$j++){
			if ($arr[$i][$j] > $maxVal){
				$maxVal = $arr[$i][$j];
			}
		}
	}
	// Diairese ola ta stoixeia toy pinaka me to maxVal
	if ($maxVal !=0)
	{
		for($i=0;$i<$count;$i++){
			for($j=0;$j<$count;$j++){
				$arr[$i][$j] = $arr[$i][$j]/$maxVal;
			}
		}
	}
}
```

Τελευταίο βήμα είναι να προσθέσουμε τον πίνακα που δεχτήκαμε ως έξοδο από την συνάρτηση NormalizeArray στον συγκεντρωτικό πίνακα απόστασης distanceMatrix. Αυτό επιτυγχάνεται μέσω της συνάρτησης aggregate, η οποία παίρνει δύο ορίσματα. Τον συγκεντρωτικό πίνακα απόστασης (by reference) και τον κανονικοποιημένο πίνακα που προέκυψε από την παραπάνω διαδικασία, arr.

```
function aggregate(&$distMatrix, $arr)
{
	$count = count($distMatrix);
	for ($i=0;$i<$count;$i++){
		for ($j=0;$j<$count; $j++){
			$distMatrix [$i][$j] += $arr[$i][$j];
		}
	}
}
```

6.2.2 Υπολογισμός πίνακων ομοιότητας για το σύνολο αγορών και τις επισκέψεις προϊόντων

Ακριβώς τα ίδια βήματα που ακολουθήθηκαν για τον υπολογισμό του πίνακα ομοιότητας του μέσου όρου αξίας των αγορών, θα ακολουθήσουμε και για τον υπολογισμό των πινάκων ομοιότητας για το σύνολο αγορών και τις επισκέψεις προϊόντων. Έτσι τροφοδοτούμε τη συνάρτηση CityBlockDistance με τους πίνακες 1 x n που υπολογίσαμε στις ενότητες 6.1.2 και 6.1.3 η οποία μας επιστρέφει δύο πίνακες n x n (orderMatrix και voteMatrix αντίστοιχα). Συνεχίζοντας με την ίδια λογική, περνάμε τους πίνακες αυτούς ως όρισμα στην συνάρτηση

NormalizeArray προκειμένουν να κανονικοποιηθούν (6.2.1) και τέλος τους κανονικοποιημένους πίνακες τους αθροίζουμε στον συγκεντρωτικό πίνακα distanceMatrix με χρήση της συνάρτησης aggregate.

6.2.3 Υπολογισμός πίνακα ομοιότητας για την βαθμολόγηση καταστήματος και προϊόντων κατηγορίας

Μία επιπλέον επεξεργασία πρέπει να υποστούν τα δεδομένα αυτά λόγω του ότι οι μεταβλητές είναι κατηγορικές (Ranking). Έτσι λοιπόν χρησιμοποιώντας τη μέθοδο Normalize Rank Transformation κανονικοποιούμε τις τιμές των πινάκων που έχουμε από τις 6.1.4 και 6.1.5. Αφού γίνει αυτό τα δεδομένα των δύο πινάκων είναι πλέον στο διάστημα [0, 1] και πλέον πρόκειται για συνεχείς τιμές οπότε μπορούμε να χρησιμοποιήσουμε την συνάρτηση CityBlockDdistance για τη μέτρηση των αποστάσεων. Το βήμα τις κανονικοποίησης των πινάκων μπορούμε να το παραλείψουμε πλέον αφού τα αρχικά μας δεδομένα ήταν ήδη κανονικοποιημένα. Τέλος και για τους δύο πίνακες κάνουμε χρήση της συνάρτησης aggregate ώστε να προστεθούν στον συγκεντρωτικό πίνακα ομοιότητας.

6.2.4 Υπολογισμός πίνακα ομοιότητας για προχωρημένους χρήστες

Στην περίπτωση αυτή έχουμε να κάνουμε με δυαδικά δεδομένα (binary data) δηλαδή 0 ή 1. Για να υπολογίσουμε την απόσταση των δυαδικών διανυσμάτων απλά θα χρησιμοποιήσουμε την μέθοδο Hamming Distance στην οποία θα τροφοδοτήσουμε τον πίνακα που προέκυψε από το 6.1.6. Η συνάρτηση HammingDist είναι η παρακάτω:

```
function HammingDist($arr)
{
        $rows = count($arr);
        $cols = count($arr[0]); // columns = ari8mos idiotitvn
        for($i=0; $i<$rows; $i++){
                for($j=0; $j<$rows; $j++){
                        $total = 0;
                        for($h=0; $h<$cols; $h++){
                                if ($arr[$i][$h] != $arr[$j][$h]){
                                        $total++;
                                }
                        }
                        $d[$i][$j] = $total/$cols;
                }
        }
        return $d;
}
```

Τέλος προσθέτουμε τον πίνακα που επιστρέφεται στον συγκεντρωτικό πίνακα, χωρίς προηγουμένως να τον κανονικοποιήσουμε αφού οι τιμές του είναι ήδη στο διάστημα [0,1].

6.2.5 Τελικός υπολογισμός πίνακα ομοιότητας

Στα προηγούμενα βήματα συγχωνεύσαμε τους έξι κανονικοποιημένους πίνακες σε ένα συγκεντρωτικό πίνακα ομοιότητας, οι οποίοι πρέπει να επισημάνουμε ότι εκτυπώνονται στην σελίδα για καλύτερη εποπτεία. Οι τιμές καθενός πίνακα ήταν στο διάστημα [0,1], όμως τώρα ο τελικός πίνακας πρέπει και αυτός να κανονικοποιηθεί προκειμένου να έχουμε σωστά αποτελέσματα. Οπότε περνάμε στην συνάρτηση NormalizeArray τον συγκεντρωτικό πίνακα ως

όρισμα και πλέον έχει προκύψει ο πίνακας ο οποίος θα τροφοδοτήσει την καρδιά του συστήματος, τον αλγόριθμο ιεραρχικής ομαδοποίησης.

6.3 Επεξήγηση αλγόριθμου ιεραρχικής ομαδοποίησης

Στο προηγούμενο υποκεφάλαιο πραγματοποιήθηκαν όλοι οι απαραίτητοι υπολογισμοί και μετασχηματισμοί των δεδομένων, προκειμένου να δημιουργηθεί ο πίνακας απόστασης. Σε αυτό το σημείο είμαστε έτοιμοι να τρέξουμε τον αλγόριθμο ιεραρχικής ομαδοποίησης.

