Ανοικτό Πανεπιστήμιο Κύπρου

Σχολή Θετικών και Εφαρμοσμένων Επιστημών

Μεταπτυχιακή Διατριβή
στα Πληροφοριακά και Επικοινωνιακά Συστήματα

Εφαρμογή Υπαρχόντων Αλγορίθμων Συστάσεων, σε Εκπαιδευτικά Σύνολα Δεδομένων, από Αποθετήρια Μαθησιακών Αντικειμένων

Νεόφυτος Νεοφύτου

Επιβλέπων Καθηγητής
Θανάσης Χατζηλάκος

Δεκέμβριος 2014

Ανοικτό Πανεπιστήμιο Κύπρου

Σχολή Θετικών και Εφαρμοσμένων Επιστημών

Εφαρμογή Υπαρχόντων Αλγορίθμων Συστάσεων, σε Εκπαιδευτικά Σύνολα Δεδομένων, από Αποθετήρια Μαθησιακών Αντικειμένων

Νεόφυτος Νεοφύτου

Επιβλέπων Καθηγητής
Θανάσης Χατζηλάκος

Η παρούσα μεταπτυχιακή διατριβή υποβλήθηκε
προς μερική εκπλήρωση των απαιτήσεων για απόκτηση

μεταπτυχιακού τίτλου σπουδών
στα Πληροφοριακά Συστήματα

από τη Σχολή Θετικών και Εφαρμοσμένων Επιστημών
του Ανοικτού Πανεπιστημίου Κύπρου

Εφαρμογή Υπαρχόντων Αλγορίθμων Συστάσεων, σε Εκπαιδευτικά Σύνολα Δεδομένων, από Αποθετήρια Μαθησιακών Αντικειμένων

Περίληψη

Τα συστήματα συστάσεων στο διαδίκτυο, έχουν ως στόχο τη συλλογή των πληροφοριών, που θα αποτυπώσουν τις προτιμήσεις του εκάστοτε χρήστη και θα του προσφέρουν την ιδανικότερη για αυτόν σύσταση. Μελετήθηκε ο τομέας των συστημάτων συστάσεων στην εκπαίδευση. Κάθε σύστημα σύστασης στηρίζεται σε έναν αλγόριθμο, ο οποίος λαμβάνει ως είσοδο, δεδομένα και προτιμήσεις του χρήστη και παράγει την ιδανικότερη για αυτόν σύσταση. Αυτό που προσπαθεί να επιτύχει το κάθε σύστημα σύστασης, είναι η σειρά των αποτελεσμάτων που θα παράξει, να έχει άμεση σχέση και συνάφεια, με αυτό που ζήτησε ο χρήστης.

Η παρούσα διπλωματική εργασία επιχειρεί μια εμβάθυνση στο τομέα των συστάσεων, μέσα από βιβλιογραφική ανασκόπηση και παράλληλα γίνεται μία προσπάθεια παρουσίασης, των τεχνικών που χρησιμοποιούνται για να συσταθούν τα συστήματα, στα οποία αναπτύσσονται. Γίνεται μια ανάλυση τριών κατηγοριών αλγορίθμων συστάσεων, των αλγορίθμων συστάσεων που εστιάζουν στο περιεχόμενο, των αλγορίθμων συνεργατικών φίλτρων και των αλγορίθμων συστάσεων ανάλυσης γράφων.

Παρουσιάζεται μια πλήρης περιγραφή των εκπαιδευτικών συνόλων δεδομένων, στα οποία εφαρμόζονται οι αλγορίθμοι συστάσεων. Μελετήθηκε η μορφή τους καθώς και ο τρόπος με τον οποίο συλλέγονται οι πληροφορίες που τα απαρτίζουν. Αναλύθηκαν οι μέθοδοι με τις οποίες αυτά διαχωρίζονται και διαιρούνται σε κλάσεις και στην συνέχεια πως επεξεργάζονται.

Χρησιμοποιήθηκε το λογισμικό recommender101, για την εφαρμογή των αλγορίθμων στα σύνολα δεδομένων MovieLens 100K, 1M, 10M, και για τη εξαγωγή των αποτελεσμάτων. Οι αλγόριθμοι που επιλέχθηκαν ώστε να αξιολογηθούν είναι ο K-NN, Slope One, Funk SVD, BPRMF Naïve Bayes.

Η εκτέλεση τους, σε διαφορετικά σύνολα δεδομένων, η δομή και το περιεχόμενο των οποίων περιγράφεται αναλυτικά στην εργασία, απέδειξε ότι οι γράφοι ανήκουν στην κατηγορία εκείνη των αλγορίθμων που σημειώνουν τα μεγαλύτερα ποσοστά. Παρουσίασαν διακυμάνσεις στην επίδοσή τους γεγονός το οποίο οφείλεται στα ιδιαίτερα χαρακτηριστικά των συνόλων δεδομένων. Αξιολογήθηκαν διάφορες μετρικές, από τα αποτελέσματα του κάθε αλγορίθμου, πάνω σε διαφορετικά σύνολα δεδομένων, για να γίνουν γνωστές οι συνθήκες υπό τις οποίες λειτουργούν καλύτερα.

Summary

Recommender systems came up having as main purpose the collection of all the available information in order to provide the user with the best suitable for him recommendation. Beyond all existing applications, the area of recommender systems seems to be invading the world of education. As described above, each and any of the recommender systems use an efficient algorithm in order to gather all the necessary information about the user and offer him as a result the most effective and appropriate solution regarding his research. The outcome of the algorithm should produce an immediate connection among user and suggestion in order for it to cover completely his necessity.

The present thesis begins with a description of the above described field as well as with what it represents. There was made a serious effort in order clarify the nature of recommender systems in combination with their present route, their acceptance from the users and their future. Through a thorough literature review there was made a serious attempt of deepening in this specific area. Attempt an analysis of three classes of recommendations algorithms, “Content-Based Recommender Algorithms”, “Collaborative Filtering Algorithms” and “Graph-Based Recommender Algorithms”.

In addition to this, the thesis contains a thorough research in the world of the datasets that basically consist the main tool of the recommender systems. Topics such as the formulation of dataset, the information included as well as the way that the information becomes part of a dataset were extensively investigated and presented. Furthermore, the present thesis tried to analyze all methods used in datasets in order to separate them and divide them into classes.

Software such as “recommender101” was used to apply a number of different algorithms over different datasets, like MovieLens 100K, 1M and 100M. The algorithms were chosen to be assessed are K-Nearest Neighbor, Naive Bayes (BPRMF), Funk SVD and Slope One. Datasets used were completely different in structure and content as their use had as main purpose the evaluation of the algorithms. Graphs proved out, to be the most reliable source, as they scored the highest percentages, in terms of prediction. Fluctuations their performance which is due to the specific characteristics of the data sets. Evaluated the results of each algorithm, on most important metrics, through different data sets, to be known the conditions under which they work best.

Ευχαριστίες

Θέλω να ευχαριστήσω θερμά τον καθηγητή μου, κ. Θανάση Χατζηλάκο, που μου έδωσε την ευκαιρία, μέσα από το θέμα της διατριβής αυτής, να ανακαλύψω κάτι καινούριο για μένα και παράλληλα να αποκτήσω την ελάχιστη εμπειρία της έρευνας και της μελέτης, στο τομέα της Πληροφορικής. Παράλληλα, ευχαριστώ θερμά τους καθηγητές του Πανεπιστημίου Πειραιά, κους Δημήτρη Σαμψών και Παναγιώτη Ζερβά για την υπόδειξη του θέματος, για τις βιβλιογραφικές τους συστάσεις, καθώς και τις συμβουλές τους. Ιδιαίτερες ευχαριστίες θέλω να εκφράσω στη Άννα Μαυρουδή, για την πολύτιμη καθοδήγηση, τις φιλικές συμβουλές και υποδείξεις της.

Τέλος νιώθω την ανάγκη να ευχαριστήσω την σύζυγο μου Χριστίνα, όπως και τα παιδιά μου, Βαλεντίνα, Μάριο και Αντρέα για την υπομονή και συμπαράσταση στην όλη μου προσπάθεια.

Περιεχόμενα

Κεφάλαιο 1

Συστήματα Συστάσεων – Εκπαιδευτικά Συστήματα Συστάσεων

Τα συστήματα συστάσεων (Recommender Systems (RSs)), είναι τεχνικές και εργαλεία λογισμικού που παρέχουν συστάσεις, για αντικείμενα, που θα χρησιμοποιηθούν για ένα χρήστη (Resnick & Varian, March 1997), (Burke, 2007), (Mahmood & Ricci, 2009). Οι συστάσεις αφορούν διάφορες διαδικασίες λήψης αποφάσεων, όπως ποια αντικείμενα θα αγοράσει ο χρήστης, τι μουσική να ακούσει, τι ειδήσεις να διαβάσει κλπ. Η λέξη αντικείμενο, είναι γενικός όρος, που χρησιμοποιείται για να δηλώσει, το τι συνιστά, το σύστημα, στον χρήστη (Ricci, et al., 2011). Ο (Burke, 2002) αναφέρει τον όρο «καθοδηγούν το χρήστη», στην επιλογή του, προτείνοντας του, χρήσιμες ή ενδιαφέροντες συστάσεις, ανάμεσα από έναν αριθμό πολλών επιλογών. Τα συστήματα αυτά, αναπτύχθηκαν, έχοντας ως στόχο την υποστήριξη των χρηστών του διαδικτύου, κατά τη διαδικασία της απόφασης (αρχικά σε εμπορικές ιστοσελίδες), μέσα από την παροχή συστάσεων (Resnick & Varian, March 1997), (Adomavicius & Tuzhilin, 2005). Αυτές οι συστάσεις, αποτελούν χρήσιμη πηγή πληροφορίας προς τον χρήστη. Δίνουν

επίσης την δυνατότητα στον χρήστη να μοιραστεί τη γνώμη του, μέσα από εμπειρίες (π.χ. μέσα από διάφορα reviews, εμπορικών ιστοσελίδων κλπ) (Hill & Terveen, 2001).

1.1 Εισαγωγή

Η τεχνολογία των συστημάτων συστάσεων εστίαζε στις δραστηριότητες του ηλεκτρονικού εμπορίου, καθώς πρότεινε επιπλέον αγορές στους καταναλωτές, ενώ παράλληλα προσπαθούσε να προσφέρει διευκόλυνση κατά την αναζήτηση πληροφοριών, αλλά και τη διαδικασία της απόφασης (Schafer, et al., 2001).

Η επιτυχία των συστημάτων αυτών, οδήγησε στην χρήση και την εφαρμογή τους και στον εκπαιδευτικό τομέα (Santos & Boticario, 2011) στον οποίο, κύριος στόχος είναι η μάθηση αλλά και η υποστήριξη της εκπαιδευτικής διαδικασίας.

Κάθε άνθρωπος έχει διαφορετικές ανάγκες αλλά και διαφορετικές απαιτήσεις σε ότι αφορά την εκπαιδευτική διαδικασία καθώς έχει διαφορετικές καταβολές, διαφορετικές προσλαμβάνουσες, διαφορετικούς στόχους αλλά και διαφορετικό χρόνο να διαθέσει. Οι εκπαιδευτικοί οργανισμοί παγκοσμίως υποστηρίζουν έως σήμερα, εκπαιδευτικά σενάρια τα οποία επικεντρώνονται στο τελικό χρήστη[1], και προσανατολίζονται σε υπηρεσίες που αφορούν το προφίλ του, δηλαδή προσαρμόζουν τις διαδικασίες τους, με βάση τις ικανότητες, τις δεξιότητες, τις εκπαιδευτικές ανάγκες του κλπ (Moreno, et al., 2009). Σύμφωνα με τους (Iorio, et al., 2006), η έρευνα συνεχίζεται στην ανεύρεση γενικότερων λύσεων, στα προβλήματα που αφορούν την εκπαιδευτική διαδικασία, έτσι ώστε να εφαρμόζονται σε ευρεία κλίμακα και να υποστηρίζουν την ανάπτυξη υπηρεσιών και να εξελίσσονται με τρόπο τέτοιο ώστε να καλύπτουν πλήρως τις ανάγκες του κάθε μαθητή.

Τα περισσότερα πανεπιστήμια χρησιμοποιούν Learning Management Systems (LMS) για να υποστηρίξουν την on line εκπαίδευση την οποία προσφέρουν (Barajas & Gannaway, 2007). Πολλά από αυτά είναι διαθέσιμα on line ή αποτελούν λογισμικά ανοιχτού κώδικα (π.χ Moodle, dotLRN, Sakai). Κάθε ένα από αυτά έχει τις ιδιαιτερότητές του, έχουν όμως όλα ως αντικειμενικό στόχο, την υποστήριξη της εκπαιδευτικής διαδικασίας και την παροχή υπηρεσιών ελέγχου για τον εκπαιδευτικό, και για το φορέα εκπαίδευσης. Τα LMS έχουν κάποια κοινά χαρακτηριστικά και παρέχουν υποστήριξη τέτοια που να

[1] Για τους σκοπούς της εργασίας, σαν τελικός χρήστης θα αναφέρεται ο εκπαιδευόμενος, ή αλλιώς ο μαθητής.

καλύπτει τα περιεχόμενα του μαθήματος αλλά και τις διαδικασίες εκπαίδευσης και μάθησης αντίστοιχα. Μπορούν π.χ να ανταλλάξουν ανάμεσα τους:

1. Πληροφορίες για το χρήστη:
    1.1. Επίπεδο γνώσεων,
    1.2. Επίπεδο δεξιοτήτων,
    1.3. Προτιμήσεις (preferences),
2. Πληροφορίες για το περιεχόμενο:
    2.1. Μεταδεδομένα για το θέμα,
    2.2. Τεχνικές πτυχές του περιεχομένου (Dagger , et al., 2007).

Σύμφωνα με τους Dagger et al (Dagger , et al., 2007), η επόμενη γενιά των LMS θα αφορά σε αρχιτεκτονικές που θα έχουν ως επίκεντρο την υπηρεσία και όχι πλέον το μαθητή και που θα αλληλεπιδρούν με άλλες διαδικτυακές εκπαιδευτικές υπηρεσίες (Muñoz-Merino, et al., 2009). Οι Dagger et al υποστηρίζουν ότι στις πλατφόρμες η-μάθησης (e-learning) της επόμενης γενιάς, ο διαχωρισμός του LMS από την λειτουργία του LCMS (Learning Content Management System), θα παρέχει υποστήριξη για μεγαλύτερη διαλειτουργικότητα. Τα συστήματα αυτά δεν θα διαμοιράζουν μόνο το περιεχόμενο και τα εκπαιδευτικά σενάρια, αλλά, και τα εργαλεία τους, τις λειτουργίες τους και τη σημασιολογία τους δυναμικά. Αυτό θα επιτρέπει στους μαθητές, μέσα από τις πλατφόρμες, να δημιουργούν προσαρμοσμένες υπηρεσίες η-μάθησης, από ένα ευρύ φάσμα υπηρεσιών. Θα επιλέγουν συνδυασμούς υπηρεσιών για τις ανάγκες τους, μέσα από διαλειτουργικές πλατφόρμες (Dagger , et al., 2007).

Τα συστήματα συστάσεων προσφέρουν μια πολύ υποσχόμενη προσέγγιση τόσο στη εκπαίδευση, όσο και στην διαδικασία της διδασκαλίας (Verbert, et al., 2011), εντοπίζοντας κατάλληλους μαθησιακούς πόρους από μια μεγάλη ποικιλία επιλογών (Ternier, et al., 2009), παρουσιάζοντας έτσι, αυξημένο ενδιαφέρον, στο τομέα της μάθησης που υποστηρίζεται από την τεχνολογία (Technologically Enhanced Learning) (Verbert, et al., 2011).

1.2 Μάθηση Υποστηριζόμενη από την Τεχνολογία – (Technologically Enhanced Learning – TEL)

Στόχος της μάθησης που υποστηρίζεται από την τεχνολογία[2], είναι να σχεδιάσει, να αναπτύξει αλλά και να δοκιμάσει τις κοινωνικοτεχνικές καινοτομίες, οι οποίες θα υποστηρίξουν τη διαδικασία της μάθησης τόσο σε ότι αφορά μαθητές, αλλά και όσο σε ότι αφορά οργανισμούς. Πρόκειται λοιπόν για έναν τομέα εφαρμογής, ο οποίος καλύπτει όλες εκείνες τις τεχνολογίες, οι οποίες συμβάλλουν στην προαγωγή της μάθησης αλλά και των εκπαιδευτικών διαδικασιών (Malone, et al., 1987).

Όπως σε κάθε τομέα που γνωρίζει ραγδαία αύξηση, έτσι και στην περίπτωση αυτή, η TEL χρειάζεται καλύτερη διαχείριση των πηγών μάθησης (Manouselis & Costopoulou, 2007). Τέτοιου είδους παραδείγματα είναι τα τεράστια αποθετήρια ψηφιακών πηγών μάθησης, που έχουν δημιουργηθεί τα περασμένα χρόνια, όπως το MERLOT[3], που έχει περισσότερες από 20.000 πηγές ή το OER-Open Educational Resources[4], που έχει 18.000 πηγές (Manouselis & Costopoulou, 2007).

Πέρα από το ίδιο το μαθησιακό περιεχόμενο, οι πηγές μπορούν επίσης να περιέχουν μονοπάτια ή ακόμη και σύνδεση με άλλους μαθητευόμενους, στο ίδιο αντικείμενο και έτσι να επιτευχθεί η συνεργατική μάθηση. Η πληθώρα αυτή, καθώς και η αλληλεπίδραση με το μεγάλο αυτό αριθμό πηγών, δίνει στους χρήστες των συστημάτων TEL, την ικανότητα, μέσα από έναν αριθμό υπηρεσιών, να αναγνωρίζουν τις κατάλληλες για αυτούς πηγές, ανάμεσα σε άπειρες επιλογές (Manouselis, et al., 2011). Απόρροια αυτού είναι η εμφάνιση των συστημάτων σύστασης και στην TEL, γεγονός το οποίο έχει στρέψει μία ιδιαίτερα μεγάλη μερίδα ερευνητών, προς αυτήν την κατεύθυνση.

Η TEL έχει άμεση σχέση, με τα δεδομένα τα οποία δημιουργούνται, από διάφορους τύπους εκπαιδευτικών ρυθμίσεων. Τέτοιοι τύποι εκπαιδευτικών ρυθμίσεων μπορεί να είναι το εκπαιδευτικό επίπεδο (π.χ. K-12 επίπεδο, ανώτερη εκπαίδευση, κατάρτιση κλπ), τυπική (π.χ μάθηση από ένα εκπαιδευτικό ίδρυμα) και άτυπη μάθηση (π.χ. οι δια

[2] ή τεχνολογικά ανεπτυγμένης μάθησης – για τους σκοπούς της παρούσας εργασίας θα χρησιμοποιείτε το ακρώνυμο του αγγλικού όρου, "Technologically Enhanced Learning": TEL

[3] ανοικτή συλλογή με απευθείας συνδέσεις διδασκαλίας και υπηρεσιών, που χρησιμοποιείται από τη διεθνή εκπαιδευτική κοινότητα, http://www.merlot.org.

[4] ανοικτή συλλογή εκπαιδευτικού υλικού που ενδέχεται να χρησιμοποιηθεί ελεύθερα χωρίς καμία επιβάρυνση. Το υλικό έχει άδεια χρήσης που αναφέρει συγκεκριμένα πως μπορεί να διαμορφωθεί, να επαναχρησιμοποιηθεί και να κοινοποιηθεί ξανά, https://www.oercommons.org.

βίου μαθητές, οι οποίοι είναι υπεύθυνοι για τη μάθησή τους με τον ρυθμό που επιθυμούν). Οι διάφοροι τύποι εκπαιδευτικών ρυθμίσεων, συχνά καλούνται μάκρο–περιεχόμενο, και αφορούν σε πιθανές ενέργειες του χρήστη, αλλά και στον τρόπο με τον οποίο αυτές μπορούν να ερμηνευθούν (Vuorikari & Berendt , 2009). Η συμμετοχή της TEL στην εκπαιδευτική διαδικασία, μπορεί να χαρακτηριστεί από την παροχή μικτής μάθησης. Ο όρος μικτή μάθηση συνδυάζει την παραδοσιακή κατά πρόσωπο μάθηση με αυτήν που υποστηρίζεται από έναν υπολογιστή. Από την άλλη, η μάθηση εξ αποστάσεως, μπορεί να υποστηριχθεί με χρήση του περιβάλλοντος της TEL είτε με σύγχρονους είτε με ασύγχρονους τρόπους. Παραδοσιακά, η εξ αποστάσεως εκπαίδευση ήταν άμεσα συνδεδεμένη με μεθόδους που αφορούσαν στον ασύγχρονο τρόπο. Πλέον, με τη χρήση του live streaming (δυνατότητα παρακολούθησης ζωντανών μεταδόσεων βίντεο και ήχου), αλλά και πολλών εργαλείων του διαδικτύου, όπως π.χ. η εικονική πραγματικότητα, διευκολύνουν την ανάπτυξη σε ότι αφορά τις σύγχρονες εξ αποστάσεως υπηρεσίες μάθησης (Manouselis, et al., 2011).

Στα πλαίσια μιας συγκεκριμένης εμπορικής εφαρμογής, η ανάπτυξη ενός συστήματος σύστασης, π.χ. αγοράς κάποιων προϊόντων, είχε συνδεθεί με έναν αριθμό αρμοδιοτήτων του χρήστη, που το σύστημα υποστήριζε (Pazzani & Billsus, 1997). Οι (Manouselis, et al., 2011) υποστηρίζουν ότι ένα σενάριο σύστασης που θα υποστηρίζει υπηρεσίες μάθησης, όπως η TEL, αποτελείται από διάφορες ιδιαιτερότητες σε ότι αφορά, το είδος[5] της μάθησης που απαιτείται. Αν συγκρίνει κάποιος μια εμπορική εφαρμογή συστάσεων, διαφέρει από αυτή ενός συστήματος μάθησης. Αντί λοιπόν, οι χρήστες, να αγοράσουν ένα προϊόν και έπειτα να το χρησιμοποιήσουν ως κτήμα τους, στην περίπτωση του συστήματος μάθησης, γίνετε μια προσπάθεια που συχνά παίρνει περισσότερο χρόνο. Οι μαθητές σπανίως επιτυγχάνουν ένα οριστικό τέλος, μετά από ένα καθορισμένο χρονικό διάστημα. Παρουσιάζονται ταυτόχρονα, διάφορες αλληλεπιδράσεις όπως π.χ. η δυσκολία σε κάποιο ερώτημα, η αμφιβολία σε κάποια απάντηση που εναπόθεσε ο μαθητής στο σύστημα, και πολλά άλλα στα διάφορα επίπεδα. Σε αυτά τα σενάρια (όπου οι μαθητές προσπαθούν να επιτύχουν σε διαφορετικά επίπεδα, με ικανότητες που διαφέρουν, σε διάφορους τομείς), ιδιαίτερα σημαντικό είναι η αναγνώριση των σχετικών εκπαιδευτικών στόχων αλλά και η παροχή υποστήριξης στους μαθητές, ώστε να πετύχουν τους στόχους αυτούς (Manouselis, et al., 2011). Από την άλλη, και σε

[5] όπως για παράδειγμα, η εισαγωγή μιας νέας μεθόδου μάθησης ή η προώθηση μιας ήδη υπάρχουσας, που μπορεί να απαιτεί εμπλουτισμό με διαφορετικούς τύπους εκπαιδευτικών πηγών κλπ.

σχέση πάντα με το περιεχόμενο, ίσως να δοθεί προτεραιότητα σε κάποιες συγκεκριμένες αρμοδιότητες του χρήστη. Το γεγονός αυτό απαιτεί συστάσεις, που ο χρονικός ορίζοντας τους, είναι κατά πολύ μεγαλύτερος από αυτές που παρέχονται για εμπορικές εφαρμογές, όπως οι συστάσεις παρόμοιων εκπαιδευτικών πηγών, της ανακεφαλαίωσης, της επανάληψης κλπ (Schneider-Hufschmidt, et al., 1993).

Η σύσταση στα πλαίσια της TEL παρουσιάζει και ιδιαιτερότητες οι οποίες βασίζονται στην πληθώρα των παιδαγωγικών θεωριών και μοντέλων. Για παράδειγμα, σε μαθητές που δεν έχουν αποκτήσει καμία προηγούμενη γνώση σε ότι αφορά ένα συγκεκριμένο τομέα, τα μαθησιακά αντικείμενα σύστασης θα πρέπει να έχουν ένα επίπεδο ελαφρώς υψηλότερο από το επίπεδο που διαθέτει έως τώρα ο μαθητής (Pazzani & Billsus, 1997).

Στο εκπαιδευτικό μοντέλο, το οποίο έχει ως επίκεντρο τον εκπαιδευτικό, επιβάλλεται να υποστηριχτούν διαφορετικές αρμοδιότητες, δυνατότητες και λειτουργίες όπως η προετοιμασία των μαθημάτων, η παράδοση ενός μαθήματος αλλά και η αξιολόγηση των γραπτών και των εργασιών. Για παράδειγμα, για την προετοιμασία ενός μαθήματος ένας εκπαιδευτικός χρειάζεται να έχει συγκεκριμένους στόχους του οποίους πρέπει να επιτύχει αλλά και συγκεκριμένες ανάγκες τις οποίες θα πρέπει να καλύψει, σχετικά πάντα και με τη μέθοδο διδασκαλίας αλλά και με το προφίλ των μαθητών. Η προετοιμασία του μαθήματος μπορεί να περιλαμβάνει αναζήτηση πληροφοριών, ανεύρεση του περιεχομένου εκείνου, που θα προσδίδει κίνητρα στους μαθητές, ανάκληση της ήδη υπάρχουσας γνώσης, αλληλεπίδραση μεταξύ των μαθητών, αλληλεπίδραση μεταξύ μαθητή-καθηγητή, χρήση τεχνολογικών εργαλείων, οπτικοποίηση αλλά και παρουσίαση των νέων ιδεών και πληροφοριών. Η παράδοση του μαθήματος μπορεί να υποστηρίζεται από διάφορες παιδαγωγικές μεθόδους, η αξιολόγηση των οποίων, έχει πάντα άμεση συνάφεια με τους στόχους που τίθενται. Τέλος μια σειρά από μεταβλητές, όπως π.χ. οι ιδιότητες των χρηστών, και διάφορες ευφυείς λειτουργίες θα πρέπει να δεσμευθούν για την παροχή προσωποποιημένων συστάσεων (Manouselis, et al., 2011). Σύμφωνα με τους (Brusilovsky, et al., 2007), Ο όρος «προσωποποίηση» στην επιστήμη των υπολογιστών αναφέρεται σε μια διαδικασία, κατά την οποία ένα διαδραστικό σύστημα (προσαρμοζόμενο σύστημα), προσαρμόζει τη συμπεριφορά του σε μεμονωμένους χρήστες με βάση τις πληροφορίες που αποκτά σχετικά μ'αυτούς και το περιβάλλον τους. Η προσωποποίηση είναι ιδιαίτερα σημαντική για την εκπαίδευση που βασίζεται στον Ιστό, για τουλάχιστον δύο γενικούς λόγους (Brusilovsky, et al., 2007):

- Οι περισσότερες εφαρμογές που βασίζεται στον Ιστό, έχουν κατασκευαστεί για να χρησιμοποιούνται από ένα πολύ ευρύ φάσμα χρηστών από οποιαδήποτε άλλη αυτόνομη εφαρμογή.
- Σε πολλές περιπτώσεις, ο μαθητής εργάζεται μόνος του στον Ιστό, πιθανώς από το σπίτι του. Η βοήθεια που ένας δάσκαλος συνήθως παρέχει προσαρμοστικά σε μια κανονική τάξη, δεν είναι διαθέσιμη.

Το σημείο κλειδί λοιπόν, στην αντιμετώπιση της υπερπληθώρας των πληροφοριών που υφίστανται σε μία κοινωνία είναι η προσωποποίηση τους. Αναμένεται λοιπόν ότι η προσωποποιημένη μάθηση (Manouselis, et al., 2011) :

- έχει τη δυναμική του να μειώσει τα κόστη παράδοσης,
- να δημιουργήσει αποτελεσματικότερα περιβάλλοντα μάθησης,
- να επιταχύνει το χρόνο αφομοίωσης,
- αλλά και να αυξήσει τα επίπεδα συνεργασίας μεταξύ των μαθητών.

1.2.1 Αναγνώριση του προβλήματος συστάσεων TEL

Σε ένα σύστημα συστάσεων, τα θέματα που ενδιαφέρουν και οι προτιμήσεις των χρηστών αντιπροσωπεύονται με διάφορες μορφές, ιδιαίτερα στα συστήματα συστάσεων τα οποία βασίζονται στη γνώμη των άλλων. Είναι λοιπόν σημαντικό να ληφθούν υπόψη οι πολλαπλοί παράγοντες ή τα κριτήρια που επηρεάζουν τις απόψεις των χρηστών, ώστε να γίνουν πιο αποτελεσματικές οι συστάσεις. Σε σχετική έρευνα, το πρόβλημα της σύστασης έχει αναγνωριστεί ως ο τρόπος, για να βοηθηθούν άτομα σε μια κοινότητα, να βρουν τις πληροφορίες ή τα προϊόντα που είναι πιο πιθανό να τους ενδιαφέρουν ή να είναι σχετικά με τις ανάγκες τους (Konstan, 2004).

Στη διεθνή βιβλιογραφία ο (Roy, 1996), θέτει το πρόβλημα της σύστασης, ως ένα πολυδιάστατο πρόβλημα. Θεωρεί ότι και το πρόβλημα των συστάσεων TEL πρέπει να διερευνηθεί και μάλιστα μπορεί να οριστεί καλύτερα αν ακολουθείται μια τέτοια πολυδιάστατη προσέγγιση, προκειμένου να εντοπιστούν τα εξής (Roy, 1996):

- Το αντικείμενο της απόφασης,
- Μια οικογένεια κριτηρίων,
- Ένα γενικό μοντέλο προτιμήσεων,
- Και μια διαδικασία υποστήριξης των αποφάσεων.

Συνεχίζοντας ο (Roy, 1996), αναφέρει ότι στη σύσταση TEL, το αντικείμενο της απόφασης είναι ένα στοιχείο s, που ανήκει στο σύνολο όλων των υποψήφιων αντικειμένων S, που εκπροσωπούν οποιοδήποτε είδος των αντικειμένων που μπορεί να συνίσταται σε έναν χρήστη. Για να εκφράσει το σκεπτικό πίσω από την απόφαση, ο (Roy, 1996), αναφέρεται στην έννοια της "problematic" απόφασης. Οι τέσσερις τύποι των κοινών problematic αποφάσεων αναφέρονται στη βιβλιογραφία σαν, Multi-Criteria Decision Making (MCDM). Σύμφωνα με τους (Adomavicius & Tuzhilin, 2011), μπορούν να θεωρηθούν έγκυροι στο πλαίσιο της σύστασης TEL (TEL Context), και είναι οι ακόλουθοι:

- Επιλογή (Choice), η οποία περιλαμβάνει την επιλογή ενός στοιχείου από ένα σύνολο υποψηφίων,
- Ταξινόμηση (Sorting), η οποία περιλαμβάνει την ταξινόμηση των στοιχείων σε προκαθορισμένες κατηγορίες,
- Κατάταξη (Ranking), το οποίο περιλαμβάνει την κατάταξη αντικειμένων από το καλύτερο προς το χειρότερο και η,
- Περιγραφή (Description), η οποία περιλαμβάνει την περιγραφή όλων των στοιχείων όσον αφορά τις επιδόσεις σύμφωνα με το κάθε κριτήριο.

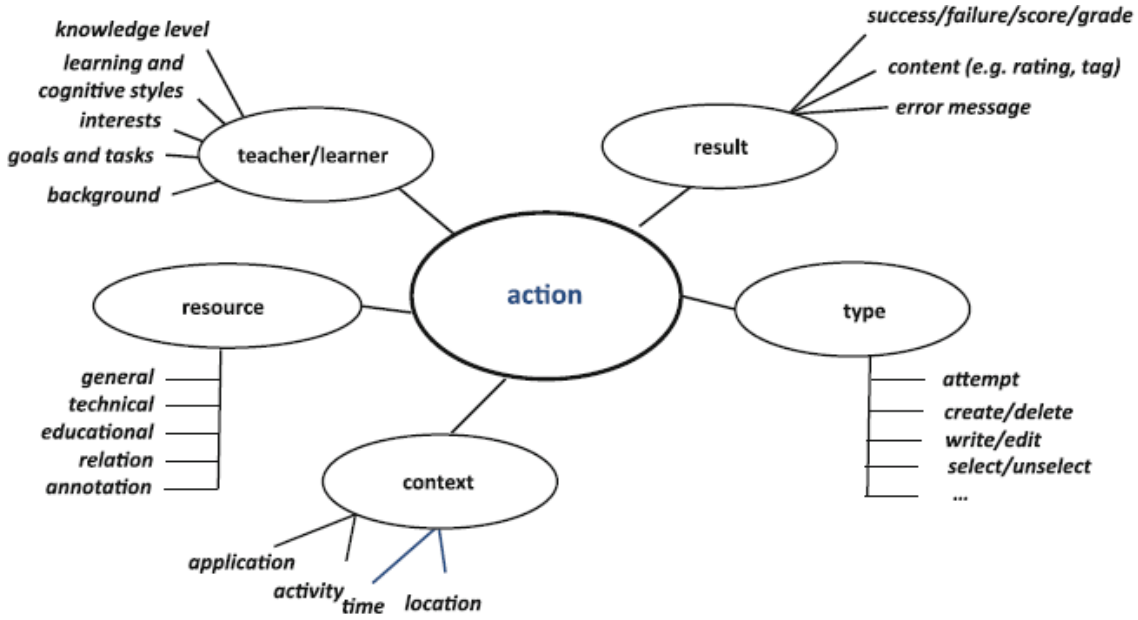
Στο MCDM, τα τέσσερα είδη κριτηρίων που επίσημα χρησιμοποιούνται είναι (Jacquet-Lagrèze & Siskos , 2001):

- Μετρήσιμο. Κριτήριο που επιτρέπει την μέτρηση ποσότητας από την κλίμακα αξιολόγησης. Όπως για παράδειγμα η μέτρηση του ορίου ηλικίας μέσα από μια κλίμακα.
- Διατεταγμένο. Κριτήριο που καθορίζει ένα διατεταγμένο σύνολο με τη μορφή μιας ποιοτικής ή περιγραφικής κλίμακας.
- Πιθανολογικό. Κριτήριο που χρησιμοποιεί κατανομές πιθανοτήτων για να αντιμετωπιστεί η αβεβαιότητα στην αξιολόγηση των εναλλακτικών λύσεων.
- Ασαφές. Κριτήριο όπου, η αξιολόγηση των εναλλακτικών λύσεων, εκπροσωπείται με τη δυνατότητά της, να ανήκει σε ένα από τα διαστήματα της κλίμακας αξιολόγησης.

Στο παρελθόν, η ανάπτυξη συστημάτων συστάσεων είχε σχέση με τον αριθμό των σχετικών καθηκόντων του χρήστη που το σύστημα συστάσεων υποστηρίζει μέσα σε κάποιο συγκεκριμένο πλαίσιο εφαρμογής. Από την άλλη πλευρά, σε σύγκριση με το

τυπικό σενάριο σύστασης στοιχείου, υπάρχουν πολλές ιδιαιτερότητες που πρέπει να ληφθούν υπόψη σχετικά με το είδος της μάθησης που είναι επιθυμητό. (Manouselis, et al., 2011). Όπως υπογραμμίζεται από τους (Romero & Ventura, 2007-b), ο τομέας TEL διαφέρει από τομείς όπως το ηλεκτρονικό εμπόριο, με διάφορους τρόπους. Στο ηλεκτρονικό εμπόριο (e-commerce), τα δεδομένα που χρησιμοποιούνται είναι συχνά απλά αρχεία καταγραφής πρόσβασης των web server, ή αξιολογήσεις των χρηστών. Το μοντέλο χρήστη και οι στόχοι των συστημάτων είναι επίσης διαφορετικά και στους δύο τομείς εφαρμογής (Drachsler, et al., 2009).

Οι (Verbert, et al., 2011) έχουν ενσωματώσει τις διάφορες κατηγορίες δεδομένων και των στοιχείων στο πλαίσιο (framework) που παρουσιάζεται στην εικόνα 1.1, για τον προσδιορισμό στοιχείων σε υπάρχοντα σύνολα δεδομένων. Το μοντέλο έχει αναπτυχθεί από τη σύνθεση υφιστάμενων έργων σχετικά με την αλληλεπίδραση δεδομένων στο πλαίσιο της TEL. Παρουσιάζει επίσης τις αλληλεπιδράσεις του μαθητή, όπως είναι η



Εικόνα 1.1: Οι μεταβλητές της TEL, όπως έχουν παρουσιαστεί από τους (Verbert, et al., 2011).

επιλογή, η αποθήκευση, η δημιουργία και η συγγραφή, στους διάφορους πόρους. (Verbert, et al., 2011)**.**

Οι (Brusilovsky, et al., 2004), εντόπισαν τις ακόλουθες κατηγορίες χαρακτηριστικών του μαθητή, με βάση την ανάλυση της υπάρχουσας βιβλιογραφίας:

- επίπεδα γνώσης,
- τους στόχους και τα καθήκοντα,
- τα ενδιαφέροντα,

- το υπόβαθρο
- τη μάθηση και
- το γνωστικό στυλ.

1.2.2 Πλαίσιο (Framework) για την Ανάλυση των Συστημάτων Συστάσεων

Αρκετά πλαίσια έχουν προταθεί στη βιβλιογραφία για την ανάλυση των συστημάτων συστάσεων (Hanani, et al., 2001). Αρχικά οι Hanani et al, διεξήγαγαν μια αναθεώρηση των ζητημάτων φιλτραρίσματος, των πληροφοριών και των συστημάτων, και παρουσίασαν ένα πλαίσιο για την ταξινόμηση των συστημάτων αυτών (Hanani, et al., 2001). Μια έρευνα που επικεντρώθηκε στις διάφορες τεχνικές σύστασης, εισάγοντας νέους τύπους συστημάτων, εκτός από το περιεχόμενο και τη συνεργασία, πραγματοποιήθηκε από τον (Burke, 2002). Στη μελέτη του περιγράφονται λεπτομερώς οι προσδιορισμένες τεχνικές σύστασης και συγκρίνονται με βάση τα οφέλη και τις ελλείψεις. Επιπλέον, οι (Montaner, et al., 2003) επικεντρώθηκαν ειδικά σε παράγοντες σύστασης και ανέλυσαν μια σειρά τέτοιων συστημάτων. Στη μελέτη τους για τα συστήματα συστάσεων οι (Adomavicius & Tuzhilin, 2005), κάνουν αξιολόγηση των διαφόρων τύπων των συστημάτων αυτών, και τα διακρίνουν με βάση το περιεχόμενο, την ικανότητα συνεργασίας τους και νέες υβριδικές μορφές τους.

Όλες οι παραπάνω μελέτες, έχουν επισημάνει σημαντικές διαστάσεις που μπορεί να χρησιμοποιηθούν για την ανάλυση και την ταξινόμηση των συστημάτων συστάσεων. Οι (Manouselis & Costopoulou, 2007) έχουν συλλέξει, έχουν επεξεργαστεί και κατηγοριοποίησαν τις διαστάσεις που προσδιόρισαν όλες αυτές οι προηγούμενες μελέτες, σε ένα προτεινόμενο πλαίσιο (Framework), για τα συστήματα συστάσεων που αφορούν την TEL, μέσα από τρεις κύριες κατηγορίες χαρακτηριστικών: *Υποστηριζόμενες εργασίες (supported tasks), προσέγγιση (approach) και λειτουργία (operation) (Εικόνα 1.2)*.

- *Υποστηριζόμενες Εργασίες (Supported Tasks):* Η κατηγορία αυτή αναφέρεται στις διαστάσεις που διακρίνουν ένα σύστημα συστάσεων ανάλογα με τα καθήκοντα των χρηστών που έχουν ως στόχο να υποστηρίξουν, όπως (Herlocker, et al., 2004):
  - ✓ Παραγωγή συστάσεων καινούριων αντικειμένων,
  - ✓ Συστάσεις άλλων μαθητών με τα ίδια ενδιαφέροντα, ή από το ίδιο εκπαιδευτικό δίκτυο (π.χ. φοιτητές ενός πανεπιστημίου),

- ✓ Παροχή εναλλακτικών διαδρομών μάθησης για την επίτευξη του μαθησιακού στόχου,
- ✓ Παραγωγή προβλέψεων στις επιδόσεις των μαθητών και στη ποιότητα διδασκαλίας των εκπαιδευτικών.

❍ *Προσέγγιση (Approach):* Η κατηγορία αυτή περιλαμβάνει τρεις διαφορετικές προοπτικές, σύμφωνα με τα στοιχεία του συστήματος που σχετίζονται με την έρευνα εξατομικευμένων προσαρμοστικών συστημάτων. Αυτά είναι (Brusilovsky, 1996):

- ✓ Μοντέλο χρήστη (User Model). Το μοντέλο χρήστη (ή προφίλ χρήστη), αναφέρεται στους τρόπους που τα χαρακτηριστικά των χρηστών εκπροσωπούνται, αποθηκεύονται και ενημερώνονται σε ένα σύστημα συστάσεων.
- ✓ Μοντέλο τομέα (Domain Model). Ομοίως με το μοντέλο του χρήστη, ένα μοντέλο τομέα απαιτείται για να εκπροσωπεί τις ιδιότητες των αντικειμένων που έχουν προταθεί.
- ✓ Εξατομίκευση (Personalisation). Αναφέρεται στις διαστάσεις που απεικονίζουν τον τρόπο, με τον οποίο το σύστημα, παρέχει τις συστάσεις του.

