ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

Τμήμα Ψηφιακών Συστημάτων

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

ΔΙΔΑΚΤΙΚΗ ΤΗΣ ΤΕΧΝΟΛΟΓΙΑΣ ΚΑΙ ΨΗΦΙΑΚΑ ΣΥΣΤΗΜΑΤΑ

ΚΑΤΕΥΘΥΝΣΗ: ΗΛΕΚΤΡΟΝΙΚΗ ΜΑΘΗΣΗ

Μεταπτυχιακή Διπλωματική Εργασία:

Παραγωγή Κοινωνικών Συστάσεων σε Διαδικτυακή Πλατφόρμα Ηλεκτρονικής Μάθησης με χρήση Τεχνολογιών Web 2.0

Κονιδάρης Δημήτριος

ΜΕ 09015

Επιβλέπων: Δουλκερίδης Χρήστος, Λέκτορας

Πειραιάς, Ιούνιος 2013

Πανεπιστήμιο Πειραιώς

Ευχαριστίες

Για την ολοκλήρωση των σπουδών μου με την περάτωση της διπλωματικής εργασίας θα ήθελα να ευχαριστήσω όλους τους καθηγητές μου στη διάρκεια του μεταπτυχιακού κύκλου σπουδών. Ειδικότερα, θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή κ. Χρήστο Δουλκερίδη, Λέκτορα, για τις γνώσεις που μου παρείχε, τις συμβουλές και την υπομονή του. Επίσης, θα ήθελα να απευθύνω θερμές ευχαριστίες στον Καθηγητή κ. Δημήτριο Σάμψων για τις γνώσεις του και για την πολύτιμη στήριξή του προς το πρόσωπό μου από την επιλογή μου στο Πρόγραμμα Μεταπτυχιακών Σπουδών έως την ολοκλήρωση των σπουδών. Εν συνεχεία θα ήθελα να εκφράσω τις ευχαριστίες μου στον Αναπληρωτή Καθηγητή κ. Συμεών Ρετάλη καθώς και στην Επίκουρη Καθηγήτρια κ. Φωτεινή Παρασκευά για όλα αυτά που μου προσέφεραν κατά τη φοίτησή μου. Ειδικότερα στην κ. Παρασκευά εκφράζω τις ευχαριστίες μου και για τις πολλές γνώσεις που μου παρείχε στην Παιδαγωγική και στη Διδακτική Μεθοδολογία. Οπωσδήποτε θα ήταν μεγάλη παράλειψη αν δεν ανέφερα τον Αναπληρωτή Καθηγητή κ. Μαρίνο Θεμιστοκλέους ο οποίος μου έδωσε χρήσιμες πληροφορίες για το ΠΜΣ και με βοήθησε στην απόφασή μου να το παρακολουθήσω. Επιπλέον, νιώθω την ανάγκη να διατυπώσω γραπτώς πλέον τις ευχαριστίες μου και προς τους συμφοιτητές μου του ΣΤ΄ κύκλου του ΠΜΣ.

Ακολούθως θα ήθελα να ευχαριστήσω τους αγαπητούς μου μαθητές του Τομέα Πληροφορικής του 1ου ΕΠΑΛ Σπάρτης οι οποίοι δοκίμασαν με ευχαρίστηση την κατασκευασθείσα πλατφόρμα αντιμετωπίζοντας την όλη διαδικασία σαν εκπαιδευτικό παιχνίδι και δεν παραπονέθηκαν για τον επιπλέον φόρτο εργασίας κατά την αξιολόγηση του συστήματος. Επίσης, ευχαριστίες οφείλονται και στους συναδέλφους καθηγητές του 1ου ΕΠΑΛ Σπάρτης για τις διευκολύνσεις που μου παρείχαν καθώς και την στήριξή τους.

Τέλος, θα ήθελα να ευχαριστήσω πάνω απ΄ όλους την αγαπημένη μου σύζυγο Βασιλική για την απέραντη υπομονή της και τη συγκλονιστική ηθική στήριξη που μου έδωσε καθ΄ όλη τη διάρκεια φοίτησης στο μεταπτυχιακό και κυρίως κατά την εκπόνηση της διπλωματικής μου εργασίας. Κατόπιν αυτών, είναι αυτονόητο ότι αφιερώνω την εργασία στην υπέροχη σύζυγό μου.

Πανεπιστήμιο Πειραιώς

Αφιερώνεται στη σύζυγό μου Βασιλική

Πανεπιστήμιο Πειραιώς

Παραγωγή Κοινωνικών Συστάσεων σε Διαδικτυακή Πλατφόρμα Ηλεκτρονικής Μάθησης με χρήση Τεχνολογιών Web 2.0

Περίληψη

Η παρούσα εργασία αναφέρεται στην κατασκευή ενός συστήματος παροχής εξατομικευμένων συστάσεων που ενσωματώνεται σε πλατφόρμα ηλεκτρονικής μάθησης η οποία αφορά το μάθημα «Ανάπτυξη Εφαρμογών σε Προγραμματιστικό Περιβάλλον» και διδάσκεται στην τελευταία τάξη Γενικών Λυκείων.

Αρχικά, παρουσιάζεται η ανάγκη ύπαρξης συστημάτων παραγωγής συστάσεων σε ιστότοπους αφού επεξηγείται αναλυτικά η εν λόγω έννοια. Εν συνεχεία δίνονται παραδείγματα ιστοτόπων που παράγουν εξατομικευμένες συστάσεις προς τους επισκέπτες τους. Ακολούθως αναλύονται οι κατηγορίες μεθόδων παραγωγής συστάσεων, καθώς και τα πλεονεκτήματα και τα μειονεκτήματά τους.

Έπειτα αναλύονται διεξοδικότερα κάποια παραδείγματα ήδη δοκιμασμένων αλγορίθμων όπως αυτών που εφαρμόζονται στο amazon, στο youtube και στο google news. Οι αλγόριθμοι αυτοί ανήκουν στην κατηγορία collaborative filtering από την οποία τελικά επελέγη αλγόριθμος που χρησιμοποιήθηκε στην πλατφόρμα. Σύμφωνα με τη λογική της κατηγορίας αυτής υπολογίζονται οι πιο όμοιοι χρήστες σε σχέση τον τρέχοντα (ενεργό) επισκέπτη της πλατφόρμας και του προτείνονται άρθρα στα οποία έχουν δώσει υψηλή βαθμολογία οι συγκεκριμένοι όμοιοι χρήστες.

Ακολουθεί μια συνοπτική περιγραφή της πλατφόρμας και αναφέρονται οι τρόποι παραγωγής συστάσεων σε αυτή, είτε είναι μερικώς προσωποποιημένοι είτε όχι. Στη συνέχεια δίνεται μια αναλυτική περιγραφή των βημάτων τροποποίησης του συστήματος, στα οποία περιλαμβάνεται και η παράθεση του αλγορίθμου που τελικώς επιλέχθηκε. Κατ' αυτόν τον τρόπο κατέστη δυνατό να μπορεί το σύστημα να παράγει εξατομικευμένες συστάσεις προς τους χρήστες του. Ακολούθως έγινε η αξιολόγηση του συστήματος το οποίο χρησιμοποιήθηκε από μαθητές Β και Γ τάξης του τομέα Πληροφορικής ΕΠΑΛ οι οποίοι έχουν καλή γνώση των αλγοριθμικών δομών. Από την αξιολόγηση προέκυψαν πολύ θετικά αποτελέσματα αφού οι μαθητές σε μεγάλο βαθμό ενθουσιάστηκαν με την πρόταση εξατομικευμένου εκπαιδευτικού υλικού και θεώρησαν ότι είχαν έναν πολύτιμο βοηθό κατά τη μελέτη τους.

Η κατασκευή της πλατφόρμας έγινε με το Σύστημα Διαχείρισης Περιεχομένου JOOMLA 1.5.22 και η παραγωγή των συστάσεων με χρήση PHP, MySQL, HTML καθώς και επεκτάσεων του JOOMLA.

Λέξεις κλειδιά: Συστήματα παραγωγής συστάσεων, συνεργατικό φιλτράρισμα, user-based, συντελεστής Pearson

Πανεπιστήμιο Πειραιώς

Social recommendations in e-learning Web site using Web 2.0 tools

Abstract

This master thesis refers to the construction of a system to provide personalized recommendations embedded in e-learning platform which deals with the subject "Application Development in Programming Environment" and is taught the last class in general high schools.

Initially, it is presented the need for recommendation engines on Internet as explained in detail in this concept. Following are examples of websites that generate personalized recommendations for their guests. Then the categories of recommender systems' methods are analyzed, as well as their advantages and disadvantages.

Moreover some examples, as amazon, youtube and google news, are analyzed that use algorithms from the category of collaborative filtering which finally we choose and develope in our e-learning website. In this category-collaborative filtering-, the similar users are calculated with the current (active) user of platform and recommend to him articles that have given high rank by specific users.

A brief description of the platform is following which describes how the recommendation is produced, either partially personalized or not. Then is given a detailed description of the steps to modify the system, including the code of the algorithm that finally selected. In this way it was possible to the system to generate personalized recommendations to users. After that the evaluation of the system took place by students of B and C class sector IT of vocational school who have good knowledge of algorithmic structures. The evaluation of the system exported very positive results. The students were largely enthusiastic about the recommended individualized educational material and felt that they had a valuable assistant in their study.

The construction of the platform was on the Content Management System JOOMLA 1.5.22 and producing recommendations using PHP, MySQL, HTML and extensions JOOMLA.

Key words: recommender systems, collaborative filtering, user- based, pearson coefficient

Πανεπιστήμιο Πειραιώς

Παραγωγή Κοινωνικών Συστάσεων σε Διαδικτυακή Πλατφόρμα Ηλεκτρονικής Μάθησης με χρήση Τεχνολογιών Web 2.0

Περιεχόμενα

Πανεπιστήμιο Πειραιώς

Πανεπιστήμιο Πειραιώς

Ευρετήριο εικόνων

Πανεπιστήμιο Πειραιώς

1.Εισαγωγή στα Συστήματα Παραγωγής Συστάσεων

1.1 Αναζήτηση πληροφοριών στο Διαδίκτυο

Κατά τη δεκαετία του 1990 ο Παγκόσμιος Ιστός έκανε δειλά δειλά τα πρώτα του βήματα όντας ουσιαστικά στα σπάργανα και δεν θύμιζε σε τίποτα το Διαδίκτυο που γνωρίζουμε σήμερα. Με τη γιγάντωση του Διαδικτύου υπήρξε συνεπακόλουθη τεράστια αύξηση της πληροφορίας που διακινείται σε αυτό. Συνεπώς η αναζήτηση πληροφοριών έγινε πολύ πιο δύσκολη και πιο σύνθετη διαδικασία. Έτσι παρουσιάστηκε η ανάγκη αντιμετώπισης του προβλήματος με τη δημιουργία μηχανών αναζήτησης. Η πρώτη μηχανή αναζήτησης, με τη μορφή που γνωρίζουμε τις μηχανές σήμερα, δημιουργήθηκε το 1993 από τον Mathew Gray και ονομάστηκε Wandex (http://encyc.org/wiki/Wandex).

Στη συνέχεια βέβαια δημιουργήθηκαν πολλές μηχανές αναζήτησης με αποτελεσματικότερους αλγορίθμους όπως η Yahoo!, η Lycos, η Altavista, κ.α.. Γνωστότερη και σημαντικότερη όλων αναδείχτηκε η Google που χρησιμοποίησε τον αλγόριθμο PageRank που δημιούργησαν οι μετέπειτα ιδρυτές της Λάρι Πέιτζ και Σεργκέι Μπριν, (http://www.Google .com/intl/el/about/company/history/)

Είναι λοιπόν δεδομένο ότι στο Διαδίκτυο μπορεί να βρει κάποιος ασύλληπτο όγκο πληροφοριών με κίνδυνο να μην δύναται να εντοπίσει αυτό που επιθυμεί. Γι' αυτό οι μηχανές αναζήτησης προσπαθούν να δώσουν το καλύτερο αποτέλεσμα χωρίς φυσικά να το καταφέρνουν πάντα. Συνέπεια αυτής της κατάστασης είναι να συνεχίζονται με αμείωτο ρυθμό οι προσπάθειες βελτιστοποίησης αλγορίθμων αναζήτησης οι οποίες δυσχεραίνονται από την προαναφερόμενη τεράστια αύξηση του όγκου των πληροφοριών που διακινείται στο Διαδίκτυο. Ενδεικτικό του μεγέθους του προβλήματος είναι το ότι ακόμα και η Google ουσιαστικά παρέχει στοιχεία μόνο για ένα ποσοστό των ιστοτόπων του Διαδικτύου που είναι όμως δύσκολο να υπολογιστεί (https://en.wikipedia.org/wiki/Google_Search).

Πανεπιστήμιο Πειραιώς

1.2 Ορισμός του προβλήματος

1.2.1 Συστήματα παραγωγής συστάσεων

Με το πέρασμα των ετών παρουσιάστηκαν περισσότερες και πιο δύσκολα αντιμετωπίσιμες ανάγκες. Συγκεκριμένα, σε πολλές περιπτώσεις δεν έπρεπε απλά να παρέχεται στους χρήστες μία λίστα δεδομένων και στοιχείων αλλά απαιτούνταν ακόμα πιο εξειδικευμένες πληροφορίες (Ansari, et al, 2000, Bogers, 2010).

Π.χ. κατά τη διαδικασία αναζήτησης δωματίων σε ξενοδοχεία μιας πόλης μέσω ενός συστήματος, θα ήταν πολύ χρήσιμο αυτό το σύστημα να μπορούσε να δίνει επιπλέον πληροφορίες για το ξενοδοχείο. Επίσης, σε ένα άλλο παράδειγμα, ο χρήστης μπορεί να ζητήσει ξενοδοχείο τριών αστέρων κοντά σε μία περιοχή του Λονδίνου. Θα είναι απολύτως επιθυμητό και ιδιαιτέρως χρήσιμο να προτείνει το σύστημα παρόμοια ξενοδοχεία (δηλαδή τριών αστέρων) εντός ή πλησίον της περιοχής αυτής απαλλάσσοντας το χρήστη από την ανάγκη να ελέγξει κάθε περίπτωση χωριστά. Τέλος, μία ακόμα κλασική περίπτωση ιστοτόπου που λειτουργεί με αυτόν τον τρόπο είναι το www.booking.com το οποίο προτείνει ξενοδοχεία βάσει της αξιολόγησης που κάνουν οι χρήστες του.

Έτσι λοιπόν προέκυψε η ανάγκη παραγωγής κοινωνικών συστάσεων δηλαδή η δυνατότητα άντλησης κάποιων πληροφοριών από ιστοτόπους που θα βοηθήσουν στη λήψη αποφάσεων. Ένα επιπλέον παράδειγμα τέτοιας περίπτωσης είναι ένας ιστότοπος που παρέχει τη δυνατότητα βαθμολόγησης ταινιών και τις προτείνει αναλόγως (Bogers, 2000). Σε αυτόν ένας επισκέπτης θέλει να δει κάποιες ταινίες και αναζητά τη βαθμολογία αυτών ούτως ώστε να βοηθηθεί στην επιλογή του. Το πρόβλημα έγκειται όχι τόσο στις ταινίες που έχουν αξιολογηθεί από κάποιους, έστω και λίγους, επισκέπτες αλλά, κυρίως, σε αυτές που δεν έχουν λάβει κάποια αξιολόγηση και, συνεπώς, είναι χαμηλά σε οποιαδήποτε βαθμολογική κλίμακα. Παράδειγμα συστήματος παραγωγής συστάσεων είναι το IMDB (Internet Movie Data Base, www.imdb.com) που προτείνει πληροφορίες για ταινίες και τηλεοπτικές σειρές.

Μία ακόμα πολύ σημαντική εφαρμογή με παραγωγή συστάσεων είναι το σύστημα των πανεπιστημίων στις ΗΠΑ που βοηθά τους φοιτητές να επιλέξουν μαθήματα από το πρόγραμμα σπουδών (Koutrika, et al, 2009 a).

Τα ανωτέρω παραδείγματα είναι μερικές μόνο περιπτώσεις από τις πάμπολλες στις οποίες καταδεικνύεται η ανάγκη παροχής συστάσεων.