6.3.1 Λεπτομέρειες αλγορίθμου

Ο αλγόριθμος καλείται με χρήση της εντολής PHP hcluster($distanceMatrix, $n, $method);. Η συνάρτηση αυτή, όπως και όλες οι προαναφερθείσες συναρτήσεις συμπεριλαμβάνονται στο αρχείο κώδικα ‘clustering.php’. Το πρώτο όρισμα της συνάρτησης hcluster είναι ο πίνακας απόστασης. Το δεύτερο όρισμα είναι η τάξη του πίνακα δηλαδή ο αριθμός των πελατών και τέλος το τρίτο όρισμα είναι η μέθοδος που θα χρησιμοποιηθεί για τον ορισμό της απόστασης μεταξύ των ομάδων. Έτσι οι τιμές που μπορεί να πάρει το τρίτο όρισμα είναι οι εξής:

- Για method = 1, ο αλγόριθμος τρέχει με την μέθοδο single-linkage
- Για method = 2, ο αλγόριθμος τρέχει με την μέθοδο average-linkage
- Για method = 3, ο αλγόριθμος τρέχει με την μέθοδο complete-linkage
- Για method = 4, ο αλγόριθμος τρέχει με την μέθοδο ward

Κατά την διάρκεια εκτέλεσης του αλγορίθμου εκτυπώνονται μετά απο κάθε ανακύκλωσή του στην οθόνη, ο ανανεωμένος πλέον πίνακας απόστασης, όλες οι ομάδες όπως έχουν προκύψει μετά από κάθε βήμα καθώς επίσης και η τιμή που επιλέχθηκε ως κριτήριο για την επιλογή της ομαδοποίησης.

6.3.2 Χρόνος εκτέλεσης

Όπως είδαμε και στο θεωρητικό υπόβαθρο (κεφ 4.6) όλοι οι αλγόριθμοι τους οποίους περιγράφουμε είναι greedy και τρέχουν σε πολυωνυμικό χρόνο. Αυτό έχει ως συνέπεια όσο αυξάνονται οι χρήστες να αυξάνεται και ο χρόνος που απαιτείται για να ολοκληρωθεί ο αλγόριθμος. Για τους 72 χρήστες ο μέσος χρόνος ολοκλήρωσης του αλγορίθμου αναλόγως και της μεθόδου ήταν 2-3 λεπτά. Όταν όμως ο αλγόριθμος έτρεχε στον απομακρυσμένο εξυπηρετητή στον οποίο ανεβάσαμε το ηλεκτρονικό κατάστημα (http://www.renko-webhosting.gr/e-shop), τότε ο μέσος χρόνος εκτέλεσης υπερέβαινε τα 5 λεπτά.

6.3.3 Εκτέλεση αλγορίθμου

Ο χρόνος εκτέλεσης του αλγορίθμου όπως επισημάνθηκε παραπάνω δεν επιτρέπει στο σύστημα να λειτουργήσει σε πραγματικό χρόνο (real time) κι ενώ ταυτόχρονα λειτουργεί το κατάστημα. Δηλαδή όταν κάποιοι χρήστες κάνουν μερικές κινήσεις οι οποίες ενδεχομένως να επηρεάζουν το αποτέλεσμα της ομαδοποίησης, δεν είναι εφικτό (ειδικά σε ένα πραγματικό σύστημα) να γίνει ο επαναϋπολογισμός των ομάδων δυναμικά (on the fly). Έτσι ο αλγόριθμος ομαδοποίησης επιτρέπεται να εκτελούνται μόνο offline. Υπεύθυνος για την εκτέλεση του clustering είναι μόνο ο διαχειριστής του συστήματος. Το μόνο που πρέπει να κάνει είναι να πληκτρολογήσει σε έναν browser την σελίδα clustering.php (π.χ. αν τρέχει το σύστημα σε τοπικό επίπεδο και τα αρχεία βρίσκονται στον virtual folder ‘e-shop’, τότε η διεύθυνση μπορεί να είναι http://localhost/e-shop/clustering.php). Φυσικά είναι στο χέρι του διαχειριστή να εφαρμόσει ένα scheduler, ώστε η σελίδα να τρέχει αυτόματα ανά τακτά χρονικά διαστήματα (π.χ. ανα εβδομάδα).

6.3.4 Έξοδος αποτελεσμάτων αλγορίθμου

Ο αλγόριθμος ιεραρχικής ομαδοποίησης που περιγράφεται παραπάνω αποθηκεύει τα αποτελέσματα που παράγονται στην βάση δεδομένων στον πίνακα clusters (Εικ 6.5). Αυτά θα χρησιμοποιηθούν αργότερα (από το Recommender system), προκειμένου να παραχθούν προτάσεις στους χρήστες και να προσαρμοστεί το σύστημα στις ανάγκες τους.

Σε κάθε ένα από τα επαναληπτικά βήματα του αλγορίθμου, για κάθε πελάτη αποθηκεύονται στη βάση πληροφορίες όπως:

- Ένας μοναδικός αριθμός (cluster_id) ο οποίος χαρακτηρίζει μοναδικά την ομάδα
- Το customer_id για του πελάτη ο οποίος ανήκει στην ομάδα
- Το τρέχον επίπεδο ομαδοποίησης (level)
- Η τιμή του επιπέδου, που στην ουσία είναι η απόσταση της ομάδας που δημιουργήθηκε.

cluster_number	customer_id	cluster_level	cluster_value
5	7	42	28
6	13	42	28
7	14	42	28
7	44	42	28
8	15	42	28

Εικ 6.5. Στιγμιότυπο του πίνακα clusters

Είναι προφανές ότι με αυτόν τον τρόπο έχουμε μία επισκόπηση για όλη την διαδρομή που ακολούθησε ο αλγόριθμος (bottom to top) από την αρχική φάση όπου κάθε χρήστης αποτελεί μια ομάδα έως το τελευταίο στάδιο όπου όλοι οι χρήστες είναι ενταγμένοι σε μία ομάδα.

6.4 Διαγραμματική αναπαράσταση αποτελεσμάτων ομαδοποίησης

Πριν την εκτέλεση του αλγορίθμου clustering, αποθηκεύονται σε ένα αρχείο κειμένου οι τιμές (διαχωρισμένες με τον χαρακτήρα '|') του τελικού πίνακα ομοιότητας, έτσι ωστέ αυτό να μπορεί να διαβαστεί από την εφαρμογή Matlab, την οποία θα χρησιμοποιήσουμε για να εξάγουμε τα δενδρογράμματα που προκύπτουν από την χρήση του αλγορίθμου.