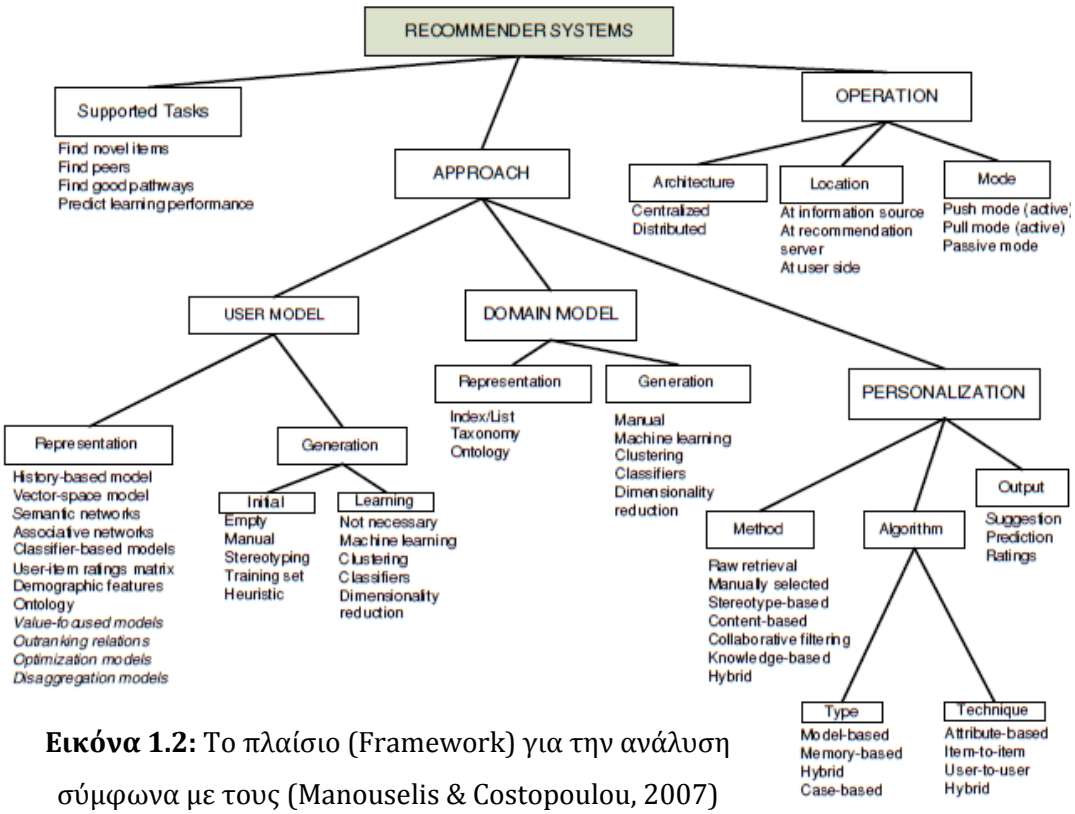
❍ *Λειτουργία (Operation):* Η κατηγορία αυτή περιλαμβάνει επίσης τρεις διαφορετικές οπτικές γωνίες που έχουν διαστάσεις οι οποίες σχετίζονται με την ανάπτυξη των συστημάτων συστάσεων:

- ✓ Αρχιτεκτονική (Architecture.). Αναφέρεται στην αρχιτεκτονική του συστήματος συστάσεων, η οποία είναι διακρίνονται συνήθως ως (Miller, et al., 2004), (Han, et al., 2004):
  - Συγκεντρωτική (Centralised)
  - Κατανεμημένη (Distributed)
- ✓ Θέση (Location). Αναφέρεται στη θέση όπου η σύσταση παράγεται και παραδίδεται. Μπορεί να ταξινομηθεί σύμφωνα με τις ακόλουθες θέσεις (Hanani, et al., 2001):
  - Στην πηγή πληροφοριών.
  - Στο διακομιστή σύστασης.

- Στο πλευρό του χρήστη.

✓ Τρόπος Λειτουργίας (Mode). Αφορά τον προσδιορισμό του ποιος κινεί τη διαδικασία σύστασης και διακρίνεται σε (Schafer, et al., 2001), (Brusilovsky, 1996):

- *Push mode*. Η σύσταση αποστέλλεται στον μαθητή, όταν αυτός δεν αλληλεπιδρά με το σύστημα. Π.χ η σύσταση του αποστέλλεται στο email του.



Εικόνα 1.2: Το πλαίσιο (Framework) για την ανάλυση σύμφωνα με τους (Manouselis & Costopoulou, 2007)

- *Pull mode.* Η σύσταση δημιουργείτε αλλά δεν αποστέλλεται στον μαθητή, εκτός κι αν το ζητήσει
- *Passive mode*. Η σύσταση δημιουργείτε σαν μέρος της λειτουργίας του κανονικού συστήματος.

1.2.3 Προκλήσεις για τα Συστήματα Συστάσεων στη TEL

Πρόσφατες εργασίες, όσον αφορά τις κοινωνικές και ψυχολογικές απαιτήσεις για το πώς οι άνθρωποι αντιδρούν και ενεργούν σύμφωνα με τα συστήματα συστάσεων, για

τις επιστήμες της μάθησης, έχουν ορίσει μερικές νέες ιδέες στον τομέα των παιδαγωγικών (Howard-Jones, et al., 2010). Πιο συγκεκριμένα, οι (Buder & Schwind, 2012) μελέτησαν μια εννοιολόγηση, που αποκλίνει από τις εργασίες για τη σύσταση του ηλεκτρονικού εμπορίου. Πιο συγκεκριμένα, επικεντρώνεται στο πώς οι μαθητές ασχολούνται με τα συνιστώμενα είδη καθώς είναι οι ίδιοι οι παραγωγοί των δεδομένων. Σε αυτή την προσέγγιση, πρέπει να εξεταστεί ένας αριθμός από σημαντικές αρχές, ώστε να εισέλθουν τα συστήματα συστάσεων στον εκπαιδευτικό τομέα. Οι περισσότερες από τις αυτές τις αρχές αντιμετωπίζονται από τη σημερινά TEL συστήματα συστάσεων και είναι οι εξής (Buder & Schwind, 2012):

- Τα συστήματα συστάσεων μεταθέτουν την ευθύνη μακριά από ειδικούς εμπειρογνώμονες.
- Η ποιότητα του περιεχομένου δεν είναι ανιχνεύσιμη σε κάθε ατομική παραγωγή.
- Τα συστήματα πρότασης προβλέπουν (και απαιτούν) τον έλεγχο του χρήστη, διευκολύνοντας έτσι την ατομική μάθηση.
- Τα συστήματα συστάσεων παρέχουν καθοδήγηση στις δραστηριότητες μάθησης.
- Τα συστήματα συστάσεων προσαρμόζονται στις ανάγκες και τις απαιτήσεις των εκπαιδευομένων.

Οι Buder και Schwind (2012) τόνισαν επίσης, ότι πρέπει να γίνουν περισσότερες εκπαιδευτικές και ψυχολογικές μελέτες για τις επιπτώσεις των συστημάτων συστάσεων σε διαφορετικούς εκπαιδευόμενους, σε διαφορετικές εργασίες μάθησης, ή σε διαφορετικά επίπεδα γνώσης. Αναφέρουν επίσης ότι οι στρατηγικές σύστασης διαφέρουν ανάλογα με τους ρόλους που οι χρήστες αναμένεται να παίξουν. Όλα αυτά καθορίζουν νέες εκπαιδευτικές απαιτήσεις, και τονίζουν την ανάγκη ενός πλαισίου ευαισθητοποίησης σε σχέση με τις γνώσεις και τις δραστηριότητες, την κριτική σκέψη, και τη μετα-γνωστική διέγερση (Buder & Schwind, 2012). Οι εξελίξεις στον σημασιολογικό ιστό, κίνησαν τα νήματα για τη δημιουργία προτύπων τέτοιων που θα περιγράφουν σημασιολογικά και θα επαναχρησιμοποιούν τα εκπαιδευτικά αντικείμενα σε διαφορετικές εκπαιδευτικές πλατφόρμες (Dagger , et al., 2007).

1.3 Σημασιολογικά Εκπαιδευτικά Συστήματα Συστάσεων (Semantic Educational Recommender Systems - SERS)

Τα συστήματα συστάσεων έχουν επιδείξει σημαντική επιτυχία σε πολλούς τομείς στους οποίους υπάρχει υπέρογκη πληροφορία (Schafer, et al., 2001), γεγονός το οποίο πυροδότησε την είσοδό τους σε εφαρμογές η-μάθησης.

Κατά την ανάπτυξη ενός εκπαιδευτικού συστήματος συστάσεων, δύο είναι οι μορφές που χρησιμοποιούνται (Drachsler, 2009):

- Από πάνω προς τα κάτω (top-down) : Η προσέγγιση αυτή είναι κατάλληλη για την η-μάθηση στην οποία η δομή, το εκπαιδευτικό υλικό και τα εκπαιδευτικά σχέδια γίνονται από επαγγελματίες (formal e-learning).
- Από κάτω προς τα πάνω (bottom-up): Η προσέγγιση αυτή είναι κατάλληλη για την η-μάθηση που είναι αυτοκατευθυνόμενη, όπου οι μαθητές κάνουν χρήση των πηγών πληροφοριών που βρίσκονται στο εκπαιδευτικό δίκτυο (non-formal e-learning).

Η εφαρμογή των συστημάτων συστάσεων στον εκπαιδευτικό τομέα και ιδιαίτερα στην η-μάθηση, απαιτεί ιδιαίτερη μελέτη σε ότι αφορά τη σχεδίαση αλλά και την υλοποίηση αλγορίθμων. Οι Drachsler et al, υποστηρίζουν ότι υπάρχει η ανάγκη διαχείρισης των δραστηριοτήτων σχεδιασμού πριν την εκτέλεση τους από τους μαθητές, αλλά και κατά τη διάρκεια συντήρησης του συστήματος (Iorio, et al., 2006). Βέβαια, τα συστήματα συστάσεων στην εκπαίδευση διαφέρουν από αυτά που χρησιμοποιούνται στο εμπόριο, καθώς, θα πρέπει να λαμβάνουν υπόψιν τους, όχι μόνο τις προτιμήσεις των εκπαιδευτικών και των μαθητών, αλλά και το διαθέσιμο υλικό που θα βοηθήσει στην επίτευξη των εκπαιδευτικών στόχων. Σημαντικό λοιπόν θέμα στα συστήματα συστάσεων είναι και η ανάγκη δημιουργίας μεγάλων datasets (σύνολα δεδομένων), τα οποία θα διευκολύνουν την ανάπτυξη και την υλοποίηση τέτοιων συστημάτων. Στην πραγματικότητα, τα συστήματα επηρεάζονται πολύ από τον τομέα από τον οποίο εξαρτούνται, έτσι λοιπόν τα χαρακτηριστικά του εκπαιδευτικού τομέα θα πρέπει να ληφθούν πάρα πολύ σοβαρά υπόψιν. Ακόμη, θα πρέπει κυρίως να εστιάζουν στα εκπαιδευτικά αντικείμενα που υπάρχουν στα διάφορα εκπαιδευτικά αποθετήρια και όχι να συλλέγουν και να κάνουν χρήση υλικού από απροσδιόριστες πηγές (Bozo, et al., 2010).

Ο ορισμός «σημασιολογικά συστήματα συστάσεων», αφορά τα συστήματα, των οποίων, οι επιδόσεις βασίζονται σε μια βάση γνώσης (Peis , et al., 2008). Αναλυτικότερα, ένα SERS αποτελείται από τα παρακάτω (Santos & Boticario, 2011):

- *Σύστημα Συστάσεων*: Πρόκειται για ένα εργαλείο το οποίο βοηθά τους χρήστες κατά τη λήψη μιας απόφασης σε περιβάλλοντα όπου η πληροφορία είναι υπέρογκη. Τα συστήματα συστάσεων κάνουν χρήση και εφαρμογή αλγορίθμων συστάσεων, οι οποίοι διευκολύνουν την αυτοματοποίηση της όλης διαδικασίας.
- *Σημασιολογία*: Επιτρέπει στους σχεδιαστές των συστημάτων συστάσεων, να παρέχουν υψηλού επιπέδου περιγραφές, οι οποίες μάλιστα ερμηνεύονται και από το ίδιο το σύστημα, κατά τη διαδικασία τόσο της επιλογής όσο και της παραλαβής της ορθότερης πρότασης, αλλά, και της αυτόματης παραγωγής νέων προτάσεων. Ακόμη, διευκολύνει την ανταλλαγή πληροφορίας ανάμεσα στα διάφορα στοιχεία του λογισμικού που συμμετέχουν στην διαδικασία συστάσεων, καθώς έχει δομηθεί με μία αρχιτεκτονική που εστιάζει την υπηρεσία.
- *Μαθησιακός χαρακτήρας*: Το πεδίο εφαρμογής ενός συστήματος είναι ο μαθησιακός τομέας, ο οποίος μάλιστα έχει κάποιες ιδιαιτερότητες, π.χ. οι συστάσεις θα πρέπει να καθοδηγούνται από τα μαθησιακά κριτήρια (όπως η διερεύνηση γνώσης, καλλιέργεια κοινωνικών δεξιοτήτων, προαγωγή συνεργατικής μάθησης), και όχι μόνο από τις προτιμήσεις των μαθητών, και πιθανά καλύπτει όλες εκείνες τις μαθησιακές δραστηριότητες όπως για παράδειγμα την ανάγνωση, την συνεργασία, την αξιολόγηση, οι οποίες είναι αρμόζουσες για κάθε μαθησιακό σενάριο η-μάθησης.

Έτσι λοιπόν, και με βάση διάφορα εκπαιδευτικά κριτήρια, τα συστήματα SERS καθοδηγούνται μέσα από τις αλληλεπιδράσεις του χρήστη, πάνω στις πλατφόρμες η-μάθησης. Έτσι, προσφέρουν εξατομικευμένες συστάσεις οι οποίες περιγράφονται σημασιολογικά και λαμβάνονται λόγω της ανταλλαγής των πληροφοριών ανάμεσα στα διάφορα στοιχεία της διαδικασίας παραγωγής, αλλά και παράδοσης των σημασιολογικών συστάσεων. Τα συστήματα SERS βασίζονται σε (Santos & Boticario, 2011):

- *Αρχιτεκτονική ανοικτού τύπου Βάσης Δεδομένων, προσανατολισμένη στις Υπηρεσίες*: Οι υπηρεσίες αυτές συντελούν στην επικοινωνία των SERS με τα LMS, διαμέσου διαφόρων διαδικτυακών υπηρεσιών, και δίνουν στα SERS τη

δυνατότητα να κάνουν χρήση των διαφόρων προτύπων για να περιγράψουν την ανταλλαγή των πληροφοριών.

- *Περιβάλλον διεπαφής μαθητή (User Interface):* Χρησιμοποιείται για να μεταφέρει στο μαθητή τις διάφορες συστάσεις με εύκολο, γρήγορο και προσιτό τρόπο. Αναπτύσσεται στο επίπεδο παρουσίασης ενός LMS. Επιτρέπει την περιγραφή της σύστασης και των στοιχείων του, όπως ενημερώθηκε από το μοντέλο συστάσεων (Το μοντέλο συστάσεων περιγράφεται στο κεφ. 1.5.1).

Όλα τα παραπάνω στοιχεία είναι σημαντικά όταν πρόκειται για την υποστήριξη μιας διαδικασίας προσανατολισμένης σε σημασιολογικές συστάσεις στην εκπαίδευση, η οποία μάλιστα κατευθύνεται από τον εκπαιδευτικό και συμπληρώνει το υπάρχον LMS (Santos, et al., 2011).

Συμπερασματικά, θα μπορούσε κανείς να υποστηρίξει ότι, τα SERS παρέχουν την υποδομή, για την παροχή συστάσεων, αλλά δεν εμπλέκονται στην κατανόηση, των σεναρίων (κυρίως top-down για formal e-learning)των συστάσεων σε ότι αφορά την η-μάθηση.

Η ανάπτυξη των κατάλληλων σημασιολογικών συστημάτων συστάσεων, γίνεται υπό τις παρακάτω προϋποθέσεις (Santos & Boticario, 2011(1)) :

- Θα πρέπει να υπάρχει ένα βασικό μοντέλο που θα χαρακτηρίζει τις συστάσεις.
- Θα πρέπει επίσης να υπάρχει μία αρχιτεκτονική που θα βασίζεται στα πρότυπα, θα προσανατολίζεται στην υπηρεσία και θα έχει ως σκοπό την αλληλεπίδραση ανάμεσα στα διάφορα στοιχεία του λογισμικού.
- Τέλος θα υπάρχει ένα περιβάλλον διεπαφής το οποίο θα παρέχει τις συστάσεις με εύκολο και εύχρηστο τρόπο.

Ένα Σύστημα Συστάσεων όπως αναφέρουν οι Santos & Boticario θα πρέπει να απαντά στα παρακάτω ερωτήματα (Santos & Boticario, 2011(2)) :

1. *Τί (What)* είναι αυτό που θα πρέπει να προταθεί ; Οι δραστηριότητες των διαθέσιμων αντικειμένων στην πλατφόρμα (π.χ. απάντηση ενός μηνύματος στο forum, συμπλήρωση ενός τεστ, ανάρτηση ενός αρχείου κειμένου).
2. *Πώς (How)* θα γίνει η εφαρμογή μιας σύστασης ; Σωστή χρήση της γλώσσας (π.χ. τυπική ή άτυπη γλώσσα, ανάλογα με τον σχεδιασμό: top-down ή bottom-up, θα εμφανίζεται στο περιβάλλον διεπαφής ή θα αποστέλλεται στο email).

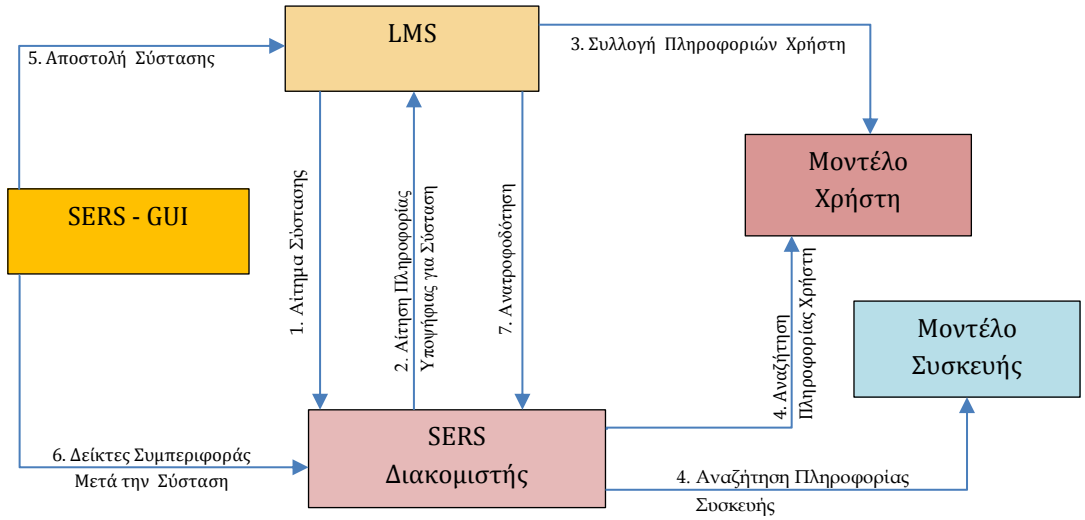
3. *Πότε (When)* θα πρέπει να προσφέρεται μια σύσταση; Σύμφωνα με τα χαρακτηριστικά του μαθητή, και του πλαισίου του μαθήματος (π.χ. κάποια πεδία του μαθητή και του πλαισίου του μαθήματος θα παίρνουν τιμές την ώρα της εκτέλεσης).
4. *Γιατί (Why)* θα προσφερθεί μια σύσταση; Η παροχή αιτιολόγησης της σύστασης. Η Παιδαγωγική βάση της σύστασης. Ο παιδαγωγικός στόχος που αναμένεται να επιτύχει ο μαθητής αν ακολουθήσει την σύσταση αυτή.
5. *Ποιά (Which)* θα είναι τα χαρακτηριστικά της σύστασης; Η περιγραφή των συστάσεων όσον αφορά τις ιδιότητές της (Semantic information) (σε ποια κριτήρια εστιάζει η σύσταση, όπως: ενεργός συμμετοχή, τεχνική βοήθεια, προσβασιμότητα, υλικό μαθήματος, κίνητρο, προφίλ, πρόοδος στη μάθηση. Παρούσα κατάσταση του μαθήματος. Πχ. Ρυθμίσεις πλατφόρμας, εκμάθηση πλατφόρμας, εκτέλεση δραστηριοτήτων μαθήματος. Προέλευση της σύστασης: προέρχεται από τον εκπαιδευτή, η πιο δημοφιλής μεταξύ των μαθητών. Ιεράρχηση των συστάσεων σε περίπτωση που υπάρχουν πολλές που ταιριάζουν με το πλαίσιο).

1.3.1 Αρχιτεκτονική Ανοικτής Βάσης που Βασίζεται στα Πρότυπα και Προσανατολίζεται στην Υπηρεσία (Open Standard-Based Service Oriented Architecture)

Η αλληλεπίδραση ανάμεσα στα σημασιολογικά συστήματα συστάσεων και στα LMS, επιτυγχάνεται διαμέσου μιας αρχιτεκτονικής προσανατολισμένης στην υπηρεσία (service-oriented architecture – SOA). Για να επιτευχθεί η απαιτούμενη λειτουργικότητα, τα SERS μπορούν να αλληλεπιδρούν και με στοιχεία όπως το μοντέλο χρήστη και το μοντέλο συσκευής. Το μοντέλο χρήστη είναι υπεύθυνο για τη συλλογή πληροφοριών σχετικών με το μαθητή, ενώ το μοντέλο συσκευής, αποθηκεύει όλη εκείνη την πληροφορία για τη συσκευή (Εικόνα 1.1). Ουσιαστικά είναι μια περιγραφή των δυνατοτήτων των συσκευών και των προτιμήσεων του χρήστη. Μπορεί να χρησιμοποιηθεί ως οδηγός για την προσαρμογή του περιεχομένου που παρουσιάζεται στη συσκευή (Σύνθετες Δυνατότητες/Προφίλ Προτιμήσεων: CC/PP (Composite Capabilities/Preference Profiles)) (Santos & Boticario, 2011). Στις 15/01/2004 η W3C (Κοινοπραξία του Παγκοσμίου Ιστού), ανακοίνωσε την έκδοση της Σύστασης: Composite Capability/Preference Profiles (CC/PP): Structure and Vocabularies 1.0 Recommendation (Σύνθετη Δυνατότητα/ Προφίλ Προτιμήσεων (CC/PP)) (W3C, 2004). Το CC/PP 1.0 είναι σύστημα για την έκφραση των δυνατοτήτων των συσκευών και τις προτιμήσεις των χρηστών, που χρησιμοποιεί το Πλαίσιο Περιγραφής Πόρων. Ένα προφίλ CC/PP χρησιμοποιείται για να καθοδηγεί την προσαρμογή του περιεχομένου και περιγράφει τις δυνατότητες των συσκευών και τις προτιμήσεις των χρηστών (W3C, 2004).

Επιπλέον, απαιτείται και η χρήση ενός εργαλείου το οποίο θα διαχειρίζεται τις συστάσεις διαμέσου ενός αλγορίθμου, και ζητά από το LMS την πληροφορία εκείνη, που είναι υποψήφια να συμμετάσχει σε μια σύσταση (Santos & Boticario, 2010). Η διαδικασία παράδοσης μιας σύστασης, ξεκινά από τον διακομιστή ενός SERS, τη στιγμή που αυτός λαμβάνει ένα αίτημα από ένα LMS. Με τη σειρά του ο διακομιστής αναζητά την πληροφορία στο μοντέλο χρήστη και στο μοντέλο συσκευής. Στη συνέχεια, γίνεται διαλογή της πληροφορίας και συγκεντρώνεται αυτή, η οποία είναι χρήσιμη και διαθέσιμη για μια σύσταση. Το περιβάλλον διεπαφής του χρήστη αναλαμβάνει την αποτύπωση των παραδοτέων συστάσεων. Εάν ο αριθμός των διαθέσιμων συστάσεων, ξεπεράσει τον αριθμό των συστάσεων που μπορούν να εμφανιστούν στην οθόνη, τότε επιλέγονται αυτά με τη μεγαλύτερη σχετικότητα, και διανέμονται διαμέσου του LMS.

Ακόμη, το LMS έχει τη δυνατότητα να παρέχει ανατροφοδότηση (feedback) στο SERS, σχετικά με τις κινήσεις που έχουν γίνει, και αν και κατά πόσο, οι συστάσεις ακολουθήθηκαν ή όχι από τους χρήστες. Οι μαθητές έχουν τη δυνατότητα αξιολόγησης του συστήματος, σχετικά πάντα με βάση την πληροφορία που έλαβαν, απαντώντας π.χ. στο εάν, αυτή του ήταν χρήσιμη την παρούσα στιγμή, αν θα τη χρησιμοποιούσε κάποια στιγμή στο μέλλον, ή εάν δεν του ήταν καθόλου χρήσιμη. Τα αποτελέσματα της αξιολόγησης αποστέλλονται στο SERS για επεξεργασία. Στην *Εικόνα 1.3* παρουσιάζονται τα πιο πάνω βήματα σε σειρά με βάση την αρίθμηση τους.



Εικόνα 1.3: Λειτουργία των συστάσεων μέσα από την αρχιτεκτονική προσανατολισμένη στην υπηρεσία (SOA), όπως ορίζεται από τους Santos & Boticario.

1.3.2 Περιβάλλον Διεπαφής Χρήστη (User Graphical Interface)

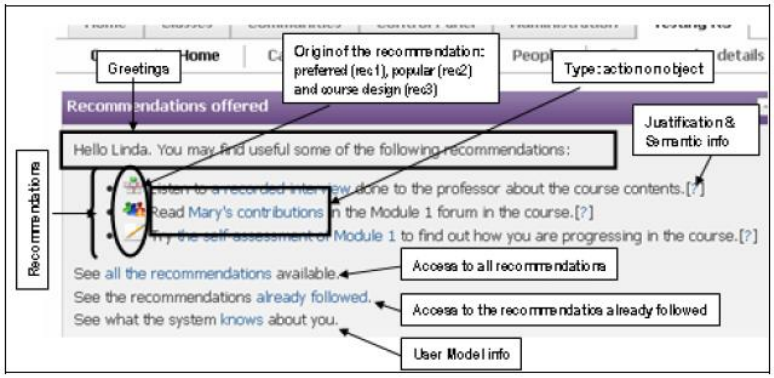
Ιδιαίτερα μεγάλης σημασίας (Romero, et al., 2007) για όλα τα συστήματα συστάσεων είναι το πώς θα παρουσιάζονται οι συστάσεις στον τελικό χρήστη. Η πιο κοινή προσέγγιση είναι η αποτύπωση μιας λίστα των πιο σχετικών συστάσεων, την οποία ακολουθεί ή όχι αν θελήσει ο χρήστης (Romero, et al., 2007). Η πληροφορία που παρουσιάζεται, είναι τα περιεχόμενα μιας σύστασης, αλλά και ο τρόπος που παρουσιάζεται η ίδια η σύσταση στο περιβάλλον διεπαφής του χρήστη (Romero, et al., 2007). Στη συνέχεια παρουσιάζεται ένα παράδειγμα μιας λίστας σύστασης, που περιλαμβάνει τρεις συστάσεις (R1, R2, R3):

"Hello Linda. You may find useful some of the following recommendations:
*-**Listen to** a recorded interview done to the professor about the course contents. **(R1 details)***
*-**Read** Mary's contributions in the Module 1 forum in the course. **(R2 details)***
*-**Try** the self-assessment of Module 1 to find out how you are progressing in the course. **(R3 details)"***

Εικόνα 1.3: Λίστα Σύστασης (Santos & Boticario, 2008).

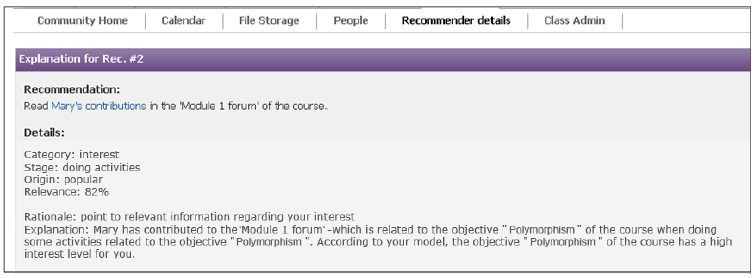
Στην *Εικόνα 1.3*, παρουσιάζεται σε περιβάλλον διεπαφής χρήστη, η λίστα σύστασης με τα ακόλουθα (Santos & Boticario, 2008):

- το χαιρετισμό (Greetings),
- τις συστάσεις (Recommendations), σε αυτή την περίπτωση υπάρχουν τρεις,
- τις ενέργειες που υποδεικνύει η σύσταση στο αντικείμενο LMS (action object), όπως π.χ. τη σύσταση 2, που υποδηλώνει στο μαθητή να διαβάσει ένα μήνυμα που δημοσιεύτηκε στο *φόρουμ της Mary*,
- την προέλευση της σύστασης (Origin of the recommendation), όπου προσδιορίζεται με μια διαφορετική εικόνα, στην αρχή της πρότασης σύστασης,
- την αιτιολόγηση και σημασιολογικές πληροφορίες (Justification & Semantic info) που συνδέονται με άλλη σελίδα, και αποτυπώνονται λεπτομερώς (*Εικόνα 1.4*).



Εικόνα 1.4: Λίστα συστάσεων μέσα από το περιβάλλον διεπαφής του χρήστη σε ένα LMS όπως παρουσιάστηκε από τους Santos & Boticario

Η *Εικόνα 1.5* αφορά τη 2η σύσταση από το προηγούμενο παράδειγμα και παρουσιάζει, μέσα από το περιβάλλον διεπαφής του χρήστη, μια επεξηγηματική σελίδα, που περιλαμβάνει λεπτομερείς πληροφορίες της σύστασης, ιδίως, της σημασιολογικής πληροφορίας, δηλαδή, την κατηγορία, το στάδιο, την προέλευση, τη συνάφεια (Category, Stage, Origin, Relevance), και την αιτιολόγηση, όπως το σκεπτικό και την επεξήγηση (Rationale, Explanation). Με τον τρόπο αυτό, δίνεται η δυνατότητα στους μαθητές, να ελέγχουν το μοντέλο σύστασης λεπτομερώς.



Εικόνα 1.5: Λεπτομερείς Πληροφορίες Σύστασης

Κεφάλαιο 2
Ανάλυση Αλγορίθμων Συστάσεων

Παρόλο που η αναζήτηση στο διαδίκτυο έχει γίνει πλέον καθημερινή ενασχόληση για μεγάλη μερίδα ανθρώπων, εξακολουθεί να υφίσταται το πρόβλημα της άμεσης και αποτελεσματικής πρόσβασης στην πληροφορία που είναι διαθέσιμη on line. Ο τεράστιος αριθμός των διαθέσιμων ιστοσελίδων, δυσχεραίνει τον εντοπισμό αυτών που είναι σχετικές με το προς αναζήτηση θέμα, ή που παρουσιάζουν κάποιο ενδιαφέρον.
(Ali & van Stam, 2004).

Υπάρχουν διάφορες προσεγγίσεις οι οποίες στοχεύουν στην επίλυση του προβλήματος, και κυρίως αφορούν στις μηχανές αναζήτησης, οι οποίες αποτελούν ένα από τα πιο χρήσιμα εργαλεία για αυτή την εργασία, ταυτόχρονα, έχουν την ίδια στιγμή, κάποια αρνητικά στοιχεία. Για παράδειγμα, κατά τη διάρκεια μιας αναζήτησης μπορούν να παράξουν έναν τεράστιο αριθμό ιστοσελίδων, ο οποίος φυσικά σε καμία περίπτωση δεν είναι «προσωποποιημένος», που σημαίνει ότι δεν έχει άμεση συνάφεια με το χρήστη (Basu, et al., 1998).

Οι λύσεις που έχουν προταθεί για τα παραπάνω αφορούν στη δημιουργία ερωτημάτων, τα οποία θα χρησιμοποιούν οι μηχανές αναζήτησης και επικεντρώνονται στο χρήστη και τις προτιμήσεις του. Σε αυτές τις λύσεις έρχονται να προστεθούν τα συστήματα σύστασης, τα οποία μάλιστα θεωρούνται χρησιμότατα εργαλεία σχετικά με την πρόσβαση στη διαθέσιμη πληροφορία (Ali & van Stam, 2004).

2.1 Εισαγωγή

Η ανάπτυξη των εργαλείων μάθησης βρίσκεται υπό ενδελεχή έρευνα τα τελευταία 50 χρόνια, και αρχικά είχε ως στόχο την ανάπτυξη λογισμικού για το εκπαιδευτικό μοντέλο, στο οποίο τον ενεργό ρόλο κατείχε ο δάσκαλος και τον παθητικό ο μαθητής (Cohen, 1995). Η κατανεμημένη τεχνητή νοημοσύνη όμως, έχει ως σκοπό την επίλυση τέτοιων προβλημάτων, με έναν πιο συνεργατικό τρόπο. Ένας από τους τομείς της κατανεμημένης τεχνητής νοημοσύνης είναι και τα πολυπρακτορικά συστήματα, στα οποία οι πράκτορες συνεργάζονται μεταξύ τους με σκοπό να επιτύχουν έναν κοινό στόχο. Οι αρχές και οι δομές πάνω στις οποίες βασίστηκαν τα πολυπρακτορικά συστήματα, παρουσιάζουν μία δυναμική για την ανάπτυξη των εκπαιδευτικών συστημάτων, λόγω του ότι τα προβλήματα που παρουσιάζουν, είναι ευκολότερο να επιλυθούν με έναν συνεργατικό τρόπο (Cohen, 1995).

Σύμφωνα με τους (Belkin & Croft, 1992) ένα πολυπρακτορικό σύστημα, είναι μια κατανεμημένη εφαρμογή του υπολογιστή, η οποία έχει δημιουργηθεί από έναν συνδυασμό αυτόνομων, ετερογενών, ασύγχρονων και έξυπνων διαδικασιών που ονομάζονται πράκτορες. Οι πράκτορες αυτοί έχουν τη δυνατότητα να συνεργάζονται μεταξύ τους, έτσι ώστε, να επιλύουν σύνθετα προβλήματα. Οι (Brusilovsky, et al., 2007) ορίζουν κάθε πράκτορα ως ένα λογισμικό, το οποίο λειτουργεί αυτόνομα και συνεχόμενα σε ένα συγκεκριμένο περιβάλλον, στο οποίο συχνά αλληλεπιδρούν και άλλοι πράκτορες, με έναν έξυπνο και ευέλικτο τρόπο, χωρίς να χρειάζονται παρέμβαση ή καθοδήγηση. Ιδεατά, κάθε πράκτορας, που λειτουργεί συνεχόμενα για ένα μεγάλο χρονικό διάστημα, θα πρέπει να είναι σε θέση να μαθαίνει διαμέσου της εμπειρίας του, και αν το περιβάλλον κάνει χρήση και άλλων πρακτόρων αυτός θα πρέπει να είναι σε θέση να επικοινωνεί και να συνεργάζεται μαζί τους σε έναν κοινό «κόσμο».

Οι παιδαγωγικοί πράκτορες σύμφωνα με τους (Brusilovsky, et al., 2007) είναι αυτοί οι οποίοι χρησιμοποιούνται για εκπαιδευτικούς σκοπούς, λειτουργούν ως πλοηγοί σε εικονικό περιβάλλον ή ως βοηθοί σε περιβάλλον εικονικής μάθησης, ώστε να βοηθήσουν τους μαθητές κατά τη διαδικασία της μάθησης. Έχουν την ιδιότητα να

απευθύνονται στο μαθητή κάνοντας χρήση φωνητικής επικοινωνίας ή μη φωνητικών χειρονομιών, όπως οι υποδείξεις. Με τον τρόπο αυτό, κάθε πράκτορας μπορεί να προσαρμοστεί στις ανάγκες του μαθητή και με τον τρόπο αυτό να λειτουργήσει ως βοήθεια και υποστήριξη.

Σύμφωνα με τους (Bobadilla, et al., 2013), με μια πιο λεπτομερής ματιά στη διαδικασία για παραγωγής ενός συστήματος συστάσεων, θα δούμε ότι βασίζεται σε ένα συνδυασμό κριτηρίων:

- Ο τύπος των διαθέσιμων στοιχείων στη βάση δεδομένων, π.χ. οι αξιολογήσεις (ratings), πληροφορίες από την εγγραφή του χρήστη, η κατάταξη των στοιχείων σύμφωνα με τα χαρακτηριστικά και τους (ranked), οι κοινωνικές σχέσεις μεταξύ των χρηστών.
- Ο αλγόριθμος φιλτραρίσματος που χρησιμοποιείται, π.χ. βασισμένος στο περιεχόμενο (content-based), συνεργατικός (collaborative), βασισμένος στη γνώση του περιεχομένου (context-aware) και υβριδικός (hybrid).
- Το μοντέλο που έχει επιλεγεί, π.χ., με βάση την άμεση χρήση των δεδομένων: Βασισμένο στη μνήμη (memory-based), ή βασισμένο στο μοντέλο (model-based).
- Οι τεχνικές που χρησιμοποιούνται: πιθανολογικές προσεγγίσεις (probabilistic approaches), δίκτυα Bayesian (Bayesian networks), ο αλγόριθμος πλησιέστερων γειτόνων (nearest neighbors algorithm), κ.λπ.
- Το επίπεδο ανεπάρκειας (sparsity level) της βάσης δεδομένων και η επιθυμητή επεκτασιμότητα (scalability).
- Οι επιδόσεις του συστήματος (κατανάλωση σε χρόνο και μνήμη).
- Η επίτευξη του επιδιωκόμενου στόχου, π.χ., προβλέψεις με υψηλού επιπέδου συστάσεις.
- Η επιθυμητή ποιότητα των αποτελεσμάτων, π.χ., η καινοτομία (novelty), η κάλυψη (coverage) και η ακρίβεια (precision).

Η έρευνα των συστημάτων συστάσεων απαιτεί τη χρήση ενός αντιπροσωπευτικού συνόλου των δημόσιων βάσεων δεδομένων που θα διευκολύνει τις έρευνες σχετικά με τις τεχνικές, τις μεθόδους και τους αλγόριθμους που αναπτύχθηκαν από τους ερευνητές. Μέσα από αυτές τις βάσεις δεδομένων, η επιστημονική κοινότητα μπορεί να αναπαράγει τα πειράματα για επικύρωση και βελτίωση των τεχνικών τους (Bobadilla, et al., 2013).

Οι Schafer, et al., στην έρευνα τους που διεξήγαγαν αναφέρουν ότι οι εσωτερικές λειτουργίες των συστημάτων συστάσεων χαρακτηρίζονται από τους αλγόριθμους

φιλτραρίσματος. Η πιο ευρέως χρησιμοποιούμενη ταξινόμηση χωρίζει τους αλγόριθμους φιλτραρίσματος σε (Schafer, et al., 2007):

- Φιλτράρισμα με βάση το περιεχόμενο (content-based filtering). Τα συστήματα με βάση το περιεχόμενο (Content-based), παρέχουν συστάσεις, αφού αναλύσουν πρώτα το περιεχόμενο όλων των σελίδων στις οποίες έχει γίνει αναζήτηση. Έπειτα αναζητούν όλες τις σελίδες εκείνες, που παρουσιάζουν αντίστοιχο περιεχόμενο (Ali & van Stam, 2004). Γίνονται συστάσεις με βάση τις επιλογές του χρήστη στο παρελθόν π.χ. σε ένα web-based e-commerce σύστημα συστάσεων, εάν ο χρήστης αγόρασε κάποιες ταινίες μυθοπλασίας στο παρελθόν, το σύστημα συστάσεων θα συστήσει μια πρόσφατη ταινία μυθοπλασίας που δεν έχει ακόμη αγοραστεί σε αυτή την ιστοσελίδα. Το φιλτράρισμα με βάση το περιεχόμενο δημιουργεί επίσης συστάσεις που χρησιμοποιούν το περιεχόμενο από τα αντικείμενα που προορίζονται για σύσταση. Ως εκ τούτου, ορισμένα περιεχόμενα μπορεί να αναλυθούν, όπως κείμενα, εικόνες και ήχοι. Από την ανάλυση αυτή, η ομοιότητα μπορεί να είναι μεταξύ αντικειμένων με βάση τα συστημένα στοιχεία που είναι παρόμοια με τα στοιχεία που ο χρήστης έχει επισκεφτεί, ακούσει, δει και σημείωσε ως θετικά (Antonopoulus & Salter, 2006).
    - Φιλτράρισμα συνεργασίας (collaborative filtering). Τα συνεργατικά συστήματα (Collaborative) που χρησιμοποιούν το φιλτράρισμα αυτό, αξιολογούν τα χαρακτηριστικά ενός χρήστη, με βάση τα χαρακτηριστικά άλλων χρηστών και έτσι τον παραπέμπουν σε σελίδες που έχουν επισκεφτεί παρόμοιοι χρήστες (Ali & van Stam, 2004). Επιτρέπει στους χρήστες να δώσουν βαθμολογίες για ένα σύνολο στοιχείων με τέτοιο τρόπο ώστε, όταν υπάρχουν επαρκείς πληροφορίες να αποθηκεύονται στο σύστημα. Έτσι προσφέρεται η δυνατότητα να κάνουμε συστάσεις για κάθε χρήστη με βάση τις πληροφορίες που παρέχονται από τους χρήστες, οι οποίοι θεωρούμε πως έχουν κοινά με αυτές τις συστάσεις (Su & Khoshgoftaar , 2009)
    - Υβριδικό φιλτράρισμα (hybrid filtering). Τέλος, τα υβριδικά συστήματα υιοθετούν και τις δύο παραπάνω πρακτικές (Ali & van Stam, 2004). Συνήθως χρησιμοποιείται ένας συνδυασμός των συστημάτων συστάσεων με το δημογραφικό φιλτράρισμα και με το φιλτράρισμα με βάση το περιεχόμενο, για να εκμεταλλευτεί τα πλεονεκτήματα της κάθε μίας από αυτές τις τεχνικές. Το υβριδικό φιλτράρισμα βασίζεται συνήθως σε βιο-εμπνευσμένες ή πιθανολογικές

μεθόδους όπως οι γενετικοί αλγόριθμοι, τα νευρωνικά δίκτυα, τα Bayesian δίκτυα και την ομαδοποίηση (clustering) (Porcel, et al., 2012).

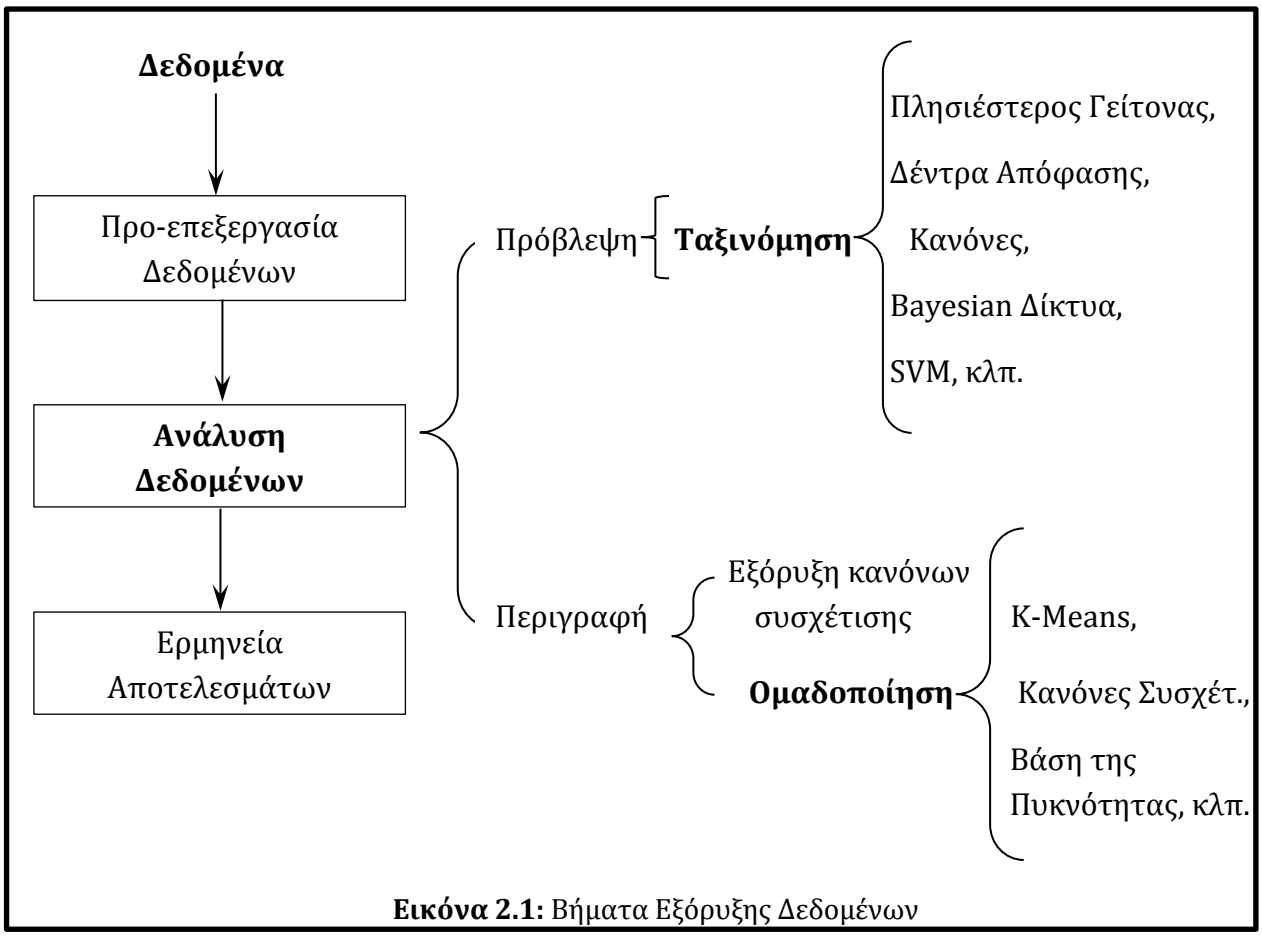
2.2 Εξόρυξη Δεδομένων για Αλγορίθμους Συστάσεων

Η εξόρυξη δεδομένων (Data Mining), μπορεί να χαρακτηριστεί σαν ένα σύνολο τεχνικών, που αναλύει μεγάλα σύνολα δεδομένων. Στην ουσία διερευνά και αναλύει μεγάλα σύνολα πρωτογενών δεδομένων, με στόχο, να παρουσιάσει συγκεκριμένες δομές και σχέσεις ανάμεσα τους. Τα περισσότερα συστήματα συστάσεων, φέρουν στο πυρήνα τους έναν αλγόριθμο, ο οποίος μπορεί να θεωρηθεί ως ένα συγκεκριμένο παράδειγμα μιας τεχνικής εξόρυξης δεδομένων (Ricci, et al., 2011).