Πανεπιστήμιο Πειραιώς

Στην παρούσα εργασία θα χρησιμοποιηθεί μία e-learning πλατφόρμα που παρέχει τη δυνατότητα δημιουργίας περιεχομένου από το χρήστη και υποστηρίζει τη χρήση ετικετών και τη βαθμολόγηση των άρθρων. Το πρόβλημα είναι ο τρόπος παραγωγής συστάσεων (recommendations) του εκπαιδευτικού υλικού προς τους χρήστες βάσει της δραστηριότητας των άλλων χρηστών. Ουσιαστικά επιδιώκεται να βοηθηθεί ο χρήστης στην επιλογή εκπαιδευτικού υλικού που θεωρείται καλό (δημοφιλές).

Θα πρέπει να σημειωθεί ότι υπάρχουν διάφορες κατηγορίες μεθόδων για την παραγωγή συστάσεων οι οποίες θα παρουσιαστούν στη συνέχεια

1.2.2 Ιστορική αναδρομή

Είναι δύσκολο να προσδιοριστεί η ακριβής ημερομηνία της πρώτης εμφάνισης της νέας ιδέας αλλά ουσιαστικά τα πρώτα συστήματα παραγωγής συστάσεων δημιουργήθηκαν στις αρχές της δεκαετίας του 1990. Πιο συγκεκριμένα, το Tapestry document filtering system θεωρείται ότι είναι το πρώτο σύστημα που χρησιμοποιεί μεθόδους collaborative filtering το 1992 (Goldberg, et al.,1992). Στο σύστημα αυτό οι χρήστες μπορούσαν να δημιουργήσουν διαφορετικά φίλτρα για εισερχόμενα μηνύματα (e-mails) ή usenet. Επίσης, το 1994 το GrouLEns research project του Πανεπιστημίου της Μινεσότα δημιούργησε ένα αυτόματο σύστημα παραγωγής συστάσεων για νέα UseNet (Resnick, et al., 1994). Ο αλγόριθμος αυτός μπορούσε να προτείνει κάποιες ειδήσεις βασιζόμενος στις βαθμολογίες των άλλων.

Την ίδια περίοδο ο Upendra Shardanard, κάνοντας τη διπλωματική του εργασία στο MIT, δημιούργησε, με την επιβλέπουσα καθηγήτρια Pattie Maes ένα σύστημα παραγωγής προσωποποιημένων συστάσεων σε μουσική ονόματι Ringo (Shardanand & Maes, 1995). Επίσης οι Grouplens, ερευνητική ομάδα του Τμήματος Επιστήμης Υπολογιστών του Πανεπιστημίου της Μινεσότα, δημιούργησαν το πρώτο διάσημο σύστημα παραγωγής συστάσεων για το γνωστό ιστότοπο Amazon.com (online βιβλιοπωλείο). Το σύστημα ήταν εύκολο να κατανοηθεί από τους χρήστες. Στην εποχή μας η Amazon παρέχει προσωποποιημένες συστάσεις (Linden, et al., 2003).

Γενικά, τα συστήματα παραγωγής συστάσεων αναλύουν τα χαρακτηριστικά του χρήστη (user profiles), το περιεχόμενο των αντικειμένων και τις συνδέσεις μεταξύ τους προσπαθώντας να προβλέψουν τη μελλοντική συμπεριφορά του χρήστη. Η βασική ιδέα είναι ότι «οι άνθρωποι που συμφώνησαν στο παρελθόν με κάτι, πιθανότατα θα συμφωνήσουν και στο μέλλον με κάτι παρόμοιο» (Resnick, et al.,

Πανεπιστήμιο Πειραιώς

1994). Αυτό που εμφανίζεται σαν αποτέλεσμα είναι μία λίστα προτεινόμενων αντικειμένων που πιθανόν να αρέσουν στο χρήστη. Αυτή η διαδικασία φυσικά μπορεί να χρησιμοποιηθεί και για εμπορικούς ή ακαδημαϊκούς σκοπούς.

Μεγάλο ενδιαφέρον και ενδεικτικό της τεράστιας σημασίας των συστημάτων παραγωγής συστάσεων έχει το βραβείο Nexflix (Netflix Prize) για τον καλύτερο collaborative filtering αλγόριθμο που αφορά πρόβλεψη βαθμολογίας ταινιών (http://en.wikipedia.org/wiki/Netflix_Prize). Το βραβείο ανέρχεται σε 1,000,000 USD (δολάρια ΗΠΑ) και δόθηκε από την Netflix στην ομάδα BellKor's Pragmatic Chaos στις 21-9-2009 για τον αλγόριθμο που δημιούργησαν. Σημειωτέον η Netflix είναι αμερικανική εταιρεία παροχής βίντεο, DVD και Blu-ray δίσκων που ιδρύθηκε το 1997 και είχε έσοδα για το 2011 που ανέρχονταν στο 1,5 δις USD.

1.2.3 Δεδομένα εισαγωγής

Τα συστήματα παραγωγής συστάσεων χρειάζονται συνήθως μεγάλο όγκο δεδομένων για να λειτουργήσουν. Αυτά τα δεδομένα αφορούν το προφίλ του χρήστη, τα ίδια τα αντικείμενα που επιλέγονται ή ψηφίζονται και πώς αυτά θα βοηθήσουν άλλους χρήστες στις επιλογές τους. Για την ολοκληρωμένη λειτουργία απαιτείται η ενεργή συμμετοχή του χρήστη. Συγκεκριμένα, ο χρήστης πρέπει να αξιολογήσει πρώτα κάποια αντικείμενα πριν το σύστημα είναι σε θέση να παράγει συστάσεις. Στις περισσότερες περιπτώσεις ζητείται από το χρήση να αξιολογήσει έναν άρθρο ή προϊόν. Ουσιαστικά, κάθε ενέργεια των χρηστών επηρεάζει τη συνολική λειτουργία του συστήματος. Μερικά συστήματα παράγουν συστάσεις σε πραγματικό χρόνο (real time) και ο χρήστης μπορεί να δει άμεσα τα αποτελέσματα των ενεργειών αφού π.χ. η αξιολόγηση που δίνει για ένα προϊόν, θα αλλάξει και τη συμπεριφορά του συστήματος δηλαδή αν θα προτείνει ή όχι το προϊόν αυτό (Prekopcsák, 2007).

Στη συνολική διαδικασία η χρησιμότητα των ετικετών (tagging) είναι αναμφισβήτητη όσον αφορά την αναζήτηση πληροφοριών. Όμως, παρουσιάστηκε η ανάγκη για μία αυτόματη μέθοδο-σύστημα που θα προτείνει περιεχόμενο στο χρήστη και θα πρέπει να μπορεί να προτείνει ένα αντικείμενο που θα ήταν δύσκολο να αναζητηθεί μέσω ενός ερωτηματολογίου (query) λόγω μη εξοικείωσης με τις κατάλληλες λέξεις-κλειδιά. Επίσης, η μέθοδος αυτή θα πρέπει να βοηθά στην ανακάλυψη σχετικών αντικειμένων για τα οποία δεν υπάρχει ο χρόνος να εντοπιστούν μέσω απλής εξερεύνησης-αναζήτησης (Prekopcsák, 2007).

Πανεπιστήμιο Πειραιώς

1.3 Στόχοι και σκοπός της εργασίας

Στην παρούσα διπλωματική εργασία θα παρουσιαστεί, όπως αναφέρθηκε ανωτέρω, ένα σύστημα παραγωγής συστάσεων σε πλατφόρμα ηλεκτρονικής μάθησης που αφορά το μάθημα «Ανάπτυξη Εφαρμογών σε Προγραμματιστικό Περιβάλλον» το οποίο διδάσκεται στην τελευταία τάξη του Γενικού και του Εσπερινού Λυκείου και εξετάζεται Πανελλαδικώς στην Τεχνολογική κατεύθυνση.

Σκοπός της εργασίας είναι να εφαρμόσει τεχνικές εξαγωγής συστάσεων και, τελικά, να προτείνει εκπαιδευτικό υλικό στους χρήστες της πλατφόρμας. Η πρόταση εκπαιδευτικών αντικειμένων θα γίνει βάσει των επιλογών του επισκέπτη της πλατφόρμας. Η καταβληθείσα προσπάθεια αποσκοπεί στο να ¨κατανοήσει¨ το σύστημα τη συμπεριφορά ενός χρήστη και να του προτείνει θέματα που άπτονται των ενδιαφερόντων του.

Συμπερασματικά, τελική επιδίωξη είναι η παροχή βοήθειας προς τους χρήστες με την πρόταση των άρθρων ούτως ώστε ο ιστότοπος να φανεί χρήσιμος. Είναι εντελώς ανεπιθύμητο να αντιμετωπίζεται η συγκεκριμένη πλατφόρμα ως ένας ακόμα ιστότοπος που προσφέρει εκπαιδευτικό υλικό χωρίς την παραμικρή καθοδήγηση στους επισκέπτες του.

Η κατασκευή της πλατφόρμας έγινε με το Σύστημα Διαχείρισης Περιεχομένου (Content Management System) JOOMLA 1.5.22 και η παραγωγή των συστάσεων με χρήση PHP, MySQL, HTML καθώς και επεκτάσεων του JOOMLA. Αναλυτική περιγραφή του ανωτέρω συστήματος υπάρχει σε επόμενο κεφάλαιο.

Πριν όμως περιγραφεί αναλυτικά το ανωτέρω σύστημα, θα παρατεθεί επισκόπηση των τεχνικών παραγωγής συστάσεων με τα πλεονεκτήματα και τα μειονεκτήματά τους καθώς επίσης θα δοθούν αναλυτικά παραδείγματα τέτοιων συστημάτων.

Πανεπιστήμιο Πειραιώς

2. Βιβλιογραφική ανασκόπηση τεχνικών παραγωγής συστάσεων και παραδείγματα

2.1 Παραδείγματα ιστότοπων με παραγωγή συστάσεων

Ακολουθούν κάποια πολύ χαρακτηριστικά παραδείγματα ιστοτόπων που χρησιμοποιούν τεχνικές παραγωγής συστάσεων

2.1.1 Περιγραφή λειτουργίας ηλεκτρονικού συστήματος

Το Amazon, το πιο δημοφιλές ίσως ηλεκτρονικό κατάστημα, ξεκίνησε τη λειτουργία του ως βιβλιοπωλείο το 1995. Ο ιδρυτής του Jeff Bezos κατάλαβε σύντομα ποια είναι τα πλεονεκτήματα του Διαδικτύου και προτίμησε να προσφέρει μια τεράστια βάση δεδομένων με εκατομμύρια τίτλους βιβλίων. Έτσι, προσείλκυσε γρήγορα ένα μεγάλο και μορφωμένο κοινό και έγινε ο φόβος και ο τρόμος των παραδοσιακών βιβλιοπωλείων που φυσικά δεν μπορούσαν να ανταγωνιστούν σε ποικιλία το Amazon (http://en.wikipedia.org/wiki/Amazon.com).

Όπως αποδείχθηκε στη συνέχεια ούτε στην εξυπηρέτηση κατάφερε κανείς να το ξεπεράσει. Το Amazon οργανώθηκε με αριστοτεχνικό τρόπο και προσέφερε από πολύ νωρίς στους πελάτες του πολλές υπηρεσίες που, ακόμη και σήμερα, ελάχιστα άλλα ηλεκτρονικά καταστήματα μπορούν να καυχηθούν ότι παρέχουν. Έτσι, απέκτησε γρήγορα εκατομμύρια πελάτες στους οποίους προσπαθεί να προωθήσει νέα προϊόντα. Σήμερα είναι κοινώς αποδεκτό ότι το πετυχημένο αυτό βιβλιοπωλείο έχει εξελιχθεί σε πολυκατάστημα που προσφέρει μεγάλη ποικιλία ειδών (βιβλία, υπολογιστές, είδη ένδυσης-υπόδησης, παιχνίδια, κ.α.) με άριστη εξυπηρέτηση και ελαχιστοποίηση των μεταφορικών εξόδων .

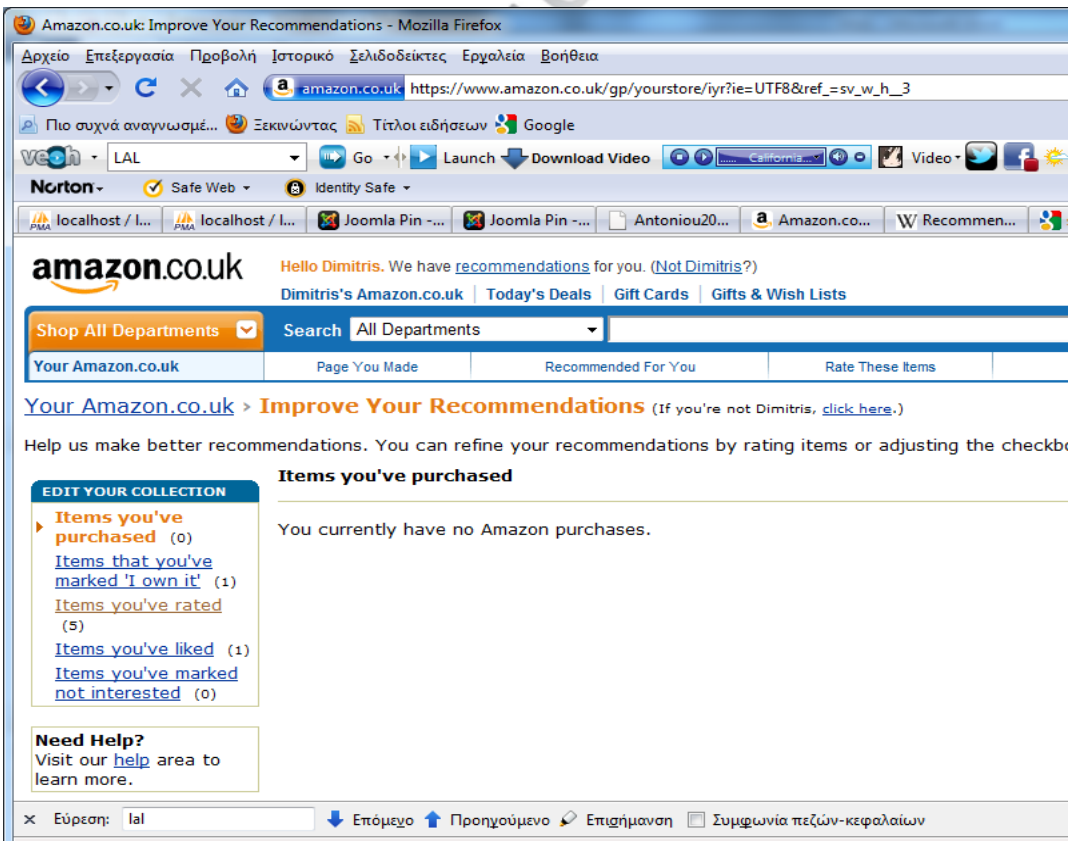
Το Amazon δέχεται εκατομμύρια επισκέψεις μηνιαίως και χρησιμοποιεί σύστημα παραγωγής προσωποποιημένων συστάσεων για να διευκολύνει τους «επισκέπτες» του. Παρόλο που είναι φιλικός σαν ιστότοπος, οι υπεύθυνοι φροντίζουν να παρέχουν προσωποποιημένες συστάσεις σε κάθε χρήστη για να μην δυσχεραίνεται στην αναζήτηση λόγω των δεκάδων χιλιάδων προϊόντων που προσφέρονται (Creel, et al., 2010).