6.4.1 Πρόγραμμα matlab για την δημιουργία δενδρογραμμάτων

Το πρόγραμμα το οποίο χρησιμοποιήθηκε για την δημιουργία των δεδρογραμμάτων στην εφαρμογή matlab είναι το ακόλουθο:

```
clear all
k=10;
a=xlsread('distance_matrix.xls',1,'A1:Bt72');   %/ read from Excel
%/X=a';
%/Y=pdist(X);
Y=squareform(a);
Z=linkage(Y,'method');                  %/ creates a hierarchical cluster tree, using the method
 clusters =cluster(Z,'maxclust',k);
t = sort(Z(:,3));
th = t(size(Z,1)+2-k);
[H,T] = dendrogram(Z,40,'colorthreshold',th);
set(H,'LineWidth',1)
title('Agglomerative Algorithm - Complete Linkage');
```

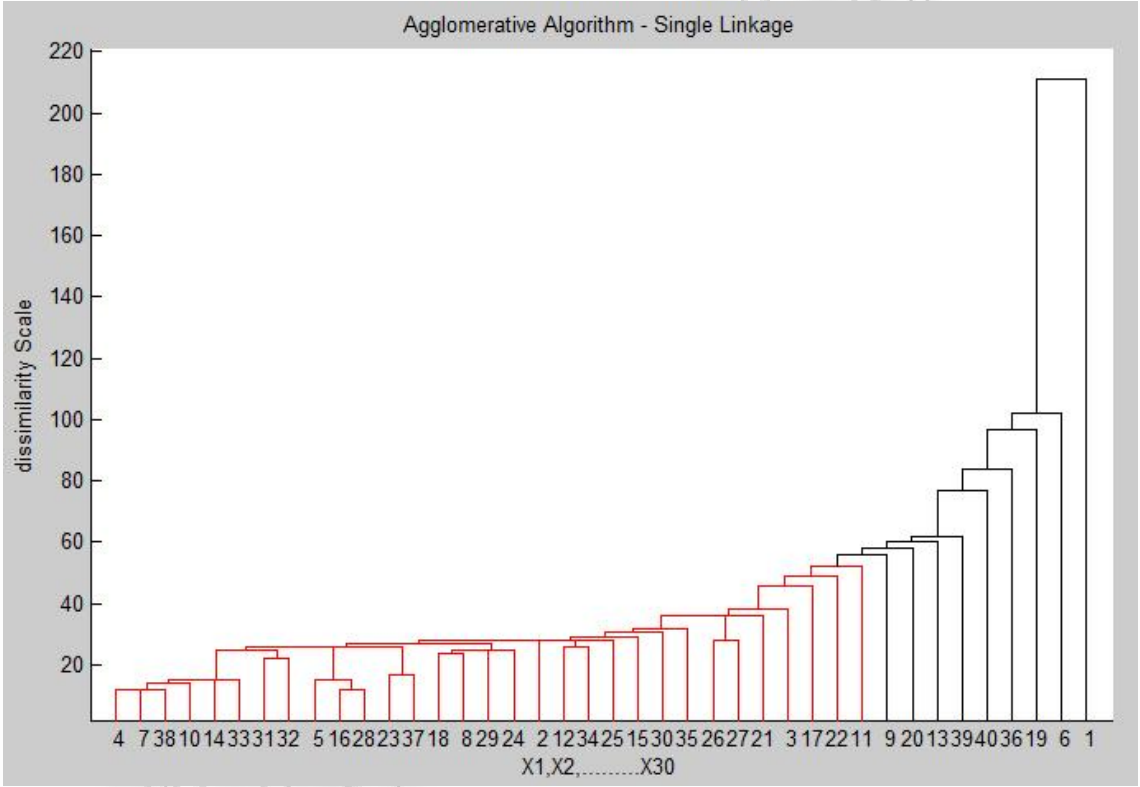
xlabel('X1,X2,.....X30');

ylabel('dissimilarity Scale');

Για να παράγουμε δενδρογράμματα για κάθε μία από τις 4 μεθόδους αλλάζουμε το όρισμα στην εντολή linkage. Στη θέση της 'method' βάζουμε single για τη μέθοδο single linkage, complete, για την μέθοδο complete linkage, average για την μέθοδο average linkage και ward για την μέθοδο ward. Το πρόγραμμα διαβάζει τα δεδομένα από ένα αρχείο Excel και εν συνεχεία, αφού δημιουργήσει τους πίνακες που χρειάζονται, εφαρμόζει την μέθοδο linkage την οποία επιλέγουμε εμείς κάθε φορά. Τέλος για την δημιουργία του δενδρογράμματος χρησιμοποιείται η εντολή dendrogram με τις κατάλληλες παραμέτρους, όπως π.χ. την παράμετρο threshold, βάσει της οποίας κόβεται και χρωματίζεται το δενδρόγραμμα.

6.4.2 Δενδρόγραμμα για την μέθοδο single-linkage

Η μέθοδος single linkage, όπως είδαμε και στο θεωρητικό υπόβαθρο (Κεφ. 4.6.1), είναι ιδιαίτερα δημοφιλής σε ερευνητικό επίπεδο, όμως στην πράξη φαίνεται να μην έχει τόσο επιθυμητά αποτελέσματα.



Εικόνα 6.6. Δενδόγραμμα μεθόδου single-linkage για 72 χρήστες

Όπως μπορούμε να διακρίνουμε στο παραπάνω δενδρόγραμμα (Εικ. 6.6), η μέθοδος single-linkage τείνει να δημιουργεί μία μεγάλη μη κυκλική ομάδα (διακρίνεται με κόκκινο χρώμα). Είναι εύκολο να διαπιστώσουμε ότι ενώ η ομαδοποίηση βρίσκεται σε αρκετά προχωρημένο στάδιο (αν κόψουμε το διάγραμμα στην τιμή 50), εντούτοις έχει δημιουργηθεί μόνο μία ομάδα (φαινόμενο chain effect). Προφανώς αυτό το είδος ομαδοποίησης δεν είναι ιδιαίτερα χρήσιμο ώστε να εξαχθούν ασφαλή συμπεράσματα για τους χρήστες.