Η διαδικασία της εξόρυξης δεδομένων συνήθως αποτελείται από 3 βήματα, που πραγματοποιούνται διαδοχικά (Amatriain, et al., 2011):

- Προ-επεξεργασία Δεδομένων,
- Ανάλυση Δεδομένων και
- Ερμηνεία Αποτελεσμάτων.

Στην **εικόνα 2.1**, παρουσιάζονται οι μέθοδοι και τεχνικές, μόνο για την ανάλυση δεδομένων, που αφορά το πρόβλημα της εξόρυξης δεδομένων, και σχετίζονται με τους



Εικόνα 2.1: Βήματα Εξόρυξης Δεδομένων

αλγορίθμους συστάσεων. Όπως έχουν παρουσιαστεί από τους (Amatriain, et al., 2011). Στη συνέχεια, σύμφωνα με την εικόνα 2.1, γίνεται αναφορά σχετικά με την Ταξινόμηση (Classification), που ανήκει στη μέθοδο της Πρόβλεψης (Prediction), και την Ομαδοποίηση (Clustering), που ανήκει στη μέθοδο της Περιγραφής (Description).

2.2.1 Ταξινόμηση - (Classification)

Ένας ταξινομητής είναι μια χαρτογράφηση μεταξύ του χώρου χαρακτηριστικών (ή του χώρου δυνατοτήτων), των στοιχείων (feature space) που θα ταξινομηθούν, και του χώρου που αντιπροσωπεύει τις τάξεις (label space). Παραδείγματος χάριν, για την ταξινόμηση ενός εστιατορίου μεταξύ δύο τάξεων, «καλό» ή «κακό», θα πρέπει στον χώρο των δυνατοτήτων, να υπάρχει μια σειρά χαρακτηριστικών που να το περιγράφουν αντίστοιχα, όπως, γρήγορο, καθαρό, μενού, τιμές, εξυπηρέτηση κλπ, έτσι ώστε η βαθμολογία να το κατατάσσει σε μια εκ των δύο κατηγοριών (Amatriain, et al., 2011).

Υπάρχουν πολλοί τύποι ταξινόμησης, αλλά οι πιο σημαντικοί είναι:

- η υπό επίβλεψη (supervised) και
- χωρίς επίβλεψη (unsupervised) ταξινόμηση.

Στην υπό επίβλεψη ταξινόμηση, υπάρχει ένα σύνολο χαρακτηριστικών, που είναι γνωστά εκ των προτέρων και υπάρχει και ένα σύνολο ταξινομημένων τάξεων που αποτελούν το σύνολο εκπαίδευσης. Αλγόριθμοι που μαθαίνουν από ταξινομητές υπό επίβλεψη θεωρούνται: ο Πλησιέστερος γείτονας (Nearest Neighbor), τα δέντρα απόφασης (Decision Trees), οι ταξινομητές που βασίζονται σε κανόνες (Ruled-based Classifiers), οι ταξινομητές Bayesian (Bayesian Classifiers), τα νευρωνικά δίκτυα (Artificial Neural Networks), οι μηχανές υποστήριξης διανυσμάτων (Support Vector Machines) κλπ (Ricci, et al., 2011).

Στην χωρίς επίβλεψη ταξινόμηση, τα χαρακτηριστικά δεν είναι γνωστά εκ των προτέρων, και τα στοιχεία, σύμφωνα με κάποια κριτήρια, οργανώνονται την στιγμή της ταξινόμησης. Σ' αυτή την κατηγορία, της ταξινόμησης χωρίς επίβλεψη, ανήκει και η Ομαδοποίηση (Clustering), (Ricci, et al., 2011), που θα περιγραφεί στην συνέχεια.

2.2.2 Ομαδοποίηση (Clustering)

Η Ομαδοποίηση, αναφέρεται και ως μη επιβλεπόμενη μάθηση. Συνίσταται στην ανάθεση αντικειμένων σε ομάδες, έτσι ώστε τα στοιχεία στις ίδιες ομάδες να μοιάζουν περισσότερο από ότι τα στοιχεία σε διαφορετικές ομάδες. Ο στόχος είναι να ανακαλυφθούν φυσικές ομάδες που υπάρχουν στα δεδομένα. Η ομοιότητα καθορίζεται

χρησιμοποιώντας ένα μέτρο απόστασης. Ο στόχος ενός αλγορίθμου ομαδοποίησης είναι να ελαχιστοποιήσει τις αποστάσεις εντός-συνόλου μεγιστοποιώντας παράλληλα τις αποστάσεις εκτός-συνόλου .

Υπάρχουν δύο κατηγορίες αλγορίθμων ομαδοποίησης (Amatriain, et al., 2011):

- Οι Ιεραρχικοί (hierarchical) και
- Οι Τμηματικοί (Partitional).

Οι τμηματικοί αλγόριθμοι ομαδοποίησης διαιρούν τα στοιχεία δεδομένων σε μη επικαλυπτόμενες ομάδες (non-overlapping clusters), έτσι ώστε κάθε στοιχείο να βρίσκεται σε μία ομάδα. Οι Ιεραρχικοί αλγόριθμοι ομαδοποίησης, διαιρούν την κάθε ομάδα, διαδοχικά, σε άλλες υπό-ομάδες αντικειμένων, παράγοντας ένα σύνολο από φωλιασμένες ομάδες, που είναι οργανωμένες σαν ένα ιεραρχικό δέντρο (Amatriain, et al., 2011).

Πολλοί αλγόριθμοι ομαδοποίησης, προσπαθούν να ελαχιστοποιήσουν την συνάρτηση, που μετρά την ποιότητα της ομαδοποίησης. Αυτή η συνάρτηση συχνά αναφέρεται ως αντικειμενική συνάρτηση, έτσι η ομαδοποίηση μπορεί να θεωρηθεί ως ένα πρόβλημα βελτιστοποίησης (optimization). Ο ιδανικό αλγόριθμος ομαδοποίησης θα εξετάσει όλες τις δυνατές τμηματοποιήσεις και εξάγει σαν αποτέλεσμα την τμηματοποίηση που ελαχιστοποιεί τη συνάρτηση της ποιότητας (Ricci, et al., 2011).

Ένα βασικό σημείο, είναι ότι η ομαδοποίηση είναι ένα δύσκολο πρόβλημα, για το οποίο η αναζήτηση βέλτιστων λύσεων συχνά δεν είναι δυνατόν! Η επιλογή συγκεκριμένου αλγορίθμου ομαδοποίησης και των παραμέτρων του εξαρτάται από πολλούς παράγοντες, συμπεριλαμβανομένου των χαρακτηριστικών των δεδομένων (Amatriain, et al., 2011).

Μέθοδοι που ανήκουν στην κατηγορία αυτή είναι (Frey & Dueck, 2007):

- Η ομαδοποίηση K-means που είναι μια μέθοδος τμηματοποίησης,
- Η ομαδοποίηση που βασίζεται στην πυκνότητα (Density-based clustering), όπως ο DBSCAN, όπου βασίζεται στην πυκνότητα των σημείων, εντός ορισμένης ακτίνας.
- Και η ομαδοποίηση που σχετίζεται με τους Κανόνες Συσχέτισης και εστιάζεται στην εξεύρεση κανόνων που θα προβλέψουν την εμφάνιση ενός στοιχείου με βάση τις εμφανίσεις των άλλων στοιχείων σε μια συναλλαγή.

Οι κύριες κατηγορίες αλγορίθμων που χρησιμοποιούνται στην εκπαιδευτική διαδικασία και αφορούν τα προαναφερθέν χωρίζονται σε τρείς βασικές κατηγορίες, και θα περιγραφούν παρακάτω. Οι κατηγορίες αυτές είναι :

- Αλγόριθμοι που εστιάζουν στο περιεχόμενο (Content-based algorithms),
- Αλγόριθμοι συνεργατικών φίλτρων (Collaborative filtering algorithms),
- Αλγόριθμοι με βάση τους γράφους (Graph-based algorithms).

2.3 Αλγόριθμοί Συστάσεων που Εστιάζουν στο Περιεχόμενο (Content-Based Recommender Algorithms)

Αυτό που ουσιαστικά προσφέρουν οι διαδικτυακές εφαρμογές, στις οποίες υπάρχει άμεση αλληλεπίδραση με το χρήστη, είναι μία λίστα με διαθέσιμες επιλογές από τις οποίες ο χρήστης δύναται να επιλέξει κάποια ή να δημιουργήσει μία σχέση αλληλεπίδρασης. Παρόλο που ο web server ουσιαστικά παρουσιάζει κώδικα σε HTML, και ο χρήστης βλέπει μία σελίδα, υπάρχει μία τεράστια βάση δεδομένων πίσω από αυτό, αλλά και μία δυναμική δημιουργία ιστοσελίδων με λίστες αντικειμένων.

Οι αλγόριθμοι συστάσεων με βάση το περιεχόμενο αναλύουν τις περιγραφές των αντικειμένων με σκοπό να αναγνωρίσουν αυτά τα οποία θα προκαλέσουν ιδιαίτερο ενδιαφέρον στο χρήστη. Λόγω του ότι τα συστήματα σύστασης ποικίλουν ανάλογα με τα αντικείμενα αναπαράστασης, στην παρούσα ενότητα θα κατηγοριοποιηθούν και οι αλγόριθμοι με βάση την αναπαράσταση αυτή (Brusilovsky, et al., 2007).

2.3.1 Αναπαράσταση Αντικειμένου

Οι αλγόριθμοι αυτοί λειτουργούν με βάση, τη δημιουργία πινάκων σε βάσεις δεδομένων. Η κάθε εγγραφή περιέχει τις τιμές για τα διάφορα χαρακτηριστικά της και διέπεται από ένα μονοσήμαντο χαρακτηριστικό, που συντελεί στη διαδικασία της αναζήτησής της. Κάθε αντικείμενο περιγράφεται από τον ίδιο αριθμό χαρακτηριστικών, ενώ υπάρχει και μία γνωστή σειρά τιμών, που μπορούν να πάρουν τα χαρακτηριστικά αυτά. Οι ιστοσελίδες που αφορούν στους αλγόριθμους αυτούς, κατασκευάζονται με βάση τα πεδία της βάσης δεδομένων, και με τρόπο τέτοιο ώστε αυτά να είναι ορατά.

Πολλές φορές, γίνεται χρήση πεδίων που περιέχουν μη δομημένα δεδομένα όπως το ελεύθερο κείμενο. Στην περίπτωση αυτή και αντίθετα με την περίπτωση των δομημένων δεδομένων, είναι αδύνατο να υπάρξει προσδιορισμός ο οποίος να καθορίζει

το ότι περιλαμβάνεται στο πεδίο. Επιπλέον, η υψηλή πολυπλοκότητα της φυσικής γλώσσας που παρίσταται, συνήθως, στα ελεύθερα πεδία και κάνει χρήση πολύσημων λέξεων ή συνωνύμων δυσχεραίνει τη διαδικασία αναζήτησης του επιθυμητού αντικειμένου στα πεδία αυτά.

Κάποια domains συνήθως παρουσιάζονται καλύτερα, εάν αποτελούνται από ημιδομημένα δεδομένα, κατά τα οποία, κάποια χαρακτηριστικά διέπονται από δομημένες τιμές, ενώ κάποια άλλα αφορούν σε ελεύθερα πεδία κειμένου.

Μία προσέγγιση σε ότι αφορά τη διαχείριση των ελεύθερων πεδίων, είναι αυτά που μπορούν να παρουσιαστούν με ένα πιο δομημένο τρόπο. Για παράδειγμα, κάθε λέξη, μπορεί να αντιμετωπιστεί ως ένα χαρακτηριστικό, το οποίο θα έχει μία Boolean τιμή, που θα καταδεικνύει εάν πρόκειται για άρθρο ή μία τιμή ακεραίου, ή θα καταδεικνύει την επανάληψη των λέξεων μέσα στην πρόταση (Brusilovsky, et al., 2007).

Πολλά από τα συστήματα που κάνουν χρήση ελεύθερου κειμένου, χρησιμοποιούν τεχνική με σκοπό να δημιουργήσουν μία δομημένη παρουσίαση, που προέρχεται από μηχανές αναζήτησης κειμένου. Στην περίπτωση αυτή, αντί να χρησιμοποιούνται ολόκληρες οι λέξεις χρησιμοποιείται μονάχα η ρίζα τους. Σκοπός της τεχνικής αυτής είναι η δημιουργία ενός όρου ο οποίος θα εμπεριέχει κοινό νόημα για πολλές λέξεις όπως υπολογίζω, υπολογιστής, υπολογισμός. Πίσω από τον όρο αυτό, κρύβεται η τιμή μίας μεταβλητής που αναπαριστά τη σημασία ή τη σχετικότητα της λέξης. Αυτό το οποίο αναμένεται από τη χρήση της τεχνικής αυτής, είναι ότι, οι όροι με το μεγαλύτερο βάρος, θα εμφανίζονται συχνότερα σε ένα κείμενο από ότι οι υπόλοιποι, και άρα, αποτελούν μία ένδειξη, πάνω στην οποία μπορεί να σχεδιαστεί η σύσταση (Quinlan, 2007).

Αντικείμενα που μπορούν να συνιστώνται σε κάποιο χρήστη περιέχουν διάφορα χαρακτηριστικά, τα οποία, ονομάζονται ιδιότητες. Για παράδειγμα, σε μια εφαρμογή σύστασης ταινίας, χαρακτηριστικά που συναντούμε για την περιγραφή της είναι: ηθοποιοί, σκηνοθέτες, είδη μουσικής, κλπ.

Όταν κάθε αντικείμενο, περιγράφεται από το ίδιο σύνολο ιδιοτήτων, και υπάρχει ένα γνωστό σύνολο τιμών των ιδιοτήτων που μπορεί να λάβει, το αντικείμενο αντιπροσωπεύεται μέσω δομημένων δεδομένων. Σε αυτήν την περίπτωση, πολλοί αλγόριθμοι μηχανικής μάθησης (Machine Learning), μπορούν να χρησιμοποιηθούν για την εκμάθηση ενός προφίλ χρήστη (Pasquale, et al., 2011).

Στα περισσότερα συστήματα φιλτραρίσματος με βάση το περιεχόμενο, οι περιγραφές των αντικειμένων, είναι χαρακτηριστικά κειμένου που εξάγονται από ιστοσελίδες,

μηνύματα ηλεκτρονικού ταχυδρομείου, κλπ, που δεν υπάρχουν χαρακτηριστικά με σαφώς καθορισμένες τιμές, έτσι, δημιουργείται μία σειρά από επιπλοκές κατά την εκμάθηση ενός προφίλ χρήστη, λόγω της ασάφειας αυτής. Το πρόβλημα είναι ότι τα παραδοσιακά προφίλ που βασίζονται σε λέξεις-κλειδιά, δεν είναι σε θέση να καταλάβουν τη σημασιολογία των ενδιαφερόντων των χρηστών, επειδή οδηγούνται κυρίως από μια λειτουργία ταιριάσματος μια συμβολοσειράς. Τέτοια προβλήματα ονομάζονται πολυσημία (λέξεις με πολλαπλό νόημα) και συνωνυμία (πολλές λέξεις με το ίδιο νόημα). Ως εκ τούτου, λόγω της συνωνυμίας, οι σχετικές πληροφορίες μπορούν να χαθούν αν το προφίλ δεν περιέχει τις ακριβείς λέξεις-κλειδιά, ενώ, λόγω της πολυσημίας, λάθος έγγραφα θα μπορούσαν να θεωρηθούν συναφή (Pasquale, et al., 2011).

Η σημασιολογική ανάλυση και η ενσωμάτωσή της σε μοντέλα εξατομίκευσης, είναι μια από τις πιο καινοτόμες και ενδιαφέρουσες προσεγγίσεις που προτείνονται στη βιβλιογραφία για την επίλυση αυτών των προβλημάτων. Η βασική ιδέα είναι η υιοθέτηση των βάσεων γνώσης, όπως λεξικά, για σχολιασμό στοιχείων που εκπροσωπούν ένα προφίλ, προκειμένου να αποκτήσουν μια «σημασιολογική ερμηνεία» οι ανάγκες των χρηστών (Ricci, et al., 2011).

2.3.2 Προφίλ Χρηστών

Στα συστήματα συστάσεων, τα προφίλ που κάνουν χρήση τις προτιμήσεις των χρηστών, χρησιμοποιούνται κατά κόρον. Τα προφίλ αυτά ενδέχεται να περιέχουν πολλούς και διαφορετικούς τύπους πληροφορίας. Οι κυριότεροι τύποι είναι (Yang & Pedersen, 1997):

1. *Το μοντέλο των προτιμήσεων του χρήστη*. Υπάρχουν πολλοί και διαφορετικοί τρόποι απεικόνισης του μοντέλου αυτού, αλλά, ο πιο συνήθης είναι μία συνάρτηση, η οποία για κάθε αντικείμενο προβλέπει την πιθανότητα ενδιαφέροντος που μπορεί να δείξει ο χρήστης.
2. *Το ιστορικό των αλληλεπιδράσεων του χρήση με το σύστημα σύστασης*. Το ιστορικό αυτό ενδεχόμενα να συμπεριλαμβάνει αντικείμενα τα οποία επισκέφθηκε παλαιότερα ο χρήστης ή και πληροφορίες σχετικές με την αλληλεπίδρασή του με το σύστημα. Άλλες πληροφορίες που μπορεί να αποθηκεύονται στο ιστορικό είναι, τα ερωτήματα που τεθήκαν από το χρήστη κατά τις διάφορες αναζητήσεις του.

Το πρόβλημα της μάθησης του προφίλ των χρηστών, παρουσιάζεται ως μια διεργασία δυαδικής κατηγοριοποίησης κειμένων. Κάθε αντικείμενο, σύμφωνα πάντοτε με τις προτιμήσεις του χρήστη, πρέπει να ταξινομηθεί σαν «ενδιαφέρον» ή «μη ενδιαφέρον». Ως εκ τούτου, το σύνολο των κατηγοριών είναι C = {c+, c-}, όπου c+ είναι η θετική κλάση (ο χρήστης ενδιαφέρεται) και c- η αρνητική (ο χρήστης δεν ενδιαφέρεται) (Pasquale, et al., 2011).

Στα συστήματα συστάσεων συνήθως χρησιμοποιούνται οι «αλγορίθμοι μάθησης» (learning algorithms), οι οποίοι, μαθαίνουν μέσα από λειτουργίες, που μοντελοποιούν τα ενδιαφέροντα του χρήστη. Αυτές οι μέθοδοι, τυπικά, απαιτούν από τους χρήστες να βαθμολογούν διάφορα αντικείμενα σε μια συγκεκριμένη ετικέτα, και αυτόματα με την διαδικασία φιλτραρίσματος, καταλήγουν στην αξιοποίηση του προφίλ του χρήστη σύμφωνα με την βαθμολόγηση του (Sebastiani, 2002), (Pasquale, et al., 2011).

Σε ένα σύστημα που διέπεται από τέτοιους αλγορίθμους, γίνεται χρήση διεπαφών, που επιτρέπουν στους χρήστες να φτιάξουν μόνοι τους τις αναπαραστάσεις των ενδιαφερόντων τους. Συχνά χρησιμοποιούνται check boxes, έτσι ώστε να δοθεί το δικαίωμα της κατευθυνόμενης επιλογής στο χρήστη. Από τη στιγμή που ο χρήστης θα σημειώσει αυτή του την προτίμηση, κινείται στον αλγόριθμο μία διαδικασία, η οποία αντιστοιχεί σε μία βάση δεδομένων, τα αντικείμενα εκείνα, που αρμόζουν στις προτιμήσεις του χρήστη και τα εμφανίζει στην οθόνη (Yang & Pedersen, 1997).

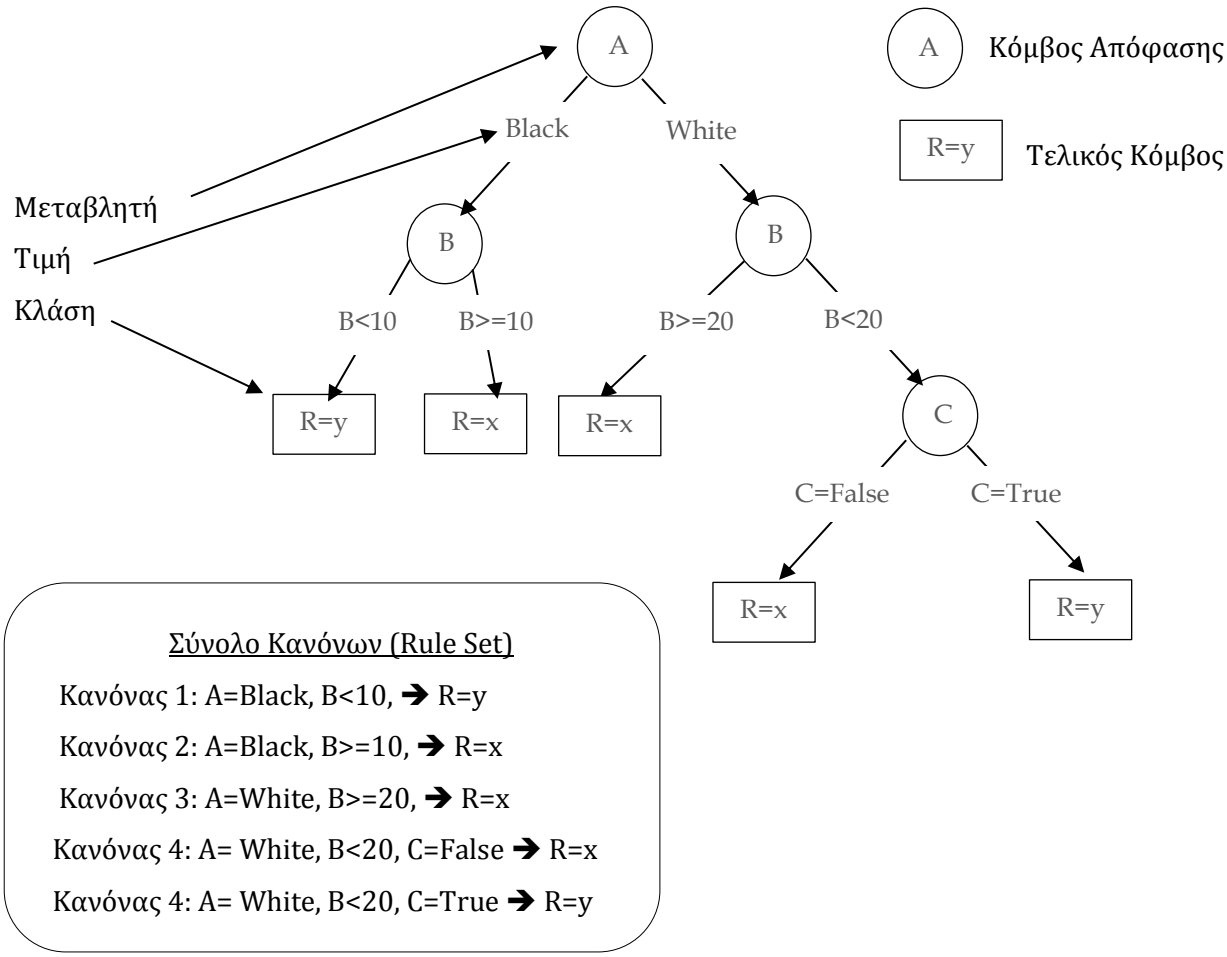
2.3.3 Δέντρα Αποφάσεων

Τα δέντρα αποφάσεων είναι ταξινομητές που έχουν την μορφή ενός δέντρου και το τελικό αποτέλεσμα του είναι μια κλάση. Τα αντικείμενα που ταξινομούνται, απαρτίζονται από κλάσεις (πεδία), που έχουν τελικές τιμές (target values). Οι κόμβοι του δέντρου μπορεί να είναι (Rokach & Maimon, 2008):

- Κόμβοι απόφασης, σε αυτούς τους κόμβους, ελέγχεται η τιμή μιας κλάσης, προκειμένου να καθοριστεί, ποια κατεύθυνση του υποδέντρου θα ακολουθηθεί.
- Τελικοί κόμβοι, που δείχνουν την τελική τιμή (target value)(Εικόνα 2.2).

Οι αλγόριθμοι αυτοί, δημιουργούν ένα δέντρο αποφάσεων, που βασίζεται σε αναδρομή των δεδομένων και στην υποδιαίρεση αυτών σε ξεχωριστές ομάδες, έως ότου η κάθε υπό ομάδα να περιέχει μονάχα μία κλάση. Το δέντρο αποφάσεων δημιουργείται, όταν χωριστούν τα δεδομένα σε υπο-ομάδες, μέχρι το σημείο όπου κάθε υπο-ομάδα να αντιστοιχεί σε μια κλάση (Cohen, 1996).

Το δέντρο σχηματίζεται από κλάσεις που έχουν την ίδια τιμή ως προς κάποιο χαρακτηριστικό. Αναπαράγεται το μοντέλο αυτό με την αναδρομή και σχηματίζεται το δέντρο απόφασης. Ένα παράδειγμα δέντρου απόφασης φαίνεται στην εικόνα 2.2.



Εικόνα 2.2: Δέντρο Απόφασης – Σύνολο Κανόνων

Το δέντρο απόφασης είναι απλή προσέγγιση μηχανικής μάθησης, που ο κάθε κόμβος απόφασης, αποτελεί μια επιλογή ανάμεσα σε πολλές εναλλακτικές λύσεις, ενώ κάθε τελικός κόμβος, μια απόφαση (ταξινόμηση). Για να ταξινομηθούν τα αντικείμενα, ακολουθούν ένα μονοπάτι του δέντρου προς τα κάτω, παίρνοντας την κάθε ακμή που αντιστοιχεί στην τιμή μιας κλάσης. Το δέντρο απόφασης είναι χρήσιμο στα προβλήματα που σχετίζονται με την κατηγοριοποίηση. Στην κατηγοριοποίηση εκτελούνται δύο βήματα. Το πρώτο είναι η δημιουργία του δέντρου απόφασης και το δεύτερο είναι η χρήση του στην Βάση Δεδομένων (Amatriain, et al., 2011).

Στην ταξινόμηση, όταν μια εγγραφή μπαίνει στο δέντρο απόφασης, ξεκινά από τον κόμβο της κορυφής και ελέγχεται ποια διαδρομή θα ακολουθηθεί, δηλαδή ποιος κόμβος-παιδί θα επιλεγεί. Ο έλεγχος αυτός επαναλαμβάνεται έως ότου να τερματιστεί

στον τελικό κόμβο. Η ταξινόμηση γίνεται με τον ίδιο τρόπο για όλες τις εγγραφές που καταλήγουν σε τελικό κόμβο. Κάθε μονοπάτι που σχηματίζεται, αντιπροσωπεύεται και από ένα κανόνα, Εικόνα 2.2, (Rokach & Maimon, 2008).

Τέλος η εκπαίδευση ενός δέντρου δημιουργείτε με την συνεχή διάσπαση των δεδομένων, βάση των μεταβλητών. Επιλέγεται πάντοτε η μεταβλητή εκείνη που διαχωρίζει καλύτερα την τελική κλάση. Στην κορυφή του δέντρου τοποθετείται η μεταβλητή που διαχωρίζει τα δεδομένα καλύτερα. Ο τερματισμός του αλγορίθμου επιτυγχάνεται όταν φτάσει σε κόμβο που δεν μπορεί πλέον να διασπάσει τα δεδομένα. Ο κόμβος αυτός αποτελεί και τον τελικό κόμβο ή φύλλο του δέντρου (Rokach & Maimon, 2008).

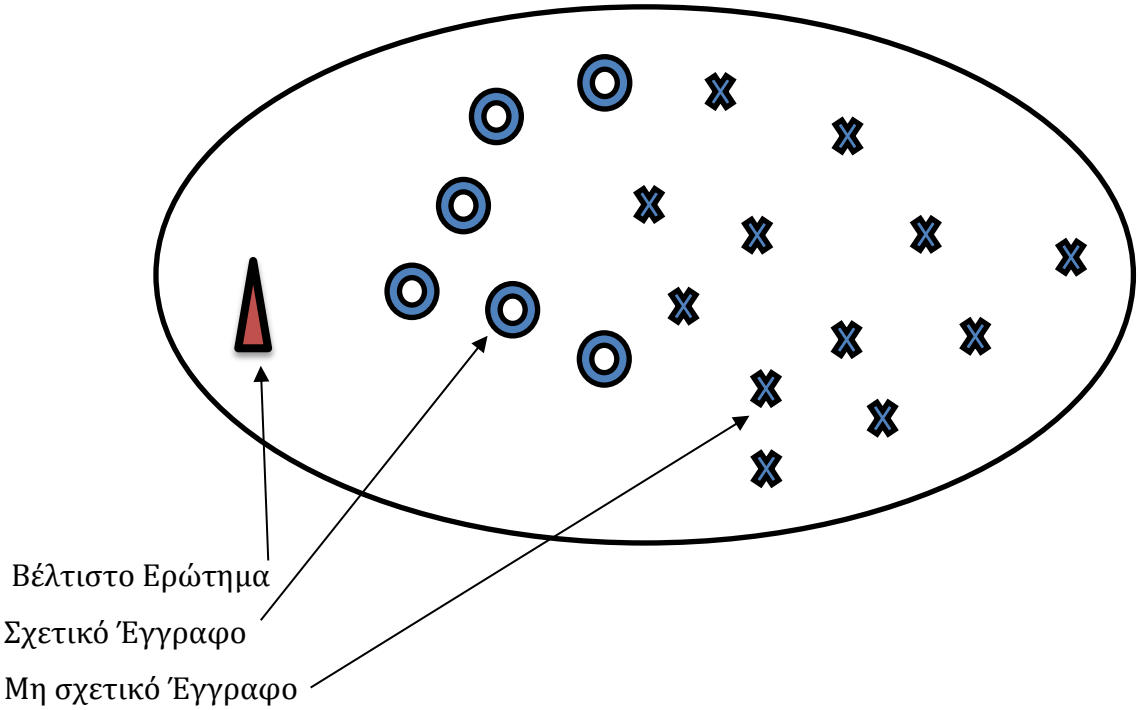
2.3.5 Ο Αλγόριθμος του Rocchio

Η επιτυχία των αλγορίθμων που κάνουν χρήση των διανυσμάτων χώρου οφείλεται στην ικανότητα του χρήστη να δημιουργεί ερωτήματα επιλέγοντας από μία ομάδα αντιπροσωπευτικών λέξεων. Η μέθοδος αυτή συντελεί στη δημιουργία ερωτημάτων, που βασίζονται σε προηγούμενες αναζητήσεις αλλά και προτιμήσεις του χρήστη. Οι μέθοδοι αυτοί συχνά αναφέρονται ως, μέθοδοι σχετικής γνώμης, βασική αρχή των οποίων, είναι, να επιτρέπουν στους χρήστες να αξιολογούν τα αντικείμενα τα οποία προσπελαύνουν, με βάση την πληροφορία την οποία λαμβάνουν. Η μορφή αυτή της λήψης σχολιασμού μπορεί να χρησιμοποιηθεί για τον επανασχεδιασμό του αρχικού ερωτήματος (Rocchio, 1971).

Ο αλγόριθμος του Rocchio είναι ευρέως χρησιμοποιούμενος σε συστήματα όπου γίνεται χρήση αλγορίθμων σχετικότητας γνώμης, που λειτουργούν με βάση το μοντέλο του διανύσματος χώρου. Ο αλγόριθμος βασίζεται στη διαμόρφωση του αρχικού ερωτήματος διαμέσου πρωτοτύπων, που αξιολογούν διαφορετικά τα σχετικά και τα μη σχετικά αντικείμενα. Η προσέγγιση αυτή δημιουργεί δύο ειδών πρωτότυπα, αθροίζοντας το σύνολο όλων των σχετικών και μη σχετικών κειμένων (Rocchio, 1971).

Η σχετικότητα της ανατροφοδότησης είναι μια τεχνική που υιοθετείται στην διαδικασία της ανάκτησης της πληροφορίας. Βοηθά τους χρήστες να βελτιώσουν, σταδιακά, διάφορα ερωτήματα με βάση τα προηγούμενα αποτελέσματα των αναζητήσεων. Αποτελείται από τις ανατροφοδοτήσεις των χρηστών και τις αποφάσεις του συστήματος σχετικά με την καταλληλότητα των πληροφοριών, πάντοτε σε σχέση

με τις ανάγκες πληροφόρησής των χρηστών. Η ιδέα της σχετικής ανατροφοδότησης (Relevance Feedback), στην ουσία τοποθετεί τον χρήστη σε μια διαδικασία ανάκτησης, βελτιώνοντας το σύνολο των αποτελεσμάτων (Pasquale, et al., 2011).



Εικόνα 2.3: Το Ερώτημα του Rocchio

Ψάχνουμε δηλαδή το διάνυσμα ενός ερωτήματος που θα μειώνει την ομοιότητα με τα μη-σχετικά κείμενα και θα αυξάνει την ομοιότητα με τα σχετικά έγγραφα, Εικόνα 2.3.

Η γενική αρχή είναι να επιτρέψει στους χρήστες να αξιολογήσουν τα έγγραφα που προτείνονται από το σύστημα σε σχέση με τις ανάγκες πληροφόρησης τους. Αυτή η μορφή της ανατροφοδότησης μπορεί στη συνέχεια να χρησιμοποιηθεί για να βελτιώσει σταδιακά το προφίλ του χρήστη ή να εκπαιδεύσει τον αλγόριθμο. Συγκεκριμένα τα βήματα της διαδικασίας είναι τα ακόλουθα (Sebastiani, 2002):

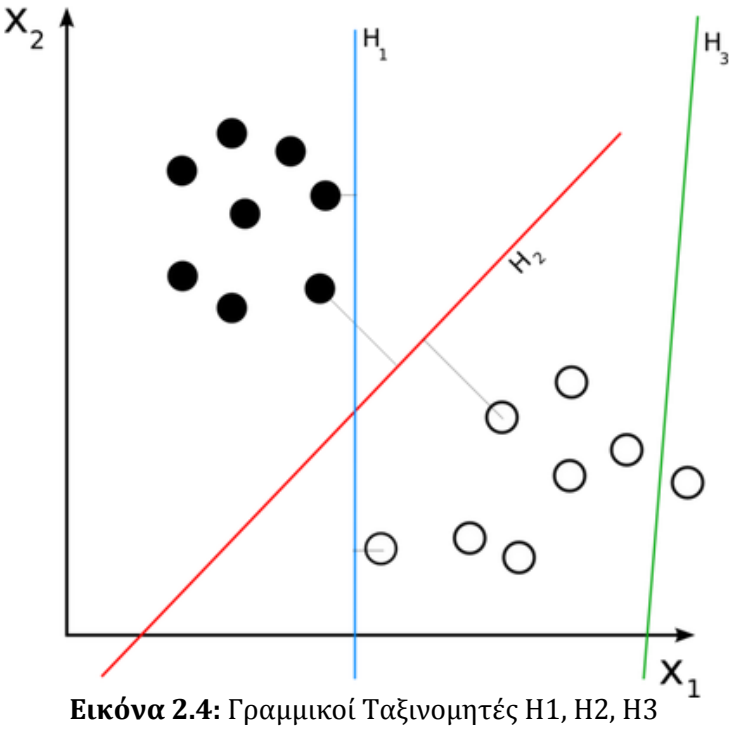
1. Χρήστης → Υποβολή Ερωτήματος
2. Σύστημα → Επιστροφή συνόλου ανακτημένων εγγράφων
3. Χρήστης → Κατάδειξη των Σχετικών και των Μη Σχετικών εγγράφων
4. Σύστημα → Υπολογισμός καλύτερης κατάταξης εγγράφων σύμφωνα με την σχετική ανατροφοδότηση του χρήστη
5. Σύστημα → Παρουσίαση αναθεωρημένου συνόλου εγγράφων

Η ανατροφοδότηση μπορεί να επιτευχθεί σε μία ή περισσότερες επαναλήψεις. Είναι δύσκολο να διαμορφωθεί ένα καλό ερώτημα, όταν δεν είναι γνωστό το σύνολο των

εγγράφων. Έτσι, με την ανατροφοδότηση μπορεί να γίνει βελτίωση του ερωτήματος (Pasquale, et al., 2011).

2.3.6 Γραμμικοί Ταξινομητές

Είναι αλγόριθμοι που διαχωρίζουν τις κλάσεις τους, χρησιμοποιώντας γραμμικές συναρτήσεις. Οι γραμμικοί ταξινομητές μπορούν να περιγραφούν σε ένα κοινό πλαίσιο αναπαράστασης, καθώς η έξοδος που παράγεται από την εκπαιδευτική διαδικασία, είναι ένα διάνυσμα n-διαστάσεων (όπως ένα κείμενο), το γινόμενο του οποίου, παράγει ένα αριθμητικό σκορ πρόβλεψης (Brusilovsky, et al., 2007).



Εικόνα 2.4: Γραμμικοί Ταξινομητές H1, H2, H3

Στην εικόνα 2.4, οι ταξινομητές H1 και H2 έχουν πετύχει στην ταξινόμηση, ενώ ο ταξινομητής H3 απέτυχε.

Η διατήρηση της αριθμητικής πρόβλεψης οδηγεί σε μία προσέγγιση γραμμικής παλινδρόμησης, ενώ με τον ορισμό ενός κατωφλίου οι συνεχείς προβλέψεις μπορούν να μετατραπούν σε διακριτές, και έτσι να δημιουργηθούν κλάσεις, που θα αποτελούν ξεχωριστές κατηγορίες. Το πλαίσιο αυτό διέπει όλους τους γραμμικούς ταξινομητές. Οι διάφοροι αλγόριθμοι όμως διαφέρουν μεταξύ τους, ως προς τις μεθόδους που χρησιμοποιούν, ώστε να «εξάγουν» από το χρήστη το διάνυσμα (Brusilovsky, et al., 2007).

Ένα σημαντικό πλεονέκτημα των αλγορίθμων αυτών είναι ότι αυτοί μπορούν να εφαρμοστούν ακόμη και on line και έτσι είναι πλήρως κατάλληλοι για εφαρμογές που λειτουργούν σε περιορισμούς πραγματικού χρόνου (Brusilovsky, et al., 2007).

2.3.7 Πιθανολογικές Μέθοδοι και Naïve Bayes

Στην εξόρυξη γνώσης μια πιθανοθεωρητική προσέγγιση είναι και το θεώρημα Bayes. Στόχος του είναι η ανεύρεση μιας πιθανής υπόθεσης, από το σύνολο υποθέσεων (Pasquale, et al., 2011).

Σύμφωνα με το θεώρημα Bayes (Yang & Pedersen, 1997):

$$P(h|D) = \frac{P(h)P(D|h)}{P(D)}$$

P(h|D): Η ζητούμενη πιθανότητα, της συγκεκριμένης υπόθεσης h, αν αποδεχτούμε το σύνολο δεδομένων D.

P(h): Η προηγούμενη γνωστή πιθανότητα της υπόθεσης h.

P(D|h): Η πιθανότητα που εκφράζει το ενδεχόμενο παρατήρησης των δεδομένων D, αν αποδεχτούμε την υπόθεση h.

P(D): Η προηγούμενη γνωστή πιθανότητα παρατήρησης των δεδομένων D.

Στον αλγόριθμο Bayes είναι σημαντικά και στατιστικά ανεξάρτητα μεταξύ τους, όλα τα ενδεχόμενα. Οι υπολογισμοί μπορούν να απλοποιηθούν αν απλοποιήσουμε (naïve) την υπόθεση. Η υπόθεση είναι γνωστή και ως «Συνθήκη Ανεξαρτησίας», γιατί υποθέτει ότι αν επιδράσει η τιμή ενός χαρακτηριστικού σε μια κλάση, αυτή θα είναι ανεξάρτητη από τις υπόλοιπες άλλες. Ο αλγόριθμος Naïve Bayes, είναι μια πιθανοθεωρητική προσέγγιση επαγωγικής μάθησης, που ανήκει στην κατηγορία των ταξινομητών Bayesian (Pasquale, et al., 2011).

Η φήμη του αλγορίθμου Naïve Bayes, αλλά και οι επιδόσεις του σε ότι αφορά τις εφαρμογές ταξινόμησης κειμένου, οδήγησαν τους ερευνητές, στο να στραφούν προς αυτόν αλλά και να τον δοκιμάσουν υπό διαφορετικές πειραματικές συνθήκες (Brusilovsky, et al., 2007).

2.4 Αλγόριθμοι Συνεργατικών Φίλτρων (Collaborative Filtering Algorithms)

Στόχος των αλγορίθμων αυτών είναι, να προβλέψουν τη χρησιμότητα των αντικειμένων για έναν συγκεκριμένο χρήστη, βασιζόμενοι πάντα σε μία βάση δεδομένων η οποία περιέχει τις προτιμήσεις μιας ομάδας, ή ενός ευρύτερου συνόλου

χρηστών. Τα συστήματα που κάνουν χρήση των αλγορίθμων συνεργατικών φίλτρων, κατηγοριοποιούνται με βάση το εάν χρησιμοποιούν άμεσες ή έμμεσες ψήφους χρηστών (Resnick & Varian, March 1997):

- Οι άμεσοι ψήφοι, συνήθως, αφορούν στην ξεκάθαρη προτίμηση ενός χρήστη και συνήθως εκδηλώνεται με κάποια βαθμολογική κλίμακα.
- Στην έμμεση, η ψηφοφορία αφορά στην σκιαγράφηση του χαρακτήρα ενός χρήστη ή των διαφόρων επιλογών του. Τα δεδομένα μπορούν να προέλθουν από ιστορικά αναζήτησης, προτιμήσεων ή άλλων τύπων πληροφορίας που έγκεινται σε συγκεκριμένα πρότυπα.