Με το σύστημα που παρέχει το Amazon ο επισκέπτης (χρήστης) μπορεί να αναζητήσει αυτό που επιθυμεί (π.χ. βιβλία) και να δει τη βαθμολογία του καθενός αντικειμένου (item). Ευνόητο είναι πως η βαθμολογία κάθε αντικειμένου έχει

Πανεπιστήμιο Πειραιώς

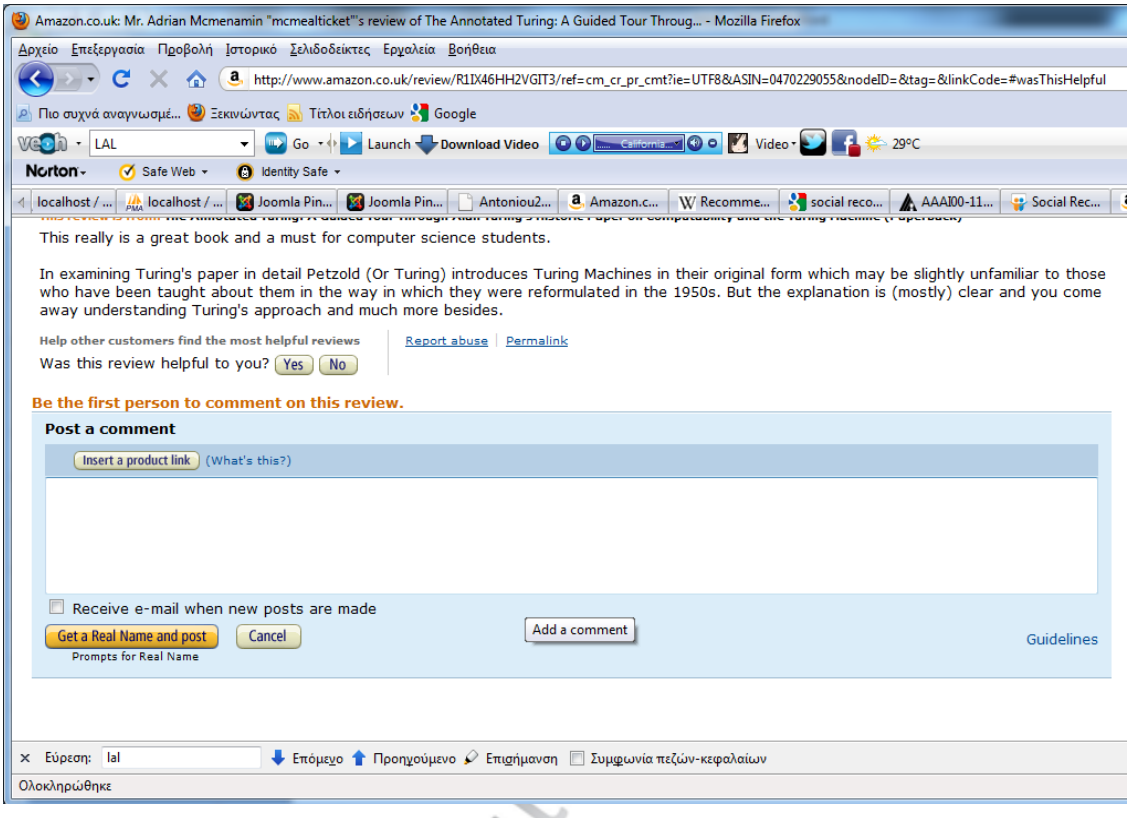
σημειωθεί από άλλους επισκέπτες. Επίσης, ο χρήστης προτρέπεται να δώσει τη δική του θετική ή αρνητική βαθμολογία σε προϊόντα για τα οποία έχει άποψη. Έτσι ο χρήστης γνωστοποιεί στο σύστημα τα ενδιαφέροντά του και τις προτιμήσεις του και, κατόπιν, το σύστημα μπορεί να του συστήσει παρόμοια προϊόντα. Π.χ αν κάποιος αξιολογήσει θετικά 2-3 βιβλία Βάσεων Δεδομένων, τότε το σύστημα θα του προτείνει αντίστοιχα βιβλία με υψηλή βαθμολογία. Αυτό έχει σαν αποτέλεσμα σημαντικό κέρδος χρόνου που δεν θα υπήρχε με τη σειριακή ή οποιουδήποτε άλλου είδους αναζήτηση όλων των τίτλων. Επιπλέον, είναι πολύ σημαντική η άποψη ενός χρήστη για ένα προϊόν αν σημειώσει σε ειδικό κουτάκι (check box) ότι το έχει ήδη αγοράσει. Τέλος, εκτός από τη βαθμολόγηση (rating) υπάρχει και η δυνατότητα χαρακτηρισμού αρεσκείας (με like) ή ως μη ενδιαφέρον (not interesting) (Linden, et al, 2003).

Σημειωτέον ότι ως αναγνωριστικό εισόδου χρησιμοποιείται ένας λογαριασμός ηλεκτρονικού ταχυδρομείου σε συνδυασμό με έναν κωδικό (password). Κατ' αυτόν τον τρόπο κάθε χρήστης είναι διακριτός και έχει το δικό του λογαριασμό. Αυτό βοηθάει ιδιαιτέρως στην παροχή συγκεντρωτικών βοηθητικών στοιχείων από το σύστημα δηλαδή ο χρήστης μπορεί ανά πάσα στιγμή να δει πόσα και ποια προϊόντα έχει βαθμολογήσει, πόσα και ποια έχει χαρακτηρίσει, κλπ. όπως φαίνεται στη συνέχεια.



Εικόνα 1. Παράδειγμα χρήσης του Amazon

Αξίζει να αναφερθεί ότι ο χρήστης μπορεί να κάνει σχολιασμό σε ένα προϊόν γράφοντας οτιδήποτε επιθυμεί γι' αυτό.



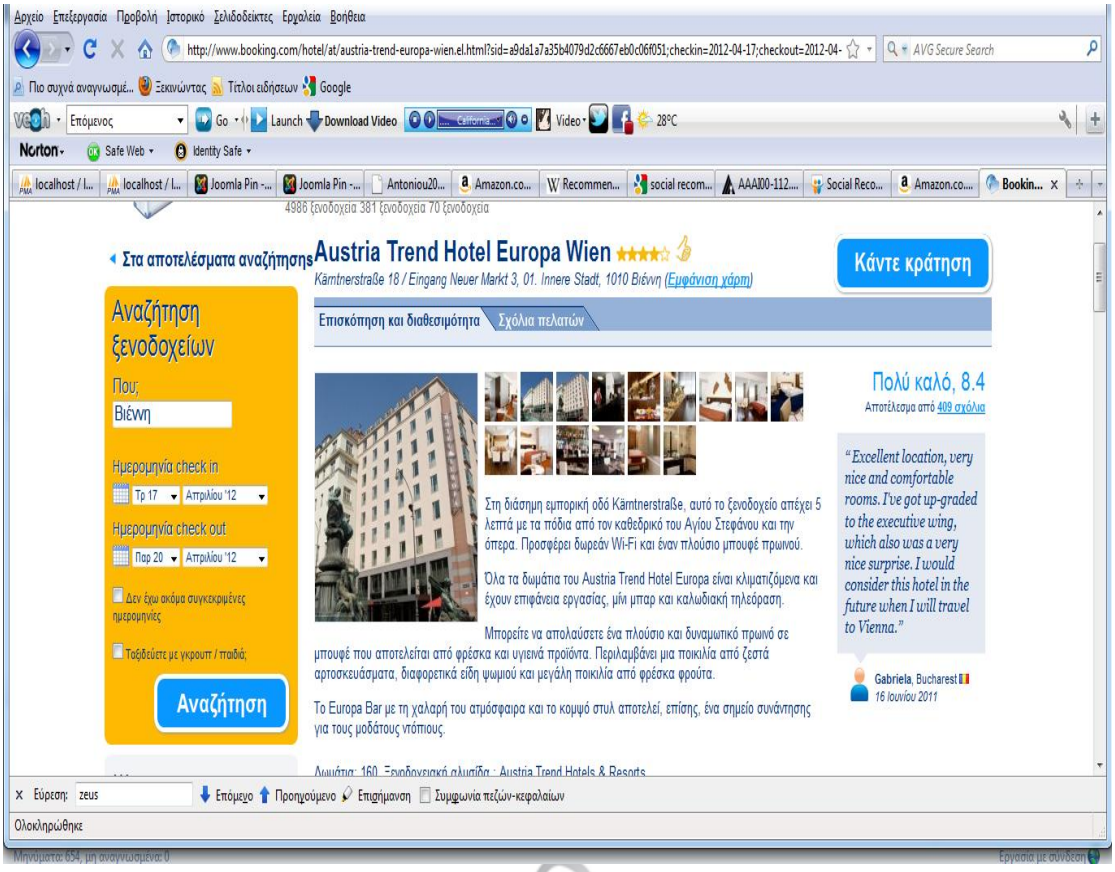
Εικόνα 2. Παράδειγμα σχολιασμού στο Amazon

Γενικά, το εν λόγω σύστημα προτρέπει το χρήστη για την παροχή προς αυτό όσο το δυνατόν μεγαλύτερου όγκου πληροφοριών γιατί αυτό βοηθάει στη βελτιστοποίηση των συστάσεων.

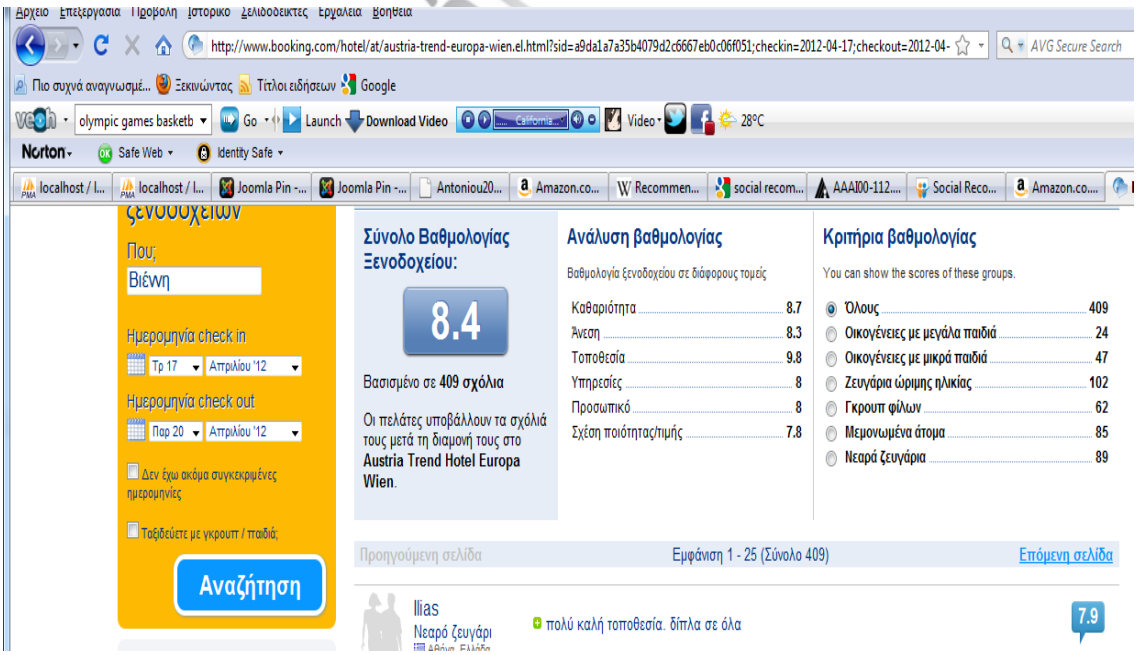
2.1.2 Παράδειγμα συστήματος κράτησης ξενοδοχείων

Ο ιστότοπος www.booking.com είναι από τους γνωστότερους και δημοφιλέστερους παγκοσμίως για την κράτηση δωματίων σε ξενοδοχεία. Ο χρήστης πρέπει να επιλέξει την τοποθεσία και τις ημερομηνίες για να λάβει έναν αριθμό προταθέντων ξενοδοχείων. Επιπλέον, ο χρήστης μπορεί να ορίσει κάποια κριτήρια όπως εύρος τιμών, κατηγορία ξενοδοχείου (πόσα αστέρια έχει), τύπο ξενοδοχείου (κλασικό, ξενώνας, διαμέρισμα, κλπ.), παροχές (Διαδίκτυο, πισίνα, γυμναστήρια, χώρος στάθμευσης, κλπ). Οι πληροφορίες που λαμβάνει ο χρήστης είναι αναλυτικές και περιλαμβάνουν ακόμα και σχόλια χρηστών όπως φαίνεται στη συνέχεια.

Πανεπιστήμιο



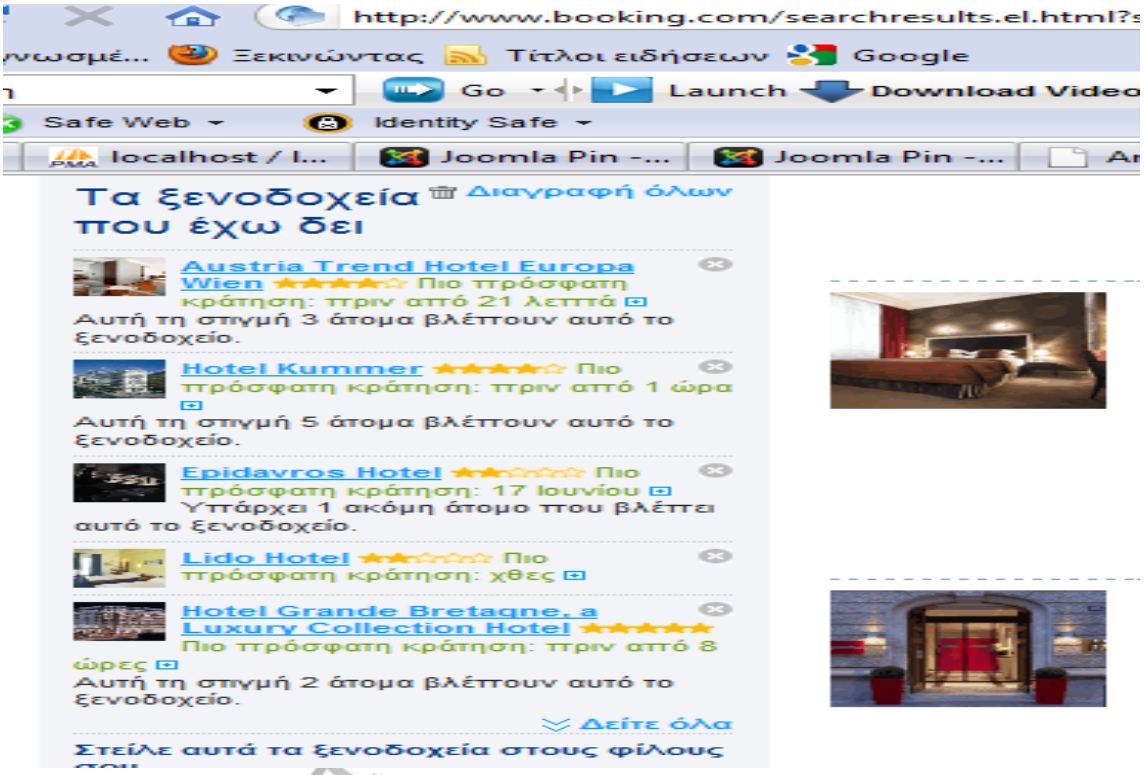
Εικόνα 3. Πρόταση ξενοδοχείου στο www.booking.com



Εικόνα 4. Ανάλυση βαθμολογίας χρηστών στο booking.com

Τα σχόλια είναι πολύ σημαντική σύσταση αφού πολλές φορές περιλαμβάνουν πολλές λεπτομέρειες προς τους υποψήφιους πελάτες. Συγκεκριμένα, όταν κάποιος κλείσει κάποιο ξενοδοχείο και μείνει σε αυτό, λαμβάνει ένα ηλεκτρονικό μήνυμα στο οποίο πρέπει να απαντήσει σε κάποιες ερωτήσεις για να προκύψει η τελική βαθμολογία. Κατά το στάδιο αυτό έχει τη δυνατότητα να προσθέσει τα σχόλια που επιθυμεί.