6.4.3 Δενδρόγραμμα μεθόδου complete-linkage

Το φαινόμενο chain effect είδαμε οτι καθιστά τη μέθοδο single-linkage ακατάλληλη για χρήση στο σύστημά μας λόγω των μεγάλων και μακροσκελών ομάδων που δημιουργεί. Αντίθετα η

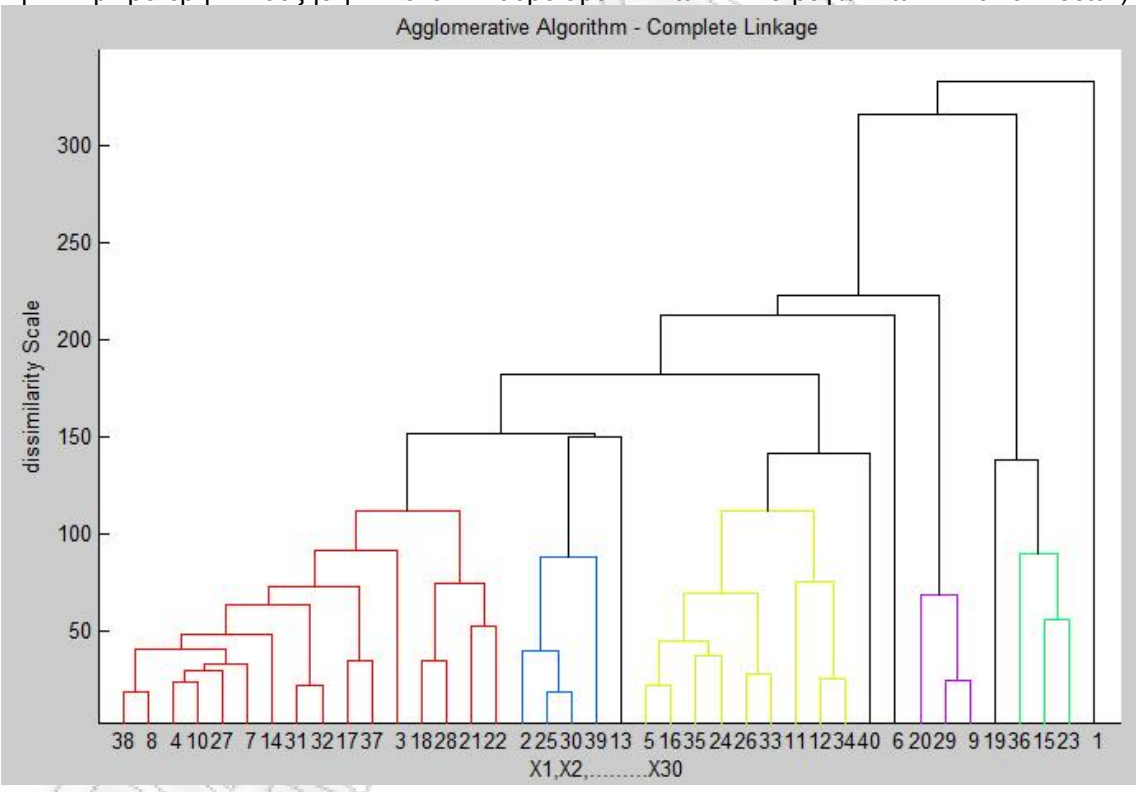
μέθοδος complete-linkage, όπως βλέπουμε παρακάτω (Εικ. 6.7) δημιουργεί πιο συμπαγής και ευδιάκριτες ομάδες. Στο υποφαινόμενο διάγραμμα είναι εμφανής η ύπαρξη 5 ομάδων ο χρωματισμός των οποίων κάνει ευδιάκριτη την διάκρισή τους (από αριστερά προς τα δεξιά διακρίνουμε την κόκκινη, την μπλε, την κίτρινη την μωβ και την πράσινη ομάδα). Ενω λοιπόν η μέθοδος αυτή δημιουργεί αισθητά καλύτερες και ισομοιρασμένες ομάδες, ωστόσο είναι ευαίσθητη στην ύπαρξη θορύβου, δηλαδή ακραίων τιμών (outliers).

6.4.4 Δενδόγραμμα μεθόδου average-linkage

Η μέθοδος average-linkage προσπαθεί να συβιβάσει τις αδυναμίες των δύο προαναφερόμενων ομάδων. Ωστόσο ενώ τα αποτελέσματα που παράγει (Εικ. 6.8) είναι εμφανώς καλύτερα από αυτά της μεθόδου single-linkage, φαίνεται να υστερεί σε σχέση με αυτά της μεθόδου complete-linkage και ο λόγος είναι η τάση που έχει να δημιουργεί μεγάλες σφαιρικές ομάδες κάτι που μπορούμε να διακρίνουμε και στο δενδρόγραμμα (δύο ομάδες μπλε και κόκκινη).

6.4.5 Δενδρόγραμμα μεθόδου Ward

Η μέθοδος Ward προσπαθεί να δημιουργήσει ομάδες με τέτοιο τρόπο ώστε να ελαχιστοποιεί το σφάλμα που προκαλείται από κάθε πιθανή ομαδοποίηση (δημιουργεί την ομάδα που προκαλεί την μικρότερη αύξηση στο άθροισμα των τετραγωνικών αποκλίσεων).



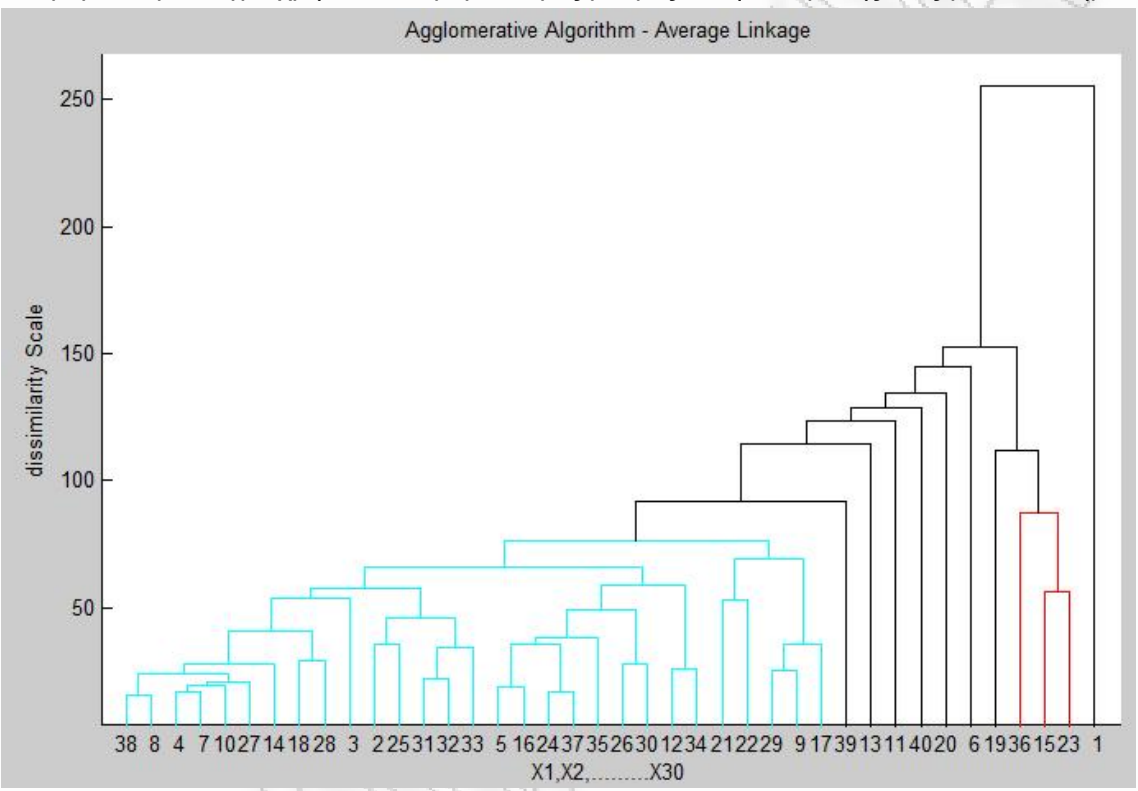
Εικόνα 6.7. Δενδόγραμμα μεθόδου complete-linkage για 72 χρήστες

Είναι ορατό από το δενδροδιάγραμμα (Εικ. 6.9) ότι πρόκειται για την μέθοδο με τα καλύτερα αποτελέσματα. Όπως διακρίνουμε από τον χρωματισμό του διαγράμματος έχει δημιουργήσει επτά ομάδες (από αριστερά προς τα δεξιά με χρώμα ροζ, κίτρινο, πράσινο, μπλε, μωβ, θαλασσί και κόκκινο) και μάλιστα σχεδόν ισοσκελισμένες, δηλαδή με παραπλήσιο αριθμό χρηστών για την κάθε ομάδα.