Οι αλγόριθμοι συνεργατικών φίλτρων εστιάζουν κυρίως στα δεδομένα που απουσιάζουν και προσπαθούν να εκμαιεύσουν όλες τις πληροφορίες σχετικά με αυτά (Resnick & Varian, March 1997).

Οι (Adomavicius & Tuzhilin, 2005), αναφέρουν σχετικά ότι, η μέθοδος των συνεργατικών φίλτρων, επεξεργάζεται τα δεδομένα με στόχο να φτιάξει το προφίλ του χρήστη με τις προτιμήσεις του. Βάση αυτού μπορεί να εκτελέσει προβλέψεις για τα πιθανά ενδιαφέροντα, αφού συσχετίσει το προφίλ του, με τα προφίλ άλλων χρηστών. Η μέθοδος αυτή στηρίζεται στην ιδέα ότι, κάποιοι χρήστες που συμφώνησαν στο παρελθόν σε ορισμένα θέματα, πιθανότατα να συμφωνήσουν και στο μέλλον σε κάποια άλλα.

Ο αλγόριθμος των συνεργατικών φίλτρων δημιουργεί ένα πρότυπο αξιολόγησης για τον κάθε χρήστη, που καταγράφει σ'αυτό, τι του άρεσε και τι όχι, από τις διάφορες επιλογές του, π.χ. ποια κείμενα επέλεξε να δεί, ή σε ποια πάτησε κλικ με το ποντίκι του, κλπ. Όταν το σύστημα ενδιαφέρεται να εκτελέσει κάποια πρόβλεψη, τότε συγκρίνει το πρότυπο αξιολόγησης του χρήστη με τα υπόλοιπα των άλλων χρηστών, ψάχνοντας τα όμοια ή τα πιο όμοια. Πολλές φορές εδώ παρατηρείται και το κριτήριο «χρήστες που χρησιμοποίησαν το Α κείμενο χρησιμοποίησαν και το Β επίσης». Έτσι εκτελούνται διάφορες συστάσεις προς τους χρήστες (Yu, et al., 2004).

2.4.1 Χαρακτηριστικά και Προκλήσεις των Συνεργατικών Φίλτρων

Στα συστήματα συνεργατικών φίλτρων, η παραγωγή υψηλής ποιότητας συστάσεων, εξαρτάται από το πόσο καλά αντιμετωπίζουν τις προκλήσεις όπως (Linden, et al., 2003), (Xiaoyuan & Khoshgoftaar, 2009):

- Ανεπάρκεια δεδομένων (Data Sparsity). Στην πραγματικότητα, πολλά συστήματα συστάσεων που χρησιμοποιούνται, αξιολογούν πολύ μεγάλα σύνολα δεδομένων, αλλιώς, ο πίνακας που περιλαμβάνει τους χρήστες και τα στοιχεία που χρησιμοποιούνται για το φιλτράρισμα συνεργασίας, θα είναι ανεπαρκείς και οι επιδόσεις των προβλέψεων θα είναι αμφισβητήσιμες. Το πρόβλημα ανεπάρκειας δεδομένων εμφανίζεται σε διάφορες καταστάσεις, συγκεκριμένα, στην περίπτωση του "cold start", όταν ένα νέος χρήστης ή στοιχείο, μόλις έχει εισέλθει στο σύστημα, και δεν υπάρχει επαρκής πληροφόρηση, τότε, είναι δύσκολο να βρεθούν παρόμοια με αυτό στο σύστημα, έως ότου νέα στοιχεία συστηθούν, από ορισμένους χρήστες που το έχουν αξιολογήσει (Adomavicius & Tuzhilin, 2005), (Yu, et al., 2004).
- Επεκτασιμότητα (Scalability). Όταν οι αριθμοί των χρηστών και αντικειμένων αυξάνονται δραματικά, οι αλγόριθμοι συνεργατικών φίλτρων, υποφέρουν από σοβαρά προβλήματα κλιμάκωσης, με τους υπολογιστικούς πόρους να κινδυνεύουν να ξεπεράσουν τα αποδεκτά επίπεδα .
- Συνωνυμία (Synonymy). Η συνωνυμία αναφέρεται στην τάση ενός αριθμού ίδιων ή πολύ παρόμοιων αντικειμένων να έχουν διαφορετικά ονόματα ή εγγραφές, όπως π.χ. η λέξη "movie" και η λέξη "film", έχουν την ίδια σημασία αλλά είναι διαφορετικές λέξεις.
- Gray Sheep. Το Gray Sheep αναφέρεται στους χρήστες των οποίων οι απόψεις δεν συμφωνούν απόλυτα ή διαφωνούν με την κάθε ομάδα ανθρώπων και ως εκ τούτου δεν επωφελούνται από το φιλτράρισμα συνεργασίας.
- Επιθέσεις Shilling (Shilling Attacks). Ο κάθε χρήστης μπορεί να προσφέρει συστάσεις. Οι άνθρωποι μπορεί να δώσουν θετικές συστάσεις για τα δικά τους αντικείμενα και αρνητικές συστάσεις για τους ανταγωνιστές τους.

Ένα σύστημα συστάσεων που παρέχει γρήγορες και ακριβείς συστάσεις, εκτός ότι προσελκύσει το ενδιαφέρον των χρηστών, δημιουργεί κι ένα δυναμικό εργαλείο στον οργανισμό που το κατέχει. Πίσω όμως από όλα αυτά κρύβεται ένας αλγόριθμος που εκτελείται σε ένα δύσκολο και πολλές φορές ανταγωνιστικό περιβάλλον (Xiaoyuan & Khoshgoftaar, 2009).

Σύμφωνα με τους (Xiaoyuan & Khoshgoftaar, 2009) οι αλγορίθμοι συνεργατικών φίλτρων μπορούν να χωριστούν σε τρεις κατηγορίες, αυτούς που βασίζονται στην

μνήμη, αυτούς που βασίζονται στο μοντέλο και τους υβριδικούς. Στη συνέχεια γίνεται μια ανάλυση των κατηγοριών αυτών.

2.4.2 Αλγόριθμοι που Βασίζονται στη Μνήμη

Γενική κατεύθυνση των αλγορίθμων συνεργατικών φίλτρων είναι να προβλέψουν την προτίμηση ενός συγκεκριμένου χρήστη μέσα από μία βάση δεδομένων. Έτσι λοιπόν η βάση αυτή περιέχει ένα σετ ψήφων $v_{i,j}$, που αντιστοιχούν στην προτίμηση που θα δείξει ο χρήστης i για το αντικείμενο j. Αν I_i είναι το σύνολο των αντικειμένων για τα οποία ο χρήστης i έδειξε προτίμηση τότε η μέση προτίμηση για το χρήστη i ορίζεται ως (Resnick & Varian, March 1997).

$$\overline{Ui} = \frac{1}{|Ij|} \sum_{j \epsilon Ii} Ui, j$$

Οι αλγόριθμοι συνεργατικών φίλτρων που βασίζονται στην μνήμη, χρησιμοποιούν το σύνολο ή ένα δείγμα της βάσης δεδομένων των χρηστών, για να δημιουργήσουν μια πρόβλεψη. Κάθε χρήστης είναι μέρος μιας ομάδας ανθρώπων με παρόμοια ενδιαφέροντα. Με τον προσδιορισμό, των λεγόμενων γειτόνων ενός νέου χρήστη (ή του ενεργού χρήστη), γίνεται μια πρόβλεψη των προτιμήσεων για νέα στοιχεία που μπορούν να παραχθούν (Sarwar, et al., May 2001).

Στους αλγόριθμους συνεργατικών φίλτρων που βασίζονται στη μνήμη, η πρόβλεψη των προτιμήσεων ενός ενεργού χρήστη γίνεται με βάση κάποια πληροφορία που ίσως ήδη υπάρχει, αλλά και ένα σετ βαρών που υπολογίζεται από τη βάση δεδομένων του χρήστη. Έτσι λοιπόν η πρόβλεψη για την προτίμηση του χρήστη ορίζεται ως ένα άθροισμα βαρών των ψήφων των άλλων χρηστών (Resnick & Varian, March 1997).

2.4.2.1 Υπολογισμός Πρόβλεψης και Σύστασης (Prediction & Recommendation)

Το πιο σημαντικό βήμα, σε ένα συνεργατικό σύστημα φιλτραρίσματος είναι να πραγματοποιηθούν προβλέψεις ή συστάσεις. Στον αλγόριθμο που βασίζεται στους γείτονες, θα επιλέγεται, ένα υποσύνολο των πλησιέστερων γειτόνων του ενεργού χρήστη, με βάση την ομοιότητα του με αυτόν, και ένα σταθμισμένο άθροισμα των αξιολογήσεων του, και θα χρησιμοποιείται για την δημιουργία προβλέψεων για τον ενεργό χρήστη (Herlocker, et al., 1999), (Xiaoyuan & Khoshgoftaar, 2009).

2.4.2.2 Διανυσματική Ομοιότητα & Ομοιότητα Υπολογισμού

Στον τομεά της ανάκλησης της πληροφορίας, οι (Breese, et al., October 1998), αναφέρουν ότι για να μετρήσουμε την ομοιότητα ανάμεσα σε δύο κείμενα, θεωρούμε, κάθε κείμενο σαν ένα διάνυσμα των συχνοτήτων των λέξεων, και υπολογίζουμε το συνημίτονο της γωνίας που σχηματίζεται από τα δύο αυτά διανύσματα. Η πιο πάνω μορφή υιοθετείτε και για το συνεργατικό φιλτράρισμα, όπου οι χρήστες παίρνουν το ρόλο των εγγράφων, οι τίτλους παίρνουν το ρόλο των λέξεων, και οι ψήφοι αναλαμβάνουν τον ρόλο των συχνοτήτων των λέξεων (Breese, et al., October 1998).

Η ομοιότητα υπολογισμού των αντικειμένων ή των χρηστών, είναι ένα κρίσιμο βήμα στους αλγορίθμους της συνεργατικής διήθησης που βασίζονται στη μνήμη. Για τα στοιχεία που βασίζονται σε αλγόριθμους συνεργατικών φίλτρων, η βασική ιδέα του υπολογισμού ομοιότητας, μεταξύ του στοιχείου i και του στοιχείου j, είναι, πρώτα να εργαστούν για τους χρήστες που έχουν δύο από αυτά τα στοιχεία, και στη συνέχεια να εφαρμόσουν έναν υπολογισμό ομοιότητας για τον προσδιορισμό της ομοιότητας, μεταξύ των δύο στοιχείων αυτών (Sarwar, et al., May 2001).

2.4.2.3 Συστάσεις Top-N (Top-N Recommendations)

Η Top-N (set of N top-ranked items) σύσταση, χρησιμοποιείται να συστήσει ένα σύνολο καταταγμένων στοιχείων που θα ενδιαφέρουν ένα συγκεκριμένο χρήστη, δηλαδή η σύσταση ενός συνόλου Ν κορυφαίων στοιχείων που θα ενδιέφεραν ένα χρήστη, π.χ. τα τρία πιο καλά βιβλία που θα μπορούσε να διαβάσει ένας χρήστης. Η Top-N τεχνική σύστασης, αναλύει τον πίνακα χρήστης-αντικείμενο για να ανακαλύψει σχέσεις μεταξύ των διαφόρων χρηστών ή αντικειμένων και τις χρησιμοποιεί για να υπολογίσει τις συστάσεις (Xiaoyuan & Khoshgoftaar, 2009).

2.4.2.4 Επεκτάσεις των αλγορίθμων που βασίζονται στη μνήμη

Ο αλγόριθμος της προκαθορισμένης ψήφου, αποτελεί μία επέκταση του αλγορίθμου της συσχέτισης. Βασίστηκε στη παρατήρηση ότι, όταν οι ψήφοι είναι σχετικά λίγες και αφορούν τόσο στον ενεργό χρήστη όσο και στον χρήστη αντιστοίχισης, ο αλγόριθμος της συσχέτισης δεν θα είναι αποδοτικός, καθώς χρησιμοποιεί μόνο την τομή των δύο παραπάνω συνόλων που σχηματίζουν οι δύο χρήστες. Εάν λοιπόν υποθέσουμε ότι

τίθεται μία προκαθορισμένη τιμή για τους τίτλους εκείνους για τους οποίους δεν υπάρχουν άμεσοι ψήφοι, τότε είναι δυνατό να σχηματιστεί η ένωση των δύο συνόλων, εισάγοντας τις προκαθορισμένες για τις ψήφους τιμές, στα κατάλληλα υπό μελέτη αντικείμενα (Resnick & Varian, March 1997).

Η προεπιλεγμένη ψήφος (Default Voting): σε πολλά συνεργατικά φίλτρα, η ομοιότητα κατά ζεύγος, υπολογίζεται μόνο αν η ψηφοφορία στα αντικείμενα έγινε και από τους δύο χρήστες (Sarwar, et al., 2000). Σύμφωνα με τους (Sarwar, et al., 2000) αυτό δεν θα είναι αξιόπιστο όταν υπάρχουν πολύ λίγες ψήφοι για τη δημιουργία τιμών ομοιότητας. Επίσης, με έμφαση στο σημείο της ομοιότητας μπορεί να παραβλευθεί η συνολική συμπεριφορά βαθμολογίας του χρήστη.

Αντίστροφη Συχνότητα Χρήστη: στις εφαρμογές όπου γίνεται χρήση των αλγορίθμων ομοιότητας διανυσμάτων, οι ακολουθίες των λέξεων διαμορφώνονται με βάση την αντίστροφη συχνότητά τους. Βασική ιδέα του αλγορίθμου αυτού είναι, να μειωθούν τα βάρη των σύνηθων λέξεων, καθώς αυτές δεν είναι χρήσιμες, σε ότι αφορά την αναγνώριση ενός θέματος ή ενός κειμένου. Την ίδια στιγμή, οι λέξεις που εμφανίζονται λιγότερο, συχνά αποτελούν σημαντικότερα χαρακτηριστικά για ένα θέμα. Με βάση λοιπόν την παραπάνω ιδέα ορίζεται η αντίστροφη συχνότητα του χρήστη, ως (Xiaoyuan & Khoshgoftaar, 2009):

$$fi = \log n/n_j$$

n_j = είναι ο αριθμός των χρηστών που προτίμησαν τελευταία το θέμα j

n = ο συνολικό αριθμός των χρηστών που καταγράφηκαν στη βάση δεδομένων.

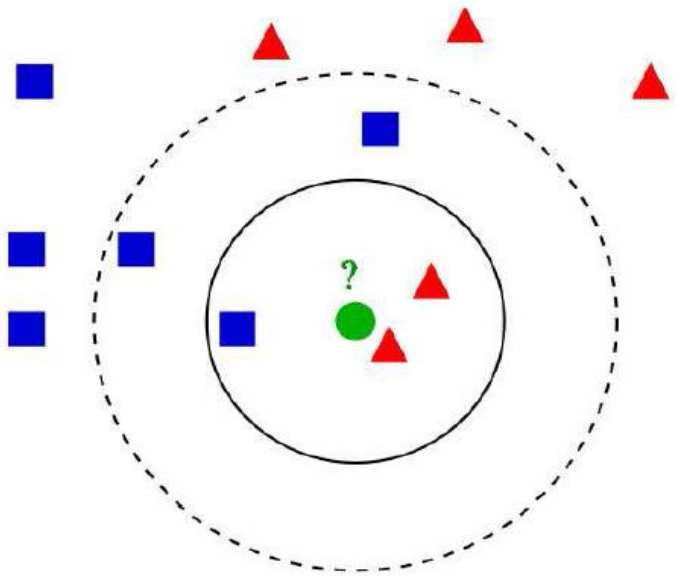
2.4.2.5 Η Μέθοδος του Πλησιέστερου Γείτονα (K-NN)

Οι μεθόδοι μάθησης που βασίζονται σε στιγμιότυπα (instance-based), αποθηκεύουν τα δεδομένα εκπαίδευσης, κάθε φορά που εισέρχετε στο σύστημα ένα νέο στιγμιότυπο για κατάταξη. Το σύνολο από συσχετιζόμενα με αυτό στιγμιότυπα, καλείται από την μνήμη, και ξεκινά η κατάταξη του νέου στιγμιότυπου. Έτσι παρέχεται μια προσέγγιση στην συναρτηση. Το σύνολο των προσεγγίσεων, δίνει πλεονέκτημα στην μέθοδο αυτή γιατί δεν αποτελούν μια πολύπλοκη λύση. Δυστυχώς όμως η ταξινόμηση νέων στιγμιοτύπων έχει υψηλό υπολογιστικό κόστος γιατί οι υπολογισμοί δεν γίνονται στην φάση της

εκπαίδευσης αλλά της ταξινόμησης (ονομάζεται και οκνηρή μάθηση). Ο αλγόριθμος k-Πλησιέστερου Γείτονα (k Nearest Neighbor) ή k-NN, είναι η πιο βασική μέθοδος μάθησης, που στηρίζεται σε στιγμιότυπα. Βασικά η τιμή της συνάρτησης για κάθε νέο στιγμιότυπο βασίζεται στις τιμές των k πλησιέστερων στιγμιοτύπων που συνθέτουν και την γειτονιά του. Κάθε δείγμα X ταξινομείται βάση των k πλησιέστερων γειτόνων του, μεγαλώνοντας σφαιρικά την περιοχή μέχρι να περιλάβει και το k δείγμα εκπαίδευσης (Bobadilla, et al., 2013), (Xiaoyuan & Khoshgoftaar, 2009).

Ο αλγόριθμος k-Πλησιέστερου Γείτονα είναι ο αλγόριθμος αναφοράς για το συνεργατικό φιλτράρισμα (Amatriain, et al., 2011). Στην εκδοχή, χρήστη προς χρήστη, ο k-NN εκτελεί τα ακόλουθα τρία βήματα για να παράγει συστάσεις για τον ενεργό χρήστη (Bobadilla, et al., 2013):

- Προσδιορίζει τους k γείτονες χρηστών (γειτονιά) για τον ενεργό χρήστη A, αφού χρησιμοποιήσει την μέτρηση ομοιότητας (similarity).
- Δημιουργεί μια σύσταση για το αντικείμενο i στον χρήστη A, εφαρμόζοντας μια προσέγγιση συνάθροισης μέσα από με τις βαθμολογίες για την γειτονιά με στοιχεία που δεν αξιολογούνται από τον χρήστη A.
- Εξάγει τις προβλέψεις από το βήμα 2 και στη συνέχεια επιλέγει τις καλύτερες N συστάσεις που παρέχουν την πιο ψηλή ικανοποίηση σύμφωνα με τους περιορισμούς του χρήστη.



Εικόνα 2.5: Παράδειγμα χρήσης αλγόριθμου k-NN, (Wikipedia, 2014)

Στην εικόνα 2.5, δίνεται ένα παράδειγμα αλγορίθμου k-NN. Ο κύκλος στο κέντρο πρέπει να ταξινομηθεί στην κλάση των τριγώνων ή στην κλάση των τετραγώνων. Αν k=3 (κύκλος με συνεχόμενη γραμμή), ανατίθεται (ταξινομείται) στην 2η κλάση γιατί υπάρχουν δύο τρίγωνα και μόνο ένα τετράγωνο μέσα από τον κύκλο. Αν k=5 (κύκλος με διακεκομμένη γραμμή), ανατίθεται στην 1η κλάση (τρία τετράγωνα εναντίον δύο τριγώνων εσσωτερικά του εξωτερικού κύκλου) (Wikipedia, 2014).

2.4.3 Αλγόριθμοι που Βασίζονται σε Μοντέλα

Οι αλγόριθμοι συνεργατικών φίλτρων, μπορούν να θεωρηθούν και ως υπολογισμοί της αναμενόμενης τιμής, μίας ψήφου δεδομένων των όσων γνωρίζουμε έως τώρα για το χρήστη. Για κάθε ενεργό χρήστη, είναι επιθυμητό να προβλεφθούν οι ψήφοι, σε αντικείμενα, τα οποία δεν έχει προσπελάσει ακόμη. Θεωρούμε ότι υπάρχει πιθανότητα, ο ενεργός χρήστης να δείξει μία συγκεκριμένη προτίμηση σε αντικείμενο προηγούμενων του προτιμήσεων (Breese, et al., October 1998).

Ο σχεδιασμός και η ανάπτυξη των προτύπων, όπως η μηχανική μάθηση, οι αλγόριθμοι εξόρυξης δεδομένων κλπ, μπορεί να επιτρέψουν στο σύστημα, να κάνει έξυπνες προβλέψεις, για εργασίες συνεργατικού φιλτραρίσματος, με βάση τα υπάρχοντα μοντέλα (model-based) (Breese, et al., October 1998) .

Πιο κάτω παρουσιάζονται διάφορες τεχνικές φιλτραρίσματος στα συνεργατικά φίλτρα που βασίζονται στο μοντέλο:

- Οι αλγόριθμοι Bayesian: ένας αλγόριθμος Bayesian είναι μια κατευθυνόμενη μη κυκλική γραφική παράσταση (DAG) με μια τριπλέτα <N,A,Θ>, όπου (Pearl, USA, 1988).:
  - ✓ κάθε κόμβος n ∈ N, παριστάνει μία τυχαία μεταβλητή,
  - ✓ κάθε κατευθυνόμενο τόξο α ∈ A μεταξύ των κόμβων είναι η πιθανολογική συσχέτιση μεταξύ των μεταβλητών, και
  - ✓ Θ είναι ένας πίνακας όρων πιθανοτήτων ποσοτικοποίησης του πόσο ένας κόμβος εξαρτάται από τους γονείς του.
- Συσταδοποιημένοι αλγόριθμοι συνεργατικών φίλτρων. (Clustering CF Algorithms). Μια συστάδα είναι μια συλλογή από αντικείμενα δεδομένων που είναι παρόμοια το ένα με το άλλο εντός της ίδιας συστάδας και είναι ανόμοια με τα αντικείμενα άλλων συστάδων (Han & Kamber, USA, 2001).

- Αλγόριθμοι συνεργατικών φίλτρων βασισμένοι στην παλινδρόμηση (Regression-Based CF Algorithms). Για αλγόριθμους συνεργατικών φίλτρων βασισμένους στη μνήμη, σε ορισμένες περιπτώσεις, δύο βαθμολογημένοι πίνακες μπορεί να είναι μακρινοί από την άποψη της Ευκλείδειας απόστασης, αλλά να έχουν πολύ υψηλή ομοιότητα με τη χρήση των μέτρων vector ή Pearson correlation. Οι μέθοδοι παλινδρόμησης που είναι καλοί στο να κάνουν προβλέψεις για τις αριθμητικές τιμές είναι χρήσιμες για την αντιμετώπιση αυτών των προβλημάτων (Xiaoyuan & Khoshgoftaar, 2009).
- MDP-Based Αλγόριθμοι συνεργατικών φίλτρων (MDP-Based CF Algorithms). Αντί να βλέπουν τη διαδικασία σύστασης ως ένα πρόβλημα πρόβλεψης, οι (Shani, et al., 2005) τη θεωρούν ως ένα διαδοχικό πρόβλημα βελτιστοποίησης και χρησιμοποιούν διαδικασίες λήψης μοντέλων Markov (MDPs), για συστήματα Συστάσεων (Shani, et al., 2005).

2.5 Αλγόριθμοι Συστάσεων Ανάλυσης Γράφων (Graph-Based Recom/der Algorithms)

Κάθε αλγόριθμος συνενώνει έναν χρήστη Χ με μία οντότητα Υ, και μπορεί να του προσφέρει μία σύσταση που είτε θα είναι αρεστή σε αυτόν είτε όχι. Με άλλα λόγια, τα κατώφλια του αλγόριθμου που τίθενται, σε ότι αφορά τις συστάσεις, και πάντα για τους αλγόριθμους που εμπεριέχουν γράφους, εμπεριέχουν πάντα μηχανισμούς μεταπηδήσεων. Οι μηχανισμοί των μεταπηδήσεων εστιάζουν περισσότερο στη σύνδεση ανάμεσα στο χρήστη και στο αντικείμενο, και όχι τόσο στην πρόβλεψη. Η φύση των συνδέσεων και των μεταπηδήσεων συντελεί άλλωστε σε σημαντικό βαθμό και στην εξήγηση των συστάσεων. Η θεωρία των γράφων, έχει χρησιμοποιηθεί και σε μαθηματικά μοντέλα, με σκοπό να εξηγήσει τις ιδιότητες των κοινωνικών δικτύων, στα οποία λειτουργούν οι αλγόριθμοι σύστασης, που βασίζονται σε γράφους (Aiello, et al., 2000).

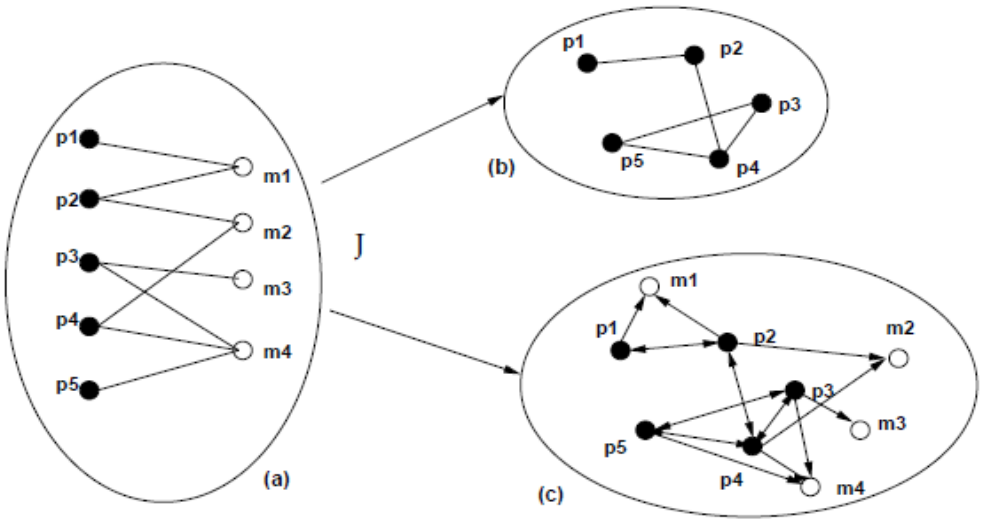
2.5.1 Κατασκευή των Συνδέσεων Μεταπήδησης

Ένα σύνολο δεδομένων σύστασης R, αποτελείται από τις βαθμολογίες που έχει δώσει μία ομάδα ατόμων για ένα συγκεκριμένο αντικείμενο. Το σύνολο δεδομένων αυτό, μπορεί να αναπαρασταθεί με έναν γράφο $G = (P \cup M, E)$ όπου (Mirza, et al., March 2003):

- P είναι το σύνολο των χρηστών,
- M των σύνολο των αντικειμένων,
- E οι αξιολογήσεις των χρηστών για το συγκεκριμένο αντικείμενο.

Το υποσύνολο M λειτουργεί ως ένας τρόπος σύνδεσης ή μεταπηδήσεων ανάμεσα στα μέλη του υποσυνόλου P. Κάθε μεταπήδηση είναι μία συνάρτηση $J: R \rightarrow S, S\underline{C}\ P \times P$ που παίρνει ως είσοδο το σύνολο σύστασης R και επιστρέφει ένα αριθμό ζευγών από στοιχεία του P. Διαισθητικά η επιστροφή αυτή σημαίνει ότι, οι κόμβοι που συμμετέχουν σε κάθε ζευγάρι μπορούν να ανευρεθούν με μία μόνο μεταπήδηση (Basu, et al., 1998).

Εάν λοιπόν ένας κόμβος B μπορεί να προσεγγιστεί από έναν κόμβο A, μόνο με μία μεταπήδηση, αν ένας τρίτος κόμβος C, μπορεί να προσεγγιστεί και αυτός από τον κόμβο B, με μία μόνο μεταπήδηση, τότε απέχει από τον A διάστημα δύο μεταπηδήσεων. Κάθε μεταπήδηση χρησιμοποιείται για τη δημιουργία του γράφου, ο οποίος συχνά ονομάζεται γράφος κοινωνικού δικτύου. Έτσι λοιπόν, ο γράφος κοινωνικού δικτύου, συνόλου δεδομένων R μίας σύστασης που εισάγεται με βάση μία δοθείσα μεταπήδηση J, είναι ένας μη κατευθυνόμενος γράφος $G_S = (P, E_S)$ όπου, οι ακμές δίνονται από το σύνολο $E_S = J(R)$.



Εικόνα 2.6: Παρουσίαση μεταπήδησης: (a) Παράδειγμα διμερούς γραφήματος, μεταξύ ανθρώπων (p) και ταινιών (m). (b) Γράφημα κοινωνικού δικτύου. (c) Συνιστών γράφος (Mirza, et al., March 2003)

Θεωρούμε ένα σύστημα συστάσεων, όπως η αξιοποίηση των κοινωνικών δεσμών (αυτά που ονομάσαμε πιο πάνω μεταπηδήσεις), ότι φέρνει πιο κοντά ένα άτομο με άλλο κόσμο που έχουν αξιολογήσει το ενδιαφέρον του για κάποιο αντικείμενο. Αν για παράδειγμα πάρουμε την περίπτωση ανθρώπων που αξιολογούν ταινίες, εικόνα 2.6, για την μοντελοποίηση αυτού θα δούμε το κοινωνικό δίκτυο των ανθρώπων ως ένα κατευθυνόμενο γράφο, όπου το συντομότερο μονοπάτι από κάποιο άτομο σε μια ταινία, μπορεί να χρησιμοποιηθεί σαν βάση για συστάσεις. Το γράφημα αυτό μπορεί να χαρακτηριστεί σαν συνιστών γράφημα. Η εικόνα 2.6 παρουσιάζει στο (b) ένα γράφημα κοινωνικού δικτύου που προκαλείται από το παράδειγμα στο (α) χρησιμοποιώντας την μεταπήδηση, και στο (c) ο συνιστών γράφος (Mirza, et al., March 2003).

Στην περίπτωση των γράφων συστάσεων, κάθε σύνολο δεδομένων σύστασης, που εισάγεται από μία συνάρτηση μεταπήδησης J, είναι ένας κατευθυνόμενος γράφος $G_r = (P \cup M, E_{sd} \cup E_{md})$ όπου (Basu, et al., 1998):

- το E_{sd} είναι μία ταξινομημένη σειρά ζευγών, που αφορά και στις δύο κατευθύνσεις του J(R),
- ενώ E_{md} είναι μία ταξινομημένη σειρά ζευγών, που αφορά στην κατεύθυνση που δείχνει προς το αντικείμενο.

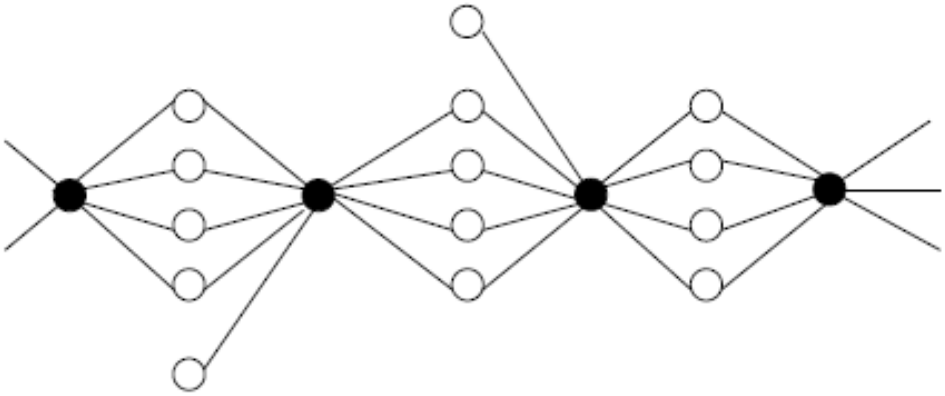
2.5.2 Βασική Προσέγγιση

Η συνάρτηση μεταπήδησης είναι ένας πολύ απλοϊκός τρόπος, για να μοντελοποιηθεί ένας αλγόριθμος σύστασης. Ένας πραγματικός αλγόριθμος σύστασης θα έπρεπε να εμπεριέχει πολλά περισσότερα στοιχεία από ότι η συνένωση κάποιων κόμβων. Οι τρόποι για να συνενωθεί ένας γράφος κοινωνικού δικτύου, με ένα γράφο σύστασης είναι πολλοί. Η βασική προσέγγιση της συνένωσης αυτής αφορά, στην κατανόηση της λειτουργίας ενός ήδη υπάρχοντος αλγορίθμου σύστασης, και της προσθήκης μεταπηδήσεων σε αυτόν, αλλά, με τρόπο αλγοριθμικό και με εργαλεία που θα προέρχονται από τη θεωρία δημιουργίας γράφων (Basu, et al., 1998).

2.5.3 Hammock

Η μεταπήδηση Hammock συνενώνει δύο χρήστες στο υποσύνολο Gs εάν αυτοί έχουν τουλάχιστον w αντικείμενα κοινά στο R (Basu, et al., 1998). Ο αριθμός των w κοινών

στοιχείων ονομάζεται πλάτος Hammock. Στην εικόνα 2.7, παρουσιάζεται μία ακολουθία μονοπατιών Hammock.



Εικόνα 2.7: Ακολουθία μονοπατιών Hammock με πλάτος w=4 (Mirza, et al., March 2003)

Τα Hammocks θεωρούνται ιδιαίτερα σημαντικά και θεμελιώδη για τα συστήματα συστάσεων. Τα πρώτα projects συστάσεων όπως το GroupLens, το Likeminds και το Firey έκαναν χρήση των Hammock μεταπηδήσεων, καταλήγοντας στον πιο άμεσο χρήστη (Kleinberg, August 2000).

Η χρήση των Hammock εγείρει έναν αριθμό αλγοριθμικών ερωτημάτων. Αφενός εάν η χρήση ενός ευρύτερου Hammock, με άλλα λόγια η χρήση των περισσότερων κοινών αναζητήσεων, θα ήταν αποτελεσματικότερη, και αφ ετέρου εάν τα συστήματα συστάσεων απαιτούν έναν ελάχιστο αριθμό ψήφων, πριν ο χρήστης να μπορέσει να χρησιμοποιήσει το σύστημα (Mirza, et al., March 2003).

2.5.4 Μοντέλα Τυχαίων Γράφων

Τα μοντέλα τυχαίων γράφων αν και παρουσιάζουν περιορισμούς, θεωρούνται τα καταλληλότερα από τα ήδη υπάρχοντα, καθώς προσφέρουν ιδιαίτερα περιγραφικά αποτελέσματα. Ένα σύνολο δεδομένων σύστασης R, χαρακτηρίζεται από το σύνολο των προτιμήσεων που εκδηλώνει κάθε χρήστης και για ένα αντικείμενο. Οι τιμές αυτές είναι εύκολο να εκμαιευθούν και άρα να χρησιμοποιηθούν, σε ότι αφορά την ανάλυση των αλγορίθμων σύστασης. Με βάση την προσέγγιση αυτή οι περισσότεροι κοινοί γράφοι δεν είναι κατάλληλοι για τα συστήματα συστάσεων, καθώς υποθέτουν ότι, όλες οι ακμές έχουν την ίδια πιθανότητα (Lu, 2000).

Το μόνο γνωστό μοντέλο που έχει καταφέρει να αντιπαρέλθει του προβλήματος αυτού είναι το μοντέλο Newman – Strogatz – Watts το οποίο χαρακτηρίζει μία οικογένεια γράφων με φάση τις κατανομές της (Lu, 2000).

2.5.5 Μοντελοποίηση των Αλγορίθμων Σύστασης

Ο γράφος σύστασης $G_r = (P \cup M, E_{sd} \cup E_{md})$ είναι κατευθυνόμενος και με βάση το μοντέλο Newman – Strogatz – Watts ορίζεται από τη συνάρτηση (Mirza, et al., March 2003):

$$G(x, y) = \sum_{j=0,k=0}^{j=\infty,k=\infty} Pjkx^j\, y^k \quad \text{όπου:}$$

- Pjk είναι η πιθανότητα μία τυχαία επιλεγμένη κορυφή να έχει εσωτερικό βαθμό j και εξωτερικό βαθμό k.

Η κατασκευή των συνδέσεων μας δίνει την πληροφορία ότι οι ακμές έχουν εξωτερικούς βαθμούς 0, ενώ ο μέσος αριθμός των τόξων που εισάγονται ή αποχωρούν από μία ακμή είναι 0. Έτσι (Mirza, et al., March 2003):

$$\sum_{jk} (j - k)pjk = \sum_{jk} (k - j)pjk = 0$$

Ο μέσος όρος μήκους του μονοπατιού μπορεί να υπολογιστεί ως:

$$lr = \frac{\log[(Np + Nm - 1)(z2 - z1) + z1^2] - \log[z1^2]}{\log[\frac{z2}{z1}]}$$

Όπου $Np + Nm$ είναι το μέγεθος του γράφου σύστασης με Nm αντικείμενα. Το μήκος lr περιλαμβάνει τόσο μονοπάτια από ανθρώπους προς αντικείμενα όσο και μονοπάτια από ανθρώπους προς ανθρώπους.

Το μέσο μήκος που αφορά μονοπάτια μόνο από ανθρώπους προς αντικείμενα μπορεί να υπολογιστεί ως:

$$lpm = \frac{(lr(Np(Np - 1) + NpNm) - lppNp(Np - 1))}{NpNm}$$

2.5.6 Προβλήματα των Συναρτήσεων Newman – Strogatz – Watts

Στην πραγματικότητα δημιουργούνται αρκετά προβλήματα από τη χρήση των παραπάνω συναρτήσεων. Πρώτον, και σε αντίθεση με τα αποτελέσματα που αφορούν στη θεωρία των τυχαίων γράφων, οι παραπάνω εξισώσεις δεν εμπεριέχουν εγγυήσεις ή επίπεδα εμπιστοσύνης (Kleinberg, August 2000). Ακόμη, αφορούν στο σύνολο των τυχαίων γράφων και έχουν μία συγκεκριμένη κατανομή, γεγονός που σημαίνει ότι θεωρούν όλους τους γράφους ισοπίθανους (Mirza, et al., March 2003).

Επιπρόσθετα, η χρήση των N_P και N_M υποθέτει ότι όλες οι ακμές είναι προσβάσιμες από οποιαδήποτε εναρκτήρια ακμή, κριτήριο το οποίο δεν ικανοποιείται πάντα. Το μοντέλο Newman – Strogatz – Watts είναι αρκετά πιο περίπλοκο από τα παραδοσιακά μοντέλα των γράφων καθώς εισάγει έναν περίπου άπειρο αριθμό παραμέτρων ενώ ταυτόχρονα υποθέτει ότι ισχύει η ίδια κατανομή για όλους τους γράφους και όλα τα μεγέθη γράφων (Basu, et al., 1998).

Κεφάλαιο 3
Σύνολα Δεδομένων-Datasets

Σε διάφορους τομείς της έρευνας, η διαθεσιμότητα των ανοικτών συνόλων δεδομένων, θεωρείται ως κλειδί για τους σκοπούς της έρευνας. Μια κοινή πρακτική στον κόσμο των συστημάτων συστάσεων, είναι, να χρησιμοποιούν τα δημόσια διαθέσιμα σύνολα δεδομένων όπως π.χ. MovieLens, Book-Crossing, ή EachMovie, προκειμένου να αξιολογήσουν διάφορους αλγορίθμους συστάσεων (Verbert, et al., 2011). Αυτά τα σύνολα δεδομένων χρησιμοποιούνται ως σημείο αναφοράς για την ανάπτυξη αλγορίθμων συστάσεων και τη σύγκρισή τους με άλλους αλγορίθμους. Στην περίπτωση που αφορά εκπαιδευτικά σύνολα δεδομένων, η εύρεση αυτών των διαθέσιμων συνόλων δεδομένων για πειραματισμό μπορεί να είναι ένα δύσκολο έργο, καθώς υπάρχουν διάφορες πηγές δεδομένων που δεν έχουν προσδιοριστεί ή δεν έχουν μελετηθεί εις βάθος (Verbert, et al., 2011).

3.1 Τα Σύνολα Δεδομένων (Datasets) της TEL

Όπως αναφέρθηκε και πιο πάνω, σημαντική προϋπόθεση για να διευκολυνθεί η έρευνα για τις τεχνολογίες σύστασης, είναι η ύπαρξη επαρκών στοιχείων από διάφορες δραστηριότητες του συστήματος και των αλληλεπιδράσεων του με τους χρήστες. Όταν η ανάλυση πραγματοποιείται για ερευνητικούς σκοπούς, και με διερευνητικό τρόπο, είναι εξίσου σημαντικό, να παράσχει στους ερευνητές επαρκή στοιχεία που προέρχονται από ένα πραγματικό ή προσομοιωμένο περιβάλλον, του στοχευόμενου τομέα. Σε έναν αυξανόμενο αριθμό επιστημονικών κλάδων, οι μεγάλες συλλογές δεδομένων αναδεικνύονται ως σημαντική κοινότητα πόρων (Chervenak, et al., 2000). Αυτά τα σύνολα δεδομένων χρησιμοποιούνται ως σημείο αναφοράς για την ανάπτυξη νέων αλγορίθμων και τη σύγκρισή τους με άλλους αλγορίθμους σε δεδομένες ρυθμίσεις. Σε σύνολα δεδομένων που χρησιμοποιούνται για τις συστάσεις αλγορίθμων, τα εν λόγω δεδομένα μπορούν για παράδειγμα να είναι ρητά (τα ratings) ή έμμεσα (τα downloads και οι ετικέτες) των δεικτών ενδιαφέροντος. Οι δείκτες αυτοί, για παράδειγμα, χρησιμοποιούνται για να βρεθούν οι χρήστες με παρόμοια ενδιαφέροντα, ως βάση για να προτείνουν στοιχεία σε ένα χρήστη (Manouselis, et al., 2010).

Το 2011, η κοινότητα συστάσεων της TEL εξακολουθούσε να εργάζεται με μικρά σύνολα δεδομένων, τα οποία δεν έγιναν διαθέσιμα στο ευρύ κοινό (Manouselis, et al., 2010). Εκείνη την εποχή, μια ειδική ομάδα του Ευρωπαϊκού δικτύου της STELLAR που ονομάζεται DATATEL (Drachsler, et al., 2010) ξεκίνησε μια πιο δομημένη ανάλυση των θεμάτων, γύρω από την ανάπτυξη, τη διανομή και τη χρήση των συνόλων δεδομένων TEL, για σχετική έρευνα, η οποία οργάνωσε την πρώτη DATATEL πρόσκληση. Μια κλήση για τα σύνολα δεδομένων TEL που καλούσε τις ερευνητικές ομάδες να υποβάλουν τα υπάρχοντα σύνολα δεδομένων από τις εφαρμογές TEL που μπορούν να χρησιμοποιηθούν ως είσοδοι στα συστήματα συστάσεων TEL (Drachsler, et al., 2010).

Τα σύνολα δεδομένων της TEL έχουν πολλές πτυχές καθώς η TEL λαμβάνει χώρα σε όλο το φάσμα της μαθησιακής διαδικασίας και έτσι δε μπορεί σε καμία περίπτωση να διαχωριστεί εάν αυτή αφορά σε τυπικές και άτυπες διαδικασίες. Και οι δύο διαδικασίες (τυπικές και άτυπες) αφορούν σε ένα διαφορετικό πλαίσιο, το οποίο πρέπει να ληφθεί υπόψιν από τα συστήματα σύστασης, με σκοπό να προσωποποιήσει την πληροφορία και να την προσφέρει στους μαθητές (Drachsler, et al., 2010).