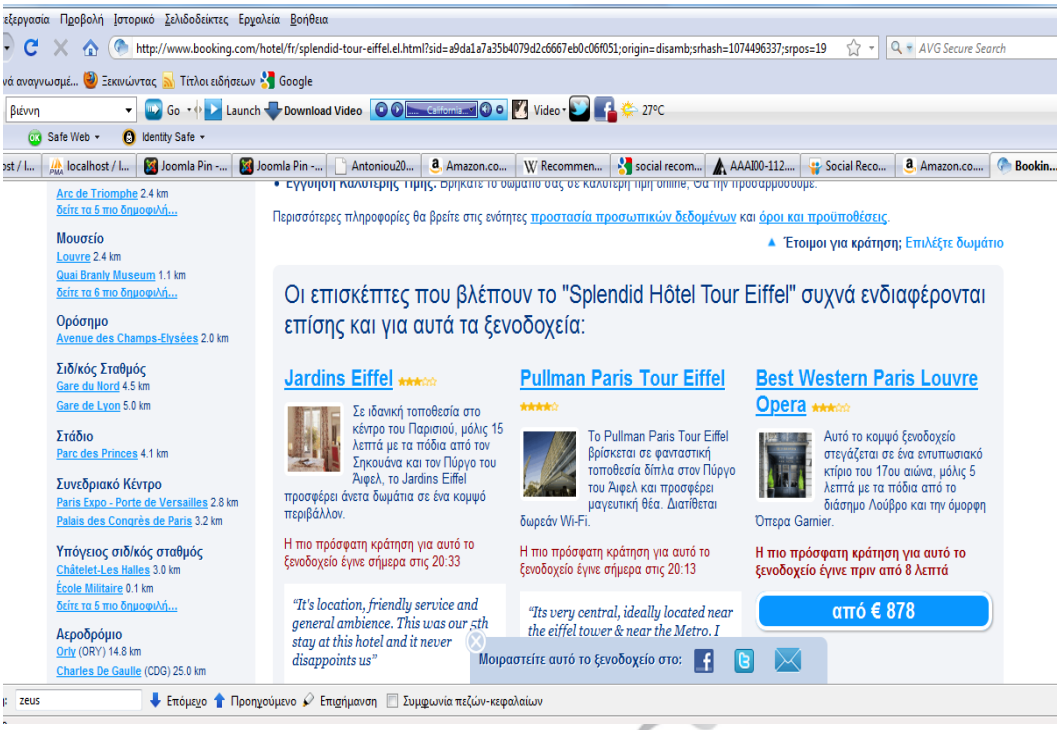
Επιπροσθέτως, ο χρήστης μπορεί να δει μία λίστα με τα ξενοδοχεία που έχει ελέγξει όπως φαίνεται στη συνέχεια.



Εικόνα 5. Λίστα ελεγχθέντων ξενοδοχείων στο booking.com

Φυσικά στην περίπτωση που γίνει μία κράτηση, ο χρήστης θα λάβει επιβεβαίωση όχι μόνο στην οθόνη του υπολογιστή του αμέσως μετά την ολοκλήρωση της κράτησης αλλά και ηλεκτρονικό μήνυμα από την booking.com.

Πιο πολύ ενδιαφέρον, στην εξεταζόμενη περίπτωση των συστημάτων παραγωγής συστάσεων, παρουσιάζει η πρόταση παρόμοιων ξενοδοχείων με αυτά που έχει ήδη εξετάσει ο χρήστης όπως φαίνεται στην επόμενη εικόνα.



Εικόνα 6. Λίστα παρόμοιων ξενοδοχείων στο booking.com

Στην περίπτωση αυτή το σύστημα εξετάζει ξενοδοχεία με παρόμοια χαρακτηριστικά (κατηγορία, τιμή, τοποθεσία, κλπ) και προτείνει στο χρήστη κάτι ανάλογο γλιτώνοντάς τον από τη χρονοβόρα διαδικασία αναζήτησης και ελέγχου ενός μεγάλου αριθμού ξενοδοχειακών μονάδων.

Πανεπιστήμιο Πε

2. 2 Μελέτη ιστότοπων με παραγωγή συστάσεων

2.2.1 Σύστημα παραγωγής συστάσεων σε Πανεπιστήμιο

Στη συνέχεια θα μελετηθεί το CourseRank (Koutrika,et al, 2009 b) που είναι ένα σύστημα παραγωγής συστάσεων σε Πανεπιστήμια των ΗΠΑ. Ως γνωστόν, στα αμερικάνικα πανεπιστήμια το σύστημα επιλογής μαθημάτων είναι πολύ διαφορετικό από αυτό της πατρίδας μας όπου ο κάθε φοιτητής εξετάζεται σε ένα, συνήθως, μεγάλο ποσοστό υποχρεωτικών μαθημάτων και έχει δικαίωμα να διαλέξει κάποια μαθήματα επιλογής από μία σχετικώς περιορισμένη λίστα για να λάβει το πτυχίο του.

Αντιθέτως, στις ΗΠΑ το σύστημα είναι ριζικά διαφορετικά. Π.χ. στο Πανεπιστήμιο Stanford (ΗΠΑ) υπάρχουν διαθέσιμα περίπου 18.600 μαθήματα (στοιχεία Ιανουαρίου 2009) από τα οποία πρέπει να επιλέξει ο φοιτητής αυτά που θεωρεί ότι είναι καλύτερα για τον ίδιο (Koutrika, et al, 2009 b). Στην περίπτωση αυτή ένα αποτελεσματικό σύστημα παραγωγής συστάσεων δεν είναι απλώς βοηθητικό αλλά αποτελεί προφανέστατα αδήριτη ανάγκη. Οι περιορισμοί που υφίστανται κατά την επιλογή των μαθημάτων δυσχεραίνουν ακόμα περισσότερο την κατάσταση. Για παράδειγμα, υπάρχουν μαθήματα που πρέπει να επιλεγούν σε συγκεκριμένες περιόδους (τετράμηνα ή εξάμηνα) και σε συγκεκριμένη σειρά. Χρησιμοποιώντας το CourseRank οι φοιτητές μπορούν να αναζητήσουν τα μαθήματα που ταιριάζουν με τα ενδιαφέροντά τους και να πάρουν προσωποποιημένες συστάσεις (Koutrika, et al, 2009 b). Αξίζει να σημειωθεί ότι το σύστημα έχει τεράστια ανταπόκριση από τους φοιτητές αφού χρησιμοποιείται από ένα ποσοστό της τάξης του 85%.

2.2.2 Γενική επισκόπηση του CourseRank

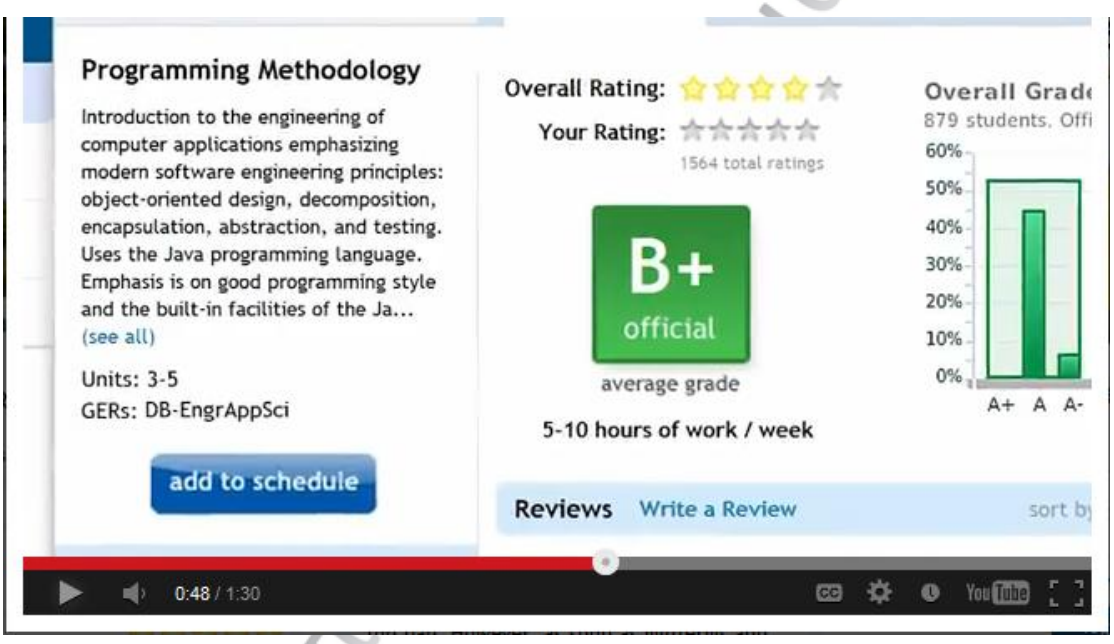
Το CourseRank έχει αρκετά χαρακτηριστικά που το διακρίνουν από άλλους ιστοτόπους που παρέχουν συστάσεις ή δίνουν κάποιας μορφής αξιολόγησης. Καταρχάς, έχει πρόσβαση όχι μόνο στα επίσημα δεδομένα του Πανεπιστημίου (περιγραφές μαθημάτων, προγράμματα, αποτελέσματα αξιολόγησης μαθημάτων που διεξάγεται από το ίδιο το Πανεπιστήμιο) αλλά και σε βάσεις δεδομένων που δημιουργούνται από δεδομένα όπως βαθμολογίες μαθημάτων, σχόλια, ερωτήσεις που συνεισφέρουν οι χρήστες (Koutrika, et al, 2009 b).

Επιπλέον έχει να αντιμετωπίσει ένα πολύ πλούσιο όγκο δεδομένων και πληροφοριών και αυτό από μόνο του είναι μία σοβαρή πρόκληση. Δηλαδή θα

Πανεπιστήμιο Πειραιώς

πρέπει να προκύπτουν χρήσιμες συστάσεις μέσα από μία τεράστια βάση δεδομένων με σημαντικά υψηλή πολυπλοκότητα (Bercovitz, et al, 2009).

Τέλος, σε αντίθεση με άλλους ιστοχώρους που είναι ανοικτοί στο ευρύ κοινό και συνήθως έχουν ένα τύπο χρήστη, το CourseRank παρέχεται στα μέλη της ακαδημαϊκής κοινότητας του Stanford που περιλαμβάνει τρεις διακριτούς τύπους χρηστών: α) Φοιτητές (προπτυχιακοί και μεταπτυχιακοί), β) Καθηγητές και λοιποί διδάσκοντες που θέλουν να ελέγχουν τα σχόλια στα μαθήματά τους και να κάνουν σύγκριση αυτών με των άλλων μαθημάτων, γ) Προσωπικό που καταχωρίζει τα απαραίτητα στοιχεία και συμβουλεύει τους φοιτητές πάνω στο σχεδιασμό του προγράμματος που θα ακολουθήσουν μέσα σε αυτό το λαβύρινθο των 18.600 μαθημάτων.



Εικόνα 7. Προτεινόμενο μάθημα μέσω του CourseRank

Στο CourseRank χρησιμοποιείται το Coursecloud το οποίο είναι ένα σύννεφο ετικετών (tag cloud) όπου οι ετικέτες (tags) είναι οι αντιπροσωπευτικότερες ή σημαντικότερες λέξεις που βρέθηκαν στα αποτελέσματα της αναζήτησης μέσω λέξης κλειδί για μαθήματα που προσφέρονται από το Πανεπιστήμιο. Για παράδειγμα, ένας φοιτητής μπορεί να ενδιαφέρεται για τάξη χορού. Πληκτρολογώντας τη λέξη κλειδί «χορός», λαμβάνει μία λίστα από μαθήματα που ταιριάζουν μαζί με μια περίληψη-περιγραφή αυτών των μαθημάτων (courses). Ουσιαστικά γίνεται αναζήτηση στη Βάση Δεδομένων του CourseRank και επιστρέφονται ακόμα και τα μαθήματα που η λέξη «χορός» υπάρχει σαν σχόλιο.

Η βαθύτερη αναζήτηση στη Βάση Δεδομένων για τη λέξη-κλειδί επιτρέπει την ανακάλυψη χρήσιμων σχέσεων μεταξύ των αποτελεσμάτων και των λέξεων-κλειδιών (Bercovitz, et al, 2009). Το Coursecloud παρέχει έννοιες που σχετίζονται με το χορό, αναφερόμενοι πάλι στο προηγούμενο παράδειγμα, όπως «παράσταση», «Λατινική Αμερική», «Ευρώπη». Οι λέξεις αυτές μπορούν να βρεθούν σε διαφορετικά σημεία της Βάσης Δεδομένων. Η βαθμολογία που παίρνουν αυτές οι λέξεις βασίζεται όχι μόνο στη σημαντικότητά τους αλλά και στη σχετική θέση στα αποτελέσματα αναζήτησης. Για παράδειγμα, η λέξη «παράσταση» μπορεί να βρεθεί στα σχόλια των χρηστών που αναφέρονται σε μαθήματα «χορού» σε ζωντανές παραστάσεις. Εννοείται ότι οι όροι στα σύννεφα δεδομένων (data clouds) είναι υπερσύνδεσμοι. Όταν ο χρήστης ψάχνει κάτι, μπορεί να επιλέξει έναν όρο από το data cloud και να οδηγηθεί σε περισσότερα αποτελέσματα (Bercovitz, et al, 2009).

2.2.3 Ευέλικτες συστάσεις στο CourseRank

Το ζητούμενο είναι να παραχθούν ευέλικτες προσωποποιημένες συστάσεις σε αντίθεση με αυτό που κάνουν παλαιότερες μέθοδοι παραγωγής συστάσεων. Στο CourseRank για παράδειγμα, οι φοιτητές μπορούν να ορίσουν τον τύπο των μαθημάτων που τους ενδιαφέρει και να ζητήσουν συστάσεις από άτομα (φοιτητές) που ανήκουν σε ομάδα με παρόμοια χαρακτηριστικά (π.χ. παρόμοια βαθμολογία). Κατ' αυτόν τον τρόπο μπορούν να λάβουν καλύτερα αποτελέσματα (Koutrika, et al, 2009 a).

Το CourseRank περιλαμβάνει μία μηχανή παραγωγής συστάσεων που είναι εύκολο να παραμετροποιηθεί και να επεξεργαστεί. Στη συγκεκριμένη περίπτωση, η μέθοδος εκφράζεται μέσα από μία υψηλού επιπέδου ροή εργασιών πάνω σε δομημένα δεδομένα και εκτελείται από μία κατάλληλη μηχανή. Στην καρδιά αυτού του συστήματος βρίσκεται ένας τελεστής (operator) που παίρνει σαν είσοδο (input) ένα σύνολο πλειάδων (εγγραφών) και τις βαθμολογεί συγκρίνοντάς τες με ένα άλλο σύνολο πλειάδων. Ο τελεστής αυτός μπορεί να καλεί συναρτήσεις όπως τη Jaccard ή την Pearson που εξετάζουν την ομοιότητα δύο συνόλων δεδομένων. Επίσης, μπορεί να χρησιμοποιεί άλλους παραδοσιακούς τελεστές όπως το SELECT και το JOIN (Koutrika, et al, 2009 a).

Ο σχεδιαστής μπορεί να δημιουργήσει εύκολα πολλαπλές παραμετροποιήσιμες ροές εργασιών για διαφορετικούς τύπους δεδομένων. Ο τελικός χρήστης μπορεί να επιλέξει αυτές και να εισαγάγει παραμέτρους στις ροές ούτως ώστε να λάβει ακριβέστερες και προσωποποιημένες συστάσεις. Αυτή η επιλογή γίνεται μέσω γραφικού περιβάλλοντος (GUI). Αξιοσημείωτο είναι ότι το CourseRank έχει μία απλή

Πανεπιστήμιο Πειραιώς

διεπαφή για την εισαγωγή απαιτήσεων. Σημειωτέον ότι οι απαιτήσεις είναι αρκετά σύνθετες λόγω της πολυπλοκότητας των δεδομένων (Koutrika, et al, 2009 a).

Συμπερασματικά, το CourseRank βοηθά τους φοιτητές να οργανώσουν το πρόγραμμά τους κατά τη διάρκεια των τεσσάρων ετών φοίτησης μέσα από την παροχή πλήθους δυνατοτήτων. Για παράδειγμα, ένα πολύ βασικό στοιχείο του είναι ότι ελέγχει τις διενέξεις (π.χ. ένα μάθημα Α πρέπει να έχει εξεταστεί επιτυχώς πριν επιλέξει ο φοιτητής τα μαθήματα Β και Γ). Επίσης, υπολογίζει δυναμικά το μέσο όρο των βαθμών των φοιτητών ανά μάθημα ούτως ώστε οι φοιτητές να έχουν ανά πάσα στιγμή διαθέσιμη τη βαθμολογία και να γνωρίζουν ποια μαθήματα είναι δυσκολότερα (Bercovitz, et al, 2009).

2.3 Τυπικός ορισμός του προβλήματος

Τα συστήματα παραγωγής συστάσεων έχουν γίνει, όπως αναφέρθηκε, ένα πολύ σημαντικό ερευνητικό πεδίο από τη δεκαετία του 1990 που πρωτοεμφανίστηκαν. Το ενδιαφέρον για την ερευνητική αυτή περιοχή είναι αυξημένο γιατί αφορά ένα πεδίο με πάρα πολλές εφαρμογές, όπως ήδη αναφέρθηκε. Παρ' όλα αυτά, χρειάζεται ακόμα τεράστια προσπάθεια για να βελτιωθούν τα συστήματα παραγωγής συστάσεων και να εφαρμοστούν σε ένα ευρύτερο πεδίο της καθημερινής ζωής αφού τα τωρινά συστήματα έχουν κάποιους περιορισμούς που θα περιγραφούν.