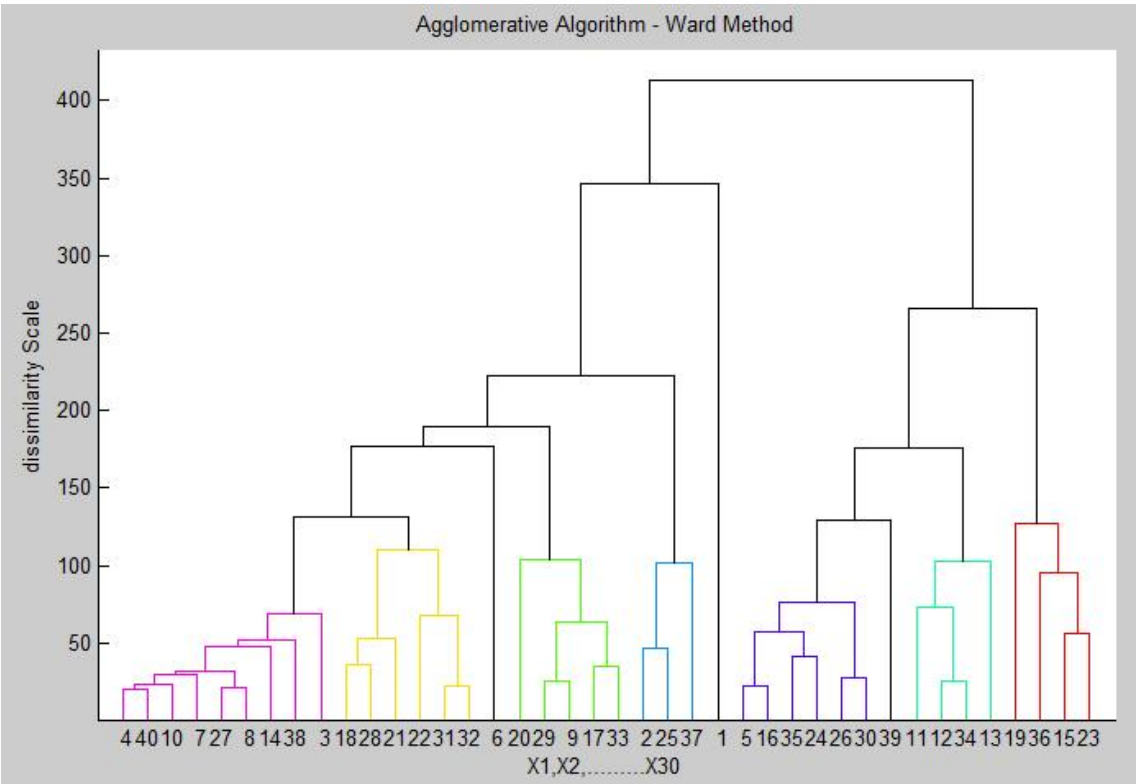
6.4.6 Συμπεράσματα που απορρέουν από τα δενδρογράμματα

Προηγουμένως παραθέσαμε τα πλεονεκτήματα και τα μειονεκτήματα της κάθε μεθόδου, συνεπικουρούμενοι από τις διαγραμματικές τους απεικονίσεις. Πρώτο συμπέρασμα που μπορεί να εξαχθεί είναι ότι πρέπει να απορρίψουμε τη μέθοδο single-linkage ως την πλέον απαράδεκτη αφού κατάφερε να δημιουργήσει μόνο μία ομάδα. Αμέσως καλύτερη είναι η μέθοδος average-linkage με τη δημιουργία 2 κυκλικών ομάδων.

Οι δύο καλύτερες μεθόδοι όπως προκύπτει από τα διαγράμματα είναι σαφώς η complete-linkage και η μέθοδος Ward. Μεταξύ αυτών των δύο, δεν είναι ξεκάθαρη κάποια σαφής υπεροχή, αν και η μέθοδος Ward φαίνεται να προκρίνεται με βραχεία κεφαλή λόγω της δημιουργία πιο ομοιόμορφων ομάδων. Ωστόσο για πιο ασφαλή συμπεράσματα, θα χρειαστεί ένα μεγαλύτερο δείγμα χρηστών και μεγαλύτερος βαθμός αλληλεπίδρασής τους με το σύστημα.



Εικόνα 6.8. Δενδόγραμμα μεθόδου average-linkage για 72 χρήστες



Εικόνα 6.9. Δενδόγραμμα μεθόδου ward για 72 χρήστες

7. ΣΥΜΠΕΡΑΣΜΑΤΑ – ΣΥΝΕΙΣΦΟΡΑ

Στο τελευταίο κεφάλαιο της παρούσας εργασίας θα επιχειρηθεί μία σύνοψη των κυριότερων σημείων που αναλύθηκαν στα προηγούμενα κεφάλαια καθώς και η συνεισφορά της σε αυτόν τον τομέα έρευνας. Επίσης θα παρατεθούν ορισμένες ιδέες για μελλοντική επέκταση της εργασίας, όπως και θα προταθούν μερικές βελτιώσεις οι οποίες θα συντελέσουν στην αποτελεσματικότερη λειτουργία του συστήματος.