Η τυπική εκπαίδευση, συνήθως είναι οργανωμένη σύμφωνα με κάποιο πρόγραμμα και συνήθως υφίσταται σε περιβάλλοντα κατευθυνόμενα από δασκάλους, με

αλληλεπιδράσεις ανθρώπων με ανθρώπους. Αντίθετα, στη *άτυπη εκπαίσευση*, η διαδικασία λαμβάνει χώρα σε διάφορες φάσεις, κατά τη διάρκεια ζωής των μαθητών, οι οποίοι δεν συμμετέχουν σε κανένα επίσημο πλαίσιο μάθησης. Οι μαθητές αυτοί κυρίως κατευθύνονται από τις προσωπικές τους ανάγκες, και είναι υπεύθυνοι για τη μάθησή τους, αλλά και για τα μονοπάτια τα οποία θα ακολουθήσουν. Το εκπαιδευτικό περιεχόμενο της άτυπης εκπαίδευσης, γίνεται ολοένα και πιο προσβάσιμο, διαμέσου πολλών και διαφορετικών πηγών του Web 2.0, όπως τα blogs ή οι πλατφόρμες διαμοιρασμού των αρχείων κλπ (Smyth, 2007).

Μέχρι στιγμής, είναι έντονη η ανάγκη της συλλογής των διαθέσιμων έως τώρα συνόλων δεδομένων, με σκοπό αυτά να χρησιμοποιηθούν στη διαδικασία της μάθησης. Μία τέτοια συλλογή, θα βοηθούσε ιδιαίτερα τους ερευνητές, ως προς το να συλλέξουν έγκυρη και εμπεριστατωμένη γνώση, σχετικά με το πώς κάποιοι συγκεκριμένοι αλγόριθμοι που χρησιμοποιούνται στα συστήματα συστάσεων, θα μπορούσαν να χρησιμοποιηθούν στην εκπαιδευτική διαδικασία (Drachsler, et al., 2010).

Για τη συλλογή των συνόλων δεδομένων TEL, ξεκίνησε πρώτη η DATATEL Challenge[6] ως μέρος του Workshop για Συστήματα Συστάσεων της TEL (Manouselis, et al., 2010). Παρόμοιο έργο εκτελείται στο Πίτσμπουργκ από το Learning Center (Δημόσιο Κέντρο Ξένων Γλωσσών). Το Δημόσιο Κέντρο Ξένων Γλωσσών Datashop[7] είναι μια αποθήκη δεδομένων που παρέχει πρόσβαση σε ένα μεγάλο αριθμό εκπαιδευτικών συνόλων δεδομένων που προέρχονται από τα ευφυή συστήματα διδασκαλίας (Manouselis, et al., 2010).

Το έργο Mulce[8] αφορά, επίσης, η συλλογή και ανταλλαγή δεδομένων αλληλεπίδρασης των εκπαιδευομένων. Η πλατφόρμα είναι διαθέσιμη στους χρήστες για να μοιραστούν, να περιηγηθούν και να αναλύσουν κοινά σύνολα δεδομένων. Κατά τη στιγμή της γραφής, 34 σύνολα δεδομένων είναι διαθέσιμα στην πύλη, συμπεριλαμβανομένου ενός συνόλου δεδομένων του έργου Virtual Math. Αυτά τα δεδομένα έχουν χρησιμοποιηθεί εκτενώς από την κοινότητα Computer Supported Collaborative Learning (CSCL) (Reffay & Betbeder, 2009).

Ένας από τους καταλύτες στην ώθηση των συστημάτων συστάσεων σε διάφορους τομείς ήταν η ύπαρξη διαθέσιμων στο κοινό δεδομένων, όπου οι σχεδιαστές και

[6] http://adenu.ia.uned.es/workshops/recsystel2010/datatel.htm
[7] http://www.learnlab.org/technologies/datashop/
[8] http://mulce-pf.univ-fcomte.fr/PlateFormeMulce/

προγραμματιστές μπορούν να χρησιμοποιούν για να δοκιμάζουν και να συγκρίνουν τις προσεγγίσεις τους. Επιπλέον, πολλές από τις προκλήσεις των δεδομένων συχνά προσελκύουν ερευνητές για συγκεκριμένα θέματα και εφαρμογές (Said, et al., New York, 2011).

3.2 Άξονες Δημιουργίας Κατάλληλων Συνόλων Δεδομένων

Η ερευνητική ομάδα GroupLens τόνισε ιδιαίτερα, τη δημιουργία, αλλά και τη χρήση, συνόλων δεδομένων, για την εξέταση της αποτελεσματικότητας κάποιων αλγορίθμων. Ιδιαίτερη έμφαση δόθηκε στις δυνατότητες τις οποίες θα πρέπει να έχει ο αλγόριθμος σύστασης, ο οποίος θα πρέπει να είναι σχεδιασμένος, ώστε να καλύπτει όλες τις λειτουργίες του συστήματος από το οποίο συγκεντρώθηκαν τα δεδομένα (Manouselis, et al., 2010).

Ένα παράδειγμα αυτού, είναι το σύστημα σύστασης MovieLens, το οποίο προσφέρει στο χρήστη τη δυνατότητα ανεύρεσης των «βέλτιστων αντικειμένων» το οποίο σημαίνει ότι, ως συστάσεις εμφανίζει στο χρήστη τα αντικείμενα εκείνα, τα οποία είναι ιδιαίτερα δημοφιλή λόγω της βαθμολογίας που έχουν λάβει. Απόρροια αυτού είναι, το σύνολο δεδομένων MovieLens, να έχει χαμηλότερες βαθμολογίες για όχι και τόσο διάσημες ταινίες (Manouselis, et al., 2010).

Απαιτούνται σύνολα δεδομένων, τα οποία θα περιέχουν ρεαλιστικές αναπαραστάσεις του συστήματος σύστασης, αλλά και όλες εκείνες τις μαθησιακές πληροφορίες, που χρειάζονται με σκοπό να αξιολογηθούν οι αλγορίθμοι σύστασης στις συγκεκριμένες περιπτώσεις. Παρακάτω περιγράφονται τρεις κύριοι άξονες κατάλληλοι, για τη δημιουργία συνόλων δεδομένων, για συστήματα σύστασης της TEL. Οι άξονες αυτοί μπορούν να βοηθήσουν, τόσο τους προγραμματιστές, όσο και τους αναλυτές συστημάτων, ώστε να δημιουργήσουν το πλέον κατάλληλο σύνολο δεδομένων, το οποίο μάλιστα θα είναι χρησιμοποιήσιμο και από άλλους ερευνητές (Manouselis, et al., 2010):

- *Δημιουργία ενός συνόλου δεδομένων το οποίο αντικατοπτρίζει ρεαλιστικά τις μεταβλητές της διαδικασίας της μάθησης.*

  Είναι σημαντικό να χρησιμοποιηθεί ένα σύνολο δεδομένων το οποίο θα είναι ρεαλιστικό αλλά και αντιπροσωπευτικό του στόχου των ρυθμίσεων που αφορούν στη μάθηση. Ένα αποτελεσματικό σύνολο δεδομένων συμπεριλαμβάνει και

αποθηκεύει πληροφορίες, που ανταποκρίνονται στο πραγματικό περιεχόμενο και την αλληλεπίδραση του χρήστη με μία μαθησιακή διαδικασία.

- *Χρήση ενός σημαντικά μεγάλου αριθμού προφίλ χρηστών.*
  Η επιτυχία των συστημάτων σύστασης συχνά οφείλεται στη διαθεσιμότητα ενός μεγάλου αριθμού προφίλ χρηστών, το οποίο συνεπακόλουθα παρέχει έναν μεγάλο αριθμό αξιολογήσεων. Εάν λοιπόν τα πραγματικά δεδομένα δεν είναι διαθέσιμα, καλό είναι να γίνει σοβαρή προσπάθεια ώστε να χρησιμοποιηθούν δεδομένα από παρόμοιες εφαρμογές ή προσομοιωμένα δεδομένα.

- *Δημιουργία συνόλων δεδομένων τέτοιων που να είναι συγκρίσιμα με άλλα.*
  Η χρήση ενός συνόλου δεδομένων από κάποιο συγκεκριμένο project με TEL, δεν σημαίνει ότι μπορεί να προσφέρει μία γενικότερη γνώση, σε ότι αφορά τη μαθησιακή διαδικασία. Έτσι, είναι λογικό να απαιτείται η δημιουργία συνόλου δεδομένων τέτοιων, τα οποία, να είναι δομημένα με τρόπο παρόμοιο με τα υπόλοιπα σύνολα δεδομένων, ειδικά σε ότι αφορά τις ίδιες μαθησιακές διαδικασίες. Μία παρόμοια δομή, διευκολύνει και τους υπόλοιπους ερευνητές ώστε να δοκιμάσουν τα συστήματα σύστασης που έχουν προτείνει, αλλά και τους αλγορίθμους τους στα ίδια δεδομένα.

Οι οδηγίες αυτές δε θα πρέπει να λαμβάνονται ως περιορισμοί, αλλά πολύ περισσότερο θα πρέπει να ενθαρρύνουν τους ερευνητές, ώστε να επιλέγουν τα πλέον κατάλληλα σύνολα δεδομένων ακόμη και αν αυτά δεν έχουν χρησιμοποιηθεί ποτέ σε συστήματα συστάσεων. Η μάθηση αποτελεί μία ιδιαίτερα προσωπική και κατευθυνόμενη από τον ίδιο τον άνθρωπο διαδικασία. Τα διάφορα δημοσιευμένα σύνολα δεδομένων, θα συντελούσαν στην αξιολόγηση της ποιότητας, αλλά και του επιπέδου υποστήριξης τους, στα συστήματα υποστήριξης των μαθητών.

3.3 Επιλογή συνόλων δεδομένων για αξιολόγηση

Πολλές βασικές αποφάσεις σχετικά με τα σύνολα δεδομένων διέπουν την επιτυχή αξιολόγηση ενός αλγόριθμου συστήματος συστάσεων. Έτσι γεννιούνται τα πιο κάτω ερωτήματα (Herlocker, et al., 2004):

- Μπορεί, η αξιολόγηση να πραγματοποιηθεί χωρίς σύνδεση σε υπάρχοντα δεδομένα που δεν απαιτούν live δοκιμές του χρήστη;

- Εάν ένα σύνολο δεδομένων δεν είναι επί του παρόντος διαθέσιμο, μπορεί να πραγματοποιηθεί αξιολόγηση σε προσομοιωμένα δεδομένα;
- Τι ιδιότητες πρέπει να έχει το σύνολο δεδομένων, προκειμένου να διαμορφώσει καλύτερα τα καθήκοντα για τα οποία αξιολογείται η σύσταση;

Μερικά παραδείγματα συμβάλουν στην αποσαφήνιση αυτών των αποφάσεων (Herlocker, et al., 2004), (Linton, et al., 1998):

- Όταν σχεδιάζεται ένας αλγόριθμος συστάσεων είναι για να συστήνει εντολές επεξεργασίας κειμένου. Κάποιος αναμένει, 5-10% των χρηστών, ή περισσότερο, να έχουν βιώσει εμπειρία/ιες, για να αξιολογήσουν. Κατά συνέπεια, δεν θα ήταν σοφό, η επιλογή αλγορίθμων συστάσεων, με βάση τα αποτελέσματα της αξιολόγησης (αξιολογήσεις αραιότητας - sparsity rating), να είναι πολύ λιγότερες από 5-10% .
- Κατά την αξιολόγηση ένας αλγόριθμος συστάσεων, όσον αφορά το πλαίσιο, «ανεύρεση χρήσιμων αντικειμένων», για σύσταση καινούριων στοιχείων, μπορεί να είναι σκόπιμο να χρησιμοποιηθούν μόνο offline αξιολογήσεις. Δεδομένου ότι ο αλγόριθμος συστάσεων, παράγει συστάσεις για τα στοιχεία που ο χρήστης δεν γνωρίζει, είναι πιθανό το σύνολο των δεδομένων να μην παρέχει αρκετές πληροφορίες για την αξιολόγηση της ποιότητας των αντικειμένων που συνίστανται. Εάν ένα στοιχείο ήταν πραγματικά άγνωστο στο χρήστη, τότε είναι πιθανό να μην υπάρχει βαθμολογία για αυτό από τον χρήστη στη βάση δεδομένων.
- Όταν αξιολογείται ένας αλγόριθμος συστάσεων, σε ένα νέο τομέα, όπου υπάρχει σημαντική έρευνα σχετικά με τη δομή των προτιμήσεων των χρηστών, ίσως είναι σκόπιμο να αναγνωρίζονται τα χαρακτηριστικά των συνθετικών συνόλων δεδομένων για περεταίρω μελέτη.

3.4 Ιδιότητες συνόλων δεδομένων

Το τελικό ερώτημα σε αυτή την ενότητα για τα σύνολα δεδομένων είναι: *"Τι ιδιότητες πρέπει να έχει το σύνολο δεδομένων, προκειμένου να γίνει καλύτερο για μια αξιολόγηση σύστασης;"* Είναι, λοιπόν, χρήσιμο να γίνει διαίρεση των ιδιοτήτων του συνόλου δεδομένων σε τρεις κατηγορίες (Herlocker, et al., 2004):

- Τα χαρακτηριστικά τομέα (Domain features), αντανακλούν τη φύση του περιεχομένου που συνιστάται, παρά οποιοδήποτε άλλο συγκεκριμένο σύστημα.
- Τα έμφυτα χαρακτηριστικά (Inherent features) αντανακλούν την φύση του συγκεκριμένου συστήματος συστάσεων από το οποίο προήλθαν τα δεδομένα (και ενδεχομένως, από τις πρακτικές συλλογής δεδομένων).
- Τα χαρακτηριστικά Δείγματος (Sample features), αντανακλούν τη διανομή ιδιοτήτων των στοιχείων, και συχνά μπορούν να χειριστούν με την επιλογή του κατάλληλου υποσύνολου ένα μεγαλύτερο σετ δεδομένων.

Τα χαρακτηριστικά του τομέα ενδιαφέροντος περιλαμβάνουν (Herlocker, et al., 2004):

- το θέμα του περιεχομένου που συνιστάται,
- τα καθήκοντα των χρηστών που υποστηρίζονται από τη σύσταση,
- την καινοτομία και την ποιότητα που χρειάζονται,
- τη σχέση κόστους / οφέλους,
- το διαχωρισμό των προτιμήσεων του χρήστη.

Τα έμφυτα χαρακτηριστικά περιλαμβάνουν πολλά χαρακτηριστικά σχετικά με τις αξιολογήσεις (Herlocker, et al., 2004):

- αν οι αξιολογήσεις είναι ρητές, έμμεσες, ή και τα δύο
- η έκταση στην οποία υπάρχουν τα στοιχεία
- οι διαστάσεις της αξιολόγησης
- η παρουσία ή η απουσία της χρονικής σήμανσης στις αξιολογήσεις.

Άλλα έμφυτα χαρακτηριστικά αφορούν τις πρακτικές συλλογής δεδομένων :

- αν καταγράφηκαν οι συστάσεις που εμφανίζονται στο χρήστη
- η διαθεσιμότητα των δημογραφικών πληροφοριών του χρήστη ή του περιεχομένου των πληροφοριών.
- οι προκαταλήψεις των χρηστών, που εμπλέκονται στη συλλογή δεδομένων.

Τα χαρακτηριστικά δείγματος ενός συνόλου δεδομένων περιλαμβάνουν πολλές από τις στατιστικές ιδιότητες που συνήθως θεωρούνται κατά την αξιολόγηση του (Miller, et al., 2004):

- η πυκνότητα των αξιολογήσεων συνολικά, η οποία μερικές φορές θεωρείται ως ο μέσος όρος του ποσοστού που έχουν αξιολογηθεί ανά χρήστη.
- ο αριθμός ή η πυκνότητα των αξιολογήσεων από τους χρήστες για τους οποίους γίνονται οι συστάσεις, αντιπροσωπεύει την εμπειρία του χρήστη στο σύστημα.

- το μέγεθος και η κατανομή ιδιοτήτων των δεδομένων. Ορισμένα σύνολα δεδομένων έχουν πολύ περισσότερα αντικείμενα παρά χρήστες, ενώ το πιο συνηθισμένο είναι οι χρήστες να είναι κατά πολύ περισσότεροι από τα αντικείμενα.

3.4.1 Παλιές και Σύγχρονες Τάσεις στα Σύνολα

Τα πιο ευρέως χρησιμοποιούμενα σύνολα δεδομένων ήταν τα EachMovie σύνολα δεδομένων (http://research.compaq.com/SRC/eachmovie/). Είναι εκτεταμένα σύνολα με πάνω από 2,8 εκατομμύρια αξιολογήσεις από 70.000 χρήστες, και περιλαμβάνουν πληροφορίες όπως οι χρονοσημάνσεις (timestamps) και τα βασικά δημογραφικά δεδομένα, για ορισμένους από τους χρήστες. Το σύνολο δεδομένων *EachMovie* (http://www.movielens.org), χρησιμοποιήθηκε σε δεκάδες έργα μηχανικής μάθησης και αλγοριθμικής έρευνας, για τη μελέτη νέων και καλύτερων τρόπων αξιολόγησης των χρηστών. Παραδείγματα τέτοιων μελετών αποτελούν τα εξής (Herlocker, et al., 2004):

- Canny, J. 2002. *"Collaborative filtering with privacy via factor analysis".* In Proceedings of the 25th Annual International ACMSIGIR Conference on Research and Development in Information retrieval. ACM, New York, 238–245.
- Domingos, P. and Richardson, M. 2003. "*Mining the network value of customers".* In Proceedings of the 7th International Conference on Knowledge Discovery and Data Mining. ACM, New York, 57–66.
- Sarwar, B. M., Karypis, G., Konstan, J. A., And Riedl, J. 2001. *"Item-based collaborative filtering recommendation algorithms".* In Proceedings of the 10th International World Wide Web Conference (WWW10).
- Reddy, P. K., Kitsuregawa, P., Sreekanth, P., And Rao, S. S. 2002. *"A graph based approach to extract a neighborhood customer community for collaborative filtering".* In Databases in Networked Information Systems, Second InternationalWorkshop. Lecture Notes in Computer Science Springer-Verlag, New York, 188–200.

Αρκετοί ερευνητές έχουν χρησιμοποιήσει το σύνολο δεδομένων Jester[9]. Περιλαμβάνει 4,1 εκατομμύρια συνεχείς αξιολογήσεις με κλίμακα από το -10,00 μέχρι και το 10,00, για 100 ανέκδοτα. Οι αξιολογήσεις έγιναν από 73.421 χρήστες. Τα δεδομένα έχουν αξιολογηθεί μεταξύ Απριλίου 1999 και Μαϊου 2003. Τα δεδομένα είναι ελεύθερα

[9] http://grouplens.org/datasets/jester/

διαθέσιμα προς χρήση έρευνας. Τα αρχεία δεδομένων είναι σε μορφή .zip, όταν αποσυμπιεστεί, είναι αρχεία Excel (.xls). Οι αξιολογήσεις είναι πραγματικές τιμές που κυμαίνονται από -10,00 μέχρι και το 10,00 (η τιμή "99" αντιστοιχεί στο "null" = "δεν έχει αξιολογηθεί"). Η πρώτη στήλη αναφέρει τον αριθμό κάθε ανέκδοτου, που βαθμολογεί ο χρήστης. Οι επόμενες 100 στήλες δίνουν τις αξιολογήσεις για τα 100 ανέκδοτα (Goldberg, et al., 2001).

3.5 Γενική Προσέγγιση για τη Δημιουργία Διαμοιραζόμενων Συνόλων Δεδομένων

Το πέρασμα των χρόνων και η τεχνογνωσία έδωσαν σημαντική ώθηση, στη δημιουργία και τη χρήση αποθετηρίων για τα σύνολα δεδομένων, με σκοπό αυτά να χρησιμοποιηθούν, για την επιστημονική κοινότητα. Παρακάτω αναλύονται οι τρεις διαφορετικές φάσεις με βάση τις οποίες διασφαλίζεται η ποιότητα ενός συνόλου δεδομένων και είναι (Drachsler, et al., 2010):

- *Η συλλογή δεδομένων,*
- *Η επεξεργασία δεδομένων,*
- *Ο διαμοιρασμός.*

Η φάση της συλλογής των δεδομένων, συχνά εντοπίζεται στη συλλογή δεδομένων Web 2.0 από μία ιστοσελίδα, με σκοπό αυτά να χρησιμοποιηθούν σε πειράματα που αφορούν στη δημιουργία συστημάτων σύστασης. Η πρώτη απόφαση που πρέπει να ληφθεί, αφορά στις πηγές, οι οποίες θα πρέπει να επιλεγούν αλλά και να χρησιμοποιηθούν. Αυτό εξαρτάται κυρίως από το σκοπό για τον οποίο θα χρησιμοποιηθεί το σύστημα σύστασης, αλλά και από το πεδίο εφαρμογής του, αν και ακόμη και οι ιστοσελίδες που αφορούν στον ίδιο τομέα μπορούν να προσφέρουν διαφορετική χρησιμότητα αλλά και διαφορετικές πληροφορίες. Έτσι λοιπόν, είναι ιδιαίτερα σημαντικό να γίνεται προσεκτική ανάλυση όλων των διαθέσιμων ιστοσελίδων με σκοπό να καθοριστεί η διαθέσιμη πληροφορία, εάν και κατά πόσο αυτή είναι χρήσιμη, και εάν πρέπει να χρησιμοποιηθεί, με σκοπό την παραγωγή των συστάσεων (Drachsler, et al., 2010).

Οι (Sarwar, et al., 2000) αναφέρουν ότι, οι διάφορες Web 2.0 ιστοσελίδες, μπορεί να διαφέρουν κατά πολύ τόσο στο μέγεθος, όσο και στα χαρακτηριστικά της βάσης χρηστών που έχουν, αλλά και στα δημογραφικά χαρακτηριστικά τους. Συνήθως τείνουν να προσελκύουν νεότερους αλλά και γνώστες της τεχνολογίας ανθρώπους σε αντίθεση

με τα σύνολα δεδομένων της απομακρυσμένης μάθησης που τείνουν να προσελκύουν μεγαλύτερους και χωρίς τεχνογνωσία ανθρώπους. Ακόμη, διαφορετικές ιστοσελίδες μπορεί να προσελκύουν και διαφορετικές κουλτούρες ανθρώπων (Sarwar, et al., 2000).

Τα χαρακτηριστικά αυτά συνήθως δεν αποτελούν κοινή γνώση και πολλές φορές μπορεί να μην συμπεριλαμβάνονται στις ερευνητικές εργασίες, μπορούν όμως να επηρεάσουν με πολλούς και δυνατούς τρόπος τα αποτελέσματα των πειραμάτων που αφορούν σε συστήματα σύστασης, και για αυτό είναι ιδιαίτερα σημαντικό να λαμβάνονται υπόψιν. Μάλιστα συνίσταται η συλλογή πολλαπλών συνόλων δεδομένων από διαφορετικές πηγές, ώστε να βελτιωθεί η γενικοποίηση των αποτελεσμάτων.

Αφού λοιπόν θα έχουν ληφθεί όλες οι παραπάνω αποφάσεις, απαιτείται η διασφάλιση της αντιπροσωπευτικότητας των συλλεγμένων δεδομένων. Οι ιστοσελίδες, και ιδιαίτερα αυτές που έχουν την μεγαλύτερη επισκεψιμότητα, συνήθως αποτελούν ιδιαίτερα δύσκολους στόχους καθώς είναι σχεδόν αδύνατο να συλλεγεί η συνολική πληροφορία που περιέχουν (Sarwar, et al., 2000). Είναι λοιπόν, ιδιαίτερα σημαντικό τα δεδομένα που συλλέγονται να αποτελούν ένα αντιπροσωπευτικό δείγμα του πληθυσμού των χρηστών μιας ιστοσελίδας, σε ότι αφορά πάντα τη γενικότερη χρήση της.

Τέλος και πριν την οριστικοποίηση της επιλογής, συνίσταται η επικοινωνία με την ίδια την ιστοσελίδα, ώστε να διαπιστωθεί εάν και κατά πόσο είναι επιθυμητός ο διαμοιρασμός των δεδομένων της.

Αμέσως μετά τη συλλογή των δεδομένων ακολουθεί *η φάση της επεξεργασίας τους*, που μπορεί να συμπεριλαμβάνει πολλά και διαφορετικά βήματα. Οποιοδήποτε σύστημα βασίζεται σε περιεχόμενο το οποίο παράγεται από χρήστες, είναι ευπαθές σε ανεπιθύμητα μηνύματα (spam), και φυσικά το web δεν αποτελεί εξαίρεση. Μία από τις πιο απλές λύσεις για να διαχειριστεί κανείς το spam, είναι η αξιολόγηση όλων των χρηστών με βάση το μέγεθος του προφίλ τους, και η αντιμετώπιση του κορυφαίου 5-10% ως χρήστες spam (Drachsler, et al., 2010).

Ένα άλλο σημαντικό θέμα, που αφορά στη συλλογή δεδομένων από τις ιστοσελίδες, είναι τα αντίγραφα. Το ιδανικό σενάριο θα ήταν, κάθε φορά που ένας χρήστης προσθέτει κάποιο περιεχόμενο που ήδη υπάρχει, σε μία ιστοσελίδα, το σύστημα να το αντιλαμβάνεται και να το αποκόπτει. Ένα αντίγραφο, είναι ασήμαντο πρόβλημα για ένα σύστημα, εάν συγκριθεί με το πρόβλημα που δημιουργείται από το spam, και οφείλεται κυρίως σε απροσεξία ή έλλειψη γνώσης. Αυτό φυσικά δε σημαίνει ότι δεν επηρεάζει και

ίσως σημαντικά τους αλγορίθμους σύστασης, καθώς για παράδειγμα πολλές φορές μπορεί να προταθούν δύο διαφορετικές εκδόσεις της ίδιας πηγής (Drachsler, et al., 2010).

Ένα τρίτο θέμα που προκύπτει από την επεξεργασία των συνόλων δεδομένων, είναι η μείωση του θορύβου, η οποία είναι και γνωστή πρακτική, σε ότι αφορά την έρευνα που γίνεται επάνω στα συστήματα συστάσεων. Γίνεται χρήση φίλτρων και τίθενται κατώφλια, όπως για παράδειγμα, η αποκοπή όλων των χρηστών οι οποίοι έχουν λιγότερα από 10 αντικείμενα στο προφίλ τους. Μία άλλη εναλλακτική, είναι η θεώρηση των προτύπων χρήστη–αντικειμένου, όπως για παράδειγμα, η επιλογή των χρηστών εκείνων που έχουν τουλάχιστον k αντικείμενα, τα οποία όμως, να έχουν προστεθεί από τουλάχιστον k χρήστες (Drachsler, et al., 2010).

Η τρίτη και τελευταία φάση περιλαμβάνει το διαμοιρασμό των πειραματικών αποτελεσμάτων αλλά και των σύνολα δεδομένων. Πέρα από τους κανόνες προστασίας, είναι σημαντικό να δοθεί προσοχή, σε ότι αφορά τη μορφή των δεδομένων που θα διαμοιραστούν στην επιστημονική κοινότητα. Συνήθως η επιλογή της μορφής τους είναι παντελώς ελεύθερη, εκτός και αν αφορά σε συγκεκριμένες απαιτήσεις εισόδου σε ένα σύστημα σύστασης. Για λόγους διαμοιρασμού όμως, το βέλτιστο είναι, αυτά να έχουν μορφή τέτοια, η οποία να μπορεί να υποστηριχτεί από ένα ανοιχτό λογισμικό. Η έμπνευση θα μπορούσε να ληφθεί από ήδη υπάρχοντα σύνολα δεδομένων όπως το MovieLens ενώ μεγάλος όγκος δεδομένων θα μπορούσε να έχει XML μορφή (Glahn, et al., 2009).

3.6 Πολιτικές που Αφορούν στη Νομική Προστασία των Συνόλων Δεδομένων

Η ιδιωτικότητα αλλά και η νομική προστασία είναι ιδιαίτερα σημαντικά θέματα σε ότι αφορά το διαμοιρασμό των συνόλων δεδομένων, καθώς η ισορροπία ανάμεσα στα ιδιωτικά δεδομένα και στα δημόσια είναι ιδιαίτερα λεπτή. Το πιο γνωστό περιστατικό ήταν η κυκλοφορία του AOL dataset το οποίο έλαβε περίπου 20 εκατομμύρια ερωτήματα από 650.000 χρήστες. Τα δεδομένα αυτά αφορούν, σε αναζητήσεις των χρηστών για την περίοδο τριών μηνών, κατά το οποίο έμειναν εκτεθειμένα, προσωπικά στοιχεία των χρηστών στο διαδίκτυο (Wikipedia, August 4, 2006).

Η προστασία της ιδιωτικότητας, αποτελεί ιδιαίτερη πρόκληση και για τα δεδομένα που μοιράζονται μέσα στα πλαίσια της TEL. Έτσι λοιπόν, πριν γίνει διαμοιρασμός ενός συνόλων δεδομένων, θα πρέπει να γίνει η αρμόδια πρόληψη, ώστε να υπάρχει η μέγιστη δυνατή ανωνυμία στα δεδομένα. Τα ονόματα των διάφορων χρηστών, θα πρέπει να είναι αφανή, και να μην υπάρχει εντοπισμός τους, σε σχέση με την ιστοσελίδα την οποία επισκέφθηκαν, και από την οποία εξήχθησαν τα δεδομένα. Οι προβληματισμοί αυτοί είναι ιδιαίτερα σημαντικοί, και σε περιπτώσεις όπου γίνεται διαμοιρασμός αποτελεσμάτων, που βασίζονται σε πειράματα με δεδομένα που έχουν συλλεχθεί από το διαδίκτυο (Wikipedia, August 4, 2006).

Παρόλο που οι συστάσεις δημιουργούνται για ανεξάρτητους χρήστες και πρέπει να είναι όσο το δυνατόν αποτελεσματικές, οι ερευνητές θα πρέπει να αναφέρονται στις γενικότερες επιρροές του συστήματος σε μία ομάδα χρηστών, γεγονός που έχει ως αποτέλεσμα την προστασία της ανωνυμίας του χρήστη. Η αναφορά σε αποτελέσματα που αφορούν σε συγκεκριμένους χρήστες, και η χρήση των δεδομένων τους, ως παράδειγμα, θα μπορούσε να θεωρηθεί μία επίθεση στην ιδιωτικότητα ενός συγκεκριμένου χρήστη. Το project APOSDLE, είχε ως στόχο της διατήρηση της ανωνυμίας των χρηστών, και την αντιστοίχησή τους με μοναδικά Ids, τα οποία χρησιμοποιούνταν με σκοπό τη σύνδεση των χρηστών με τα σημαντικά για αυτούς δεδομένα. Τα πραγματικά δεδομένα των χρηστών αποθηκεύονται σε ένα ξεχωριστό σύστημα με περιορισμένη πρόσβαση. Επιπλέον, τα χρήσιμα δεδομένα θα μπορούσαν να κρυπτογραφηθούν πριν αποθηκευτούν σε μία βάση δεδομένων (Vuorikari, et al., January 2008).

Παρόλο που η προσέγγιση αυτή είναι ιδιαίτερα ασφαλής, σε ότι αφορά την προστασία των δεδομένων κατά το χρόνο εκτέλεσης, ίσως να μην είναι η κατάλληλη, σε ότι αφορά την εξαγωγή των δεδομένων, από μία βάση και τη χρήση τους σε ένα σύνολο δεδομένων. Χωρίς να υπάρχει πρόσβαση στα κρυπτογραφημένα δεδομένα, είναι δύσκολο να γίνει οποιαδήποτε σύγκριση ανάμεσα στα συστήματα συστάσεων (Vuorikari, et al., January 2008).

Με στόχο να υπερπηδηθούν αυτά τα εμπόδια, έχουν δημιουργηθεί διάφορες πολιτικές, που προστατεύουν τα προσωπικά δεδομένα, αλλά ταυτόχρονα επιτρέπουν την πρόσβαση για σκοπούς καθαρά ερευνητικούς, αλλά και για την ανάπτυξη των συστημάτων σύστασης. Σε ότι αφορά τα μη επίσημα σύνολα δεδομένων, η προστασία των προσωπικών δεδομένων, είναι δευτερευούσης σημασίας, καθώς βασίζονται σε

δεδομένα του ιστού, ή σε ανοιχτές εκπαιδευτικές πηγές που εντοπίζονται ελεύθερες στο διαδίκτυο. Σύμφωνα με τις πολιτικές προστασίας, ο ιδιοκτήτης ενός συνόλου δεδομένων, θα πρέπει να συμφωνεί με αυτούς που έχουν συνδράμει στη δημιουργία του, σχετικά με την επεξεργασία την οποία θα υποστεί (Wikipedia, August 4, 2006). Διαφορετικά, δεν επιτρέπεται στον ιδιοκτήτη η περαιτέρω επεξεργασία των δεδομένων.

Παρόλο που τα δεδομένα που δημοσιεύονται στο διαδίκτυο ανήκουν στο δημόσιο τομέα, η συλλογή και ο διαμοιρασμός τους βρίσκονται ακόμη στην γκρίζα ζώνη. Η αντιμετώπιση του θέματος αυτού απαιτεί συνεργασία αυτού που θα διαμοιράσει το σύνολο δεδομένων, με αυτόν που κατέχει τα αρχικά δεδομένα.

3.7 Μορφές Ανταλλαγής των Συνόλων Δεδομένων

Κατά την ανταλλαγή των δεδομένων, ανάμεσα στους ερευνητές, απαιτείται μία πολύ ξεκάθαρη περιγραφή της μορφής των δεδομένων, η οποία είναι απαραίτητη, ώστε να είναι ξεκάθαρο το ποια δεδομένα χρησιμοποιούνται για το σύστημα σύστασης και τη διαδικασία της μάθησης (Drachsler, et al., 2010).

Για να παράγει ένα σύστημα σύστασης, την κατάλληλη σύσταση χρειάζεται τόσο την άμεση όσο και την έμμεση τροφοδότηση με σχόλια ώστε να φτάσει στο επιθυμητό αποτέλεσμα. Η τροφοδότηση αυτή μπορεί να έρθει σε διάφορες μορφές. Για παράδειγμα στα συστήματα σύστασης που βασίζονται στο περιεχόμενο μπορεί να είναι αναφορές σε ένα προϊόν ή λέξεις με τις οποίες, οι χρήστες περιγράφουν ένα αντικείμενο.

Ένα σημαντικό θέμα που αφορά στην εξαγωγή, της τροφοδότησης σχολίων από το χρήστη, σχετικά με τα σύνολα δεδομένων, σε ένα συγκεκριμένο περιβάλλον εφαρμογής, είναι η δήλωση της τροφοδότησης που συλλέγεται αλλά και η μορφή της (Vuorikari, et al., January 2008). Για παράδειγμα ένα εξαγόμενο σύνολο δεδομένων, θα πρέπει να δηλώνει εάν η πληροφορία που υπάρχει σε αυτό προέρχεται από ψηφοφορία, επίσκεψη ή tags. Επιπλέον, θα πρέπει να δηλώνει την ακριβή δομή της συλλεγόμενης τροφοδότησης (Manouselis & Vuorikari, 2009).

Με βάση όλα τα παραπάνω η πλέον αποδεκτή μορφή την οποία θα μπορούσε να διατίθεται ένα σύνολο δεδομένων, ώστε αυτό να είναι πλήρως αποδεκτό, θα μπορούσε να είναι CSV ή XML (Manouselis & Vuorikari, 2009). Για να διευκολυνθεί η διαδικασία της επεξεργασίας των πληροφοριών συνίσταται τα σύνολα δεδομένων, να τη χωρίζουν σε διαφορετικά αρχεία, καθένα από τα οποία θα περιέχει πληροφορία διαφορετικού

τύπου. Έτσι λοιπόν διαφορετικά αρχεία θα πρέπει να συγκαταλέγονται σε φακέλους που θα περιέχουν την παρακάτω πληροφορία (Drachsler, et al., 2010):

- ID χρήστη : ID του χρήστη που παρείχε τα δεδομένα,
- ID αντικειμένου : ID του αντικειμένου στο οποίου αφορούν τα δεδομένα,
- Περιεχόμενο / Τιμή : Η πραγματική πληροφορία σχετικά με το αντικείμενο.

Για τα σύνολα δεδομένων, τα οποία περιέχουν σύνθετες ιδιότητες, οι φάκελοι των αρχείων, θα πρέπει να περιέχουν αρχεία, που να περιέχουν την παρακάτω μορφή (Drachsler, et al., 2010):

- ID χρήστη : ID του χρήστη που παρείχε τα δεδομένα.
- ID αντικειμένου : ID του αντικειμένου στο οποίου αφορούν τα δεδομένα.
- Περιεχόμενο / Τιμή της ιδιότητας 1 : Η πραγματική πληροφορία σχετικά με την ιδιότητα 1 του αντικειμένου.
- Περιεχόμενο / Τιμή της ιδιότητας Ν : Η πραγματική πληροφορία σχετικά με την ιδιότητα Ν του αντικειμένου.
- Επιπλέον πληροφορία όπως οι χρονικές στιγμές που μπορούν να συμπεριληφθούν στο αρχείο.

Η διαλειτουργικότητα ανάμεσα στα διάφορα συστήματα και η επαναχρησιμοποίηση της πληροφορίας απαιτεί τη χρήση συγκεκριμένων σχημάτων πληροφοριών. Αυτό σημαίνει ότι πέρα από τα αρχεία των συνόλων δεδομένων που παράγονται και διαμοιράζονται, απαιτείται και μία περιγραφή των ιδιοτήτων τους που να τα συνοδεύει. Τα δεδομένα που εμπεριέχονται συνολικά στα σύνολα δεδομένων θα πρέπει να έχουν, αλλά και να περιγράφουν, τα παρακάτω χαρακτηριστικά (Drachsler, et al., 2010).

- Εφαρμογή / Περιβάλλον : Μία σύντομη περιγραφή της εφαρμογής με την οποία συλλέχθηκαν τα δεδομένα και το σχετικό πλαίσιο.
- Υπεύθυνος συνόλου δεδομένων: Είναι άνθρωπος με τον οποίο γίνεται η επικοινωνία σε ότι αφορά το σύνολο δεδομένων.
- Νομική προστασία / Πολιτική ανοιχτής πρόσβασης : Αφορά το εάν και κατά πόσο το σύνολο δεδομένων είναι δημόσια διαθέσιμο, ή υπόκειται σε κάποιους κανονισμούς.

- Συλλογή / Επεξεργασία δεδομένων: Αφορά τον τρόπο και τις μεθόδους που ακολουθήθηκαν κατά τη συλλογή των δεδομένων
- Κάλυψη του συνόλου δεδομένων: Απαντά στο ερώτημα του εάν και κατά πόσο το σύνολο δεδομένων, αποτελεί μία πλήρη αντανάκλαση του συνολικού συστήματος ή αφορά μόνο σε συγκεκριμένους χρήστες, ή συγκεκριμένες περιόδους.
- Στατιστικά που αφορούν στην εφαρμογή: Συνολικό αριθμός χρηστών, αντικειμένων, ψήφων, αναθεωρήσεων κλπ.
- Στατιστικές του συνόλου δεδομένων: Συνολικός αριθμός των αρχείων που καταχωρήθηκαν, στατιστικά ανά αρχείο και άλλες παρόμοιες πληροφορίες.

3.8 Learning & Knowledge Analytics (LAK)

Η ανάγκη για καλύτερη συλλογή, ανάλυση και επεξεργασία των δεδομένων στα πλαίσια της TEL έχει εκφραστεί από πολλούς επιστήμονες και μάλιστα μεταφράστηκε ως μία επείγουσα ανάγκη σε ότι αφορά το LAK (Learning and Knowledge Analytics) η οποία συζητήθηκε σε πολλά συνέδρια τα τελευταία έτη (Verbert, et al., 2012).

Ανάμεσα σε άλλα, έντονη μελέτη υπάρχει σχετικά με την ανάλυση των δεδομένων του μαθητή, και την αναγνώριση των προτύπων με σκοπό πάντα, να προβλεφθούν τα αποτελέσματα της μάθησης αλλά και να προταθούν σχετικές πηγές ή να ανιχνευθούν λάθη στα πρότυπα ή επιρροές στους μαθητές.

Ο Siemens το 2010 (Siemens, August 9, 2010) περιέγραψε το LAK ως «χρήση των έξυπνων δεδομένων, που παράγονται από τους μαθητές, μέσα από ανάλυση των μοντέλων, με σκοπό την ανίχνευση πληροφορίας και κοινωνικών συνδέσεων, για την πρόβλεψη και την παροχή συμβουλών σε ότι αφορά τη μάθηση». Το πρώτο συνέδριο σχετικά με το LAK στέφθηκε από ιδιαίτερη επιτυχία γεγονός το οποίο κατέδειξε ότι η οπτικοποίηση της πληροφορίας, η ανάλυση των κοινωνικών δικτύων και οι τεχνικές εξόρυξης εκπαιδευτικών δεδομένων προσφέρουν ιδιαίτερα ενδιαφέρουσες προοπτικές σε αυτόν τον τομέα. Παρόλο που οι διάφορες τεχνικές που εφαρμόστηκαν διαφέρουν ως προς το περιεχόμενο ή τους επιδιωκόμενους στόχους, έχουν ως κύριο αντικείμενο την αναγνώριση των αναγκών των χρηστών αλλά και την υποστήριξη των αναγκών αυτών με τη χρήση έξυπνων και προσαρμοστικών συστημάτων (Verbert, et al., 2012).

Η έρευνα σε ότι αφορά τα web analytics, τις μηχανές αναζήτησης, αλλά και τα συστήματα συστάσεων αποτελεί ένα άριστο παράδειγμα του κατά πόσο η συγκέντρωση των δεδομένων, μπορεί να χρησιμοποιηθεί ώστε να βελτιώσει τα όσα έως τώρα προσφέρονται στους χρήστες (Verbert, et al., 2012). Τα τελευταία χρόνια έχουν αναπτυχθεί διάφορα συστήματα συστάσεων, έξυπνα συστήματα διδασκαλίας αλλά και οπτικά αναλυτικά συστήματα με σκοπό να χρησιμοποιηθούν κατά τη διάρκεια της εκπαίδευσης, δυστυχώς όμως πολλές φορές έχουν παραμείνει στα «χέρια» των δημιουργών τους ή ακόμη περισσότερες αδυνατούν να περάσουν στο επόμενο στάδιο της δημιουργίας τους (Drachsler, et al., 2010).

Ένα ιδιαίτερα σημαντικό στοιχείο το οποίο θα διευκόλυνε ιδιαίτερα την έρευνα σε αυτόν τον τομέα είναι η ύπαρξη των εκτενών επισκοπήσεων σε ότι αφορά τα διαθέσιμα σύνολα δεδομένων, η οποία θα έδινε στους ερευνητές ένα πλήθος διαθέσιμων πηγών με το οποίο θα μπορούσαν να πειραματιστούν αλλά και να αναλύσουν τις ιδιότητες εκείνες, που αρμόζουν καλύτερα στα πειράματά τους. Τα σύνολα δεδομένων αυτά, χρησιμοποιούνται ως αποθετήρια για την ανάπτυξη νέων αλγορίθμων και τη σύγκρισή τους με ήδη υπάρχοντες, σε συγκεκριμένα περιβάλλοντα, ενώ μπορούν να λειτουργήσουν και ως δείκτες σχετικότητας. Οι δείκτες αυτοί θα μπορούσαν για παράδειγμα να εντοπίσουν χρήστες με παρόμοια ενδιαφέροντα, με σκοπό να προτείνουν κάποια αντικείμενα στο χρήστη (Drachsler, et al., 2010).