Καταρχάς το κύριο πρόβλημα που παρουσιάζεται στα συστήματα παραγωγής συστάσεων είναι, όπως έχει αναφερθεί ανωτέρω, η πρόγνωση-πρόβλεψη βαθμών για αντικείμενα (items) που δεν έχουν ελεγχθεί-εξεταστεί-βαθμολογηθεί από κανέναν χρήστη-επισκέπτη. Είναι ευκόλως κατανοητό ότι αντικείμενα που έχουν λάβει κάποιας μορφής βαθμολογίας, μπορούν να χρησιμοποιηθούν στη συνολική διαδικασία για την τελική πρόταση προς το χρήστη. Το μείζον πρόβλημα είναι αυτό που αναφέρθηκε δηλαδή ότι μέσα σε ένα τεράστιο όγκο δεδομένων, τα περισσότερα αντικείμενα (π.χ. βιβλία, CD, ταινίες, κ.α.) πιθανότατα δεν έχουν ελεγχθεί από κανένα χρήστη. Το ερώτημα που ανακύπτει εύλογα είναι: «Πώς θα είναι δυνατόν να προβλεφθούν βαθμολογίες για αυτά τα αντικείμενα και να προταθούν αναλόγως στο χρήστη;» (Adomavicius & Tuzhilin, 2005)

Εξετάζοντας το πρόβλημα πιο αυστηρά, ονομάζουμε C το σύνολο όλων των χρηστών (c ο χρήστης) και S (s το αντικείμενο) το σύνολο όλων των δυνατών αντικειμένων (βιβλία, ταινίες, εστιατόρια, ξενοδοχεία, κ.α.) που μπορούν να προταθούν. Επιπροσθέτως, ονομάζεται u η συνάρτηση χρησιμότητας που μετράει τη χρησιμότητα του αντικειμένου s στο χρήστη c. Αυτό που επιδιώκεται είναι η

Πανεπιστήμιο Πειραιώς

επιλογή των αντικειμένων που μεγιστοποιούν τη συνάρτηση χρησιμότητας για το χρήστη c. Σημειωτέον ότι το S μπορεί να είναι τεράστιο σε μέγεθος με εκατομμύρια αντικείμενα (Adomavicius & Tuzhilin, 2005). Επίσης, είναι ευνόητο ότι ο αριθμός των χρηστών μπορεί να είναι πολύ μεγάλος και να ανέρχεται σε εκατομμύρια.

Στα συστήματα συστάσεων η χρησιμότητα αντιπροσωπεύεται συνήθως από έναν αριθμό που δείχνει πώς ένας χρήστης αξιολογεί ένα αντικείμενο. Τα αντικείμενα (αναφερόμαστε πάντα σε αυτά που δεν έχουν αξιολογηθεί από τον ενδιαφερόμενο χρήστη) που τελικώς θα προταθούν είναι αυτά που η βαθμολογία τους είναι η μέγιστη μέσα σε ένα σύνολο υποψηφίων προτεινόμενων. Εναλλακτικά μπορεί να προταθούν τα Ν καλύτερα στη βαθμολογία αντικείμενα. Στη συνέχεια θα δειχθεί πώς γίνεται η παραγωγή συστάσεων μέσω συγκεκριμένων μεθόδων.

2.4 Κατηγορίες μεθόδων παραγωγής συστάσεων

Διακρίνονται σε τέσσερις κύριες κατηγορίες

1. Content based (βασισμένες στο περιεχόμενο) κατά την οποία θα προταθούν στο χρήστη αντικείμενα παρόμοια με αυτά που επέλεξε στο παρελθόν.
2. Collaborative (συνεργατικές) κατά την οποία θα προταθούν στο χρήστη αντικείμενα που επιλέγουν άτομα με παρόμοια ενδιαφέροντα
3. Knowledge based (βασισμένες στη γνώση) που δεν χρειάζεται να χρησιμοποιηθούν δεδομένα βαθμολογιών αλλά απαιτούν καλή γνώση των χαρακτηριστικών των αναζητούμενων προϊόντων
4. Υβριδικές που συνδυάζουν τις τρεις ανωτέρω προσεγγίσεις

Στη βιβλιογραφία μπορούν να εντοπιστούν και άλλες κατηγορίες οι οποίες, αν και εντάσσονται στις προαναφερθείσες, μελετώνται ξεχωριστά λόγω σημαντικότητας. Στην περίπτωσή μας η Knowledge based μέθοδος θα μπορούσε να θεωρηθεί υποκατηγορία της Content based. Παρόλα αυτά στην παρούσα εργασία ακολουθείται ο διαχωρισμός βάσει του βιβλίου Recommender Systems: An introduction των Jannach, Zanker et al, 2011. Για την κατηγορία Collaborative χρησιμοποιείται πολύ συχνά και ο όρος collaborative filtering (συνεργατική διήθηση ή συνεργατικό φιλτράρισμα), που θα χρησιμοποιηθεί πολλάκις και στην παρούσα μελέτη.

Πανεπιστήμιο Πειραιώς

2.4.1 Μέθοδοι Content based

2.4.1.1 Γενική περιγραφή

Στις Content based μεθόδους, το σύστημα προσπαθεί να «καταλάβει» τις ομοιότητες μεταξύ υπαρχόντων αβαθμολόγητων αντικειμένων και αντικειμένων που έχουν λάβει υψηλή βαθμολογία από το χρήστη στο παρελθόν. Π.χ. αν κάποιος έχει αξιολογήσει με υψηλό βαθμό τους «Πειρατές της Καραϊβικής No1», είναι απολύτως λογικό να του προταθούν και οι υπόλοιπες ταινίες με το ανάλογο θέμα δηλαδή «Πειρατές της Καραϊβικής No2, No3 και No4». Επίσης, ένα σύστημα συστάσεων σε ταινίες μπορεί να χρησιμοποιήσει κριτήρια όπως πρωταγωνιστές, σκηνοθέτης, σεναριογράφος, κλπ.

Η content based προσέγγιση έχει τις απαρχές της στην ανάκτηση πληροφοριών (Baeza-Yates & Ribeiro-Neto, 1999, Salton, 1989). Τα συστήματα αυτού του τύπου παράγουν συστάσεις κυρίως για εφαρμογές κειμένου (text based) που το περιεχόμενό τους περιγράφεται με λέξεις-κλειδιά (keywords). Για παράδειγμα, η content based εφαρμογή Fab system (Balabanovic & Shoham, 1997) προτείνει ιστοσελίδες χρησιμοποιώντας τις εκατό σημαντικότερες λέξεις. Η σημαντικότητα μία λέξης σε ένα κείμενο καθορίζεται με κάποια μέτρα στάθμισης που μπορούν να οριστούν με διάφορους τρόπους.

2.4.1.2 Αναλυτική περιγραφή της μεθόδου content based

Η μέθοδος content based προσπαθεί να προσδιορίσει τα αντικείμενα που ταιριάζουν καλύτερα με τις προτιμήσεις του χρήστη. Γενικά ο πιο απλός τρόπος να περιγραφεί ένας κατάλογος αντικείμενων είναι να υπάρξει μια σαφής λίστα χαρακτηριστικών για κάθε αντικείμενο.

Για να παραχθούν συστάσεις, τα συστήματα content-based συνήθως αξιολογούν το πόσο ‘ισχυρά’ όμοια είναι τα αντικείμενα που ακόμα δεν έχει δει ο χρήστης, με τα αντικείμενα που έχει δει στο παρελθόν και του άρεσαν. Αυτή η ομοιότητα μπορεί να μετρηθεί με διάφορους τρόπους. Μια προσέγγιση είναι ο υπολογισμός της ομοιότητας ή της επικάλυψης των συσχετιζόμενων λέξεων κλειδιά (Jannach, et al., 2011). Επιπλέον, μια συνήθης μετρική ομοιότητας είναι ο συντελεστής Dice. Αν, για παράδειγμα, έχουμε μια λίστα βιβλίων, τότε κάθε βιβλίο B_i περιγράφεται από ένα

Πανεπιστήμιο Πειραιώς

σύνολο λέξεων κλειδιά keywords(B_i). Ο συντελεστής Dice μετρά την ομοιότητα μεταξύ των βιβλίων b_i και b_j ως εξής:

$$\frac{2 \times |keywords(b_i) \cap keywords(b_j)|}{|keywords(b_i)| + |keywords(b_j)|}$$

Ανάλογα με το πρόβλημα, έχουν αναπτυχθεί αρκετοί τρόποι εύρεσης της ομοιότητας, που υπάρχουν στη διεθνή βιβλιογραφία.

Ιστορικά, οι content-based αλγόριθμοι αναπτύχθηκαν για να προτείνουν, όπως προαναφέρθηκε, εφαρμογές κειμένου (μηνύματα, e-mail και ειδήσεις). Η προσέγγιση αυτή δεν αφορά τη διατήρηση μιας λίστας από χαρακτηριστικά και 'μετα-πληροφορίες' αλλά χρησιμοποιεί μια λίστα από σχετικές λέξεις κλειδιά του αρχείου.

Σε μια πρώτη και, ίσως, αρκετά επιπόλαιη προσέγγιση, θα μπορούσαμε να δημιουργήσουμε μια λίστα από όλες τις λέξεις που εμφανίζονται σε όλα τα αρχεία και να περιγράψουμε κάθε αρχείο με ένα Boolean διάνυσμα. Τότε η τιμή 1 (ένα) σημαίνει ότι η λέξη παρουσιάζεται στο αρχείο ενώ η τιμή 0 (μηδέν) δείχνει μη ύπαρξη. Στην περίπτωση αυτή η περιγραφή του προφίλ του χρήστη θα γινόταν με παρόμοιο τρόπο: με ένα (1) θα συμβολίζαμε το ιδιαίτερο ενδιαφέρον του χρήστη για μια λέξη κλειδί. Έτσι το κατάλληλο ταίριασμα άρθρου-χρήστη θα επιτυγχανόταν μετρώντας την επικάλυψη των λέξεων κλειδιών. Σε μια τέτοια προσέγγιση τα προβλήματα είναι προφανή. Αρχικά, υποθέτουμε ότι όλες οι λέξεις έχουν την ίδια βαρύτητα, κάτι το οποίο δεν ισχύει. Επιπλέον, θα βρεθεί μεγαλύτερη επικάλυψη μεταξύ του προφίλ του χρήστη και ενός άρθρου το οποίο είναι μεγαλύτερο σε μέγεθος (Jannach, et al., 2011).

2.4.1.3 Επίλυση προβλημάτων της content based

Για να λυθούν αυτά τα προβλήματα, τα αρχεία περιγράφονται χρησιμοποιώντας την κωδικοποίηση TF-IDF (term frequency-inverse document frequency). Τα έγγραφα με την κωδικοποίηση αυτή μπορούν να περιγραφούν ως διανύσματα Ευκλείδειου χώρου πολλών διαστάσεων (Salton, et al. 1975). Η TF (term frequency: συχνότητα όρου) περιγράφει το πόσο συχνά εμφανίζεται ένας όρος σε ένα έγγραφο, υποθέτοντας ότι οι σημαντικές λέξεις εμφανίζονται συχνότερα. Αρκετές φορές γίνεται κανονικοποίηση της συχνότητας αυτής για να αποφευχθεί υψηλή συσχέτιση σε μεγάλα άρθρα (Chakrabarti, 2002).

Πανεπιστήμιο Πειραιώς

Η IDF (inverse document frequency: αντίστροφη συχνότητα εγγράφου), από την άλλη, είναι ένα μέτρο που συνδυάζεται με την TF. Στόχος της είναι η μείωση του βάρους των λέξεων κλειδιά που εμφανίζονται πολύ συχνά σε όλα τα έγγραφα. Η ιδέα πίσω από την προσπάθεια μείωσης είναι ότι οι συχνές, ως προς την εμφάνιση, λέξεις στα άρθρα δεν βοηθούν στο να γίνει διάκριση αυτών.

Τα διανύσματα TF-IDF είναι συνήθως πολύ μεγάλα και διάσπαρτα. Στην προσπάθεια να γίνουν συμπαγή και χωρίς περιττές πληροφορίες χρησιμοποιούνται και άλλες τεχνικές, όπως βλέπουμε κάτωθι (Jannach, et al., 2011):

- Η μέθοδος *stop words and stemming*. Βάσει αυτής απομακρύνονται λέξεις που είναι απαραίτητες στο να συνταχθεί σωστά νοηματικώς μια πρόταση. Τέτοιες λέξεις είναι: άρθρα, αντωνυμίες, προθέσεις.
- Η μέθοδος *size cutoffs που* αποτελεί άλλη μια μέθοδο αναπαράστασης ενός εγγράφου που αφαιρεί το «θόρυβο» και παρουσιάζει μόνο τις *n* αντιπροσωπευτικότερες λέξεις.
- Η τεχνική *Phrases.* Με τον τρόπο αυτό είναι δυνατό να χρησιμοποιούνται ολόκληρες φράσεις σαν όροι. Οι φράσεις πολλές φορές είναι πιο περιγραφικές για ένα άρθρο από κάποιες λέξεις χωριστές, άρα ικανότερες να χρησιμοποιηθούν ως λέξεις κλειδιά.

2.4.1.4 Περιορισμοί-Μειονεκτήματα

Στα συστήματα content based (Balabanovic & Shoham, 1997, Shardanand & Maes, 1995) υπάρχουν μερικοί περιορισμοί όπως:

Α. Δύο αντικείμενα που ανήκουν σε παρόμοια σύνολα, είναι αρκετά δυσδιάκριτα. Τα συστήματα αυτά δεν μπορούν ξεχωρίσουν ένα καλογραμμένο άρθρο από ένα κακογραμμένο αν αυτά έχουν πολλές κοινές λέξεις-κλειδιά. Αυτό γίνεται να προσπάθεια να βελτιωθεί με τις μεθόδους της παραγράφου 2.4.1.3.

Β. Δίνονται συστάσεις μόνο για παρόμοια αντικείμενα. Π.χ. αν κάποιος δεν έχει εμπειρία από ελληνική κουζίνα, τότε σίγουρα δεν θα του προταθεί κανένα ελληνικό εστιατόριο όσο καλό κι αν είναι.

Γ. Μπορεί να προτείνουν κάτι που είναι υπερβολικά όμοιο και πρακτικά δεν έχει καμία χρησιμότητα αφού θα ήταν προφανής επιλογή.

Δ. Ο χρήστης πρέπει να έχει αξιολογήσει αρκετά αντικείμενα πριν το σύστημα «καταλάβει» τις προτιμήσεις του και του δώσει αξιόπιστες συστάσεις.

Πανεπιστήμιο Πειραιώς

Συμπερασματικά, ένα σύστημα πρέπει να προτείνει ένα σχετικά ευρύ φάσμα επιλογών και όχι ένα ομογενοποιημένο σύνολο δεδομένων. Π.χ. αν σε κάποιον αρέσει μία ταινία ενός σκηνοθέτη, δεν είναι απαραίτητο ότι θα του αρέσουν όλες οι ταινίες αυτού του καλλιτέχνη.

2.4.2 Collaborative μέθοδοι

Σε αυτές τις μεθόδους γίνεται προσπάθεια να παραχθούν συστάσεις βάσει της βαθμολογίας που δίνουν χρήστες με παρόμοια ενδιαφέροντα δηλαδή χρήστες που είναι αρκετά «όμοιοι». Π.χ. όμοιοι μπορούν να χαρακτηριστούν αυτοί που έχουν αξιολογήσει με τον ίδιο βαθμό μία ταινία ή παρόμοιες ταινίες. Επίσης, όμοιοι χρήστες μπορούν να θεωρηθούν αυτοί που έχουν παραπλήσια δημογραφικά χαρακτηριστικά όπως είναι η ηλικία, το φύλο, η μόρφωση, ο τόπος προσωρινής ή μόνιμης κατοικίας (Pazzani, 1999). Το πρώτο collaborative σύστημα είναι το Grundy (Rich, 1979) που χρησιμοποιούσε στερεότυπα για να χτίσει ατομικά μοντέλα(προφίλ) για κάθε χρήστη και πρότεινε στον καθένα σχετικά βιβλία.