7.1 Συνεισφορά

Μετά από εκτεταμένη έρευνα, τα περισσότερα επιστημονικά συγγράμματα και ειδικά τα πιο σύγχρονα, σχετικά με το θέμα τα οποίο εξετάζουμε, ασχολούνται ως επί το πλείστον με την μοντελοποίηση των προϊόντων δηλαδή είναι συστήματα item-based. Εμείς επιχειρήσαμε μια user-based προσέγγιση του θέματος βασιζόμενοι μόνο στις κινήσεις των χρηστών στο ηλεκτρονικό μας κατάστημα.

Επίσης στην πλειοψηφία τους τα περισσότερα item-based συστήματα το μόνο χαρακτηριστικό πάνω στο οποίο βασίζουν την μοντελοποίησή τους, είναι η ψήφος των χρηστών για κάθε προϊόν. Αντίθετα, στην παρούσα εργασία δημιουργήσαμε τα διανυσματικά προφίλ των χρηστών από πολλά διαφορετικά χαρακτηριστικά γνωρίσματα, (συμπεριλαμβανομένης και της ψήφου), σχηματίζοντας με αυτόν τον τρόπο μία πιο σφαιρική άποψη για κάθε χρήστη.

Αν και τα ηλεκτρονικά καταστήματα προϊόντων πληροφορικής είναι πολυ διαδεδομένα τα τελευταία χρόνια, εντούτοις δεν βρήκαμε κάποια σύγγραμματα που να ασχολούνται αποκλειστικά με τη συγκεκριμένη κατηγορία καταστημάτων. Τα περισσότερα πραγματικά συστήματα που εξετάσαμε εμπορεύονται ταινίες, βιβλία, είδη ρουχισμού, μουσική κ.α. Η ιδιαιτερότητα που πιθανώς έχουν τα προϊόντα πληροφορικής είναι ότι ανανεώνονται με ραγδαία ταχύτητα και διαθέτουν τεράστια ποικιλομορφία. Έτσι είναι λίγο δύσκολο να δουλέψει για παράδειγμα το μοντέλο Netflix [Κεφ. 2.3.9], το οποίο επιχειρεί να διαφημίσει όλο το εύρος των

προϊόντων του και κυρίως των παλαιότερων. Δηλαδή είναι πιο δύσκολο να 'πλασάρεις' στους χρήστες ένα παλαιότερο προϊόν πληροφορικής, τη στιγμή που καθημερινά αυτά ανανεώνονται, από το να πλασάρεις ένα παλιό βιβλίο ή ταινία. Η έννοια του 'κλασσικού' στην πληροφορική δεν συναντάται τόσο συχνά ιδιαίτερα στην κατηγορία του hardware. Έτσι ακόμα και αν οι χρήστες έχουν βαθμολογήσει ένα προϊόν με άριστα, αν αυτό είναι αρκετά παλιό και πιθανώς ξεπερασμένο, τότε μικρή σημασία έχει να το προτείνεις στον χρήστη.

Τέλος συγκρίναμε τέσσερις δημοφιλής μεθόδους ιεραρχικής ομαδοποίησης, η οποία ναι μεν είναι μια μέθοδος η οποία δύσκολα θα εφαρμοστεί για τη μοντελοποιήση χρηστών σε καταστήματα που έχουν χιλιάδες ή εκατομμύρια πελάτες (γιατί είναι ασύμφορη υπολογιστικά), ωστόσο θεωρούμε ότι η μοντελοποίηση που περιγράψαμε (φυσικά με κάποιες βελτιώσεις – προσθήκες), μπορεί να αποτελέσει μια αξιόλογη λύση για μικρομεσαίες επιχειρήσεις καθώς είναι και συμφέρουσα οικονομικα (χρήση μόνο open-source εργαλείων).

7.2 Συμπεράσματα

Το κυριότερο συμπέρασμα που μπορεί να εξαχθεί από την παρούσα εργασία έχει να κάνει με την σύγκριση των τεσσάρων διαφορετικών μεθόδων ιεραρχικής ομαδοποίησης που εφαρμοστηκαν (Κεφ 6). Όπως είδαμε η μέθοδος Ward προκρίνεται ως η καλύτερη λύση καθώς τείνει να οδηγεί στον σχηματισμό ισοσκελισμένων και και ισομερώς καταμερισμένων ομάδων.

Αμέσως μετά έπεται η μέθοδος complete-linkage, η οποία δημιουργεί συμπαγείς ομάδες αλλά όχι τόσο ισοσκελισμένες όσο η Ward, έχοντας ως μειονέκτημα την ευαισθησία σε ακραίες τιμές (outliers).

Τελευταίες στην αξιολόγηση μέθοδοι, είναι κατά σειρά η average-linkage, με χειρότερη όλων την μέθοδο single-linkage. Η πρώτη έχει την τάση να δημιουργεί ομάδες με μικρές αποκλίσεις λόγω του ότι στον υπολογισμό της απόστασης λαμβάνουν μέρος όλα τα σημεία μιας ομάδας και έχει την τάση να επηρεάζεται από τα ακραία σημεία. Αντίθετα η μέθοδος single-linkage πάσχει από το σύνδρομο chain-effect το οποίο ευθύνεται για τη δημιουργία μεγάλων μη σφαιρικών ομάδων.

Ένα άλλο συμπέρασμα το οποίο μπορεί να εξαχθεί είναι ότι οι αλγόριθμοι της ιεραρχικής ομαδοποίησης που εξετάσαμε είναι υπολογιστικά ασύμφοροι για να χρησιμοποιηθούν σε real-time συστήματα με πολύ μεγάλο όγκο χρηστών. Ωστόσο σε μικρότερα συστήματα και πιθανώς με τη χρήση κατανεμημένων συστημάτων, θα μπορούσε να εφαρμοστεί χωρίς ιδιαίτερα προβλήματα.

7.3 Βελτιώσεις – Μελλοντικά πλάνα

Σε αυτή την ενότητα θα παρουσιάσουμε μερικές σκέψεις για βελτιώσεις ή προσθήκες που μπορούν να γίνουν ώστε να βελτιωθεί το μελλοντικό κατάστημα.

7.3.1 Διεπαφή διαχείρησης της ομαδοποίησης

Μία ιδέα για καλύτερη διαχείρηση της λειτουργίας της ομαδοποίησης είναι η κατασκευή μιας διεπαφής, από την οποία ο διαχειριστής του συστήματος να μπορεί να θέτει ορισμένες παραμέτρους όπως:

- Ποια μέθοδο ιεραρχικής ομαδοποίησης να εφαρμόσει.