Κατά το 4ο (2010, Βαρκελώνη)ACM συνέδριο (The ACM Conference series on Recommender Systems - http://recsys.acm.org/), σχετικά με τα συστήματα συστάσεων, αλλά και το 5ο πανευρωπαϊκό συνέδριο σχετικά με την TEL, το Σεπτέμβριο του 2010, ζητήθηκε από τους ερευνητές να καταθέσουν, όλα τα έως εκείνη τη στιγμή διαθέσιμα σύνολα δεδομένων, που χρησιμοποιούνταν σε εφαρμογές της TEL (Govaerts, et al., 2010). Ιδιαίτερη προσπάθεια σε αυτόν τον τομέα γίνεται και στο PSLC (Pittsburgh Science of Learning Center). Το PSLC DataShop είναι ένα αποθετήριο δεδομένων που παρέχει πρόσβαση σε έναν μεγάλο αριθμών εκπαιδευτικών συνόλων δεδομένων που προέρχονται από έξυπνα συστήματα διδασκαλίας. (Stamper, et al., 2010). Η ιστοσελίδα του PSLC DataShop (https://pslcdatashop.web.cmu.edu), παρέχει δύο βασικές υπηρεσίες στην κοινότητα της επιστήμης της εκπαίδευσης, ένα αποθετήριο ερευνητικών δεδομένων, και μια σειρά από εργαλεία ανάλυσης και επεξεργασίας των δεδομένων.

Στις μέρες μας έχουν καταγραφεί περισσότερα από 270 σύνολα δεδομένων, τα οποία περιέχουν τις ενέργειες 58 εκατομμυρίων μαθητών, και πολλοί είναι εκείνοι οι ερευνητές που τα χρησιμοποίησαν με σκοπό να προβλέψουν τις αποδόσεις και τις επιδόσεις των μαθητών (Siemens, August 9, 2010). Μάλιστα, υπάρχει και μία πλατφόρμα διαθέσιμη στην οποία μπορεί να γίνει διαμοιρασμός αλλά και ανάλυση των συνόλων δεδομένων. Στο portal αυτό υπάρχουν 34 διαθέσιμα σύνολα δεδομένων, ανάμεσά τους και το σύνολο του "*Virtual Math Team project",* το οποίο αφορά στη χρήση on line συνεργατικών περιβαλλόντων, που έχουν ως στόχο να υποστηρίξουν την μάθηση των μαθηματικών, σε παιδιά μέχρι 12 χρόνων. Τα σύνολα δεδομένων αυτά, έχουν χρησιμοποιηθεί εκτενώς και από την κοινότητα CSCL (Computer Supported Collaborative Learning), καθώς και από ερευνητές που είχαν ως στόχο τη μελέτη του τομέα, που αφορά στην αφομοίωση των γλωσσών από παιδιά (Stahl, 2009).

Πολλοί είναι οι οργανισμοί εκείνοι που έχουν συνεισφέρει και έχουν παράσχει σύνολα δεδομένων, τα οποία μάλιστα περιγράφουν τη δομή των οργανισμών, τη δομή των τάξεων που προσφέρουν, τις εκπαιδευτικές πηγές αλλά και τις σχέσεις μεταξύ των ανθρώπων. Επιπλέον, μέσα στο διαμοιραζόμενο υλικό συμπεριλαμβάνονται διάφορα σχήματα αλλά και λεξικά που περιγράφουν την εσωτερική δομή ενός ακαδημαϊκού ινστιτούτου, τις σχέσης και τις ροές των δραστηριοτήτων στα κοινωνικά δίκτυα αλλά και τις εκπαιδευτικές πηγές. Τα σχήματα και τα λεξικά αυτά, προσφέρουν ιδιαίτερα ενδιαφέρουσες προοπτικές, σχετικά με το διαμοιρασμό και την επαναχρησιμοποίηση της εκπαιδευτικής αλληλεπίδρασης των δεδομένων, που είναι πάντα σχετική με την έρευνα της LAK.

Το DataCite.org είναι ένας οργανισμός που ενθαρρύνει τους χρήστες να καταγράψουν τα ερευνητικά σύνολα δεδομένων, και να προσθέσουν αναγνωριστικά σε αυτά, ώστε να είναι εύκολα διαχειρίσιμα ως ερευνητικά αντικείμενα. Στην ιστοσελίδα του το DataCite (http://www.datacite.org/), αναφέρει ότι είναι ένας μη-κερδοσκοπικός οργανισμός που ιδρύθηκε στο Λονδίνο την 1η Δεκεμβρίου 2009 και ο στόχος τους είναι η δημιουργία ευκολότερης πρόσβασης σε νόμιμα ερευνητικά δεδομένα.

Το δίκτυο Dataverse, είναι μία εφαρμογή ανοιχτού λογισμικού που αφορά στη δημοσίευση και στην ανακάλυψη ερευνητικών δεδομένων (King, 2007). Το δίκτυο αυτό έχει εγκατασταθεί στο πανεπιστήμιο του Harvard, ενώ η ιστοσελίδα τους (http://thedata.org/), αναφέρει ότι έχει ως στόχο, την αποθήκη και τη μακροπρόθεσμη διατήρηση, μέσα από καλές αρχειακές πρακτικές, ερευνητικών δεδομένων.

Κεφάλαιο 4
Εφαρμογή Αλγορίθμων Συστάσεων σε Σύνολα Δεδομένων

Στο κεφάλαιο αυτό της εργασίας, θα γίνει μία αναφορά στα καθήκοντα του χρήστη σε ένα σύστημα συστάσεων, και στις ιδιότητες των συστημάτων συστάσεων.

Στη συνέχεια επιχειρείται χρήση διαφόρων αλγορίθμων και την εφαρμογή τους σε σύνολα δεδομένων που είναι αποδεκτά από τη διεθνή βιβλιογραφία ως εκπαιδευτικά. Ειδικότερα με χρήση ενός προγράμματος που βασίζεται στη γλώσσα προγραμματισμού java, έγινε προσομοίωση τεσσάρων αλγορίθμων υπό το πρίσμα διάφορων μετρικών αλλά και με την επιρροή διαφόρων παραμέτρων. Εκμαιεύθηκε μία σειρά αποτελεσμάτων με στόχο την αξιολόγηση των αλγορίθμων αυτών αλλά και της συμπεριφοράς τους υπό τις παραπάνω επιρροές. Παρακάτω παρατίθεται μία

περιγραφή των αλγορίθμων, των μετρικών που χρησιμοποιήθηκαν καθώς επίσης και μία πλήρης παρουσίαση των αποτελεσμάτων που διεξήχθησαν.

4.1 Καθήκοντα Χρήστη στα Συστήματα Συστάσεων

Για να αξιολογηθεί σωστά ένα σύστημα συστάσεων, είναι σημαντικό να κατανοήσουμε τους στόχους και τα καθήκοντα για τα οποία χρησιμοποιείται. Έτσι, λοιπόν, θα επικεντρωθούμε στους στόχους και τα καθήκοντα των τελικών χρηστών (σε αντίθεση με τους στόχους των εμπόρων και των άλλων ενδιαφερομένων μερών του συστήματος). Τα καθήκοντα έχουν αντληθεί από την ερευνητική βιβλιογραφία. Κάθε καθήκον, παρουσιάζει κάποιες επιπτώσεις του, στην αξιολόγηση. Ενώ τα καθήκοντα έχουν προσδιοριστεί ως σημαντικά, με βάση την εμπειρία από έρευνες συστημάτων συστάσεων και από την επισκόπηση δημοσιευμένων ερευνών, αναγνωρίζοντας ότι η λίστα είναι ελλιπής. Καθώς οι ερευνητές και προγραμματιστές προχωρούν σε νέους τομείς σύστασης, αναμένουμε ότι είναι χρήσιμο να συμπληρώνουν αυτή τη λίστα και/ή να τροποποιήσουν τα καθήκοντα αυτά σύμφωνα με τους συγκεκριμένους αυτούς τομείς. Ο στόχος είναι κατά κύριο λόγο να προσδιοριστεί το ανεξάρτητο πεδίο περιγραφής καθηκόντων, που θα βοηθήσει στη διάκριση μεταξύ διαφορετικών μέτρων αξιολόγησης. Τα καθήκοντα συνοψίζονται στα εξής (Herlocker, et al., 2004) :

- Σχολιασμός Πλαισίου (Annotation in Context). Το σενάριο συστάσεων φιλτράρει τις αναρτήσεις, για να αποφασιστεί ποια αξίζει να διαβαστεί.
- Ανεύρεση καλών αντικειμένων (Find Good Items). Είναι μια ταξινομημένη λίστα με αντικείμενα που έχουν συσταθεί.
- Ανεύρεση όλων των καλών αντικειμένων (Find All Good Items). Οι περισσότερες διαδικασίες σύστασης επικεντρώνονται σε ορισμένα καλά αντικείμενα, ενώ θα έπρεπε να αποδεσμεύουν πολλά άχρηστα.
- Σύσταση ακολουθίας (Recommend Sequence). Στη προκειμένη περίπτωση δεν δίνεται σύσταση ενός αντικειμένου, αλλά η σύσταση μιας ακολουθίας αντικειμένων ή εργασιών.
- Άσκοπη Περιήγηση (Just Browsing). Οι αλγορίθμοι συστάσεων συνήθως αξιολογούνται με το πόσο καλά βοηθούν τον χρήστη στο να πάρει μια απόφαση. Υπάρχουν όμως χρήστες που βρίσκουν ευχαρίστηση, να περιηγούνται στον ιστό, απλά από συνήθεια. Για τις περιπτώσεις αυτές, η ακρίβεια των αλγορίθμων μπορεί να είναι λιγότερο σημαντική από τη διεπαφή, ή ευκολία του χρήσης, και το επίπεδο και τη φύση των πληροφοριών που παρέχονται.

- Ανεύρεση Αξιόπιστων χρηστών (Find Credible Recommender). Αρκετοί χρήστες δεν εμπιστεύονται τα μηχανήματα. Κάποιοι άλλοι αλλάζουν το προφίλ τους για να δουν την αντίδραση του συστήματος συστάσεων.
- Βελτίωση του Προφίλ (Improve Profile). Οι χρήστες αξιολογούν, επειδή πιστεύουν ότι βελτιώνει το προφίλ τους, βελτιώνοντας έτσι και την ποιότητα των συστάσεων που λαμβάνουν.
- Εκφράζοντας τον εαυτό μας (Express Self). Ορισμένοι χρήστες μπορεί να μην ενδιαφέρονται για τις συστάσεις. Αυτό που είναι σημαντικό γι αυτούς είναι να τους επιτρέπετε η αξιολόγηση. Σε κάποιους άλλους αρέσει να υπάρχει και κάποιος βαθμός ανωνυμίας. Αυτή η έκφραση του χρήστη παρέχει περισσότερα δεδομένα, τα οποία μπορούν να βελτιώνουν την ποιότητα των συστάσεων.
- Βοηθώντας άλλους (Help Others). Ορισμένοι άλλοι χρήστες νιώθουν ευχαρίστηση συμβάλλοντας στη αξιολόγηση των συστημάτων συστάσεων, γιατί πιστεύουν στα κοινωνικά οφέλη με την συμβολή τους.

4.2 Ιδιότητες Συστημάτων Συστάσεων

Ένα μεγάλο σύνολο ιδιοτήτων μπορούν να ληφθούν υπόψιν όταν θα αποφασιστεί η σύσταση που θα υιοθετηθεί. Όπως διαφορετικές εφαρμογές έχουν διαφορετικές ανάγκες, έτσι και ο σχεδιαστής του συστήματος πρέπει να αποφασίσει σχετικά με τις σημαντικές ιδιότητες που θα μετρήσουν για την συγκεκριμένη εφαρμογή.

Μερικές από τις ιδιότητες είναι (Herlocker, et al., 2004), (Hijikata, et al., 2009), (Herlocker, et al., 2000), (Smyth & McClave, 2001), (O'Mahony, et al., 2004), (Shani, et al., 2005), (George, 2005) :

4.2.1 Προτίμηση χρήστη (User Preference)

Στην έρευνα, μας ενδιαφέρει το πρόβλημα επιλογής, όπου θα πρέπει να επιλέξουμε έναν από μια σειρά υποψήφιων αλγορίθμων. Μια προφανής επιλογή είναι να εκτελέσουμε μια μελέτη χρηστών (μέσα σε θέματα) και να ζητήσουμε από τους συμμετέχοντες να επιλέξουν ένα από τα συστήματα (Hijikata, et al., 2009). Αυτή η αξιολόγηση δεν περιορίζει τα θέματα με συγκεκριμένα κριτήρια, και είναι γενικά ευκολότερο για τους ανθρώπους να πάρουν τέτοιες αποφάσεις από το να δώσουν αποτελέσματα για την

εμπειρία τους. Στη συνέχεια, μπορούμε να επιλέξουμε το σύστημα που είχε το μεγαλύτερο αριθμό ψήφων.

4.2.2 Ακρίβεια Πρόβλεψης (Prediction Accuracy)

Η ακρίβεια πρόβλεψης είναι μακράν το πιο πολυσυζητημένο κριτήριο στα συστήματα συστάσεων. Στη βάση της συντριπτικής πλειοψηφίας των συστημάτων συστάσεων βρίσκεται ένας κινητήρας πρόβλεψης. Ο κινητήρας αυτός μπορεί να προβλέψει τις απόψεις των χρηστών πάνω από τα στοιχεία (π.χ. αξιολογήσεις ταινιών) ή την πιθανότητα χρήσης (π.χ. αγορά). Μια βασική υπόθεση σε ένα σύστημα συστάσεων είναι ότι ένα σύστημα το οποίο παρέχει περισσότερες ακριβείς προβλέψεις θα προτιμηθεί από το χρήστη. Έτσι, πολλοί ερευνητές επικεντρώνονται στο να βρούν αλγόριθμους που να παρέχουν καλύτερες προβλέψεις. Η ακρίβεια πρόβλεψης είναι τυπικά ανεξάρτητη από τη διεπαφή χρήστη, και μπορεί έτσι να είναι μετρήσιμη σε ένα offline πείραμα.

4.2.3 Κάλυψη (Coverage)

Καθώς η ακρίβεια πρόβλεψης ενός συστήματος σύστασης, ειδικά σε συνεργατικά συστήματα φιλτραρίσματος, σε πολλές περιπτώσεις αυξάνεται με την ποσότητα των δεδομένων, ορισμένοι αλγόριθμοι μπορούν να προβούν σε συστάσεις με υψηλή ποιότητα, αλλά μόνο για ένα μικρό μέρος των στοιχείων που έχουν τεράστιες ποσότητες δεδομένων. Ο όρος της κάλυψης μπορεί να αναφέρεται σε αρκετές διακριτές ιδιότητες του συστήματος, όπως item space coverage, user space coverage, cold start.

4.2.4 Αυτοπεποίθηση (Confidence)

Η αυτοπεποίθηση στη σύσταση μπορεί να οριστεί ως η αυτοπεποίθηση του συστήματος στις συστάσεις ή τις προβλέψεις (Herlocker, et al., 2000). Όπως σημειώσαμε παραπάνω, το συνεργατικό φιλτράρισμα συστάσεων τείνει να βελτιώσει την ακρίβειά του, καθώς η ποσότητα των δεδομένων πάνω από τα στοιχεία μεγαλώνει. Ομοίως, η εμπιστοσύνη στην προβλεπόμενη ιδιοκτησία συνήθως αναπτύσσεται ταυτόχρονα με την ποσότητα των δεδομένων. Σε πολλές περιπτώσεις, ο χρήστης μπορεί να επωφεληθεί από την παρατήρηση αυτών των βαθμολογιών αυτοπεποίθησης (Herlocker, et al., 2000). Όταν το σύστημα αναφέρει ένα χαμηλό επίπεδο αυτοπεποίθησης σε ένα συνιστώμενο σημείο,

ο χρήστης μπορεί να μελετήσει περαιτέρω το στοιχείο πριν από τη λήψη μιας απόφασης.

4.2.5 Εμπιστοσύνη (Trust)

Ενώ η εμπιστοσύνη είναι η εμπιστοσύνη του συστήματος στις αξιολογήσεις του, ωστόσο η εμπιστοσύνη που αναφέρεται εδώ είναι η εμπιστοσύνη του χρήστη στο σύστημα σύστασης. Για παράδειγμα, αυτό μπορεί να είναι ευεργετικό για το σύστημα να συστήσει μερικά στοιχεία που ο χρήστης γνωρίζει ήδη και έχει εγκρίνει. Με αυτό τον τρόπο, παρόλο που τα κέρδη των χρηστών δεν έχουν αξία από τη σύσταση αυτή, παρατηρείται ότι το σύστημα παρέχει εύλογες συστάσεις, οι οποίες μπορούν να αυξήσουν την εμπιστοσύνη στις συστάσεις αγνώστων στοιχείων. Ένας άλλος συνηθισμένος τρόπος ενίσχυσης της εμπιστοσύνης στο σύστημα είναι η εξήγηση των συστάσεων που το σύστημα παρέχει. Η εμπιστοσύνη στα συστήματα καλείται επίσης και ως η αξιοπιστία του συστήματος.

4.2.6 Καινοτομία (Novelty)

Νέες συστάσεις είναι οι συστάσεις για τα στοιχεία που ο χρήστης δεν γνωρίζει (Konstan, et al., 2006). Σε εφαρμογές που απαιτούν νέα σύσταση, μια προφανής και εύκολη εφαρμογή προσέγγισης είναι να φιλτράρει τα στοιχεία που ο χρήστης ήδη έχει ή χρησιμοποιεί. Ωστόσο, σε πολλές περιπτώσεις οι χρήστες δεν θα αναφέρουν όλα τα στοιχεία που έχουν χρησιμοποιήσει στην παρελθόν. Έτσι, αυτή η απλή μέθοδος είναι ανεπαρκής για να φιλτράρει όλα τα στοιχεία που ο χρήστης ήδη γνωρίζει.

4.2.7 Τυχαίες Ανακαλύψεις (Serendipity)

Serendipity είναι ένα μέτρο του πόσο εκπλάγηκε ο χρήστης στις επιτυχημένες προτάσεις που του έγινε από το σύστημα (συστάσεις που δεν τις περίμενε). Φυσικά στις τυχαίες συστάσεις μεγαλώνει το serendipity. Το επιθυμητό είναι να ισοζυγίζεσαι η ακρίβεια με το serendipity.

4.2.8 Διαφορετικότητα (Diversity)

Η διαφορετικότητα ορίζεται γενικά ως το αντίθετο της ομοιότητας. Σε ορισμένες περιπτώσεις, προτείνοντας ένα σύνολο όμοιων στοιχείων μπορεί να μην είναι τόσο χρήσιμο για τον χρήστη, διότι μπορεί να χρειαστεί περισσότερος χρόνος για να διερευνήσει το φάσμα των αντικειμένων. Σκεφτείτε για παράδειγμα μια σύσταση για διακοπές (Smyth & McClave, 2001), όπου το σύστημα θα πρέπει να συστήσει πακέτα διακοπών. Παρουσιάζοντας μια λίστα με 5 συστάσεις, όλες για την ίδια θέση, μεταβάλλοντας μόνο την επιλογή του ξενοδοχείου, μπορεί να μην είναι τόσο χρήσιμες όσο προτείνοντας 5 διαφορετικές θέσεις. Ο χρήστης μπορεί να δει τις διάφορες συνιστώμενες θέσεις και να ζητήσει περισσότερες λεπτομέρειες σχετικά με ένα υποσύνολο από τις θέσεις που είναι κατάλληλες γι 'αυτόν.

4.2.9 Χρησιμότητα (Utility)

Πολλές ιστοσελίδες ηλεκτρονικού εμπορίου χρησιμοποιούν ένα σύστημα σύστασης προκειμένου να βελτιώσου τα έσοδά τους, π.χ., ενισχύοντας το cross-sell. Σε τέτοιες περιπτώσεις, ο κινητήρας σύστασης μπορεί να κριθεί από τα έσοδα που θα αποφέρει για την ιστοσελίδα (Shani, et al., 2005). Σε γενικές γραμμές, μπορούμε να ορίσουμε διάφορους τύπους των λειτουργιών κοινής ωφέλειας που η σύσταση προσπαθεί να βελτιστοποιήσει. Για τέτοιες συστάσεις, η μέτρηση της χρησιμότητας, ή της αναμενόμενης χρησιμότητας των συστάσεων μπορεί να είναι πιο σημαντική από τη μέτρηση της ακρίβειας των συστάσεων.

4.2.10 Κίνδυνος (Risk)

Σε ορισμένες περιπτώσεις, μια σύσταση μπορεί να συνδέεται με ένα δυνητικό κίνδυνο. Για παράδειγμα, όταν συνιστώνται αποθέματα για την αγορά, οι χρήστες μπορεί να επιθυμούν να αποστρέψουν τον κίνδυνο, προτιμώντας τα αποθέματα που έχουν χαμηλότερη αναμενόμενη αύξηση, αλλά και χαμηλότερο κίνδυνο κατάρρευσης. Από την άλλη πλευρά, οι χρήστες μπορεί να αναζητήσουν τον κίνδυνο, προτιμώντας τα αποθέματα που έχουν δυνητικά υψηλό, ακόμη και αν είναι λιγότερο πιθανό, το κέρδος. Σε τέτοιες περιπτώσεις, μπορεί να θέλουμε να μην αξιολογηθεί μόνο η (αναμενόμενη) αξία που παράγεται από μια σύσταση, αλλά και να ελαχιστοποιηθεί ο κίνδυνος.

4.2.11 Ανθεκτικότητα (Robustness)

Η ανθεκτικότητα είναι η σταθερότητα της σύστασης στην παρουσία των πλαστών πληροφοριών (O'Mahony, et al., 2004), που συνήθως εισάγεται επίτηδες προκειμένου να επηρεάσει τη σύσταση. Καθώς όλο και περισσότεροι άνθρωποι βασίζονται σε συστήματα συστάσεων, μπορούν να επηρεάσουν το σύστημα προκειμένου να αλλάξει την βαθμολογία ενός στοιχείου ώστε να είναι επικερδής για έναν ενδιαφερόμενο. Για παράδειγμα, ο ιδιοκτήτης ενός ξενοδοχείου μπορεί να επιθυμεί να ενισχύσει την ικανότητα του ξενοδοχείου του. Αυτό μπορεί να γίνει με προφίλ ψεύτικων χρηστών που αξιολογούν το ξενοδοχείο θετικά, είτε με ψεύτικους χρήστες που αξιολογούν τους ανταγωνιστές αρνητικά.

4.2.12 Προστασία Προσωπικών Δεδομένων (Privacy)

Σε ένα συνεργατικό σύστημα φιλτραρίσματος, ένας χρήστης γνωστοποιεί πρόθυμα τις προτιμήσεις του για στοιχεία του συστήματος με την ελπίδα να πάρει χρήσιμες συστάσεις. Ωστόσο, είναι σημαντικό για τους περισσότερους χρήστες που οι προτιμήσεις τους παραμένουν απόρρητες, κανένας τρίτος να μη μπορεί να χρησιμοποιήσει το σύστημα σύστασης για να μάθει κάτι σχετικά με τις προτιμήσεις κάποιου συγκεκριμένου χρήστη.

4.2.13 Προσαρμοστικότητα (Adaptivity)

Συστήματα πραγματικής σύσταση μπορούν να λειτουργήσουν σε ένα περιβάλλον όπου η συλλογή στοιχείων αλλάζει γρήγορα. Ίσως το πιο προφανές παράδειγμα των συστημάτων αυτών είναι η σύσταση των ειδήσεων ή συναφών ιστοριών σε εφημερίδες (George, 2005). Σε αυτό το σενάριο οι ιστορίες μπορεί να είναι ενδιαφέρουσες μόνο για ένα σύντομο χρονικό διάστημα και στη συνέχεια να γίνονται ξεπερασμένες.

4.3 Μεθοδολογία

Πρωταρχικός στόχος είναι η εφαρμογή υπαρχόντων αλγορίθμων συστάσεων, σε εκπαιδευτικά σύνολα δεδομένων, από αποθετήρια μαθησιακών αντικειμένων. Αφού αναλύθηκαν λοιπόν, πιο πάνω, οι τρείς κατηγορίες αλγορίθμων:

- Αλγόριθμοί συστάσεων που εστιάζουν στο περιεχόμενο (Content-Based Recommender Algorithms),
- Αλγόριθμοι συνεργατικών φίλτρων (Collaborative Filtering Algorithms),
- Αλγόριθμοι συστάσεων ανάλυσης γράφων (Graph-Based Recommender Algorithms),

στη συνέχεια θα γίνει μια προσπάθεια μελέτης των επιδόσεων των αλγορίθμων, ακολουθώντας τα εξής βήματα:

- Παρουσίαση και ανάλυση των εκπαιδευτικών συνόλων δεδομένων,
- Παρουσίαση του πλαισίου εργασίας (framework), Recommender 101,
- Ανάλυση των αλγορίθμων που χρησιμοποιούνται στο πλαίσιο εργασίας,
- Μετρικές που αξιολογήθηκαν,
- Αξιολόγηση πειραμάτων - Αποτελέσματα.

4.3.1 Παρουσίαση και Ανάλυση των Εκπαιδευτικών Συνόλων Δεδομένων

Η ιστοσελίδα https://movielens.org/ MovieLens, είναι μια μη εμπορικού χαρακτήρα ιστοσελίδα, με εξατομικευμένες συστάσεις για ταινίες, που τρέχει κάτω από την επίβλεψη του ερευνητικού εργαστηρίου του πανεπιστημίου της Μινεσότα.

Η GroupLens συνέλεξε και διέθεσε τα σύνολα δεδομένων αξιολόγησης, από την ιστοσελίδα MovieLens. Τα σύνολα αυτά τα πήρε σε διάφορες χρονικές περιόδους, ανάλογα με το μέγεθος του σετ. Τα σύνολα δεδομένων MovieLens, αποτελούν από τα πλέον αποδεκτά σύνολα, σε ότι αφορά το εκπαιδευτικό περιεχόμενο. Αποτελείται από ένα μεγάλο αριθμό ταινιών καθώς και τις αξιολογήσεις χρηστών που έχουν δοθεί σχετικά με αυτές. Υπάρχει σε τρεις διαφορετικές εκδόσεις (GroupLens, 2014):

1η Έκδοση, των 100k, που αποτελείται από (GroupLens, 2014):

- 100.000 αξιολογήσεις μεταξύ του 1-5 από 943 χρήστες στις 1682 ταινίες,
- κάθε χρήστης έχει βαθμολογήσει τουλάχιστον 20 ταινίες,

- παρέχει απλές δημογραφικές πληροφορίες για τους χρήστες όπως ηλικία, φύλο, επάγγελμα, Τ.Τ.
- Τα δεδομένα έχουν συλλεγεί μέσω της ιστοσελίδας MovieLens (movielens.umn.edu) μέσα σε περίοδο επτά μηνών, από την 19η Σεπτεμβρίου, 1997 μέχρι τις 22 Απριλίου 1998. Στοιχεία από χρήστες που είχαν λιγότερες από 20 αξιολογήσεις ή δεν είχαν πλήρεις δημογραφικές πληροφορίες αφαιρέθηκαν από αυτό το σύνολο δεδομένων.

Τα πιο σημαντικά αρχεία που συναντά κάποιος και στην έκδοση 100K είναι τα εξής (GroupLens, 2014):

- Αρχείο u.data. Σ'αυτό βρίσκονται τα δεδομένα 100000 αξιολογήσεων από 943 χρήστες σε 1682 αντικείμενα. Κάθε χρήστης έχει βαθμολογήσει τουλάχιστον 20 ταινίες. Οι χρήστες και τα στοιχεία αριθμούνται διαδοχικά από το 1. Τα δεδομένα είναι τυχαία και το σετ αποτελείται (GroupLens, 2014): user id | item id | rating | timestamp
- Αρχείο u.info. Περιλαμβάνει τον αριθμό των χρηστών, των αντικειμένων, και τις βαθμολογίες από το σετ u.data.
- Αρχείο u.item. Περιέχει πληροφορίες για τα αντικείμενα (ταινίες) που αποτελούνται από τα εξής πεδία: movie id | movie title | release date | video release| date | IMDb URL | unknown | Action | Adventure | Animation| Children's | Comedy | Crime | Documentary | Drama| Fantasy | Film-Noir | Horror | Musical | Mystery | Romance| Sci-Fi | Thriller | War | Western |
- Αρχείο user.dat. Περιέχει πληροφορίες για τους χρήστες: UserID | Gender | Age | Occupation | Zip-code.

2η Έκδοση, του 1M, που αποτελείται από αρχεία που περιέχουν 1.000.209 ανώνυμες αξιολογήσεις μεταξύ του 1-5, σε περίπου 3.900 ταινίες, και πραγματοποιούνται από 6.040 MovieLens χρήστες που εντάχθηκαν στο MovieLens το 2000 (GroupLens, 2014).

Όλες οι αξιολογήσεις βρίσκονται στο αρχείο rating.dat και έχουν την μορφή: UserID|MovieID|Rating|Timestamp, όπου:

- Διάστημα τιμών UserIDs είναι τιμές μεταξύ 1 και 6040,
- Διάστημα τιμών MovieIDs είναι τιμές μεταξύ 1 και 3952,
- Η αξιολόγηση έγινε σε κλίμακα 5αστέρων (μόνο σε ολόκληρα αστέρια),
- Χρονική Σήμανση Timestamp,
- Κάθε χρήστης έχει τουλάχιστον 20 αξιολογήσεις,

Οι πληροφορίες του χρήστη βρίσκονται στο αρχείο users.dat, και έχουν την ακόλουθη μορφή: UserID|Gender|Age|Occupation|Zip-code

- Όπου «Gender» επιλέγουν το φύλο με "M" άρρεν και "F" για θήλυ.
- Για την ηλικία επιλέγουμε ένα από τα ακόλουθα διαστήματα:

  1: "κάτω των 18"
  18: "18-24"
  25: "25-34"
  35: "35-44"
  45: "45-49"
  50: "50-55"
  56: "56+"

- Για την απασχόληση, μια από τις πιο κάτω επιλογές:

  0: "other" ή αν δεν έχει καθοριστεί
  1: "academic/educator"
  2: "artist"
  3: "clerical/admin"
  4: "college/grad student"
  5: "customer service"
  6: "doctor/health care"
  7: "executive/managerial"
  8: "farmer"
  9: "homemaker"
  10: "K-12 student"
  11: "lawyer"
  12: "programmer"
  13: "retired"
  14: "sales/marketing"
  15: "scientist"
  16: "self-employed"
  17: "technician/engineer"
  18: "tradesman/craftsman"
  19: "unemployed"
  20: "writer"

- Οι πληροφορίες για τις ταινίες βρίσκονται στο αρχείο, "movies.dat" και έχουν την ακόλουθη μορφή: MovieID|Title|Genres,
- Οι τίτλοι είναι ταυτόσημοι με τους τίτλους που παρέχονται από το IMDB (συμπεριλαμβανομένου και του έτους κυκλοφορίας),
- Τα είδη επιλέγονται από τις ακόλουθες κατηγορίες:

  Action
  Adventure
  Animation

Children's
Comedy
Crime
Documentary
Drama
Fantasy
Film-Noir
Horror
Musical
Mystery
Romance
Sci-Fi
Thriller
War
Western

3η Έκδοση, των 10M, όπου το σύνολο αυτό περιέχει 10,000,054 αξιολογήσεις και 95,580 ετικέτες που έχουν εφαρμοστεί σε 10,681 ταινίες από 71,567 χρήστες (GroupLens, 2014).

Οι χρήστες επιλέχθηκαν τυχαία για την ένταξη. Όλοι οι χρήστες που επιλέχτηκαν είχαν βαθμολογήσει τουλάχιστον 20 ταινίες. Σε αντίθεση με τα προηγούμενα σύνολα δεδομένων MovieLens, δεν συμπεριλαμβάνονται δημογραφικές πληροφορίες. Κάθε χρήστης αντιπροσωπεύεται από ένα αναγνωριστικό μόνο. Τα δεδομένα υπάρχουν σε τρία αρχεία, movies.dat, ratings.dat και tags.dat (GroupLens, 2014).

Όλες οι αξιολογήσεις βρίσκονται στο αρχείο ratings.dat. Κάθε γραμμή του αρχείου αποτελεί μία αξιολόγηση ενός χρήστη για μια ταινία, και έχει την ακόλουθη μορφή: UserID|MovieID|Rating|Timestamp

- Η αξιολόγηση, "Rating" έγινε σε κλίμακα 5 αστέρων (συμπεριλαμβανομένου και μισού αστεριού),
- Χρονική Σήμανση Timestamp,

Όλες οι ετικέτες περιέχονται στο αρχείο tags.dat. Κάθε γραμμή του αρχείου αποτελεί μία ετικέτα που εφαρμόζεται σε μια ταινία από έναν χρήστη, και έχει την ακόλουθη μορφή (GroupLens, 2014): UserID|MovieID|Tag|Timestamp

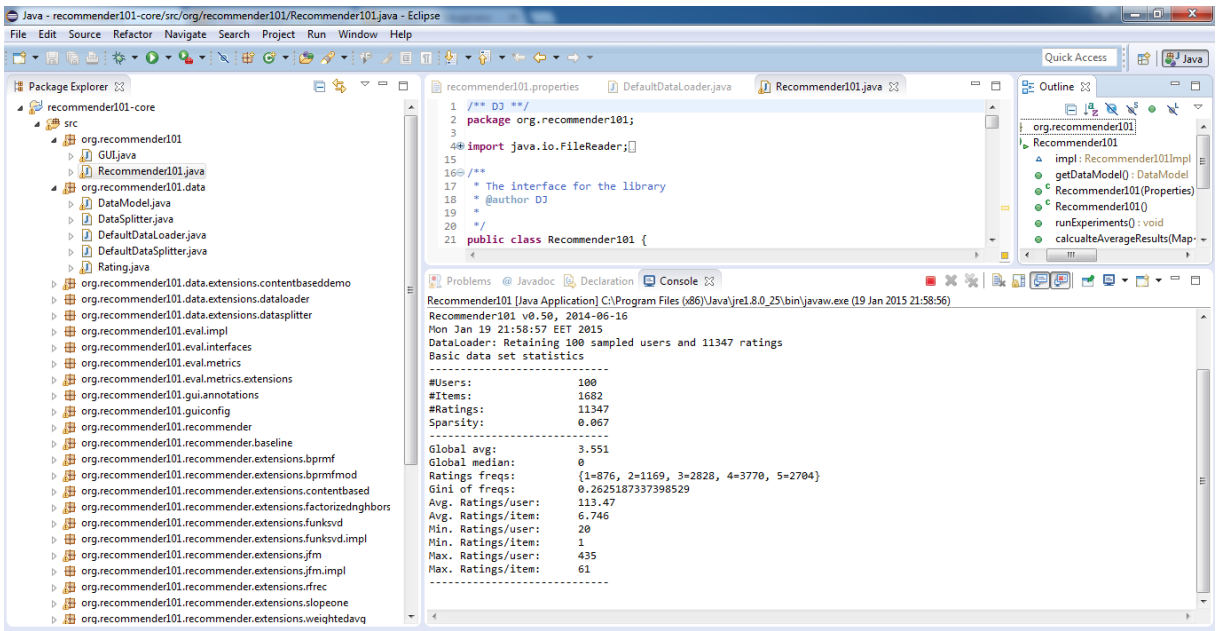
- Οι ετικέτες "Tags", είναι μεταδεδομένα για τις ταινίες που δημιουργούνται από τον χρήστη. Κάθε ετικέτα είναι συνήθως μια λέξη ή σύντομη φράση. Η έννοια, η αξία και ο σκοπός μιας συγκεκριμένης ετικέτας καθορίζεται από κάθε χρήστη.
- Οι τίτλοι είναι ταυτόσημοι με τους τίτλους που παρέχονται από το IMDB (συμπεριλαμβανομένου και του έτους κυκλοφορίας),

- Τα είδη επιλέγονται από τις ακόλουθες κατηγορίες:

Action	Film-Noir
Adventure	Horror
Animation	Musical
Children's	Mystery
Comedy	Romance
Crime	Sci-Fi
Documentary	Thriller
Drama	War
Fantasy	Western

4.3.2 Παρουσίαση του Πλαισίου Εργασίας (framework), Recommender 101

Το λογισμικό που χρησιμοποιήθηκε ονομάζεται Recommender 101 (Jannach, et al., 2013), έχει δημιουργηθεί από τους Jannach, Lerche, Gedikli και Bonnin, και είναι υλοποιημένο στη γλώσσα προγραμματισμού Java. Η τελευταία έκδοση, Recommender101 v0.50, κυκλοφόρησε 16/06/2014. Πρόκειται για εύχρηστο framework, που δίνει τη δυνατότητα διεξαγωγής μίας σειρά πειραμάτων, off-line, καθώς παρέχει στο χρήστη μία σειρά μετρικών αλλά και στρατηγικών αξιολόγησης των αλγορίθμων. Παράλληλα, επιτρέπει στο χρήστη τόσο να επεκτείνει τους ήδη υπάρχοντες αλγορίθμους και μετρικές όσο και να προτείνει δικούς του σε συνδυασμό πάντα με τα ήδη υπάρχοντα. Η υλοποίηση του framework έχει γίνει στο πρόγραμμα Eclipse και μέσα από αυτό γίνεται και η εκτέλεση του (Jannach, et al., 2013).



Εικόνα 4.1: Το κεντρικό παράθυρο του framework Recommender 101 μέσα από την eclipse.

Η δομή του framework χωρίζεται σε διάφορα επίπεδα. Ξεκινώντας από το ψηλότερο μπορεί κανείς να διακρίνει τέσσερις φακέλους οι οποίοι περιέχουν και τις συνολικές ρυθμίσεις του προγράμματος (Jannach, et al., 2013):

- Ο φάκελος conf περιέχει τα αρχεία με όλες τις ρυθμίσεις του προγράμματος.
- Στο φάκελο data βρίσκονται αποθηκευμένα τα διάφορα datasets.
- Στο φάκελο Lib εμπεριέχονται οι βιβλιοθήκες που χρησιμοποιούνται.
- Στο φάκελο src βρίσκεται η υλοποίηση της εφαρμογής.

Η προσαρμογή των διαφόρων παραμέτρων γίνεται από το configuration file (reccommender101.properties), που βρίσκεται στον κατάλογο conf. Οι παράμετροι που μπορούν να προσαρμοστούν είναι:

1. Η κλάση που θα φορτώσει και θα αναλύσει τα δεδομένα, όπως επίσης η διαδρομή και το αρχείο δεδομένων.

   DataLoaderClass=org.recommender101.data.DefaultDataLoader:filename=data/movielens/MovieLens100kRatings.txt

2. Η βαθμολογική κλίμακα σύμφωνα με τα δεδομένα που θα αναλυθούν

   GlobalSettings.minRating = 1
   GlobalSettings.maxRating = 5

3. Καθορισμός της ελάχιστης βαθμολόγησης

   GlobalSettings.listMetricsRelevanceMinRating = 5

4. Η κλάση που θα χωρίσει τα δεδομένα σε εκπαιδευτικά και για τεστ

   DataSplitterClass=org.recommender101.data.DefaultDataSplitter

5. Η κλάση που δείχνει τους αλγορίθμους που θα αξιολογηθούν

   AlgorithmClasses=
   org.recommender101.recommender.extensions.funksvd.FunkSVDRecommender,
   org.recommender101.recommender.extensions.slopeone.SlopeOneRecommender,
   org.recommender101.recommender.extensions.factorizednghbors.FactorizedNeighborhoodRecommender

6. Οι μετρικές που θα υπολογίζονται

   org.recommender101.eval.metrics.Precision,\
   org.recommender101.eval.metrics.Recall,\
   org.recommender101.eval.metrics.F1,\

org.recommender101.eval.metrics.NDCG,\
org.recommender101.eval.metrics.MRR,\
org.recommender101.eval.metrics.MAE,\
org.recommender101.eval.metrics.RMSE

Η επεξεργασία των συνόλων δεδομένων στην εφαρμογή γίνεται από το configuration file (org.reccommender101.data), που βρίσκεται στον κατάλογο src. Οι παράμετροι που μπορούν να προσαρμοστούν είναι:

- από τις κλάσεις
  - DataModel.java και
  - Rading.java
- και η διαμόρφωσή τους από τις κλάσεις
  - DataSplitter.java και
  - DefaultDataLoader.jav.

Η εικόνα 4.2 παρουσιάζει τη συνολική δομή της εφαρμογής.

Σύμφωνα με την εικόνα 4.2, οι ακόλουθες ενέργειες λαμβάνουν χώρα κατά την εκτέλεση (Jannach, et al., 2013):

1. Αρχικά ελέγχεται κατά πόσον ένα εξωτερικό αρχείο ρυθμίσεων έχει περάσει παραμετρικά ή όχι και φορτώνονται κατά περίπτωση.

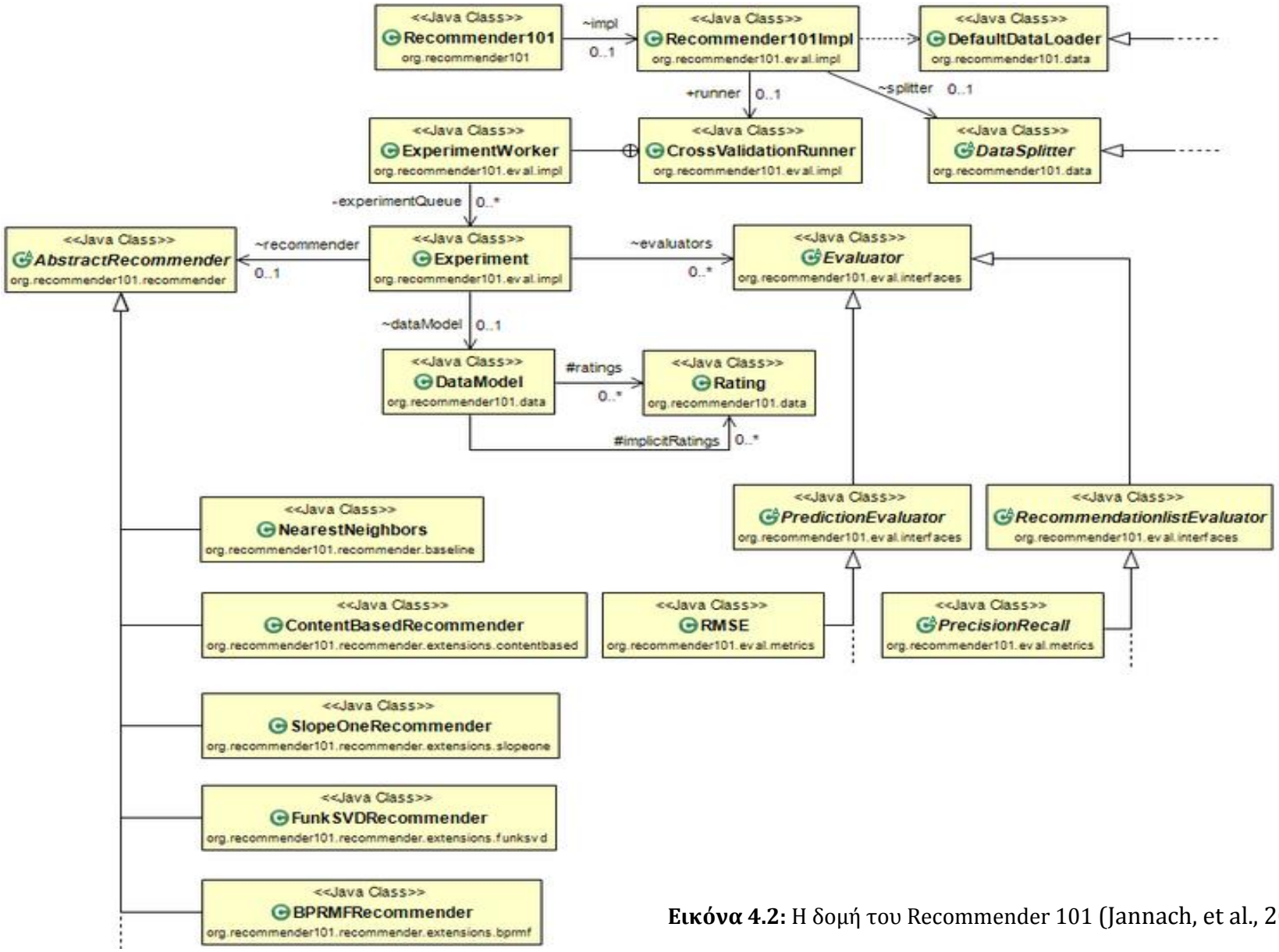
2. Στη συνέχεια, γίνεται λήψη του συνόλου δεδομένων, και τοποθετείται στο φάκελο δεδομένων data.