Οι αλγόριθμοι collaborative μεθόδων χωρίζονται σε δύο κατηγορίες:

- memory-based (ή heuristic-based) και
- model-based (Breese, et al, 1998)

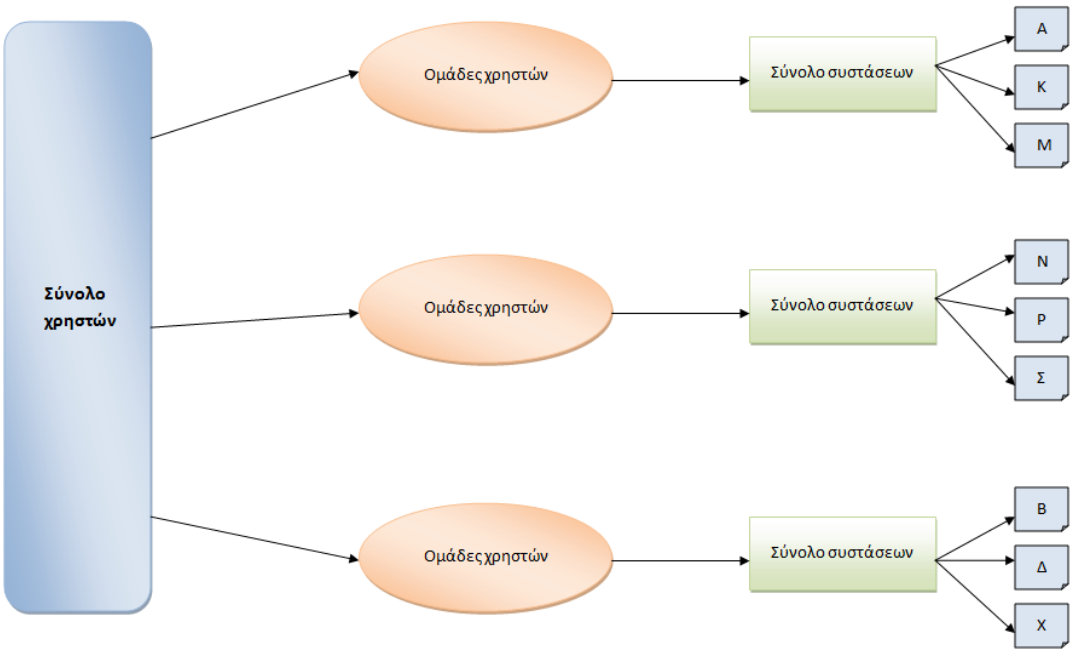
Οι memory-based αλγόριθμοι (Breese, et al, 1998, Delgado & Ishii, 1999) είναι ουσιαστικά ευριστικοί αλγόριθμοι, δηλαδή δίνουν προσεγγιστικές λύσεις, που κάνουν προβλέψεις βασισμένοι στη συλλογή των ήδη βαθμολογημένων αντικειμένων. Οι τιμές των αβαθμολόγητων αντικειμένων υπολογίζονται συνήθως σαν ένα σύνολο των βαθμών των πιο όμοιων χρηστών για αυτά τα αντικείμενα. Σε πολλές προσεγγίσεις αυτής της κατηγορίας η ομοιότητα δύο χρηστών βασίζεται στη βαθμολογία των κοινών αντικειμένων που έχουν αξιολογήσει. Για παράδειγμα, αν ο χρήστης Α έχει βαθμολογήσει τα αντικείμενα Χ, Ψ, Ω και ο χρήστης Β τα αντικείμενα Λ, Χ, Ψ τότε, προφανώς, η ομοιότητα θα υπολογιστεί από τα αντικείμενα Χ και Ψ.

Αντιθέτως, οι model based αλγόριθμοι χρησιμοποιούν τις υπάρχουσες βαθμολογίες για να «μάθουν» ένα μοντέλο το οποίο θα κάνει τις προβλέψεις. Γενικά, σε αυτήν την περίπτωση χρησιμοποιούνται μοντέλα πιθανοτήτων (Getoor & Sahami, 1999).

Οι collaborative μέθοδοι μειονεκτούν στο ότι καινούργια αντικείμενα δεν έχουν καθόλου βαθμολογία καθώς και ότι πρέπει να εξαχθούν προβλέψεις για πολύ περισσότερα αντικείμενα από αυτά που έχουν ήδη βαθμολογία. Ουσιαστικά θα

Πανεπιστήμιο Πειραιώς

πρέπει να εξαχθούν αποτελέσματα από ένα σχετικά μικρό μέρος του συνολικού αριθμού αντικειμένων (Herlocker, et al. 2004).



Εικόνα 8. Παραγωγή συστάσεων με τη χρήση μεθόδων collaborative filtering

2.4.3 Knowledge based μέθοδοι

2.4.3.1 Γενικό πλαίσιο

Τα knowledge based συστήματα παραγωγής συστάσεων βοηθούν στην αντιμετώπιση τυχόν προβλημάτων που προκύπτουν από τις μεθόδους content based, χωρίς όμως αυτό να σημαίνει πως είναι πανάκεια ή πως δεν έχουν και αυτά τα δικά τους προβλήματα.

Σε γενικό πλαίσιο οι knowledge based μέθοδοι βασίζονται στη λεπτομερειακή γνώση των χαρακτηριστικών των αντικειμένων, γι' αυτό και ισχυρίζεται ότι θα μπορούσαν να ενταχθούν στις content based μεθόδους. Τα knowledge based συστήματα χωρίζονται σε δύο κατηγορίες:

- Τα βασιζόμενα σε περιορισμούς (constraint-based) και

- τα βασιζόμενα σε υποθέσεις (cased-based systems, Bridge, et al. 2005, Burke 2000).

Και οι δύο αυτές κατηγορίες είναι παρόμοιες και βασίζονται στο ότι ο χρήστης πρέπει να καθορίσει τις απαιτήσεις και το σύστημα να βρει μια λύση. Στην περίπτωση που δεν είναι δυνατό κάτι τέτοιο, ο χρήστης πρέπει να επαναπροσδιορίσει και να αλλάξει τις απαιτήσεις του. Οι προτάσεις αυτές διαφέρουν στον τρόπο που χρησιμοποιούν τις γνώσεις. Στην περίπτωση case-based συστήματος, οι προτάσεις εστιάζουν στην ανάκτηση παρόμοιων αντικειμένων στη βάση διαφόρων τύπων μέτρων ομοιότητας εν αντιθέσει με τα constraint-based συστήματα στα οποία οι προτάσεις στηρίζονται σε ένα ρητά καθορισμένο σύνολο από κανόνες για παραγωγή συστάσεων.

Στα constraint-based συστήματα (Felfernig & Burke 2008, Felfernig, et al. 2006–07, Zanker, et al. 2010) το σύνολο των αντικειμένων που προτείνονται είναι καθορισμένο από την αναζήτηση προϊόντων που πληρούν τους κανόνες αυτούς. Τα υπάρχοντα χαρακτηριστικά του αντικειμένου βρίσκονται κατηγοριοποιημένα και ο χρήστης δηλώνει τις δικές του απαιτήσεις. Για παράδειγμα, σε ένα ιστότοπο πώλησης ψηφιακών φωτογραφικών μηχανών υπάρχουν χαρακτηριστικά των προϊόντων όπως οπτικό ζουμ, δυνατότητα βιντεοσκόπησης, ήχος, ανάλυση και τιμή. Δίνεται λοιπόν στον χρήστη η δυνατότητα να επιλέξει το εύρος των τιμών αυτών. Τότε το knowledge based σύστημα του προτείνει την κατάλληλη φωτογραφική μηχανή ανάλογα με τις προτιμήσεις/περιορισμούς του.

2.4.3.2 Συστήματα βασιζόμενα σε Περιορισμούς

Ένα βασιζόμενο σε περιορισμούς σύστημα μπορεί να παρουσιαστεί ως ένα πρόβλημα ικανοποίησης περιορισμών (constraint satisfaction problem), όπως αναφέρουν οι Felfering & Burke, 2008 καθώς και οι Zanker, et al. 2010.

Ένα κλασικό τέτοιο πρόβλημα περιγράφεται από μια τριπλέτα (V, D, C), όπου:

- V είναι το σύνολο των μεταβλητών
- D το σύνολο των πεπερασμένων πεδίων ορισμού για αυτές τις μεταβλητές
- C το σύνολο των περιορισμών που περιγράφουν τους συνδυασμούς των τιμών των μεταβλητών που μπορούν να πάρουν ταυτόχρονα.

Μια λύση στο πρόβλημα αυτό αντιστοιχίζει μια τιμή σε κάθε μεταβλητή του V με τέτοιο τρόπο έτσι ώστε όλοι οι περιορισμοί να ικανοποιούνται.

Πανεπιστήμιο Πειραιώς

Οι Felfering & Burke (2008) και οι Zanker, et al. (2010) θεωρούν ότι ένα constraint-based σύστημα παραγωγής συστάσεων μπορεί να κατασκευαστεί με τέτοιου είδους φορμαλισμό και να εξαγάγει ένα σύστημα το οποίο τυπικά θα συμπεριλαμβάνει δυο διαφορετικά σύνολα μεταβλητών, δηλαδή $V = V_C \cup V_{PROD}$, όπου το πρώτο (V_C : Customer properties) περιγράφει τις δυναμικές/πιθανές απαιτήσεις του πελάτη, ενώ το δεύτερο (V_{PROD} : Product properties) τις ιδιότητες των προϊόντων σε μια συλλογή.

Με παρόμοια λογική, υπάρχουν τρία διαφορετικά σύνολα περιορισμών δηλαδή $C= C_R \cup C_F \cup C_{PROD}$, όπου ορίζουν ποια προϊόντα πρέπει να προταθούν σε ένα πελάτη σε συγκεκριμένη κατάσταση απαιτήσεων. Το C_R (Compatibility constraints) αποτελεί το σύνολο των περιορισμών ως προς τις επιλογές του χρήστη. Στο προαναφερθέν παράδειγμα αν ο πελάτης επιλέξει το χαρακτηριστικό του ήχου, τότε αναγκαστικά η φωτογραφική μηχανή θα έχει τη δυνατότητα της βιντεοσκόπησης. Το C_F (Filter Conditions), από την άλλη, ορίζει ποια προϊόντα πρέπει να επιλεχθούν κάτω από συγκεκριμένες συνθήκες. Με άλλα λόγια, τα φίλτρα αυτά ορίζουν τις σχέσεις μεταξύ των μεταβλητών V_C και V_{PROD}. Τέλος, το C_{PROD} (Product Constraints) ορίζει τα διαθέσιμα προϊόντα στη συλλογή. Ουσιαστικά κάθε λύση του προβλήματος ικανοποίησης περιορισμών αντιστοιχεί σε μια συνεπή πρόταση (Zanker et al., 2010).

2.4.3.3 Συστήματα βασισμένα σε υποθέσεις (case based systems)

Στα case-based συστήματα, τα αντικείμενα προτείνονται χρησιμοποιώντας μέτρα ομοιότητας που περιγράφουν σε τι είδους μέγεθος οι ιδιότητες του αντικειμένου ταιριάζουν με κάποιες από τις δοθείσες απαιτήσεις του χρήστη. Αυτά τα μέτρα ομοιότητας ονομάζονται μέτρα απόστασης ομοιότητας (distance similarity). Στη βιβλιογραφία υπάρχουν τέτοιου είδους τύποι που κάνουν τους αναγκαίους υπολογισμούς (Lorenzi & Ricci, 2005).

Τα συστήματα παραγωγής προτάσεων που ανήκουν στην κατηγορία knowledge-based είναι ουσιαστικά σχεδιασμένα για συγκεκριμένες καταστάσεις και απαιτούν από το χρήστη να γνωρίζει αυτό ακριβώς που θέλει, πράγμα το οποίο ανάλογα με το πρόβλημά μας μπορεί να είναι πλεονέκτημα ή και σοβαρότατο μειονέκτημα.

Πανεπιστήμιο Πειραιώς

2.4.4 Υβριδικές μέθοδοι

Κάθε μια από τις παραπάνω μεθόδους έχει τα πλεονεκτήματα και τα μειονεκτήματά της. Για να καταφέρουν οι ερευνητές να συνδυάσουν τη δύναμη των προηγούμενων μεθόδων ανέπτυξαν υβριδικούς αλγορίθμους που περιλαμβάνουν τους προηγούμενους αλγορίθμους.

Ταξινομούνται ως ακολούθως, ανάλογα με την αρχική μας κατηγοριοποίηση σε μεθόδους (Tran & Cohen, 2000, Jannach, et al. 2011):

- Εφαρμόζονται οι προηγούμενες μέθοδοι (content based, collaborative και knowledge) και συνδυάζονται τα αποτελέσματα.
- Ενσωμάτωση content based χαρακτηριστικών σε collaborative προσεγγίσεις.
- Ενσωμάτωση collaborative χαρακτηριστικών σε content based προσεγγίσεις.
- Ενσωμάτωση knowledge χαρακτηριστικών σε content based προσεγγίσεις.
- Ενσωμάτωση knowledge χαρακτηριστικών σε collaborative προσεγγίσεις.
- Ενσωμάτωση content based χαρακτηριστικών σε knowledge προσεγγίσεις.
- Ενσωμάτωση collaborative χαρακτηριστικών σε knowledge based προσεγγίσεις.
- Κατασκευή ενοποιημένου μοντέλου που συνδυάζει και τις τρεις προσεγγίσεις.

Η μία διάσταση του προβλήματος που απασχολεί έναν δημιουργό αλγορίθμων είναι το είδος του προβλήματος καθώς και οι απαιτήσεις του ως προς τα δεδομένα. Όπως φαίνεται στην εικόνα 9 τα προβλήματα μπορούν να κατηγοριοποιηθούν σε τέσσερις τύπους.

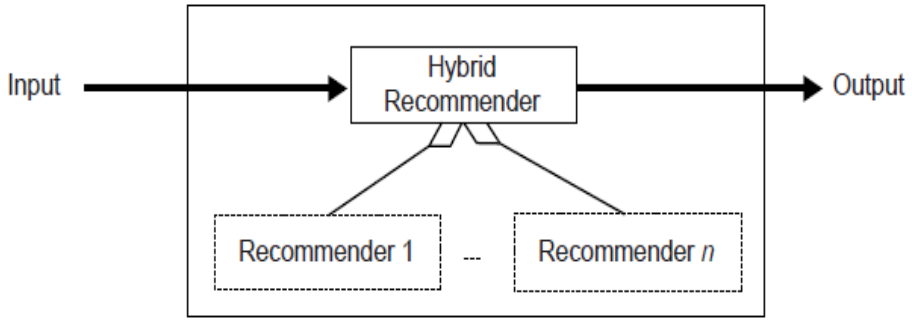
Paradigm	User profile and contextual parameters	Community data	Product features	Knowledge models
Collaborative	Yes	Yes	No	No
Content-based	Yes	No	Yes	No
Knowledge-based	Yes	No	Yes	Yes

Εικόνα 9. Κατηγοριοποίηση προβλημάτων

Η άλλη διάσταση που απασχολεί την ανάπτυξη ενός υβριδικού αλγορίθμου είναι η μέθοδος που θα χρησιμοποιηθεί έτσι ώστε ο αλγόριθμος να συνδυάζει δύο ή

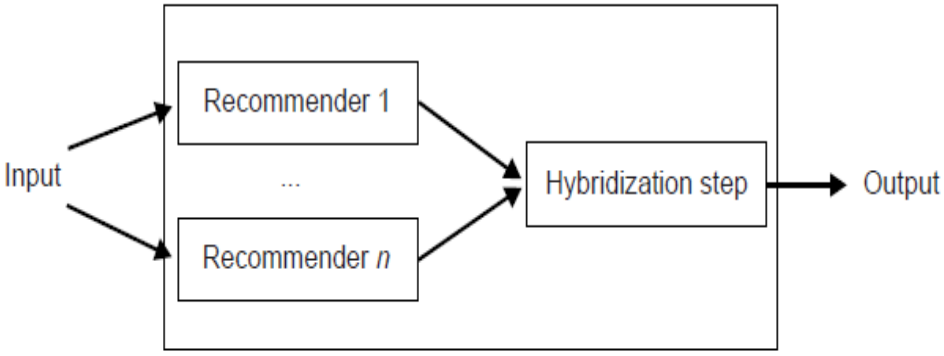
περισσότερα συστήματα παραγωγής συστάσεων. Οι μέθοδοι αυτές μπορούν να κατηγοριοποιηθούν σε τρεις βασικούς σχεδιαστικούς άξονες:

- τη μονολιθική,
- την παράλληλη και
- την σωληνωτή, όπως φαίνεται στα παρακάτω σχήματα:



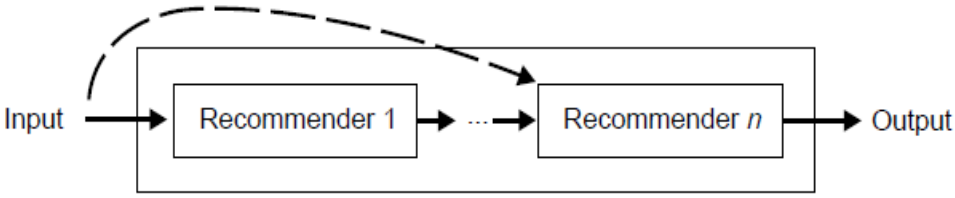
Εικόνα 10. Μονολιθική Υβριδική Σχεδίαση

Στη μονολιθική σχεδίαση τα διαφορετικά συστήματα παραγωγής συστάσεων, που θα χρησιμοποιηθούν, συνδυάζονται σε ένα και μόνο ένα υβριδικό σύστημα το οποίο δέχεται δεδομένα και εξάγει αποτελέσματα. Ο συνδυασμός αυτών των συστημάτων σε ένα σύστημα απαιτεί ειδική σχεδίαση.