- Να καθορίσει στις πόσες ομάδες θα τερματίσει ο αλγόριθμος (threshold)
- Να επιλέξει ποιες μετρικές μεθόδοι θα εφαρμοστούν αναλόγως με την κατηγορία δεδομένων (πχ. Συνεχείς τιμές, κατηγορικές κτλ)
- Να καθορίσει την μορφή των αποτελεσμάτων (εξαγωγή σε Excel, ascii κτλ)
- Να μπορεί να καθορίσει συστηματική εκτέλεση του αλγόριθμου χωρίς να παρεμβαίνει ο ίδιος
- Να έχει άμεση εποπτεία των εξαγόμενων αποτελεσμάτων σε περιβάλλον web

7.3.2 Συμπερίληψη περισσότερων χαρακτηριστικών γνωρισμάτων

Η διαδικασία του να συμπεριληφθούν περισσότερα χαρακτηριστικά γνωρίσματα στην διαδικασία της ομαδοποίησης μπορεί να συνεισφέρει στην αύξηση της ποιότητας των αποτελεσμάτων. Μερικά από αυτα μπορεί να είναι τα εξής:

- Να προστεθούν modules για τη διαχείριση των δημοφιλέστερων μέσων κοινωνικής δικτύωσης (Facebook, twitter).Έτσι όχι μόνο έχουμε ένα επιπλέον χαρακτηριστικό γνώρισμα για κάθε χρήστη, αλλα βοηθάμε και στην διάδοση νέων του καταστήματος μέσω των δικτύων αυτών.
- Αποθήκευση και επεξεργασία των αναζητήσεων που πραγματοποιεί ένας χρήστης με σκοπό την εξαγωγή χρήσιμης πληροφορίας από αυτές
- Προσθήκη Wish-List
- Προσθήκη δυνατότητας για να μπορούν οι χρήστες να εκφράζουν το σχόλιό τους οι χρήστες για κάθε προϊόν (reviews) και εύρεση μεθόδων για λεκτική ανάλυσή τους (textual analysis)

7.3.3 Βελτίωση και προσθήκη αλγορίθμων

Η προσθήκη επιπλέον αλγορίθμων (π.χ αλγόριθμος BIRCH) θα βοηθήσει στην αξιολόγηση και σύγκριση αποτελεσμάτων μεταξύ περισσότερων επιλογών, συμβάλλοντας στην εξαγωγή ασφαλέστερων συμπερασμάτων.

Επίσης βελτιώσεις μπορούν να γίνουν μειώνοντας τις κλήσεις SQL προς τη βάση (μείωση roundtrips) και χρησιμοποιώντας stored-procedures ή άλλες δυνατότητες που προσφέρουν τα σύγχρονα RDBMS.

Ακόμα για την αποφυγή των roundtrips στον server, μπορούμε να αποθηκεύουμε περισσότερα προσωρινά δεδομένα στη μεταβλητή session του browser (με ιδιαίτερη προσοχή γιατί θα τεθούν θέματα ασφαλείας).

Τέλος μπορούμε να σκεφτούμε την χρησιμοποιήσουμε μερικών δομών δεδομένων (διπλές συνδεδεμένες λίστες, R-trees κ.α) για την αποθήκευση των ενδιάμεσων αποτελεσμάτων του αλγόριθμου ομαδοποίησης, προκειμένου να βελτιώσουμε την πολυπλοκότητα του σε χρόνο και χώρο.

7.4.4 Προσθήκη item-based και content ή rule-based μεθόδων

Η προσθήκη μίας ή περισσότερων item-based μεθόδων, όπως π.χ. Pearson coorelation ή slope one, θα βοηθούσε στην σύγκριση με τις model-based μεθόδους που έχουμε ήδη εφαρμόσει. Επίσης η προσθήκη ενός content based ή rule based αλγορίθμου θα βοηθούσε στην αντιμετώπιση καταστάσεων cold-start, δηλαδή καταστάσεων όπου δεν υπάρχουν δεδομένα για κάποιον νέο χρήστη ή για έναν επισκέπτη που δεν έχει εγγραφεί.

Βιβλιογραφία

[1] D.N. Sotiropoulos, G.A. Tsihrintzis, A. Savvopoulos, and M. Virvou, A Comparison of Customer Data Clustering Techniques in an e-Shopping Application , in: Proceedings of 2nd International Workshop on Web Personalization, Recommender Systems and Intelligent User Interfaces, Ireland, 2006

[2] J. Ben Schafer, Joseph Konstan, John Riedl, Recommender Systems in E-Commerce, In EC '99 Proceedings of the 1st ACM conference on Electronic commerce

[3] M. Virvou, A. Savvopoulos, An intelligent TV-shopping Application that provides Recommendations, 19th IEEE International Conference on Tools with Artificial Intelligence ICTAI 2007, 29-31 October 2007, Patras ,Greece.

[4] Μ. Βίρβου, 'Σημειώσεις για το μάθημα Μοντελοποίηση Χρηστών'

[5] Μ. Βίρβου, 'Σημειώσεις για το μάθημα Ειδικά θέματα Τεχνολογίας Λογισμικού'

[6] Bill Kules, User Modeling for Adaptive and Adaptable Software Systems Department of Computer Science University of Maryland, College Park, MD 20742 USA

[7] M. Virvou, A. Savvopoulos, D. N. Sotiropoulos, G. A. Tsihrintzis : Constructing Stereotypes for an Adaptive e-Shop using AIN-based Clustering, International Conference on Adaptive and Natural Computing Algorithms ICANNGA 2007, 11-14 April 2007

[8] [SFW99] Spiliopoulou, M., Faulstich, L. C. and Wilkler, K.: *A data miner analyzing the navigational behavior of Web users*, Proceedings of the Workshop on Machine Learning in User Modeling of the ACAI99, Chania, Greece, pp.54-64, 1999

[9] [PPK+00] Paliouras, G., Papatheodorou, C., Karkaletsis, V. and Spyropoulos, C. D.:*Clustering the Users of Large Web Sites into Communities*, Proceedings of International Conference on Machine Learning (ICML), Stanford, California, pp. 719-726, 2000

[10] B. M. Sarwar, G. Karypis, J. A. Konstan, and J.Riedl, Item-based collaborative _ltering recommendation algorithms, in Proceedings of the 10th International World Wide Web Conference (WWW10), Hong Kong, May 2001.