3. Ακολούθως, δημιουργείται ένα αντικείμενο του τύπου Recommender101.Impl, αρχικοποιείται και φορτώνεται σ'αυτό το configuration. Τα πειράματα υπολογίζονται με τις μεθόδους:

- runExperiments( )
- getLastResults( )
- printSortedEvaluationResults ( )

Τέλος, η μέθοδος printSortedEvaluationResults( ) εξάγει τα αποτελέσματα.

Το Recommender101.Impl είναι η κεντρική κλάση του πειράματος στο Recommender101. Περιλαμβάνει τις ρυθμίσεις, το σύνολο δεδομένων, τις μετρικές και τις συστάσεις. Καλείται είτε μέσω εκτέλεσης του Recommender101.java ή



Εικόνα 4.2: Η δομή του Recommender 101 (Jannach, et al., 2013)

χρησιμοποιούν τις μεθόδους helper μέσα από το Recommender101.java. Οι Μετρικές και αλγόριθμοι είναι συνδεδεμένα με τα στιγμιότυπα του πειράματος τα οποία μπορεί να επεξεργαστούν ταυτόχρονα από την ExperimentWorker (Jannach, et al., 2013).

Το πακέτο org.recommender101.recommender περιέχει την εφαρμογή των διαφόρων αλγορίθμων. Υπάρχουν οι βασικές μεθόδοι των αλγορίθμων σύστασης καθώς και πιο σύνθετες επεκτάσεις. Όλοι οι αλγόριθμοι συστάσεων καλούνται από την κλάση Abstract.Recommender.

Οι μετρικές στο Recommender101 βασίζονται πάντα στον class Evaluator που βρίσκεται στο org.recommender101.eval.interfaces. Καλούνται από τον Evaluator για κάθε αξιολόγηση που πρέπει να εκτελέσουν. Υπάρχουν δύο κλασεις μετρικών που είναι διαθέσιμες στο Recommender101 (Jannach, et al., 2013):

- Η πρώτη κλάση είναι ο PredictionEvaluator (για πρόβλεψη). Αξιολογεί την ποιότητα των προβλέψεων. Τυπικά παραδείγματα είναι τα MAE ή το RMSE.
- Η δεύτερη κλάση είναι ο RecommendationlistEvaluator (για σύσταση). Αξιολογεί την ποιότητα μιας ολόκληρης λίστας π.χ. Precision, Recall, κλπ (ακρίβεια, ανάκληση).

Το Recommender101 για να ξεκινήσει την επεξεργασία του, χρειάζεται σαν είσοδο σύνολα δεδομένων. Η μορφή των δεδομένων που διαχειρίζεται ο Recommender101, ορίζονται στη κλάση DataModel.java και Rating.java, και επεξεργάζονται από την κλάση DataSplitter και DefaultDataLoader, που βρίσκονται στο πακέτο org.recommender101.data (Jannach, et al., 2013).

Το framework του Recommender101 βρίσκεται αναρτημένο στο διαδίκτυο στην ιστοσελίδα:

http://ls13-www.cs.uni-dortmund.de/homepage/recommender101/index.shtml#running

(Jannach, et al., 2013)1

4.3.3 Ανάλυση Αλγορίθμων που Χρησιμοποιούνται στο Πλαίσιο Εργασίας

Στο πλαίσιο εργασίας χρησιμοποιούνται οι πιο κάτω αλγορίθμοι για να δοκιμαστούν με εκπαιδευτικά σύνολα δεδομένων.

4.3.3.1 Ο Αλγόριθμος του Πλησιέστερου Γείτονα (Nearest Neighbor)

Όπως έχει αναφερθεί και στο κεφάλαιο 2.4.2.5, ο αλγόριθμος k-Πλησιέστερου Γείτονα (k Nearest Neighbor) ή k-NN, είναι η πιο βασική μέθοδος μάθησης, που στηρίζεται σε στιγμιότυπα. Οι μεθόδοι μάθησης που βασίζονται σε στιγμιότυπα (instance-based), αποθηκεύουν τα δεδομένα εκπαίδευσης, κάθε φορά που εισέρχεται στο σύστημα ένα νέο στιγμιότυπο για κατάταξη. Το σύνολο από συσχετιζόμενα με αυτό στιγμιότυπα, καλείται από την μνήμη, και ξεκινά η κατάταξη του νέου στιγμιότυπου. Έτσι παρέχετε μια τοπική προσέγγιση στην συνάρτηση. Το σύνολο των τοπικών αυτών προσεγγίσεων δίνει πλεονέκτημα στην μέθοδο αυτή γιατί δεν αποτελούν μια πολύπλοκη λύση. Η τιμή της συνάρτησης για κάθε νέο στιγμιότυπο βασίζεται στις τιμές των k πλησιέστερων στιγμιοτύπων που συνθέτουν και την γειτονιά του. Κάθε δείγμα Χ ταξινομείται βάση των k πλησιέστερων γειτόνων του, μεγαλώνοντας σφαιρικά την περιοχή μέχρι να περιλάβει και το k δείγμα εκπαίδευσης (Bobadilla, et al., 2013).

Ο αλγόριθμος των πλησιέστερων γειτόνων αποτελεί τη βάση πολλών εκπαιδευτικών αλγορίθμων αλλά και μεθόδων. Η βασική αρχή πίσω από τον αλγόριθμο αυτό έγκειται στην ανακάλυψη ενός αριθμού δειγμάτων εκπαίδευσης (training), που βρίσκονται πιο κοντά (σε απόσταση), με το νέο σημείο, αλλά και την παραγωγή μίας πρόβλεψης από αυτά. Η βασική αρχή λειτουργίας του αλγορίθμου, έγκειται στην ανεύρεση ενός προκαθορισμένου αριθμού δειγμάτων εκπαίδευσης (training), τα οποία την ίδια στιγμή θα βρίσκονται στην πιο κοντική απόσταση με το νέο σημείο και να υπολογίσει τις αποστάσεις αυτές. Ο αριθμός των δειγμάτων μπορεί να καθορίζεται από το χρήστη ή και να ποικίλει ανάλογα με την πυκνότητα των στοιχείων που βρίσκονται εντός του χώρου. Παρά την απλότητά του ο αλγόριθμος αποδείχθηκε να είναι ιδιαίτερα επιτυχής σε ένα μεγάλο αριθμό προβλημάτων τόσο ταξινόμησης όσο και αναδρομή (Xiaoyuan & Khoshgoftaar, 2009).

Παρακάτω παρουσιάζεται με βήματα η εκτέλεση του αλγορίθμου:

1. Ξεκίνα από μία αυθαίρετη κορυφή και θεώρησέ την ως την τρέχουσα κορυφή

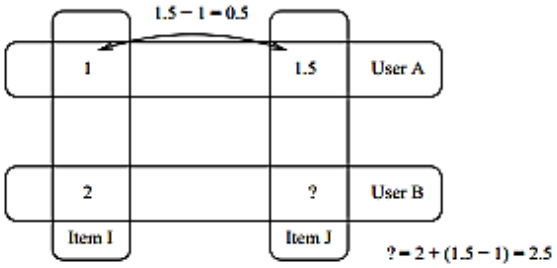
2. Βρες την κοντινότερη ακμή που ενώνει την τρέχουσα κορυφή με την κορυφή V την οποία δεν έχεις ακόμη επισκεφθεί
3. Θέσε ως τρέχουσα κορυφή τη V
4. Θεώρησε ότι πλέον την έχεις επισκεφθεί
5. Εάν όλες οι κορυφές του γράφου είναι επισκέψιμες τερμάτισε
6. Πήγαινε στο βήμα 2

Η συχνότητα της επίσκεψης των κορυφών είναι και το αποτέλεσμα του αλγορίθμου.

Ο αλγόριθμος αυτός όπως περιγράφηκε και παραπάνω είναι ιδιαίτερα εύκολο να υλοποιηθεί. Πολλές φορές όμως μπορεί να παραλείψει τις κοντινότερες διαδρομές οι οποίες είναι εύκολο να εντοπιστούν, ακόμη και με γυμνό μάτι, λόγο της «άπληστης» φύσης του. Στη χειρότερη περίπτωση, ο αλγόριθμος θα εκτελέσει μία διαδρομή που θα είναι πολύ μεγαλύτερη από τη βέλτιστη (Bobadilla, et al., 2013).

4.3.3.2 Ο Αλγόριθμος Slope One

Πρόκειται για έναν αλγόριθμο που παρουσιάστηκε το 2005 από τους Lemire & Maclachlan, και ανήκει στην κατηγορία των αλγορίθμων συνεργατικών φίλτρων (Lemire & Maclachlan, 2005). Πρόκειται για έναν πολύ απλό αλγόριθμο ο οποίος βασίζεται κυρίως στις αξιολογήσεις (ratings), γεγονός που τον καθιστά εύκολό στο να χρησιμοποιηθεί και ταυτόχρονα ιδιαίτερα αποτελεσματικό σε σχέση με τους πιο σύνθετους αλγορίθμους. Ο slope one προτάθηκε με στόχο να βελτιώσει την απόδοση, αλλά και να διευκολύνει την εφαρμογή των αλγορίθμων, σε αξιολογήσεις που βασίζονται στα αντικείμενα (item–based rating). Για το λόγο αυτό χρησιμοποιεί μία απλούστερη μορφή αναδρομής που ορίζεται από τον τύπο ($f(x) = x + b$) και χρησιμοποιεί μία ελεύθερη παράμετρο. Η παράμετρος αυτή δεν αποτελεί τίποτα παραπάνω από τη μέση διαφορά ανάμεσα στις αξιολογήσεις (ratings) δύο αντικειμένων, και έχει αποδειχθεί, να είναι πολύ πιο ακριβής από τη γραμμική αναδρομή, σε κάποιες περιπτώσεις. Το παράδειγμα που βρίσκεται πιο κάτω παρουσιάζει μία εκτέλεση του αλγορίθμου (Lemire & Maclachlan, 2005).



Εικόνα 4.3: Παράδειγμα αξιολόγησης και πρόβλεψης από τον αλγόριθμο Slope One (Wikipedia, 2014)

Όπως παρουσιάζεται και στην εικόνα 4.3:

- Ο χρήστης A αξιολόγησε με 1 το αντικείμενο I και με 1,5 το αντικείμενο J.
- Ο χρήστης B αξιολόγησε με 2 το αντικείμενο I και δεν αξιολόγησε το J.
- Ο Slope One απάντησε με 2,5 γιατί: 2+(1,5-1) = 2,5.

Για δοθέντα λοιπόν αριθμό n αντικειμένων, το μόνο που απαιτείται για τον slope one είναι να αποθηκεύσει το μέσο όρο των διαφορών και τον αριθμό των κοινών αξιολογήσεων (ratings), για κάθε ένα από τα n2 ζεύγη αντικειμένων. Εάν λοιπόν έχουμε n αντικείμενα, m χρήστες και n αξιολογήσεις (ratings), η διαφορά ανάμεσα στα ζεύγη των αντικειμένων θα είναι ως και n(n-1)/2 μονάδες αποθήκευσης και έως n2 χρονικά βήματα (Lemire & Maclachlan, 2005).

4.3.3.3 Ο Αλγόριθμος Funk Singular Value Decomposition (Funk SVD)

Στη γραμμική άλγεβρα , η ανάλυση σε ιδιάζουσες τιμές είναι μία παραγοντοποίηση ενός πίνακα με πραγματικά ή μιγαδικά στοιχεία, με πολλές χρήσιμες εφαρμογές στη θεωρία σημάτων και τη στατιστική. Η ανάλυση ενός m×n πίνακα M, με πραγματικά ή μιγαδικά στοιχεία, σε ιδιάζουσες τιμές είναι μια παραγοντοποίηση της μορφής:

$M=U\Sigma V^*$ πραγματικός ή μιγαδικός ορθομοναδιαίος πίνακας. Τα διαγώνια στοιχεία Σi,i του Σ είναι γνωστά ως ιδιάζουσες τιμές του M. Οι m στήλες του U και οι n στήλες του V όπου U είναι ένας m×m πραγματικός ή μιγαδικός ορθομοναδιαίος πίνακας, Σ ένας m×n ορθογώνιος διαγώνιος πίνακας με μη αρνητικές τιμές στην διαγώνιο και V* (ο συζυγής ανάστροφος του V, ή απλά ο ανάστροφος του V αν ο V είναι πραγματικός), ένας n×n ονομάζονται αριστερά-ιδιάζοντα διανύσματα και δεξιά-ιδιάζοντα διανύσματα του M, αντίστοιχα. Η ανάλυση σε ιδιάζουσες τιμές και η ανάλυση σε ιδιοτιμές είναι στενά συνδεδεμένες. Δηλαδή:

- Τα αριστερά-ιδιάζοντα διανύσματα του M είναι τα ιδιοδιανύσματα του MM*.
- Τα δεξιά-ιδιάζοντα διανύσματα του M είναι τα ιδιοδιανύσματα του M*M.
- Οι μη-μηδενικές ιδιάζουσες τιμές του M (που εμφανίζονται στις διαγώνιες θέσεις του Σ) είναι οι τετραγωνικές ρίζες των μη μηδενικών ιδιοτιμών του M*M και του MM*.

$$\underset{m\times n}{X} = \underset{m\times r}{U}\ \underset{r\times r}{S}\ \underset{r\times n}{V^{\mathsf{T}}}$$

Μετάφραση από την ιστοσελίδα:

http://en.wikipedia.org/wiki/Singular_value_decomposition

4.3.3.4 Ο Αλγόριθμος BPRMF (Naïve Bayes)

Ο Naïve Bayes είναι ένας ταξινομητής ο οποίος βασίζεται στην παραδοχή ότι η τιμή ενός συγκεκριμένου χαρακτηριστικού δεν έχει καμία σχέση με την παρουσία ή την απουσία κάποιου άλλου χαρακτηριστικού . Για παράδειγμα ένα φρούτο μπορεί να θεωρηθεί μήλο εάν αυτό είναι κόκκινο, στρογγυλό και έχει συγκεκριμένη διάμετρο. Ο αλγόριθμος θεωρεί ότι καθένα από αυτά τα χαρακτηριστικά λειτουργεί ανεξάρτητα στην πιθανότητα το φρούτο αυτό να είναι ένα μήλο. Το σημαντικό πλεονέκτημα του αλγορίθμου αυτού είναι ότι, απαιτεί μονάχα ένα μικρό αριθμό training δεδομένων για να εκτιμήσει τις παραμέτρους που είναι απαραίτητες για την ταξινόμηση και καθώς όλες οι μεταβλητές θεωρούνται ανεξάρτητες μεταξύ τους δεν χρειάζεται σε καμία περίπτωση να υπολογιστεί ολόκληρος ο πίνακας συνδιακύμανσης (Brusilovsky, et al., 2007).

Η συνάρτηση με βάση την οποία λειτουργεί ο αλγόριθμος δίνεται παρακάτω:

$$classify(f_1, \dots, f_n = argmax\, p(C = c) \prod_{i=1}^{n} p(F_i = f_i | C = c)$$

4.3.4 Μετρικές που αξιολογήθηκαν

Για να αξιολογηθούν οι παραπάνω αλγόριθμοι, αλλά και για να ληφθούν τα ανάλογα αποτελέσματα σχετικά με αυτούς, κρίθηκε απαραίτητη η χρήση και η εφαρμογή

κάποιων μετρικών πάνω σε αυτούς. Με βάση λοιπόν τον παραπάνω στόχο εφαρμόστηκαν οι ακόλουθες μετρικές.

Ακρίβεια πρόβλεψης (Accuracy)

Φυσικά και πρόκειται για την πλέον ενδιαφέρουσα μετρική σε ότι αφορά τα συστήματα συστάσεων καθώς και αυτήν που έχει συζητηθεί περισσότερο σε βιβλιογραφικές αναφορές. Βασική προϋπόθεση ενός συστήματος συστάσεων αποτελεί το γεγονός ότι αυτό θα προσφέρει έναν ικανοποιητικό αριθμό προβλέψεων που θα ικανοποιούν τα προτιμήσεις του χρήστη και έτσι λοιπόν η έρευνα στρέφεται προς την αναζήτηση του πλέον αποδοτικού αλγόριθμου που θα οδηγεί στην παραπάνω κατεύθυνση. (Herlocker, et al., 2004).

Μέσο απόλυτο σφάλμα (Mean absolute error)

Η μετρική αυτή καταμετρά την απόκλιση που υπάρχει ανάμεσα σε μία προβλεπόμενη βαθμολογία και στη βαθμολογία που θα έχει δώσει πραγματικά ο χρήστης. Η μετρική αυτή χρησιμοποιείται σε περιπτώσεις όπου ο χρήστης δεν ενδιαφέρεται ιδιαίτερα για τα όποια λάθη παρουσιαστούν, σε αντικείμενα που έχουν υψηλές βαθμολογίες ή θα έπρεπε να βαθμολογηθούν υψηλά. Είναι πλέον κατάλληλη για περιπτώσεις όπου το αποτέλεσμα που επιστρέφει στο χρήστη αφορά μόνο σε αντικείμενα που βρίσκονται στις υψηλότερες θέσεις της κατάταξης. Με βάση τα παραπάνω, ενδεχόμενα να μην είναι εξέχουσας σημασίας πόσο ακριβής είναι οι προβλέψεις για αντικείμενα για τα οποία ο χρήστης δεν παρουσιάζει κανένα ενδιαφέρον. Πέραν της απεικόνισης της ακρίβειας των προβλέψεων σε κάθε βαθμολογία το μέσο απόλυτο λάθος παρουσιάζει ακόμη δύο συγκριτικά πλεονεκτήματα. Αρχικά, ο υπολογιστικός μηχανισμός του είναι απλός και εύκολος στο να κατανοηθεί και δεύτερον έχει ιδιαίτερα αξιόπιστες στατιστικές ιδιότητες που συνιστούν στην εξέταση της διαφοράς των μέσων απολύτων λαθών ανάμεσα σε δύο συστήματα (Hijikata, et al., 2009).

Ακρίβεια και ανάκληση (Precision & Recall)

Σε πολλά από τα συστήματα συστάσεων που χρησιμοποιούνται δεν προσφέρεται πρόβλεψη των απαιτήσεων του χρήστη αλλά η πρόταση αντικειμένων τα οποία θα μπορούν να χρησιμοποιηθούν από αυτόν. Για παράδειγμα, κάποιος χρήστης μπορεί να

μην επέλεξε ή δεν χρησιμοποίησε κάποιο αντικείμενο, λόγω του ότι δεν είχε γνώση της ύπαρξης του, και αφού αυτό παρουσιάστηκε από το σύστημα συστάσεων να επέλεξε να το χρησιμοποιήσει. Υπό κανονικές συνθήκες αναμένεται ένα trade off ανάμεσα σε αυτές τις μετρικές. Έτσι λοιπόν αν παρατηρηθεί μία σχετική αύξηση στην ανάκληση, την ίδια στιγμή παρατηρείται μία σχετική μείωση στην ακρίβεια. Σε εφαρμογές όπου ο αριθμός των συστάσεων είναι προκαθορισμένος η μετρική της ακρίβειας παρουσιάζεται να είναι πιο χρήσιμη. Αντίθετα, σε εφαρμογές όπου ο αριθμός των συστάσεων που παρουσιάζεται στο χρήστη δεν είναι προκαθορισμένος, προτιμάται να αξιολογούνται οι αλγόριθμοι με βάση το εύρος των μηκών των λιστών σύστασης από ότι τα σταθερά μήκη (George, 2005).

Κανονικοποιημένο αθροιστικό κέρδος (NDCG)

Σε αρκετές περιπτώσεις οι εφαρμογές παρουσιάζουν στους χρήστες μία λίστα συστάσεων είτε οριζόντια είτε κάθετη αποτελούμενη από μία σειρά συστάσεων την οποία θα πρέπει αυτοί να διατρέξουν και σε κάποιες περιπτώσεις ενδεχόμενα και να μεταβούν σε επόμενη σελίδα ή σελίδες ώστε να βρουν το αντικείμενο που τους ενδιαφέρει. Αυτό το οποίο κυρίως ενδιαφέρει στις εφαρμογές αυτές είναι η κατηγοριοποίηση των αντικειμένων να γίνεται με βάση την προτίμηση του χρήστη και για το λόγο αυτό υπάρχουν δύο προσεγγίσεις (Shani, et al., 2005):

- Στην πρώτη περίπτωση αξιολογείται το αποτέλεσμα είτε με βάση το πόσο κοντά έρχεται στη σειρά των προτιμήσεων του χρήστη και
- Στην δεύτερη περίπτωση με την καταμέτρηση της χρησιμότητας των συστάσεων που παράγει το σύστημα για ένα χρήστη.

Η χρησιμότητα που παράγεται για κάθε σύσταση είναι η χρησιμότητα του αντικείμενου που συστάθηκε μειωμένη κατά ένα παράγοντα που εξαρτάται από τη θέση της σύστασης στη λίστα συστάσεων. Θεωρείται ότι οι χρήστες διατρέχουν μία λίστα συστάσεων από την αρχή της προς το τέλος της και έτσι η χρησιμότητα μειώνεται αυξητικά καθώς αυτή πλησιάζει το τέλος της λίστας. Υπό αυτό το πρίσμα, η πιθανότητα να παρατηρηθεί ένα συγκεκριμένο αντικείμενο σε μία λίστα εξαρτάται καθαρά από τη θέση στην οποία αυτό παρουσιάζεται και όχι από το ίδιο το αντικείμενο που προτείνεται (Shani, et al., 2005).

Κάλυψη πρόβλεψης (Prediction Coverage)

Η κάλυψη ενός συστήματος είναι η μετρική εκείνη η οποία κάνει μία ακριβή μέτρηση του εύρους των αντικειμένων σε ένα σύστημα με βάση τα οποία αυτό μπορεί να κάνει προβλέψεις ή συστάσεις. Τα συστήματα με χαμηλότερη κάλυψη είναι λιγότερο χρήσιμα σε χρήστες καθώς παρουσιάζουν περιορισμό σε ότι αφορά τη βοήθεια που μπορούν να παράσχουν στο χρήστη. Η μετρική της κάλυψης είναι ιδιαίτερα σημαντική στην ανεύρεση όλων εκείνων των «καλών αντικειμένων», καθώς τα συστήματα που δεν τη χρησιμοποιούν δε μπορούν να αξιολογήσουν πολλά από τα αντικείμενα, άρα αδυνατούν να εντοπίσουν και όλα εκείνα τα αντικείμενα που αφορούν στις προτιμήσεις του χρήστη. Η κάλυψη χαρακτηρίζεται άμεσα από την ερώτηση «Ποιο είναι το ποσοστό εκείνο των αντικειμένων για το οποίο το σύστημα μπορεί να κάνει προβλέψεις;» και ορίζεται ως κάλυψη πρόβλεψης (Shani, et al., 2005).

Ο πλέον εύκολος τρόπος ώστε να μετρηθεί η κάλυψη είναι η επιλογή ενός τυχαίου δείγματος χρήστη/αντικειμένου, η αίτηση μιας πρόβλεψης για κάθε ένα τέτοιο ζευγάρι και η μέτρηση του ποσοστού για το οποίο ήταν δυνατή η πρόβλεψη (Hijikata, et al., 2009).

4.3.5 Αξιολόγηση πειραμάτων - Αποτελέσματα

Στην παρακάτω ενότητα παρουσιάζονται τα αποτελέσματα που εκμαιεύθηκαν από την εφαρμογή των αλγορίθμων, στο σύνολο δεδομένων με χρήση πάντα συγκεκριμένων μετρικών. Διενεργήθηκε μία σειρά πειραμάτων σε διαφορετικά μεγέθη του συνόλου δεδομένων MovieLens υπό διαφορετικό ποσοστό αραιότητας (ανεπάρκειας δεδομένων – data sparsity) των δεδομένων κάθε φορά. Ακόμη, στο πείραμα συμμετείχε και ο αριθμός των χρηστών που πήρε τις τιμές 50, 100 και 500. Τα αποτελέσματα που λήφθηκαν συγκεντρώθηκαν σε πίνακες και παρουσιάζονται παρακάτω τόσο για την κάθε μετρική ξεχωριστά όσο και για τους αλγόριθμους που χρησιμοποιήθηκαν.

4.3.5.1 Ακρίβεια πρόβλεψης (Accuracy)

MovieLens 100K						
	Weighted Average	Popularity and Average	BPRMF	Funk SVD	Nearest Neighbor	Slope One
Sparsity	Χρήστες 50					
0.3	68%	64,7%	64,5%	67%	66%	65,7%
0.7	68,2%	65%	65,7%	68%	66,8%	66,8%
1	68,25%	65,2%	65,3%	90.9%	66,9%	69%
	Χρήστες 100					
0.3	67%	68%	64,6%	68.4%	67,8%	68%
0.7	67,2%	68,2%	64,4%	67%	66,6%	68,6%
1	67,7%	68,7%	66,9%	69.6%	69,2%	69,5%
	Χρήστες 500					
0.3	68%	69,7%	76,4%	74%	68,7%	69%
0.7	67,6%	70%	76%	73%	68,4%	68,6%
1	68,5%	70,2%	76,8%	73.9%	69,4%	69,7%

MovieLens 1M						
	Weighted Average	Popularity and Average	BPRMF	Funk SVD	Nearest Neighbor	Slope One
Sparsity	Χρήστες 50					
0.3	70%	70,5%	87,8%	95.6%	12,3%	0%
0.7	71,4%	71%	85,9%	95.5%	7,1%	0%
1	71,7%	72,7%	89,7%	98.7%	7,2%	0%
	Χρήστες 100					

0.3	74%	73%	88,2%	97.85	23%	1,3%
0.7	73,6%	73,2%	87,4%	95.5%	15,1%	0,3%
1	74%	73,6%	87,2%	97%	13,7%	0%
	Χρήστες 500					
0.3	74,7%	74%	87,9%	96.9%	47,1%	1,2%
0.7	73%	73,6%	87,9%	96.8%	53%	1,6%
1	75,2%	74,8%	87,9%	98%	57%	0%

MovieLens 10M						
	Weighted Average	Popularity and Average	BPRMF	Funk SVD	Nearest Neighbor	Slope One
Sparsity	Χρήστες 50					
0.3	72%	73,7%	86,9%	94.4%	6,8%	0%
0.7	72,6%	74%	86%	94.5%	16,9%	0%
1	72,8%	74,2%	90%	96.9%	9,6%	0%
	Χρήστες 100					
0.3	74%	75%	90,7%	97.2%	23,7%	0%
0.7	75%	75,3%	85,6%	97%	18,4%	0%
1	76,2%	75,8%	84,9%	99%	17,1%	0%
	Χρήστες 500					
0.3	76%	76%	86,7%	96.5%	50,9%	0%
0.7	76,2%	74%	86,1%	95.8%	49%	0%
1	76,7%	77%	87,1%	96.7%	83,6%	0%

Οι αλγόριθμοι των γράφων παρουσιάζονται να είναι οι πλέον σταθεροί και οι πλέον αποδοτικοί στην πολύ σημαντική αυτή μετρική της ακρίβειας της πρόβλεψης. Όποιο και

αν είναι το μέγεθος του συνόλου δεδομένων που εξετάζεται, ο αριθμός των χρηστών ή η αραιότητα των δεδομένων αυτοί παρουσιάζονται σταθεροί και μάλιστα καταγράφουν και αρκετά υψηλά ποσοστά. Ακόλουθοι είναι οι αλγόριθμοι με βάση το περιεχόμενο ενώ τέλος οι αλγόριθμοι των συνεργατικών φίλτρων, ενώ αρχικά καταγράφουν σημαντικά θετικά ποσοστά, παρουσιάζουν μία αστάθεια και μία πτώση στην απόδοσή τους καθώς το μέγεθος του συνόλου των δεδομένων και ο αριθμός των χρηστών αυξάνεται.

4.3.5.2 Απόλυτο Μέσο Λάθος (Mean Absolute Error)

MovieLens 100K						
	Weighted Average	Popularity and Average	BPRMF	Funk SVD	Nearest Neighbor	Slope One
Sparsity	Χρήστες 50					
0.3	100%	100%	-	82.1%	80,6%	65,7%
0.7	100%	100%	-	80.6%	79,7%	66,8%
1	100%	100%	-	90.9%	88,7%	66,9%
	Χρήστες 100					
0.3	79,9%	87,7%	-	74.9%	74%	68%
0.7	79,2%	87,8%	-	78.9%	77,8%	66,6%
1	79,5%	87,8%	-	78.7%	77,5%	69,5%
	Χρήστες 500					
0.3	75,4%	83,1%	-	74%	74,7%	69%
0.7	75,7%	82,9%	-	72.9%	73,6%	68,6%
1	75,8%	83,2%	-	73.9%	75%	69,7%

MovieLens 1M						
	Weighted Average	Popularity and Average	BPRMF	Funk SVD	Nearest Neighbor	Slope One
Sparsity	Χρήστες 50					
0.3	76%	100%	-	86,5%	63,2%	0
0.7	76,2%	100%	-	94,2%	63,2%	0
1	76,7%	100%	-	95,3%	63,5%	0
	Χρήστες 100					
0.3	77%	86,9%	-	89,1%	89,1%	1,3%
0.7	77,8%	87,1%	-	95,5%	100%	3%
1	77,9%	87,4%	-	96,4%	100%	0
	Χρήστες 500					
0.3	75,4%	82,7%	-	93%	96,3%	1,2%
0.7	75,7%	82,9%	-	88,3%	84,4%	1,6%
1	75,2%	83,1%	-	98%	90,8%	0%

MovieLens 10M						
	Weighted Average	Popularity and Average	BPRMF	Funk SVD	Nearest Neighbor	Slope One
Sparsity	Χρήστες 50					
0.3	75,6%	100%	-	90,7%	78,9%	0%
0.7	75,1%	100%	-	90,4%	100%	0%
1	74,7%	100%	-	92%	100%	0%
	Χρήστες 100					

0.3	76,2%	86,7%	-	84,3%	82,2%	0%
0.7	76,1%	87%	-	92,4%	73%	0%
1	78%	87,4%	-	89,6%	77,9%	0%
	Χρήστες 500					
0.3	74,3%	83,6%	-	94,8%	92,9%	0%
0.7	75%	83,9%	-	91,6%	93,6%	0%
1	74,5%	83,8%	-	88,1%	83,6%	0%

Στην περίπτωση του απόλυτου μέσου λάθους, τα δεδομένα για τους αλγόριθμους και την κατηγορία στην οποία ανήκει ο καθένας τροποποιούνται αισθητά. Οι αλγόριθμοι με βάση το περιεχόμενο παρουσιάζουν σταθερά υψηλά ποσοστά μέσου λάθους, γεγονός το οποίο σημαίνει ότι οι αξιολογήσεις που προβλέπουν, απέχουν αισθητά από τις πραγματικές αξιολογήσεις τις οποίες θα έδινε ο μέσος χρήστης. Στην περίπτωση των γράφων, η μετρική αυτή παρουσιάζεται να είναι προβληματική καθώς ο αλγόριθμος BPRMF αποτυγχάνει να δώσει σαφή αποτελέσματα, ενώ τέλος στα συνεργατικά φίλτρα και ειδικά σε έναν από τους δύο αλγορίθμου τα ποσοστά παρουσιάζονται να είναι ιδιαίτερα χαμηλά.

4.3.5.3 Ακρίβεια και ανάκληση (Precision and Recall)

MovieLens 100K – Precision						
	Weighted Average	Popularity and Average	BPRMF	Funk SVD	Nearest Neighbor	Slope One
Sparsity	Χρήστες 50					
0.3	62%	61,3%	60,9%	62,4%	62,5%	62,5%
0.7	62,3%	62,1%	62,4%	64,6%	64,8%	64,9%
1	64%	63%	65,3%	61,7%	62%	61,8%
	Χρήστες 100					
0.3	64,3%	62%	64,6%	64,7%	64,7%	65%

0.7	64,5%	62%	62,4%	62,8%	62,6%	62,7%
1	64,6%	62,2%	65,3%	61,7%	61,8%	62,1%
	Χρήστες 500					
0.3	75,5%	43%	64,6%	66%	65,4%	65,7%
0.7	74,9%	42,8%	60,1%	65,7%	65,2%	65,5%
1	75,6%	42,9%	66,9%	65,5%	64,9%	69,7%

MovieLens 100K – Recall						
	Weighted Average	Popularity and Average	BPRMF	Funk SVD	Nearest Neighbor	Slope One
Sparsity	Χρήστες 50					
0.3	67,%	52,4%	76,5%	77.3%	74,3%	74,5%
0.7	67,7%	52,7%	69,5%	71.3%	68,9%	68,9%
1	67,8%	53%	72,5%	73.9%	72,8%	72,7%
Χρήστες 100						
0.3	72,2%	70,9%	69,5%	72.5%	71,1%	71,5%
0.7	72,3%	71,3%	69,4%	71.8%	70%	70,9%
1	74%	71,1%	77,6%	79.9%	78,6%	79%
	Χρήστες 500					
0.3	75,5%	70,5%	69,5%	72,7%	72,3%	72,6%
0.7	77%	70,8%	69%	72,1%	71,9%	72,1%
1	76,2%	71,2%	71,8%	74,8%	74,5%	74,8%

MovieLens 1M – Precision						
	Weighted Average	Popularity and Average	BPRMF	Funk SVD	Nearest Neighbor	Slope One
Sparsity	Χρήστες 50					
0.3	68,3%	53%	90,3%	91.7%	13,8%	0%
0.7	68,6%	54,2%	89,6%	91.4%	8,2%	0%
1	68,2%	55%	97%	97.5%	7,8%	0%
	Χρήστες 100					
0.3	72,5%	70%	94,7%	95.8%	26%	0,09%
0.7	73,2%	71,2%	92,6%	93.9%	17,1%	0,02%
1	73,5%	71,5%	91,6%	93%	15,9%	0%
	Χρήστες 500					
0.3	76,3%	70,8%	92,7%	93,9%	52%	0,09%
0.7	76,9%	70,9%	92,6%	93,8%	58,6%	1,2%
1	77,2%	70,9%	95,4%	96,2%	53,2%	0,07%

MovieLens 1M – Recall						
	Weighted Average	Popularity and Average	BPRMF	Funk SVD	Nearest Neighbor	Slope One
Sparsity	Χρήστες 50					
0.3	67,5%	62,1%	85,5%	100%	11,4%	0%
0.7	67,9%	63,25	82,6%	100%	6,3%	0%
1	68,2%	62,7%	83,9%	100%	6,8%	0%

	Χρήστες 100					
0.3	70,3%	71,5%	82,5%	100%	20,5%	0,95
0.7	71,2%	71,7%	83%	100%	13,6%	0,2%
1	70,9%	72,3%	83,2%	100%	12,1%	0
	Χρήστες 500					
0.3	75,2%	75,2%	83,7%	100%	43,1%	0,09%
0.7	74,8%	76,4%	83,7%	100%	48,3%	1,2%
1	76,2%	76,3%	81,7%	100%	24,2%	0,07%

MovieLens 10M – Precision						
	Weighted Average	Popularity and Average	BPRMF	Funk SVD	Nearest Neighbor	Slope One
Sparsity	Χρήστες 50					
0.3	66,3%	68%	86,9%	89,4%	8%	0%
0.7	62,8%	62,3%	87,7%	94,5%	18,4%	0%
1	64,9%	68,4%	90%	94%	10,3%	0%
	Χρήστες 100					
0.3	71,6%	72,5%	93,9%	94,6%	25,8%	0%
0.7	71,55%	72,7%	92,7%	94,2%	21,2%	0%
1	72,3%	72,9%	90%	92,1%	20,6%	0%
	Χρήστες 500					
0.3	75,4%	73,6%	91,8%	93,2%	56,9%	12%
0.7	75,6%	73,9%	90,2%	91,9%	54,7%	21%
1	75,7%	74,2%	91,3%	92,4%	58,1%	17%

MovieLens 10M – Recall						
	Weighted Average	Popularity and Average	BPRMF	Funk SVD	Nearest Neighbor	Slope One
Sparsity	Χρήστες 50					
0.3	69,7%	70,3%	86,2%	100%	6%	0%
0.7	69,9%	71,2%	84,6%	100%	15,6%	0%
1	69,9%	71,5%	83,7%	100%	9%	0%
	Χρήστες 100					
0.3	75,4%	75,6%	87,8%	100%	21,9%	0%
0.7	75,7%	76,5%	79,5%	100%	16,3%	0%
1	75,8%	76,7%	80,3%	100%	14,8%	0%
	Χρήστες 500					
0.3	76,5%	78,1%	82,2%	100%	46%	8,22%
0.7	76,6%	78,2%	82,3%	100%	44,4%	8,12%
1	76,6%	78,3%	86,4%	100%	49,5%	9,3%

MovieLens 100K						
	Weighted Average	Popularity and Average	BPRMF	Funk SVD	Nearest Neighbor	Slope One
Sparsity	Χρήστες 50					
0.3	68,3%	72,1%	77,2%	81.2%	80%	79,7%
0.7	68,7%	72,3%	76,4%	78.6%	77,5%	78%
1	68,9%	72,3%	75%	78.6%	78,2%	78,5%
	Χρήστες 100					
0.3	71,6%	75,4%	75,7%	81%	80,3%	80,4%

0.7	72,3%	75,7%	75,2%	79.8%	79,3%	79,4%
1	72,6%	75,9%	77%	81.3%	81,1%	81,2%
	Χρήστες 500					
0.3	75,2%	77,8%	76,4%	82,2%	81,7%	82,3%
0.7	75,4%	77,6%	76%	81,8%	81,6%	82,1%
1	75,5%	77,9%	76,8%	82,7%	82,1%	82,5%

Προχωρώντας στις επόμενες δύο μετρικές, αυτές της ακρίβειας και της ανάκλησης, παρατηρούνται και καταγράφονται ποσοστά αλλά και επιδόσεις ανάλογες με αυτές της πρώτης μετρικής, δηλαδή αυτής της ακρίβειας πρόβλεψης. Οι γράφοι παρουσιάζονται να είναι και πάλι οι πλέον αποτελεσματικοί και να καταγράφουν σταθερά υψηλά ποσοστά που σε πολλές περιπτώσεις μάλιστα είναι και 100% καθώς το μέγεθος του συνόλου και ο αριθμός των χρηστών αυξάνεται. Ακολουθούν και πάλι οι αλγόριθμοι περιεχομένου οι οποίοι παρουσιάζουν και αυτοί μία σταθερότητα στις επιδόσεις τους λαμβάνοντας όμως χαμηλότερα ποσοστά επιτυχίας. Τέλος, ακολουθούν οι αλγόριθμοι συνεργατικών φίλτρων, οι οποίοι ενώ σημειώνουν και αυτοί υψηλά ποσοστά και μικρά σύνολα δεδομένων, παρατηρούνται να παρουσιάζουν αστάθεια και μείωση των ποσοστών όσο τα σύνολα των δεδομένων αυξάνονται στο μέγεθός τους.

4.3.5.4 Κανονικοποιημένο Αθροιστικό Κέρδος (NDCG)

MovieLens 1M						
	Weighted Average	Popularity and Average	BPRMF	Funk SVD	Nearest Neighbor	Slope One
Sparsity	Χρήστες 50					
0.3	72,4%	74,6%	97,9%	98,6%	98%	-
0.7	72,7%	74,9%	96,9%	97,3%	98,2%	-
1	72,9%	75,3%	99,6%	97,8%	98%	-

	Χρήστες 100					
0.3	73,5%	77,2%	98,5%	99,3%	26,7%	4%
0.7	73,7%	77,8%	98,4%	98,9%	89,1%	-
1	74,2%	77,1%	97,9%	98,5%	12,1%	-
	Χρήστες 500					
0.3	75,6%	80,1%	98,5%	98,9%	95,2%	85,7%
0.7	75,9%	81,2%	98,4%	98,9%	96,1%	86,1%
1	76,2%	81,5%	99,3%	99,2%	97%	86,75

MovieLens 10M						
	Weighted Average	Popularity and Average	BPRMF	Funk SVD	Nearest Neighbor	Slope One
Sparsity	Χρήστες 50					
0.3	75,6%	80,1%	97,2%	97,7%	97,5%	-
0.7	76,3%	80,3%	98,4%	98,9%	100%	-
1	76,3%	80,3%	98,9%	99%	100%	-
	Χρήστες 100					
0.3	77,1%	85,2%	98,9%	99%	91,3%	-
0.7	77,2%	85,2%	98,1%	98,9%	96,8%	-
1	77,1%	85,7%	98,4%	98,5%	93,7%	-
	Χρήστες 500					
0.3	75,4%	86,3%	76,7%	98,9%	94,8%	82,2%
0.7	75,7%	86,6%	98,2%	98,3%	93,2%	81,2%
1	76,2%	86,9%	98,4%	98,9%	95,5%	49,5%

Στην περίπτωση του κανονικοποιημένου κέρδους τα δεδομένα τροποποιούνται κατά κάποιο τρόπο καθώς και τα συνεργατικά φίλτρα παρουσιάζουν ιδιαίτερα υψηλά ποσοστά επιτυχίας τόσο για μικρά όσο και για μεγαλύτερα σύνολα δεδομένων. Ναι μεν ο αλγόριθμος slope one παρουσιάζεται προβληματικός στις επιδόσεις του, από την άλλη όμως ο αλγόριθμος nearest neighbor παρουσιάζει ιδιαίτερα σταθερά και υψηλά ποσοστά επιτυχίας. Οι γράφοι, παρουσιάζονται και πάλι αξιόπιστοι και αποτελεσματικοί με τα ποσοστά που σημειώνουν να είναι στην πλειοψηφία τους μεγαλύτερα του 90% ενώ ακολουθούν οι αλγόριθμοι περιεχομένου, με τη σύνηθη και για αυτούς σταθερότητα και ποσοστά περίπου από 70% έως 90%.