Εικόνα 11: Παράλληλη Υβριδική Σχεδίαση

Στην παράλληλη σχεδίαση τα διαφορετικά συστήματα λειτουργούν παραλλήλως δεχόμενα το καθένα διαφορετικές εισόδους. Τα αποτελέσματα αυτών, βάσει κατάλληλου υβριδικού μηχανισμού, συνθέτουν την τελική πρόταση.



Εικόνα 12. Σωληνωτή Υβριδική Σχεδίαση

Στη σωληνωτή σχεδίαση εφαρμόζεται μια σταδιακή διαδικασία κατά την οποία το πρώτο σύστημα δέχεται δεδομένα, παράγει αποτελέσματα τα οποία αποτελούν δεδομένα εισόδου για το επόμενο. Έτσι, ακολουθιακά, το τελευταίο σύστημα εξάγει την τελική σύσταση.

Πανεπιστήμιο Πειραιώς

3 Παρουσίαση αλγορίθμων παραγωγής συστάσεων κατηγορίας collaborative filtering

3.1 Εισαγωγή

Στη συνέχεια παρουσιάζονται διάφοροι αλγόριθμοι που χρησιμοποιούνται για την εξαγωγή συστάσεων με μεθόδους collaborative filtering (συνεργατικό φιλτράρισμα ή διήθηση). Στη διεθνή βιβλιογραφία υπάρχουν και άλλοι αλγόριθμοι που δεν θα μας απασχολήσουν στην παρούσα εργασία. Αρχικά παρουσιάζεται ο συμβολισμός που χρησιμοποιείται και ύστερα παρατίθενται οι αλγόριθμοι.

Με $U = \{ u_1, u_2, \dots, u_m \}$ θεωρούμε το σύνολο των χρηστών, ενώ το σύνολο των αντικειμένων με $I = \{ i_1, i_2, \dots, i_n \}$.

Κάθε χρήστης $u_i \in U$ έχει βαθμολογίες $I_u \subseteq I$, ενώ κάθε αντικείμενο, αντίστοιχα έχει βαθμολογηθεί από ένα υποσύνολο $U_i \subseteq U$, των χρηστών. Ο ενεργός χρήστης, του οποίου αναζητείται η πρόβλεψη συμβολίζεται με u_a.

Με *R* συμβολίζουμε το σύνολο των πιθανών βαθμολογιών. Στην παρούσα εργασία χρησιμοποιείται η κλίμακα 1 έως 5, άρα *R*= { 1, 2, 3, 4, 5 }. Με V συμβολίζουμε τον πίνακα του χρήστη με τις βαθμολογίες του. Για κάθε στοιχείο του V, έχουμε $v_{u_i} \in R \cup \emptyset$, δηλαδή έχουμε τη βαθμολογία του χρήστη $u \in U$ για το αντικείμενο $i \in I$, ενώ με το κενό σύνολο Ø, εννοούμε ότι ο χρήστης δεν έχει ψηφίσει το συγκεκριμένο αντικείμενο. Με $p_{u_i} \in R \cup \emptyset$ συμβολίζουμε την πρόβλεψη, που προκύπτει από τον εκάστοτε αλγόριθμο collaborative filtering για το αντικείμενο $i \in I$ του χρήστη $u \in U$.

Το σύνολο των βαθμολογιών του χρήστη είναι το εξής $v_u = \{v_{u_i} \in V \ / \ i \in I_u\}$, ενώ το σύνολο των βαθμολογιών του αντικειμένου είναι $v_i = \{v_{u_i} \in V \ / \ u \in U_i\}$. Τέλος με $\overline{v_u}$ συμβολίζουμε το μέσο όρο βαθμολογίας του χρήστη και με $\overline{v_i}$ το μέσο όρο βαθμολογίας του αντικειμένου (Cacheda, et al, 2011).

3.2 Αλγόριθμοι τύπου User-based (Neighborhood-based)

Αυτού του τύπου οι αλγόριθμοι είναι ιδιαίτερα δημοφιλείς. Ακολουθούν μια διαδικασία τριών βημάτων (Cacheda, et al, 2011):

- Υπολογισμός της ομοιότητας μεταξύ του ενεργού χρήστη και των υπολοίπων χρηστών.
- Επιλογή ενός υποσυνόλου των χρηστών (μιας γειτονιάς) βάσει της ομοιότητάς τους με τον ενεργό χρήστη.

Πανεπιστήμιο Πειραιώς

- Υπολογισμός της πρόβλεψης χρησιμοποιώντας τις βαθμολογίες της γειτονιάς.

Για την υλοποίηση κάθε βήματος έχουν προταθεί αρκετές στρατηγικές, οι οποίες παρουσιάζονται.

Για το πρώτο βήμα, αυτό του υπολογισμού της ομοιότητας μεταξύ των χρηστών έχουμε τα εξής:

Α. Συντελεστής συσχέτισης του Pearson (Pearson correlation coefficient).

Η συσχέτιση των χρηστών υπολογίζεται με βάση τον παρακάτω συντελεστή:

$$s(a,u) = \frac{\sum_{i \in I_a \cap I_v}(v_{a_i} - \overline{v_a})(v_{u_i} - \overline{v_u})}{\sqrt{\sum_{i \in I_a \cap I_v}(v_{a_i} - \overline{v_a})^2 \sum_{i \in I_a \cap I_v}(v_{u_i} - \overline{v_u})^2}}$$

Β. Περιορισμένος συντελεστής Pearson (Constrained Pearson)

Το 1995 οι Shardanand και Maes πρότειναν τη χρησιμοποίηση της κεντρικής τιμής της βαθμολογίας των χρηστών (για παράδειγμα για την κλίμακα 1 έως 5, την τιμή 3) αντικαθιστώντας έτσι το μέσο όρο των βαθμολογιών του κάθε χρήστη. Επομένως ο συντελεστής συσχέτισης Pearson για την κλίμακα 1 έως 5, γίνεται:

$$s(a,u) = \frac{\sum_{i \in I_a \cap I_v}(v_{a_i} - 3)(v_{u_i} - 3)}{\sqrt{\sum_{i \in I_a \cap I_v}(v_{a_i} - 3)^2 \sum_{i \in I_a \cap I_v}(v_{u_i} - 3)^2}}$$

Γ. Διανυσματική ομοιότητα συνημιτόνου (Vector similarity, cosine)

Αυτό είναι ένα άλλο μέτρο ομοιότητας για τους χρήστες, το οποίο αντιμετωπίζει τους χρήστες σαν διανύσματα των βαθμολογιών των αντικειμένων. Ύστερα υπολογίζει το συνημίτονο, μέσω του εσωτερικού γινομένου, μεταξύ των διανυσμάτων δύο χρηστών. Μια τιμή κοντά στο ένα (1) σημαίνει ομοιότητα, ενώ μια τιμή κοντά στο μηδέν (0) σημαίνει αντίθεση.

$$s(a,u) = \sum_{j \in I} \frac{v_{a_j}}{\sqrt{\sum_{k \in I_a} v_{a_k}^2}} \frac{v_{u_j}}{\sqrt{\sum_{k \in I_u} v_{u_k}^2}}$$

Πανεπιστήμιο Πειραιώς

Δ. Μέση τετραγωνική διαφορά (Mean squared difference)

Έτσι μετράται η ομοιότητα των χρηστών μέσω της τετραγωνικής διαφοράς των αντικειμένων που έχουν βαθμολογηθεί από όλους.

$$msd(a,u) = \frac{\sum_{i \in I_a \cap I_v} (v_{a_i} - v_{u_i})^2}{|I_a \cap I_u|}$$

Στη συνέχεια, οι χρήστες των οποίων η διαφορά είναι μεγαλύτερη από ένα κάτω όριο L απορρίπτονται, ενώ η ομοιότητα των υπολοίπων υπολογίζεται από τον τύπο:

$$s(a,u) = \frac{L - msd(a,u)}{L}$$

Ε. Σταθμικός συντελεστής Pearson (Weighted Pearson)

Αυτή η μέτρηση είναι εμπνευσμένη από το βαθμό εμπιστοσύνης της γειτονιάς. Αν δύο χρήστες έχουν λίγα αντικείμενα κοινά και οι βαθμολογίες των αντικείμενων αυτών συμπίπτουν, τότε κανένα από τα προαναφερθέντα μέτρα δεν τους θεωρεί όμοιους. Καθώς ο αριθμός των αντικειμένων αυτών αυξάνεται, η αιτία που συμπίπτουν οι βαθμολογίες είναι ότι οι χρήστες είναι πάρα πολύ όμοιοι και δεν είναι απλή σύμπτωση. Έτσι η εμπιστοσύνη στο μέτρο της ομοιότητας αυξάνεται, καθώς αυξάνεται ο αριθμός των κοινών αντικειμένων (Cacheda, et al, 2011).

$$s(a,u) = \begin{cases} s_{pearson}(a,u)^{\frac{|I_a \cap I_u|}{50}} & |I_a \cap I_u| < 50 \\ s_{pearson}(a,u) & αλλιώς \end{cases}$$

Για το δεύτερο βήμα του αλγορίθμου, οφείλουμε να επιλέξουμε γειτονιά. Παρουσιάζουμε δύο εναλλακτικές:

- Συσχέτιση ορίου (Correlation threshold)

Πανεπιστήμιο Πειραιώς

- Μεγαλύτερο αριθμό γειτόνων (max number of neighbors)

Με βάση την πρώτη μέθοδο, επιλέγουμε τους χρήστες των οποίων η ομοιότητα με τον ενεργό χρήστη ξεπερνά ένα δοθέν όριο. Με βάση τη δεύτερη μέθοδο, γίνεται επιλογή των N χρηστών που είναι αρκετά όμοιοι με τον ενεργό.

Κατά το τρίτο στάδιο του αλγορίθμου user-based, γίνεται ο υπολογισμός της πρόβλεψης, βάσει των βαθμολογιών των γειτόνων που επιλέχθηκαν κατά το δεύτερο βήμα. Παρουσιάζουμε δύο τρόπους:

- Σταθμική συσχέτιση (Weighted by correlation)
- Z-βαθμολογική κανονικοποίηση (Z-score normalization)

Βάσει της πρώτης, όσο ένας χρήστης είναι 'περισσότερο' όμοιος με τον ενεργό, τόσο η βαθμολογία του είναι πιο ακριβής και χρησιμοποιείται για την πρόβλεψη.

Με βάση τη δεύτερη μέθοδο αυτή της Z-βαθμολογικής κανονικοποίησης, η πρόβλεψη γίνεται με τον παρακάτω τύπο. Σημειώνεται δε, ότι με τον τύπο αυτό η κανονικοποίηση που πραγματοποιείται, αφορά τους χρήστες που ψηφίζουν μόνο αρνητικά ή μόνο θετικά τα αντικείμενα (Cacheda, et al, 2011).

$$p_{a_i} = \overline{v_a} + \sigma_a \frac{\sum_{u \in Neigh_a} \left[\left(\frac{v_{u_i} - \overline{v_u}}{\sigma_u} \right) s(a,u) \right]}{\sum_{u \in Neigh_a} s(a,u)}$$

Πανεπιστήμιο Πειραιώς

3.3 Αλγόριθμοι τύπου Item-based

Οι αλγόριθμοι αυτού του τύπου είναι παρόμοιοι με τους αλγορίθμους User-based, αλλά αντί της αναζήτησης γειτονιάς μεταξύ των χρηστών, ερευνούμε για όμοια αντικείμενα. Όπως και στους αλγορίθμους User-based (Sarwar, et al, 2001), έτσι και εδώ υπάρχουν διαφορετικές στρατηγικές υπολογισμού της ομοιότητας των αντικειμένων. Άρα, και ο κλασσικός τύπος για το συντελεστή συσχέτισης μεταφέρεται ως εξής (Cacheda, et al, 2011):

$$s(i,j) = \frac{\sum_{u \in U}(v_{u_i} - \overline{v_u})(v_{u_j} - \overline{v_u})}{\sqrt{\sum_{u \in U}(v_{u_i} - \overline{v_u})^2 \sum_{u \in U}(v_{u_j} - \overline{v_u})^2}}$$

Αφού υπολογιστεί η ομοιότητα μεταξύ διαφορετικών αντικειμένων, γίνεται η επιλογή των N καλύτερων γειτόνων. Για να υπολογιστεί η πρόβλεψη, αθροίζουμε τις βαθμολογίες του ενεργού χρήστη με αυτές των γειτόνων σταθμισμένες με την ομοιότητα με του αντικείμενου, του οποίου την πρόβλεψη αναζητούμε (Cacheda, et al, 2011).

$$p_{a_j} = \frac{\sum_i (s(i,j) v_{a_i})}{\sum_i |s(i,j)|}$$

Από τα πλεονεκτήματα αυτού του τύπου των αλγορίθμων σε σχέση με τους user-based είναι ότι τείνουν να είναι πιο στατικοί και ότι η γειτονιά μπορεί να υπολογιστεί εκτός σύνδεσης (offline).

Πανεπιστήμιο Πειραιώς

3.4 Ένωση ομοιότητας (Similarity Fusion)

Οι αλγόριθμοι αυτού του τύπου εκμεταλλεύονται περισσότερες πληροφορίες του συστήματος και συμφέρουν σε περιπτώσεις που υπάρχουν πολύ λίγες βαθμολογίες αφού μπορούν να δώσουν ακριβέστερες προβλέψεις.

Οι αλγόριθμοι αυτοί συνδυάζουν τις:

- βαθμολογίες του αντικειμένου από τους χρήστες που είναι όμοιοι με τον ενεργό (SUR)
- βαθμολογίες του ενεργού χρήστη σε όμοια αντικείμενα (SIR)
- βαθμολογίες των όμοιων αντικειμένων από τους όμοιους χρήστες (SUIR)

Η πρόβλεψη δίνεται από τον παρακάτω τύπο:

$$p_{u_i} = \sum_{r \in R} P\left(v_{u_i} = r \,|SUR, SIR, SUIR\right)$$
$$= \left(\sum_{r \in R} P\left(v_{u_i} = r \,|SUIR\right)\delta\right) + \left(\sum_{r \in R} P\left(v_{u_i} = r \,|SUR\right)\delta\,(1-\lambda)\right)$$
$$+ \left(\sum_{r \in R} P\left(v_{u_i} = r \,|\, SIR\right)(1-\delta)(1-\lambda)\right)$$

Πανεπιστήμιο Πειραιώς

3.5 Διάγνωση προσωπικότητας (Personality Diagnosis)

Αυτή η μέθοδος σχετίζεται με την παρατήρηση ότι ένας χρήστης μπορεί να βαθμολογεί με διαφορετικό τρόπο ίδιο αντικείμενο είτε λόγω διάθεσης είτε λόγω άλλων ψυχολογικών παραμέτρων. Το προφίλ ή η προσωπικότητα του χρήστη, αντιστοιχίζοντας τα με τις βαθμολογίες του, πρέπει να συνοδεύονται με τις παρεκκλίσεις αυτές. Οι αλλαγές αυτές υπολογίζονται ως πιθανότητα της κατανομής Gaussian noise (Cacheda, et al, 2011, Pennock, et al, 2000). Η πιθανότητα αυτή δίνεται από τον τύπο:

$$P\left(v_{ij} = x \mid v_{ij}^{true} = y\right) \simeq e^{\frac{-(x-y)^2}{2\sigma^2}}$$

Χρησιμοποιώντας αυτούς τους αλγορίθμους υπολογίζεται η πιθανότητα του ενεργού χρήστη να έχει την ίδια προσωπικότητα-συμπεριφορά με τους άλλους χρήστες. Αυτές οι πιθανότητες χρησιμοποιούνται για να υπολογιστεί η πιθανότητα κατανομής των βαθμολογιών ενός συγκεκριμένου αντικειμένου. Επομένως η πρόβλεψη είναι η πιο πιθανή βαθμολογία.