[11] Tom Morrison, Uwe Aickelin, An Artificial Immune System as a Recommender for Web Sites, In Proceedings of the 1st Internal Conference on ARtificial Immune Systems (ICARIS-2002), pp 161-169, Canterbury, UK, 2002.

[12] Η. Κοντοδιός, Προσαρμοστικό σύστημα δημιουργίας προτάσεων σε ηλεκτρονικό κατάστημα.

[13] Bing Liu, Web data mining: exploring hyperlinks, contents, and usage data

[14] Marden J. I., 1995, Analyzing and Modeling Rank Data, Chapman & HLI, London

[15] Everit, B.S., 1978, Graphical Techniques for Multivariate data

[16] Badrul Saruar, George Karypis, Joseph Konstan and John Riedl, Item-based Collaborative Filtering Recommendation Algorithms

[17] Hadi Khosravi Farsani, and Mohammadali Nematbakhsh, A Semantic Recommendation Procedure for Electronic Product Catalog

[18] Al Mamunur Rashid George Karypis John Riedl, Inuence in Ratings-Based Recommender Systems: An Algorithm-Independent Approach

[19] P. Resnick, N. Iacovou, M. Sushak, P. Bergstrom, and J. Riedl, Grouplens: An open architecture for collaborative _ltering of netnews, in Proceedings of CSCW 1994, ACM SIG Computer Supported Cooperative Work, 1994.

[20] Εμμ. Α. Γιακουμάκης, Τεχνολογία λογισμικού, Τόμος Β, Εκδόσεις Α. Σταμούλη 1993.

[21] J. S. Breese, D. Heckerman and C. Kadie, Empirical analysis of predictive algorithms for collaborative _ltering, in Proceedings of the Fourteenth Annual Conference on Uncertainty in AI, July 1998.

[22] Karamolegkos, P.N., Patrikakis C.Z., Doulamis N.D, Tragos E.Z., User - Profile based Communities Assessment using Clustering Methods in Personal, Indoor and Mobile Radio Communications, 2007, PIMRC 2007. IEEE 18th International Symposium

[23] J. Ben Schafer, Joseph A. Konstan, John Riedl E-Commerce Recommendation Applications

[24] P. Domingos and M. Richardson, Mining the Network Value of Customers, Proceedings of the Seventh International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, 2001. ACM Press

[25] L. Kaufman and P.J Rousseeuw. Finding Groups in Data, Wiley 1990.

[26] J. H. Ward. Hierarchical groupings to optimize an objective function. Journal of the American Statistical Association, 58:234 244, 1963

[27] C. Fraley. Algorithms for Model-Based Gaussian Hierarchical Clustering, 1996

[28] George Karypis. Evaluation of Item-Based Top-N Recommendation Algorithms, in CIKM ’01 Proceedings of the tenth international conference on Information and knowledge management

[29] R. Burke and J. Sandvig, Model-Based Collaborative Filtering as a Defense Against Profile Injection Attacks, In Proceedings of the 21st National Conference on Artificial Intelligence (AAAI ’06), Boston, Massachusetts, July 16-20, 2006

[30] Εμμ. Σκορδαλάκη, Εισαγωγή στην τεχνολογία λογισμικού, 1991

[31] Y. H Wu and A. L. P. Chen, Index Structures of User Profiles for Efficient Web Page Filtering Services, Proceedings of IEEE Conference on Distributed Computing Systems, pp. 644-651, April 2000.

[32] T. W. Yan and H. Garcia-Molina, Index Structures for Selective Dissemination of Information under the Boolean Model, ACM Transactions on Database Systems, 19(2): 332-364, 1994.

[33] M. J. Pazzani, Daniel Billsus, Content-Based Recommendation Systems, Lecture Notes in Computer Science Volume 4321/2007

[34] Bergstrom, S. Raberg, L. (2003). Adopting The Rational Unified Process: Success With RUP. Addison-Wesley

[35] Booch, G. Jacobson, I. Rumbaugh, J. (1999). The Unified Software Development Process. Addison-Wesley

[36] Yanchu Zhang, Guandong Xu, Xiaofang Zhou, A Latent Usage Approach for Clustering Web Transaction and Building User Profile

[37] Michael W. Trosset, Representing Clusters: K-Means Clustering, Self-Organizing Maps, and Multidimensional Scaling, February 20, 2008

[38] Daniel Lemire, Anna Maclachlan, Slope One Predictors for Online Rating-Based Collaborative Filtering, In SIAM Data Mining (SDM'05), Newport Beach, California, April 21-23, 2005.

[39] Umardand Shripad Manikrao, T.V.Prabhakar, Dynamic Selection of Web Services with Recommendation System, International Conference on Next Generation Web Services Practices, August 2005, Seoul, Korea.

[40] Petros Drineas, Iordanis Kerenidis, and Prabhakar Raghavan,Competitive recommendation systems. In Proc. of the thiry fourth annual ACM symposium on Theory of computing

[41] Philip Bonhard, Clare Harries, John McCarthy, M. Angela Sasse, Accounting for Taste: Using Profile Similarity to Improve Recommender Systems, In CHI 2006 Proceedings, Social Computing 2

[42] James Bennett, Stan Lanning, The Netflix Prize

[43] http://www.netflixprize.com

[44] Herlocker, J, Konstan, J., Terveen, L., and Riedl, J. Evaluating Collaborative Filtering Recommender Systems. *ACM Transactions on Information Systems* 22 (2004), ACM Press

[45] Distances between Clustering, Hierarchical Clustering, 36-350, Data Mining, 14 September 2009

[46] Μ. Βαρζιγιάννης – Μ. Χαλκίδη, Εξόρυξη γνώσης από βάσεις δεδομένων, 2003

[47] Ε. Κιουντούζης, Ανάλυση και σχεδιασμός Πληροφοριακών Συστημάτων, 2003

[48] Ν. Αβούρης, Εισαγωγή στην επικοινωνία ανθρώπου – υπολογιστή, Εκδόσεις δίαυλος, 2000

[49] Ι. Βλαχάβας, Π. Κεφάλας, Ν. Βασιλειάδης, Φ. Κοκκορας, Η. Σακελλαρίου, Τεχνητή Νοημοσύνη Γ' Έκδοση, Εκδόσεις Γκιούρδα, 2006

[50] Ε. Ι. Γιαννακουδάκης, Σχεδιασμός και διαχείριση Βάσεων Δεδομένων, Εκδόσεις Ευγ. Μπένου, 1999

[51] http://people.revoledu.com

[52] http://www.php.net

[53] http://www.mysql.com