4.3.5.5 Κάλυψη πρόβλεψης (Prediction Coverage)

	MovieLens 100K					
	Weighted Average	Popularity and Average	BPRMF	Funk SVD	Nearest Neighbor	Slope One
Sparsity	Χρήστες 50					
0.3	100%	100%	0%	99%	47,4%	100%
0.7	100%	100%	0%	100%	46,3%	100%
1	100%	100%	0%	100%	48,3%	100%
	Χρήστες 100					
0.3	100%	100%	0%	99.2%	95,2%	100%
0.7	100%	100%	0%	100%	97,9%	100%
1	100%	100%	0%	100%	96,4%	100%
	Χρήστες 500					
0.3	100%	100%	0%	99%	100%	100%
0.7	100%	100%	0%	99,2%	100%	100%
1	100%	100%	0%	99,7%	1005	100%

MovieLens 1M						
	Weighted Average	Popularity and Average	BPRMF	Funk SVD	Nearest Neighbor	Slope One
Sparsity	Χρήστες 50					
0.3	100%	100%	0%	98%	45,7%	100%
0.7	100%	100%	0%	98,5%	48,4%	100%
1	100%	100%	0%	99%	46,3%	100%
	Χρήστες 100					
0.3	100%	100%	0%	99%	98,5%	100%
0.7	100%	100%	0%	99,1%	99,3%	100%
1	100%	100%	0%	99,5%	99,7%	100%
	Χρήστες 500					
0.3	100%	100%	0%	99,8%	100%	100%
0.7	100%	100%	0%	99,8%	100%	100%
1	100%	100%	0%	100%	100%	100%

MovieLens 10M						
	Weighted Average	Popularity and Average	BPRMF	Funk SVD	Nearest Neighbor	Slope One
Sparsity	Χρήστες 50					
0.3	100%	100%	0%	98%	52,1%	100%
0.7	100%	100%	0%	99%	53,2%	100%
1	100%	100%	0%	99,2%	53,6%	100%
	Χρήστες 100					
0.3	100%	100%	0%	99,7%	97,4%	100%

0.7	100%	100%	0%	100%	96,8%	100%
1	100%	100%	0%	100%	95,4%	100%
	Χρήστες 500					
0.3	100%	100%	0%	100%	100%	100%
0.7	100%	100%	0%	100%	100%	100%
1	100%	100%	0%	100%	100%	100%

Σε ότι αφορά τη μετρική της ακρίβειας της πρόβλεψης και οι τρεις κατηγορίες αλγορίθμων παρουσιάζονται να είναι πλέον αποδοτικές καθώς καταγράφουν ιδιαίτερα σημαντικά ποσοστά επιτυχίας για όλα τα μεγέθη συνόλων δεδομένων και για όλους τους αριθμούς των χρηστών. Στην πρώτη θέση της κατάταξης έρχονται οι αλγόριθμοι με βάση το περιεχόμενο, ακολουθούν στην περίπτωση αυτή τα συνεργατικά φίλτρα και τέλος οι γράφοι καθώς ο αλγόριθμος BPRMF παρουσιάζει και πάλι προβληματική συμπεριφορά και καταγράφει ποσοστά της τάξεως του 0%.

Κεφάλαιο 5
Επίλογος - Συμπεράσματα

Στα πλαίσια της μεταπτυχιακής διατριβής αυτής έγινε μια προσπάθεια περιγραφής του τομέα των συστημάτων συστάσεων (recommender systems), καθώς και των τεχνικών που χρησιμοποιούν τα περιβάλλοντα στα οποία αναπτύσσονται, μέσα από βιβλιογραφική αναφορά. Η αναφορά αυτή επεκτάθηκε και στον τομέα των συνόλων δεδομένων (datasets), τα οποία αποτελούν τη βάση δημιουργίας των συστημάτων σύστασης. Μελετήθηκε η μορφή τους, ο τρόπος με τον οποίο συλλέγονται και ο τρόπος με τον οποίο γίνεται η αξιολόγησή τους.

5.1 Ανακεφαλαίωση

Η εφαρμογή των συστημάτων συστάσεων στον εκπαιδευτικό τομέα και ιδιαίτερα στην η-μάθηση, απαιτεί ιδιαίτερη μελέτη σε ότι αφορά τη σχεδίαση αλλά και την υλοποίηση αλγορίθμων. Τα συστήματα συστάσεων στην η-μάθηση διαφέρουν από αυτά στο η-εμπόριο, γιατί, θα πρέπει να λαμβάνουν υπόψιν τους:

- τις προτιμήσεις των εκπαιδευτικών και των μαθητών,
- το διαθέσιμο υλικό που θα βοηθήσει στην επίτευξη των εκπαιδευτικών στόχων.

Σημαντικό λοιπόν θέμα στα συστήματα συστάσεων είναι και η ανάγκη δημιουργίας μεγάλων datasets (σύνολα δεδομένων), τα οποία θα διευκολύνουν την ανάπτυξη και την υλοποίηση τέτοιων συστημάτων.

Τα «σημασιολογικά συστήματα συστάσεων», αφορούν τα συστήματα, των οποίων, οι επιδόσεις βασίζονται σε γνώση που προέρχεται από εξατομικευμένες συστάσεις οι οποίες περιγράφονται σημασιολογικά και λαμβάνονται λόγω της ανταλλαγής των πληροφοριών ανάμεσα στα διάφορα στοιχεία της διαδικασίας παραγωγής, αλλά και παράδοσης των σημασιολογικών συστάσεων. Η ανάπτυξη των κατάλληλων σημασιολογικών συστημάτων συστάσεων, θα πρέπει να γίνεται υπό τις παρακάτω προϋποθέσεις (Santos & Boticario, 2011(1)) :

- Θα πρέπει να υπάρχει ένα βασικό μοντέλο που θα χαρακτηρίζει τις συστάσεις.

- Θα πρέπει επίσης να υπάρχει μία αρχιτεκτονική που θα βασίζεται στα πρότυπα, θα προσανατολίζεται στην υπηρεσία και θα έχει ως σκοπό την αλληλεπίδραση ανάμεσα στα διάφορα στοιχεία του λογισμικού.
- Τέλος θα υπάρχει ένα περιβάλλον διεπαφής το οποίο θα παρέχει τις συστάσεις με εύκολο και εύχρηστο τρόπο.

Τα συστήματα σύστασης, τα οποία μάλιστα θεωρούνται χρησιμότατα εργαλεία σχετικά με την πρόσβαση στη διαθέσιμη πληροφορία. Συνήθως αυτά ανήκουν σε τρείς κατηγορίες:

- Αλγόριθμοι που εστιάζουν στο περιεχόμενο (Content-based algorithms), παρέχουν συστάσεις, αφού αναλύσουν πρώτα το περιεχόμενο όλων των σελίδων στις οποίες έχει γίνει αναζήτηση. Έπειτα αναζητούν όλες τις σελίδες εκείνες, που παρουσιάζουν αντίστοιχο περιεχόμενο.
- Αλγόριθμοι συνεργατικών φίλτρων (Collaborative filtering algorithms), αξιολογούν τα χαρακτηριστικά ενός χρήστη, με βάση τα χαρακτηριστικά άλλων χρηστών και έτσι τον παραπέμπουν σε σελίδες που έχουν επισκεφτεί παρόμοιοι χρήστες.
- Αλγόριθμοι με βάση τους γράφους (Graph-based algorithms), όπου συνενώνει έναν χρήστη Χ με μία οντότητα Υ, και μπορεί να του προσφέρει μία σύσταση που είτε θα είναι αρεστή σε αυτόν είτε όχι.

Σημαντικό στην ανάπτυξη ενός συστήματος συστάσεων είναι η συγκρότηση ενός αξιόλογου αποθετηρίου δεδομένων, τα οποία θα περιέχουν ρεαλιστικές αναπαραστάσεις του συστήματος σύστασης, αλλά και όλες εκείνες τις μαθησιακές πληροφορίες, που χρειάζονται, με σκοπό να ανταποκρίνονται οι αλγόριθμοι συστάσεων.

Μπορούν να τεθούν άξονες ώστε να βοηθήσουν, στη δημιουργία του πλέον κατάλληλου συνόλου δεδομένων (Manouselis, et al., 2010):

- Να αντικατοπτρίζει ρεαλιστικά τις μεταβλητές της διαδικασίας της μάθησης,
- Χρήση ενός σημαντικά μεγάλου αριθμού προφίλ χρηστών,
- Δημιουργία συνόλων δεδομένων τέτοιων που να είναι συγκρίσιμα με άλλα.

5.2 Σύνοψη Συμπερασμάτων

Στο κεφάλαιο 4 έγινε μια παράθεση αλγορίθμων συστάσεων, που χρησιμοποιούνται σήμερα σχετικά με τα σύνολα δεδομένων. Στόχος ήταν η αξιολόγηση τους και η καταλληλότητα της χρήσης τους στα εκπαιδευτικά σύνολα δεδομένων.

Μία σειρά αποτελεσμάτων, που εξήχθησαν από τη χρήση των αλγορίθμων ήταν η προσπάθεια κατανόησης της χρήσης των αλγορίθμων σε ότι αφορά τα σύνολα δεδομένων.

Η συλλογή και μελέτη των συνόλων δεδομένων, από διάφορα αποθετήρια είναι πολύ επίμονη εργασία ιδιαίτερα στον διαχωρισμό των διαφόρων στοιχείων τους, προκειμένου να εφαρμοστούν οι κατάλληλοι αλγορίθμοι συστάσεων. Η σύγκριση και αξιολόγηση των συστάσεων που παρέχουν στους τελικούς χρήστες είναι το σημαντικότερο στοιχείο καταλληλότητας τους. Έτσι μπορούν να αναδειχθούν τα πλεονεκτήματα και μειονεκτήματα αυτών, με βάση τις ιδιαιτερότητες του εκάστοτε συνόλου δεδομένων.

Πιο αναλυτικά στη συνέχεια δίνονται τα αποτελέσματα ανάλογα με τις ιδιότητες των συνόλων δεδομένων.

Ακρίβεια πρόβλεψης (Accuracy)

1. Οι αλγόριθμοι των γράφων παρουσιάζονται να είναι οι πιο σταθεροί και αποδοτικοί. Ανεξάρτητα με το μέγεθος του συνόλου δεδομένων, ο αριθμός των χρηστών ή η αραιότητα των δεδομένων, αυτοί παρουσιάζονται σταθεροί και μάλιστα καταγράφουν και αρκετά υψηλά ποσοστά.
2. Ακόλουθοι είναι οι αλγόριθμοι με βάση το περιεχόμενο.
3. Τέλος οι αλγόριθμοι των συνεργατικών φίλτρων, ενώ αρχικά καταγράφουν σημαντικά θετικά ποσοστά, παρουσιάζουν μία αστάθεια και μία πτώση στην απόδοσή τους καθώς το μέγεθος του συνόλου των δεδομένων και ο αριθμός των χρηστών αυξάνεται.

Μέσο Απόλυτο Σφάλμα (Mean Absolute Error)

1. Οι αλγόριθμοι με βάση το περιεχόμενο παρουσιάζουν σταθερά υψηλά ποσοστά, γεγονός το οποίο σημαίνει ότι οι αξιολογήσεις που προβλέπουν, απέχουν αισθητά από τις πραγματικές αξιολογήσεις τις οποίες θα έδινε ο μέσος χρήστης.

2. Στην περίπτωση των γράφων, η μετρική αυτή παρουσιάζεται να είναι προβληματική καθώς ο αλγόριθμος BPRMF αποτυγχάνει να δώσει σαφή αποτελέσματα.
3. Στα συνεργατικά φίλτρα και ειδικά σε έναν από τους δύο αλγορίθμους τα ποσοστά παρουσιάζονται να είναι ιδιαίτερα χαμηλά.

Ακρίβεια και Ανάκληση (Precision and Recall)

Καταγράφονται ποσοστά αλλά και επιδόσεις ανάλογες με αυτές της πρώτης μετρικής, δηλαδή αυτής της ακρίβειας πρόβλεψης.

1. Οι γράφοι παρουσιάζονται να είναι και πάλι οι πλέον αποτελεσματικοί και να καταγράφουν σταθερά υψηλά ποσοστά που σε πολλές περιπτώσεις μάλιστα είναι και 100% καθώς το μέγεθος του συνόλου και ο αριθμός των χρηστών αυξάνεται.
2. Ακολουθούν και πάλι οι αλγόριθμοι περιεχομένου οι οποίοι παρουσιάζουν και αυτοί μία σταθερότητα στις επιδόσεις τους λαμβάνοντας όμως χαμηλότερα ποσοστά επιτυχίας.
3. Τέλος, ακολουθούν οι αλγόριθμοι συνεργατικών φίλτρων, οι οποίοι ενώ σημειώνουν και αυτοί υψηλά ποσοστά και μικρά σύνολα δεδομένων, παρατηρούνται να παρουσιάζουν αστάθεια και μείωση των ποσοστών όσο τα σύνολα των δεδομένων αυξάνονται στο μέγεθός τους.

Κανονικοποιημένο Αθροιστικό Κέρδος (NDCG)

1. Τα συνεργατικά φίλτρα παρουσιάζουν ιδιαίτερα υψηλά ποσοστά επιτυχίας τόσο για μικρά όσο και για μεγαλύτερα σύνολα δεδομένων. Ναι μεν ο αλγόριθμος slope one παρουσιάζεται προβληματικός στις επιδόσεις του, από την άλλη όμως ο αλγόριθμος nearest neighbor παρουσιάζει ιδιαίτερα σταθερά και υψηλά ποσοστά επιτυχίας.
2. Οι γράφοι, παρουσιάζονται και πάλι αξιόπιστοι και αποτελεσματικοί με τα ποσοστά που σημειώνουν να είναι στην πλειοψηφία τους μεγαλύτερα του 90%.
3. Οι αλγόριθμοι περιεχομένου, με τη συνήθη και για αυτούς σταθερότητα και ποσοστά περίπου από 70% έως 90%.

Κάλυψη Πρόβλεψης (Prediction Coverage)

Και οι τρεις κατηγορίες αλγορίθμων παρουσιάζονται να είναι πλέον αποδοτικές καθώς καταγράφουν ιδιαίτερα σημαντικά ποσοστά επιτυχίας για όλα τα μεγέθη συνόλων δεδομένων και για όλους τους αριθμούς των χρηστών.

1. Στην πρώτη θέση είναι οι αλγόριθμοι με βάση το περιεχόμενο,
2. Τα συνεργατικά φίλτρα στη δεύτερη.
3. Οι γράφοι καθώς ο αλγόριθμος BPRMF παρουσιάζει και πάλι προβληματική συμπεριφορά και καταγράφει ποσοστά της τάξεως του 0%.

Αυτό το οποίο τελικά διαπιστώνεται, έπειτα από τον πειραματισμό και την εξέταση των δεδομένων, καθώς και την εφαρμογή των αλγορίθμων, είναι ότι η κάθε κατηγορία αλγόριθμου, επηρεάζεται σημαντικά από τα διάφορα χαρακτηριστικά του dataset στο οποίο θα εφαρμοστεί. Ναι μεν οι γράφοι παρουσιάζονται αποτελεσματικότεροι στο σύνολό τους, παρόλα αυτά όμως για διαφορετικό dataset διαφορετικός θα είναι και ο αλγόριθμος με τη μεγαλύτερη επιτυχία. Ως κατηγορίες αλγορίθμων ασφαλώς φαίνονται να είναι με τη πιο πάνω σειρά που παραθέτουμε (στην αποτελεσματικότητα), τα χαρακτηριστικά όμως του dataset δεν παύουν και αυτά να παίζουν σημαντικό ρόλο και να καθορίζουν τις επιδόσεις.

Συμπερασματικά:

- Οι γράφοι είναι πιο γρήγοροι αλγόριθμοι, αλλά όχι για όλα τα dataset.
- Τα χαρακτηριστικά του κάθε dataset είναι πολύ σημαντικά για την επίδοση του αλγορίθμου.
- Ο κάθε αλγόριθμος έχει τις δικές του επιδώσεις.
- Το κάθε dataset όταν εφαρμόζεται σε διαφορετικό αλγόριθμο έχει διαφορετικά αποτελέσματα.
- Σε ότι αφορά τα πλεονεκτήματα, οι γράφοι είναι γρήγοροι και αποδοτικοί και για μεγάλα dataset, οι αλγόριθμοι με βαση το περιεχόμενο δεν παρουσιάζουν ιδιαίτερα πλεονεκτήματα γιατί κατέληξαν και στην τελευταία θέση και τα συνεργατικά φίλτρα ναι μεν είναι λίγοτερο αποτελεσματικοί από τους γράφους, λειτουργούν πολύ καλά όμως στα μεγάλα datasets.

5.3 Επεκτάσεις

Αναμφισβήτητα, ο τομέας των συστάσεων, στη η-μάθηση, ήταν και θα είναι ένα από τα βασικότερα ερευνητικά πεδία στον τομέα της πληροφορικής. Αυτό που θα μπορούσε να μελετηθεί περεταίρω θα ήταν η μελέτη, η σχεδίαση και κατασκευή ενός εργαλείου συστάσεων, στο περιβάλλον κάποιου ανοικτού συστήματος LMS (Learning Management System), όπως π.χ. των LAMS, έτσι ώστε να προτείνει στο μαθητή, συστάσεις υλικού από το δικό του αποθετήριο. Με αυτό τον τρόπο θα μπορούσε άμεσα να χρησιμοποιηθεί το αποθετήριο των LAMS, χωρίς ιδιαίτερα προβλήματα συγχρονισμού, μεταφοράς δεδομένων, προσβασιμότητας κλπ, κάτι που θα παρουσίαζε άμεσα αποτελέσματα, κυρίως όσο αφορά την επαναχρησιμοποίηση του υλικού. Το αποθετήριο των LAMS θα μπορούσε να μετατραπεί σε ένα καλό εκπαιδευτικό σύνολο δεδομένων.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Adomavicius, G. & Tuzhilin, A., 2005. *Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions.* s.l., IEEE Trans. Knowl. Data Eng.,17, 734–749 .

Adomavicius, G. & Tuzhilin, A., 2011. *Context-aware recommender systems. Recommender Systems Handbook, by F. Ricci, L. Rokach, B. Shapira.* USA, Springer, pp. 217–253.

Aiello, W., Chung, F. & Lu, L., 2000. *A Random Graph Model for Massive Graph.* s.l., In STOC'00, Proceedings of the ACM Symposium on Theory of Computing, pages 171-180. ACM Press.

Ali, K. & van Stam, W., 2004. *TiVo: Making Show Recommendations Using a Distributed Collaborative Filtering Architecture..* Seattle, WA, In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 394-401.

Amatriain, X., Jaimes, A., Nuria, O. & Pujol, J. M., 2011. *Data Mining Methods for Recommender Systems.* Spain: s.n.

Antonopoulus, N. & Salter, J., 2006. *Cinema screen recommender agent: combining collaborative and content-based filtering.* s.l., IEEE Intelligent Systems 35– 41.

Barajas, M. & Gannaway, G., 2007. *Implementing e-learning in the traditional higher education institution.* s.l., Higher Educ. Eur., 32, 111–119.

Basu, C., Hirsh , H. & Cohen , W., 1998. *Recommendation as Classification: Using Social and Content-Based Information in Recommendation.* Madison, WI, . In: Proceedings of the 15th National Conference on Artificial Intelligence, 714-720.

Belkin, N. & Croft, B., 1992. *Information Filtering and Information Retrieval: Two Sides of the Same Coin!.* s.l., Communications of the ACM 35(12) 29-38.

Billsus, D., Pazzani , M. & Chen, J., 2002. *A Learning Agent for Wireless News Access.* s.l., In: Proceedings of the International Conference on Intelligent User Interfaces 33-36.

Bobadilla, J., Ortega, F., Hernando, A. & Gutierrez, A., 2013. *Recommender System Survey.* Madrid, Spain, Knowledge-Based Systems 46, 109-132.

Bozo, J., Alarcón , R. & Iribarra, S., 2010. *Recommending Learning Objects According to a Teachers' Contex Model. In Sustaining TEL: From Innovation to Learning and Practice, Lecture Notes in Computer Science.* Berlin, Germany, Springer:Volume 6383, pp. 470–475.

Breese, J., Heckerman, D. & Kadie, C., October 1998. *Empirical analysis of predictive algorithms for collaborative filtering.* Radmon, Technical Report MSR-TR-98-12.

Brusilovsky, P., 1996. *Methods and techniques of adaptive hypermedia. User Model. User-Adapt.* s.l., Interact. 6(2–3), 87–129.

Brusilovsky, P., 2003. *Adaptive and Intelligent Web-based Educational Systems.* s.l., International Journal of Artificial Intelligence in Education, pp. 156-169.

Brusilovsky, P., Karagiannidis, C. & Sampson, D., 2004. *Layered evaluation of adaptive learning systems. Int. J. Continuing Eng. Educ. Lifelong Learn. Spec. Issue AdaptivityWeb Mob. Learn. Serv. 14(4/5), 402–421.* s.l., Inderscience Pub.

Brusilovsky, P., Kobsa, A. & Nejdl, W., 2007. *The Adaptive Web, LNCS 4321, pp. 325 – 341, 2007.* Berlin Heidelberg, © Springer-Verlag .

Buder, J. & Schwind, C., 2012. *Learning with personalized recommender systems: a psychological view.* s.l., Comput. Hum. Behav. 28, 207–216.

Burke, R., 2002. *Hybrid recommender systems: Survey and experiments,* User-Model. User-Adapt. Interact: s.n.

Burke, R., 2007. *Hybrid web recommender system. In Adaptive Web.* Berlin/Heidelberg, Springer.

Butoianu, V. και συν., 2010. *User context and personalized learning: a federation of contextualized attention metadata.* s.l., Journal of Universal Computer Science, Vol. 16, pp.2252–2271.

Chervenak, A., Foster, I., Kesselman, C. & Salisbury, C., 2000. The data grid: towards an architecture for the distributed management and analysis of large scientific data sets. Issue J. Netw. Comput. Appl. 23(3), 187–200.

Cohen, W., 1995. *Fast Effective Rule Induction.* Tahoe City, CA, In: Proceedings of the Twelfth International Conference on Machine Learning, 115-123.

Cohen, W., 1996. *Learning Rules that Classify E-mail.* s.l., In: Papers from the AAAI Spring Symposium on Machine Learning in Information Access, 18-25.

Dagger , D., O'Connor, A., Lawless, S. & Walsh, E., 2007. *Service-oriented e-learning platforms. From Monolithic systems to flexible services.* s.l., IEEE Internet Comput., 3, 28–35.

Drachsler, H., 2009. *Navigation Support for Learners in Informal Learning Networks.* Heerlen, The Netherlands, Open Universiteit Nederland.

Drachsler, H., 2010. Issues and considerations regarding sharable data sets for recommender systems in technology enhanced learning. Procedia Comput. Sci. 1(2)(2849–2858 (2010). doi:10. 1016/j.procs.2010.08.010).

Drachsler, H., 2009. *Effects of the ISIS Recommender System for navigation support in selforganized learning networks.* s.l., J. Educ. Technol. Soc. 12, 122–135.

Drachsler, , 2010. *Issues and Considerations regarding Sharable Data Sets for Recommender Systems in Technology Enhanced Learning.* 1st Workshop on Recommender Systems for Technology Enhanced Learning, RecSysTEL.

Frey, B. J. & Dueck, D., 2007. *Clustering by passing messages between data points.* 307 επμ. s.l.:Science.

García, E., Romero, C., Ventura, S. & De Castro, C., 2008. *Collaborative recommender system using distributed rule mining for improving web-based adaptive courses.* s.l., s.n.

Garcia, E., Romero, C., Ventura, S. & Castro, C. D., 2009. *An architecture for making recommendations to courseware authors using association rule mining and collaborative filtering, User Modeling and User-Adapted Interaction.* s.l., 19.1-2: 99-132.

George, T., 2005. *A scalable collaborative filtering framework based on co-clustering,* s.l.: In: Fifth IEEE International Conference on Data Mining, pp. 625–628.

Glahn, C., Specht, M. & Koper, R., 2009. *Visualisation of Interaction Footprints for Engagement in Online Communities.* s.l., Educational Technology & Society, vol. 12 no. 3, pp. 44–57.

Goldberg, K., Roeder, T., Guptra, D. & Perkins, K., 2001. *Eigenstate: A constant-time collaborative filtering algorithm,* s.l.: Inf. Retr. 4, 133–151.

Govaerts, S., Verbert, K., Klerkx, J. & Duval, E., 2010. *Visualizing activities for self-reflection and awarenes.* s.l., Lecture Notes in Computer Science, 6483, 91-100.

GroupLens, R., 2014. *Social Computing Research at the University of Minnesota.* [Ηλεκτρονικό] Available at: http://grouplens.org/datasets/movielens/ [Πρόσβαση 10 12 2014].

Hanani, U., Shapira, B. & Shoval, P., 2001. *Information filtering: overview of issues, research and systems..* s.l., User Model. User-Adapt. Interact. 11, 203–259 .

Han, J. & Kamber, M., USA, 2001. Data Mining: Concepts and Techniques. Στο: s.l.:s.n.

Han, P., Xie, B., Yang, F. & Shen, R., 2004. *A scalable P2P recommender system based on distributed collaborative filtering.* s.l., Expert Syst. Appl. 27, 203–210.

Herlocker, J. L., Konstan, J. A., Borchers, A. & Riedl, J., 1999. *An algorithmic framework for performing collaborative filtering.* s.l.:in Proceedings of the Conference on Research and Development in Information Retrieval (SIGIR '99), pp. 230– 237.

Herlocker, J. L., Konstan, J. A. & Riedl, J. T., 2000. *Explaining collaborative filtering recommendations,* New York, NY, USA: In: CSCW '00: Proceedings of the 2000 ACM conference on Computer supported cooperative work, pp. 241–250. ACM, DOI http://doi.acm.or.

Herlocker, J. L., Konstan, J. A. & Terveen, L. G., 2004. *Evaluating collaborative filtering recommender systems.* s.l., ACM Trans. Inform. Syst. 22(1), 553.

Hijikata, Y., Shimizu, T. & Nishida, S., 2009. *Discovery-oriented collaborative filtering for improvinguser satisfaction,* New York, NY, USA: In: IUI '09: Proceedings of the 13th international conference on Intelligentuser interfaces, pp. 67–76. ACM, DOI h.

Hill, W. & Terveen, L. G., 2001. *Beyond Recommender Systems: Helping People Help Each Other. In HCI in the New Millennium,* Boston, MA, USA: Carroll, J., Ed.; Addison Wesley.

Howard-Jones, P., Ott, M., Van Leeuwen, T. & De Smedt, B., 2010. *Neuroscience and technology enhanced learning.* s.l., ARV White paper.

Iorio, A. D., 2006. *Automatically producing accessible learning objects.* s.l., Educ. Technol. Soc., 9, 3–16.

Jacquet-Lagrèze, E. & Siskos , Y., 2001. *Preference disaggregation: 20 years of MCDA experience.* s.l., European Journal of Operational Research, 130 (2), 233–245.

Jannach, D., Lerche, L., Gedikli, F. & Bonnin, G., 2013. *What recommenders recommend - An analysis of accuracy, popularity, and sales diversity effects.* Rome, Italy, , 21st International Conference on User Modeling, Adaptation and Personalization (UMAP 2013).

King, G., 2007. *An introduction to the dataverse network as an infrastructure for data sharing.* s.l., Sociological Methods Research, 36(2), 173-199.

Kleinberg, J., August 2000. *The Small-World Phenomenon: An Algorithmic Perspective.* s.l., Nature, Vol. 406(6798).

Konstan, J. A., 2004. *Introduction to recommender systems: algorithms and evaluation.* s.l., ACM Trans. Inf. Syst. 22(1), 1–4 .

Konstan, J. A., 2006. *Lessons on applying automated recommender systems to information-seeking tasks,* s.l.: In: AAAI .

Lemire, D. & Maclachlan, A., 2005. *Slope One Predictors for Online Rating-Based Collaborative Filtering*. s.l., s.n.

Linden, G., Smith, B. & York, J., 2003. *"Amazon.com recommendations: item-to-item collaborative filtering," IEEE Internet Computing.* vol. 7, no. 1, pp. 76–80 επιμ. s.l.:s.n.

Linton, F., Charron, A. & Joy, D., 1998. *OWL: A recommender system for organziation-wide learning,* s.l.: In Proceedings of the 1998 Workshop on Recommender Systems 65–69.

Lu, J., 2004. *A Personalized e-Learning Material Recommender System, Proceedings of the 2nd International Conference on Information Technology for Application.* s.l., (ICITA), ISBN 0-646-42313-4.

Lu, L., 2000. *The Diameter of Random Massive Graphs.* s.l., In SODA'00, Proceedings of the Twelth Annual ACM-SIAM Symposium on Discrete Algorithms, pages 912-921. ACM/SIAM Press.

MacWhinney, B., 1996. *The CHILDES System. American Journal of Speech-Language Pathology, 5, 5-14.* s.l., s.n.

Mahmood, T. & Ricci, F., 2009. *Improving recommender systems with adaptive convensional stradegies. In: C. Cattuto, G. Ruffo, F. Menczer (eds.) Hypertext, pp. 73–82.* s.l., ACM.

Malone, T. και συν., 1987. *Intelligent information sharing systems.* s.l., Communications of the ACM, 30(5):390-402.

Manouselis , N. & Vuorikari, R., 2009. *What if Annotations were Reusable: A Preliminary Discussion. In: M. Spaniol, Q. Li, R. Klamma & R. W. H. Lau (eds.).* Aachen, Germany, Proceedings of the 8th International Conference on Web-based Learning (ICWL), pp. 255.

Manouselis, N., Drachsler, H., Verbert, K. & Santos, O. C., 2010. RecSysTEL preface. Issue Procedia Comput. Sci. 1(2), 2773–2774 .

Manouselis, N. & Costopoulou, C., 2007. *Analysis and Classification of Multi-Criteria Recommender Systems. World Wide Web: Internet and Web Information Systems.* s.l., , Special Issue on Multi-channel Adaptive Information Systems on the World Wide Web, 10(4):415-441.

Manouselis, N., Drachsler, H., Verbertd, K. & Duval, E., 2010. *Recommender Systems for Learning.* pp 1-30 επιμ. Us: Springer.

Manouselis, N., Drachsler, H., Vuorikari, R. & Hummel, H., 2011. *Recommender Systems in Technology Enhanced Learning. In: Kantor P, Ricci F, Rokach L, Shapira B (eds).* Springer επιμ. US: Recommender Systems Handbook, pp. 387-415.

Manouselis, N. & Vuorikari, R., 2009. *What if Annotations were Reusable: A Preliminary Discussion. In: M. Spaniol, Q. Li, R. Klamma & R. W. H. Lau (eds.).* Aachen, Germany, Proceedings of the 8th International Conference on Web-based Learning , pp. 255 (ICWL).

Manouselis, N., Vuorikari, R. & VanAssche, F., 2010. Collaborative recommendation of e-learning resources. an experimental investigation. J. Comput. Assist. Learn. 26(4), 227–242(26(4), 227–242).

Masters, J., Dhyastha, T. & Shakouri, A., 2008. *ExplaNet: A Collaborative Learning Tool and Hybrid Recommender System for Student-Authored Explanations.* s.l., Interactive Learning Research, 19(1), 51-74.

Masters, J., Dhyastha, T. & Shakouri, A., 2008. *ExplaNet: A Collaborative Learning Tool and Hybrid Recommender System for Student-Authored Explanations (2008).* s.l., Interactive Learning Research, 19(1), 51-74.

Miller, B. N., Konstan, J. A. & Riedl, J., 2004. *Pocket Lens: toward a personal recommender system.* s.l., ACM Trans. Inf. Syst. 22(3), 437–476.

Mirza, B. J., Keller, B. J. & Ramakrishnan, N., March 2003. *Studying Recommendation Algorithms by Graph Analysis.* s.l., Journal of Intelligent Information Systems, Volume 20 Issue 2, , Pages 131-160.

Montaner, M., Lopez, B. & de la Rosa, J. L., 2003. *A taxonomy of recommender agents on the internet.* s.l., Artif. Intell. Rev. 19, 285–330.

Moreno, G., Martinez-Normand, L. & Boticario, J. G., 2009 . *Research on standards supporting A2UN@: Adaptation and accessibility for All in higher education.* s.l., CEUR Workshop Proc., 495, 1–10.

Muñoz-Merino, P. J., Delgado-Kloos, C. & Fernández-Naranjo, J., 2009. *Enabling interoperability for LMS educational services.* s.l., Comput. Stand. Interfaces, 31, 484–498.

O'Mahony, M., Hurley, N. & Kushmerick, N., 2004. *Collaborative recommendation,* s.l.: A robustness analysis. ACM Trans. Internet Technol. 4(4), 344–377. DOI http://doi. acm.org/10.1145/1031114.1031116.

Pasquale, L., Marco, d. G. & Giovanni, S., 2011. *Content-based Recommender Systems: State of the Art and Trends.* Italy, DOI 10.1007/978-0-387-85820-3_3.

Pazzani, M. & Billsus, D., 1997. *Learning and Revising User Profiles: The Identification of Interesting Web Sites.* s.l., Machine Learning, 27:313-331.

Pearl, J., USA, 1988. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, Morgan Kaufmann, . Στο: s.l.:s.n.

Peis , E., Morales-del-Castillo, J. M. & Delgado-Lópe, 2008. *Analysis of the state of the topic.* s.l., Hipertext.net.

Porcel, C., Tejeda-Lorente, A., Martinez, M. A. & Herrera-Viedma, E., 2012. *A hybrid recommender system for the selective dissemination of research resources in a technology transfer office.* s.l., Information Sciences 184 (1) 1–19.

Quinlan, J., 1993. *Programs for Machine Learning.* s.l., Morgan Kauffman, 235-240.

Quinlan, J. R., 2007. *Centre for Advanced Computing Sciences.* Sydney, New South Wales Institute of Technology.

Reffay, C. & Betbeder, M., 2009. Sharing corpora and tools to improve interaction analysis. In Proceedings of EC-TEL '09, LNCS, vol. 5794, ed. by U. Cress, V. Dimitrova, M. Specht (Springer-Verlag, Berlin, Heidelberg).

Resnick, P. & Varian, H. R., 1997. *Recommender Systems.* s.l., Commun. ACM. 40, 56–58.

Resnick, P. & Varian, H. R., March 1997. *Recommender Systems.* s.l., Communcation of the ACM /Vol. 40, No. 3, 56–58.

Ricci, F., Rokach, . L., Shapira, B. & Kantor, . P. B., 2011. *Recommender Systems Handbook.* New York: Springer.

Rocchio, J., 1971. *Relevance Feedback in Information Retrieval. In: G. Salton (ed.). The SMART System: Experiments in Automatic Document Processing.* NJ, Prentice Hall 313-323.

Rokach, L. & Maimon, O., 2008. *Data Mining with Decision Trees: Theory and Applications.* s.l.:World Scientific Publishing.

Romero, C. & Ventura, S., 2007-b. *Educational data mining: A survey from 1995 to 2005.* s.l., Expert Systems with Applications, 22, 135-146.

Romero, C., Ventura, S., Delgado, J. A. & De Bra , P., 2007. *Personalised Links Recommendation Based on Data Mining in Adaptive Educational Hypermedia Systems.* Crete, Greece, In Proceedings of Second European Conference on Technology Enhanced Learning, ECTEL.

Roy, B., 1996. *Multicriteria Methodology for Decision Aiding.* Dordrecht, Kluwer Academic Publishers.

Said, A., Berkovsky, S., De Luca, E. W. & Hermanns, J., New York, 2011. Challenge on context-aware movie recommendation. Τόμος : CAMRa2011. in Proceedings of the 5th ACM conference on Recommender systems (RecSys '11), .

Salton, G., 1989. *Automatic Text Processing.* Boston, MA, USA, Addison-Wesley Longman Publishing Co., Inc. .

Santos, O. C. & Boticario, J. G., 2011. *Requirements for Semantic Educational Recommender Systems in Formal E-Learning Scenarios.* s.l., Algorithms, 4, 131-154; doi:10.3390/a4030131.

Santos, O. C. & Boticario, J. G., 2008. *Users' Experience with a Recommender System in an Open Source Standards-Based LMS. In Proceedings of 4th Symposium of the WG HCI and UE of the Austrian Computer Society – Usability and HCI for Education and Work (USAB 2008).* Graz, Austria, pp. 185–204.

Santos, O. C. & Boticario, J. G., 2010. *Usability methods to elicit recommendations for semantic educational recommender systems.* s.l., IEEE Learn. Technol. Newsl. 11–12.

Santos, O. C. & Boticario, J. G., 2011(1). *Requirements for Semantic Educational Recommender Systems in Formal E-Learning Scenarios.* s.l., Algorithms, 131-154; doi:10.3390/a4030131.

Santos, O. C. & Boticario, J. G., 2011(2). *TORMES methodology to elicit educational oriented recommendations.* s.l., Lect. Notes Artif. Intell., 6738, 541–543.

Santos, O. C. & Boticario, J. G., 2011. *Educational Recommender Systems and Techniques: Practices and Challenges.* Hershey, PA, USA, IGI Publisher, in press.

Santos, O. C., Mazzone, E., Aguilar, M. J. & Boticari, 2011. *Designing a user interface to managing recommendations for virtual learning communities.* s.l., Int. J. Web Based Commun, in press.

Sarwar, B., Karypis, G., Konstan, J. & Riedl, J., 2000. *Analysis of Recommendation Algorithms for e-Commerce. In: Jhingran, A., Mason, J. M., Tygar, D..* New York, (eds) Proceedings of the 2nd ACM Conference on Electronic Commerce, pp. 158-167.

Sarwar, B. M., Karypis, G., Konstan, J. A. & Riedl, J., May 2001. *Itembased collaborative filtering recommendation algorithms.* s.l.:in Proceedings of the 10th International Conference on World Wide Web (WWW '01), pp. 285–295.

Schafer, J. B., Frankowski, D., Herlocker, J. & Sen, S., 2007. *Collaborative filtering recommender systems, in: P. Brusilovsky, A. Kobsa, W. Nejdl (Eds.), The Adaptive Web.* s.l., pp. 291–324 (Chapter 9).

Schafer, J. B., Konstan, J. A. & Riedl, J., 2001. *E-commerce recommendation applications.* s.l., Data Mining and Knowledge Discovery, 5: 115–152.

Schneider-Hufschmidt, M., Kuhme, T. & Malinowski, U., 1993. *Adaptive user interfaces: Principles and practice. Human Factors in Information Technology.* North-Holland, Amsterdam, (eds).

Sebastiani, F., 2002. *Machine Learning in Automated Text Catigorization. ACM Computing Surveys.* 34(1) επιμ. s.l.:s.n.

Shani, G. & Gunawardana, A., 2011. *Evaluating Recommendation Systems,* s.l.: Springer Science+Business Media.

Shani, G., Heckerman, D. & Brafman , R. I., 2005. AnMDP-based recommender system. Τόμος Journal of Machine Learning Research, vol. 6, pp. 1265–1295.

Shani, G., Heckerman, D. & Brafman, R. I., 2005. *An mdp-based recommender system,* s.l.: Journal of Machine Learning Research 6, 1265–1295.

Siemens, G., August 9, 2010. *What are Learning Analytics?.* [Ηλεκτρονικό] Available at: http://www.elearnspace.org/blog/2010/08/25/what-are-learning-analytics [Πρόσβαση 12 03 2014].

Smyth, B. & McClave, P., 2001. *Similarity vs. diversity,* s.l.: In: ICCBR, pp. 347–361.

Smyth, B., 2007. *Case-based Recommendation. In: Brusilovsky P, Kobsa A, Neidl W(eds) The AdaptiveWeb: Methods and Strategies of Web Personalization.* Lecture Notes in Computer Science, Vol. 4321, pp. 342-376 επιμ. Berlin, Heidelberg, New York: Springer-Verlag.

Stahl, G., 2009. *Studying virtual math teams.* New York, NY, Springer.

Stamper, J. C. και συν., 2010. *PSLC DataShop: A data analysis service for the learning science community. In V. Aleven et al. (Eds.).* Berlin, Proceedings of Intelligent Tutoring Systems (pp. 455-456). Springer.

Steed, C., Sept, 2002. "*Why personalized is the way ahead for learning*" *IT Training.* s.l., s.n.

Su, X. & Khoshgoftaar , T. M., 2009. *A survey of collaborative filtering techniques, Advance in Artificial Intelligence 2009.* s.l., 1-19.

Ternier, S., 2009. *The ariadne infrastructure for managing and storing metadata.* s.l., IEEE Internet Computing, 13(4):18{25.

Tsunoda, M., July 2005. *Javawock: A Java Class Recommender System Based on Collaborative Filtering.* s.l., Proc. of 17th International Conference on Software Engineering and Knowledge Engineering (SEKE2005), pp.491-497.

Verbert, K., 2011. *Dataset-driven Research for Improving Recommender System for Learning.* New York, Proceedings of the 1st Learning Analytics & Knowledge Conference (pp. 44-53).

Verbert, K., Manouselis, N., Drachsler, H. & Duval, E., 2012. *Dataset-Driven Research to Support Learning and Knowledge Analytics.* s.l., Educational Technology & Society, 15 (3), 133–148.

Vuorikari , R. & Berendt , B., 2009. *Study on contexts in tracking usage and attention metadata in multilingual Technology Enhanced Learning.* s.l., Lecture Notes in Informatics, pp. 181, 1654-1663.

Vuorikari, R., Manouselis, N. & Duval, E., January 2008. *Using Metadata for Storing, Sharing, and Reusing Evaluations in Social Recommendation: the Case of Learning Resources. In: Go D.H. & Foo S. (Eds.)*

Social Information Retrieval Systems: Emerging Technologies and ApApplications for Searching the Web Effecti. Hershey, Idea Group Publishing.

W3C, 2004. *Η Κοινοπραξία του Παγκοσμίου Ιστού Εκδίδει το CC/PP 1.0 ως Σύσταση του W3C.* [Ηλεκτρονικό]
Available at: http://www.w3c.gr/press/pressreleases/2004/01/ccpp-pressrelease.el.html
[Πρόσβαση 12 1 2015].

Wikipedia, 2014. *From Wikipedia, the free encyclopedia.* [Ηλεκτρονικό]
Available at: http://en.wikipedia.org/wiki/Slope_One
[Πρόσβαση 25 11 2014].

Wikipedia, 2014. *Wikipedia "k-nearest neighbors algorithm".* [Ηλεκτρονικό]
Available at: http://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm
[Πρόσβαση 28 07 2014].

Wikipedia, August 4, 2006. *AOL search data leakl. Accessed online on 30 June 2010 at.* [Ηλεκτρονικό]
Available at: http://en.wikipedia.org/wiki/AOL_search_data_scandal
[Πρόσβαση 10 03 2014].

Xiaoyuan, S. & Khoshgoftaar, T. M., 2009. *A Survey of Collaborative Filtering Techniques.* USA: s.n.

Yang, Y. & Pedersen, J., 1997. *A Comparative Study on Feature Selection in Text Categorization.* Nashville, TN, In: Proceedings of the Fourteenth International Conference on Machine Learning, p.412-420.

Yu, K. και συν., 2004. *Probabilistic memory-based collaborative filtering.* vol. 16, no. 1, pp. 56–69 επιμ. s.l.:IEEE Transactions on Knowledge and Data Engineering.

Zaiane, O. R., 2002. *Building a Recommender Agent for e-Learning Systems, in: ICCE '02: Proceedings of the International Conference on Computers in Education.* Washington, DC, USA, IEEE Computer Society, ISBN 0-7695-1509-6, 55.