Πανεπιστήμιο Πειραιώς

3.6 Παρουσίαση αλγορίθμων της κατηγορίας collaborative

Στη συνέχεια θα παρατεθούν κάποιοι αλγόριθμοι παραγωγής συστάσεων κατηγορίας collaborative που εφαρμόζονται σε πολύ γνωστούς ιστοτόπους

3.6.1 Αλγόριθμος συστάσεων του Amazon

Στο Amazon (Linden, et al, 2003) χρησιμοποιούνται ειδικοί αλγόριθμοι για να προσωποποιήσουν (εξατομικεύσουν) συστάσεις στους διάφορους χρήστες. Το ηλεκτρονικό κατάστημα αλλάζει πολύ γρήγορα βασισμένο στις επιθυμίες των πελατών είτε πρόκειται για ένα προγραμματιστή που αναζητά αντίστοιχα βιβλία είτε για μια μητέρα που ψάχνει παιχνίδια για το παιδί της.

Οι αλγόριθμοι παραγωγής συστάσεων που αφορούν το ηλεκτρονικό εμπόριο λειτουργούν μέσα σε ένα ιδιαιτέρως απαιτητικό περιβάλλον. Για παράδειγμα:

- Ένας μεγάλος έμπορος λιανικής έχει τεράστια ποσά δεδομένων, εκατομμύρια πελατών και εκατομμύρια διαφορετικά προϊόντα.
- Πολλές εφαρμογές απαιτούν αποτελέσματα σε πραγματικό χρόνο και μάλιστα σε λιγότερο από ένα ή δύο δευτερόλεπτα
- Οι νέοι πελάτες έχουν πολύ περιορισμένη πληροφόρηση βασισμένη σε λίγες αγορές και σε βαθμολογία προϊόντων
- Παλαιότεροι πελάτες έχουν ακριβώς το αντίθετο πρόβλημα δηλαδή υπερπληθώρα πληροφόρησης βασισμένων σε πολλά προϊόντα και αγορές.
- Τα δεδομένα που σχετίζονται με τον εκάστοτε πελάτη είναι εξαιρετικά ευμετάβλητα. Π.χ. η επίσκεψη του ιστοτόπου για ένα προϊόν δίνει νέα διαφορετικά στοιχεία στον αλγόριθμο ο οποίος θα πρέπει να ανταποκριθεί αναλόγως.

Υπάρχουν τρεις κοινές προσεγγίσεις για την επίλυση του ανωτέρω προβλήματος:

- παραδοσιακές μέθοδοι collaborative filtering
- clusters model
- search based μεθόδους

Γενικά οι αλγόριθμοι παραγωγής συστάσεων ξεκινούν βρίσκοντας έναν αριθμό πελατών που έχουν αγοράσει προϊόντα και τα έχουν βαθμολογήσει. Αυτά τα προϊόντα είναι υπερσύνολο των προϊόντων, αγορασμένων και βαθμολογημένων, του ενεργού χρήστη. Στη συνέχεια θα προτείνει κάποια προϊόντα που ανήκουν σε αυτό το υπερσύνολο και δεν έχουν εξεταστεί από τον ενεργό χρήστη

Πανεπιστήμιο Πειραιώς

Στη συνέχεια θα γίνει σύγκριση των προσεγγίσεων αυτών με τον αλγόριθμο *item to item collaborative filtering* δηλαδή αυτόν που χρησιμοποιεί το Amazon.

3.6.1.1 Παραδοσιακές μέθοδοι collaborative filtering (user based)

Οι μέθοδοι αυτές (Linden, et al, 2003) αναπαριστούν έναν πελάτη ως ένα N-διάστατο διάνυσμα όπου N είναι ο αριθμός των προϊόντων. Τα στοιχεία του διανύσματος έχουν θετική ή αρνητική βαθμολογία. Ο αλγόριθμος πολλαπλασιάζει τα στοιχεία του διανύσματος με τους αντίστροφους αριθμούς των πελατών προσπαθώντας να κάνει πιο σχετικά τα λιγότερα γνωστά προϊόντα. Για όλους σχεδόν τους πελάτες, αυτό το διάνυσμα είναι υπερβολικά αραιό. Ο αλγόριθμος παράγει συστάσεις βασισμένος σε λίγους πελάτες που είναι πιο όμοιοι με τον ενεργό χρήστη. Διευκρινίζεται ότι ενεργός χρήστης είναι αυτός για τον οποίο θα δοθούν προβλέψεις. Η μέτρηση της ομοιότητας γίνεται με τη χρήση του συνημιτόνου της γωνίας των δύο διανυσμάτων, τα οποία αφορούν δύο χρήστες. Μία συνήθης τεχνική του αλγορίθμου είναι να βαθμολογεί κάθε προϊόν σύμφωνα με το πόσοι όμοιοι χρήστες το αγόρασαν.

Η υπολογιστική πολυπλοκότητα του αλγορίθμου είναι μεγάλη, της τάξης του O(NM) στη χειρότερη περίπτωση όπου M είναι ο αριθμός των πελατών και N ο αριθμός των προϊόντων. Παρ' όλα αυτά, επειδή τα διανύσματα είναι πολύ αραιά, δηλαδή δεν έχουν πολλές τιμές, η πολυπλοκότητα του αλγορίθμου τείνει προς O(N+M). Ακόμα κι έτσι, όμως, για μεγάλο όγκο δεδομένων της τάξεως των 10 εκατομμυρίων πελατών του και του ενός εκατομμυρίου προϊόντων, ο αλγόριθμος αντιμετωπίζει σοβαρά προβλήματα επιδόσεων και κλιμάκωσης.

Είναι δυνατό να βελτιωθεί η επίδοση του αλγορίθμου με τη μείωση του δείγματος. Συγκεκριμένα, μπορούμε να μειώσουμε τους πελάτες επιλέγοντας μέσα από ένα αισθητά μικρότερο δείγμα και απορρίπτοντας τα πολύ δημοφιλή προϊόντα ή αυτά με πολύ χαμηλή βαθμολογία. Επιπλέον, μπορεί να μειωθεί κι άλλο το δείγμα των προϊόντων με διάφορες άλλες μεθόδους, που δεν χρήζουν αναφοράς, αλλά το τίμημα είναι βαρύτατο γιατί η ποιότητα των συστάσεων θα είναι πολύ χαμηλή. Πρώτον, αν ο αλγόριθμος εξετάζει μόνο ένα μικρό δείγμα πελατών, οι επιλεγέντες πελάτες θα είναι λιγότερο όμοιοι με τον τελικό χρήστη. Δεύτερον, αν ο αλγόριθμος απορρίπτει τα πολύ δημοφιλή ή τα αντιδημοφιλή προϊόντα, τελικώς αυτά δεν θα προταθούν ποτέ.

Πανεπιστήμιο Πειραιώς

3.6.1.2 Μοντέλα Διασποράς (Cluster Models)

Τα μοντέλα διασποράς (Linden, et al, 2003) προσπαθούν να βρουν τους όμοιους χρήστες διαιρώντας τους πελάτες σε πολλά τμήματα και αντιμετωπίζουν το έργο σαν ένα πρόβλημα ταξινόμησης. Ο στόχος του αλγορίθμου είναι να ταιριάξει το χρήστη με το τμήμα που περιέχει τους πιο όμοιους πελάτες. Έπειτα χρησιμοποιεί τις αγορές και τις βαθμολογίες αυτών των πελατών για την παραγωγή συστάσεων. Η αρχική επιλογή πελάτη για την ένταξή του σε ένα τμήμα γίνεται τυχαία. Στη συνέχεια γίνεται επαναληπτικά η πρόσθεση πελατών στα τμήματα με κάποια πρόβλεψη για τη δημιουργία νέων τμημάτων. Με λίγα λόγια, ο αλγόριθμος δημιουργεί τα τμήματα, υπολογίζει την ομοιότητα του χρήστη με το διάνυσμα που περιγράφει κάθε τμήμα (segment), επιλέγει το τμήμα με τη μεγαλύτερη ομοιότητα και ταξινομεί το χρήστη αναλόγως.

Τα μοντέλα διασποράς έχουν καλύτερη online κλιμάκωση και επίδοση από παραδοσιακές μεθόδους collaborative filtering διότι συγκρίνουν το χρήστη με ένα συγκεκριμένο αριθμό τμημάτων παρά με ολόκληρη τη βάση. Όμως και σε αυτήν την περίπτωση η ποιότητα των παραγόμενων συστάσεων είναι χαμηλή αφού οι συσχετιζόμενοι πελάτες εντός των τμημάτων δεν είναι απαραιτήτως και οι πιο σχετικοί.

3.6.1.3 Μέθοδοι βασισμένες στην αναζήτηση ή στο περιεχόμενο

Αυτές οι μέθοδοι αντιμετωπίζουν το πρόβλημα σαν ένα πρόβλημα σχετικών προϊόντων και δεν πρέπει να συγχέονται με τις κλασικές μεθόδους content based που αναλύθηκαν στην παράγραφο 2.4.1. Δοθέντων των προϊόντων και των βαθμών ενός χρήστη, ο αλγόριθμος κατασκευάζει ένα ερωτηματολόγιο αναζήτησης για να βρει άλλα δημοφιλή προϊόντα από τον ίδιο συγγραφέα, καλλιτέχνη ή σκηνοθέτη ή με παρόμοιες λέξεις κλειδιά. Π.χ αν κάποιος αγοράσει τη συλλογή ταινιών «Ο Νονός» είναι πιθανό το σύστημα να προτείνει ταινίες του Φράνσις Φόρντ Κόπολα, του Μάρλον Μπράντο, ή με παρόμοια θεματολογία (π.χ μαφία). Αν ο χρήστης έχει διεκπεραιώσει λίγες αγορές ή βαθμολογήσεις, τότε ο αλγόριθμος λειτουργεί ικανοποιητικά. Όμως, για χρήστες με χιλιάδες αγορές είναι πρακτικά πολύ δύσκολο να δημιουργηθεί ερωτηματολόγιο συμπεριλαμβάνοντας όλα αυτά τα προϊόντα. Συνεπώς, η ποιότητα των συστάσεων δεν είναι υψηλή. Σημειωτέον ότι οι συστάσεις αυτών των αλγορίθμων πολλές φορές είναι γενικές (π.χ. προτείνουν όλες τις ταινίες μιας κατηγορίας) ή πολύ περιορισμένες (προτείνουν όλα τα βιβλία ενός συγγραφέα).

Πανεπιστήμιο Πειραιώς

3.6.1.4 Αλγόριθμος Amazon (item to item collaborative filtering)

Στην περίπτωση αυτή ο αλγόριθμος ταιριάζει κάθε μία από τις αγορές του χρήστη και τις βαθμολογίες του με παρόμοια αντικείμενα και τα συνδυάζει σε μία λίστα συστάσεων. Καταρχάς, για να καθορίσει τα πιο όμοια αντικείμενα, ο αλγόριθμος κατασκευάζει έναν πίνακα όμοιων αντικειμένων (προϊόντων) βρίσκοντας αντικείμενα τα οποία οι πελάτες «τείνουν» στο να τα αγοράσουν. Θα μπορούσε να κατασκευαστεί ένας πίνακας που να συσχετίζει κάθε αντικείμενο με όλα τα υπόλοιπα αλλά αυτό δεν θα ήταν αποδοτικό γιατί οι περισσότερες θέσεις των διανυσμάτων θα ήταν χωρίς τιμή.

Είναι προτιμότερο να υπολογιστεί η ομοιότητα ανάμεσα σε ένα αντικείμενο (προϊόν) και όλα τα σχετικά αντικείμενα. Η ομοιότητα μπορεί να υπολογιστεί με διάφορους τρόπους αλλά ένας συνηθισμένος είναι με τη μέτρηση του συνημιτόνου όπως περιγράφηκε ανωτέρω στις παραδοσιακές μεθόδους στην παράγραφο 3.6.1.1. Η πολυπλοκότητα του αλγορίθμου είναι υψηλή στη χειρότερη περίπτωση, της τάξης $O(N^2M)$. Στην πράξη, όμως, είναι πλησιέστερη της τάξης του O(NM) επειδή πολλοί πελάτες έχουν διεκπεραιώσει λίγες αγορές.

Δοθέντος ενός πίνακα με παρόμοια αντικείμενα, ο αλγόριθμος εντοπίζει τα αντικείμενα που είναι παρόμοια με καθένα από τα αγορασθέντα και βαθμολογηθέντα προϊόντα του ενεργού χρήστη. Στη συνέχεια συγκεντρώνει αυτά τα αντικείμενα και προτείνει τα πιο δημοφιλή ή τα πιο σχετικά.

Ο αλγόριθμος του Amazon διαχειρίζεται έναν τεράστιο όγκο δεδομένων, σχεδόν 100 εκατ. πελάτες, και αυτό τον κάνει να υπερτερεί έναντι των άλλων αφού έχει αποδειχτεί, πέραν των άλλων, ότι είναι πολύ πιο επεκτάσιμος από τους υπόλοιπους.

Το κλειδί στη στο αλγόριθμο που χρησιμοποιεί το Amazon (Linden, et al, 2003) είναι η δημιουργία των πινάκων με τα όμοια αντικείμενα offline σε αντίθεση με την παραδοσιακή user-based μέθοδο. Επιπλέον, αφού ο αλγόριθμος προτείνει πολύ σχετικά αντικείμενα, η ποιότητα των συστάσεων είναι άριστη ακόμα και σε μικρό όγκο δεδομένων που βασίζεται σε 2-3 αντικείμενα. Συμπερασματικά ο αλγόριθμος του Amazon (item to item collaborative filtering) είναι ένας εξαιρετικός αλγόριθμος παραγωγής προσωποποιημένων συστάσεων.

Πανεπιστήμιο Πειραιώς

3.6.2 Αλγόριθμος youtube

3.6.2.1 Εισαγωγή

Το YouTube είναι ο δημοφιλέστερος και μεγαλύτερος διαδικτυακός τόπος μεταφόρτωσης και αναπαραγωγής βίντεο. Δημιουργήθηκε το Φεβρουάριο του 2005 και το Νοέμβριο του 2006 ονομάστηκε από το περιοδικό Time "Invention of the Year 2006" (Εφεύρεση της χρονιάς 2006). Τον Οκτώβριο του 2006, η εταιρεία αγοράστηκε από την Google με ανταλλαγή μετοχών αξίας 1,65 δισεκατομμυρίων δολαρίων ΗΠΑ και σήμερα λειτουργεί ως θυγατρική της Google. Υπολογίζεται ότι κάθε λεπτό αναφορτώνονται («ανεβαίνουν») στο YouTube περίπου 60 ώρες βίντεο. Ο αριθμός των ημερήσιων προβολών ανέρχεται στα 4 δισεκατομμύρια ενώ οι προβολές του 2011 έφτασαν στον εξωπραγματικό αριθμό του ενός τρισεκατομμυρίου. Ενδεικτικά αναφέρεται ότι για να παρακολουθηθούν όλα τα βίντεο του YouTube χρειάζεται συνεχόμενη παρακολούθηση αρκετών αιώνων (Cheng, et al 2008).

Από αυτό είναι φανερό ότι δεν γίνεται να λειτουργήσει ομαλώς και σωστά το YouTube χωρίς σύστημα παραγωγής συστάσεων αφού αν κάποιος επισκέπτης του ιστοτόπου ήθελε να παρακολουθήσει μια σειρά σχετικών βίντεο, θα χανόταν στο λαβυρινθώδες τοπίο αναρίθμητων βίντεο. Καταβάλλεται λοιπόν μεγάλη προσπάθεια ούτως ώστε οι συστάσεις να είναι προσωποποιημένες (Davidson, et al, 2010).

Για την παραγωγή των συστάσεων συνδυάζονται μέθοδοι καθαρής αναζήτησης με ερωτηματολόγιο (π.χ. δίνεται όσο το δυνατόν πληρέστερη και ταυτόχρονα συντομότερη περιγραφή του αναζητούμενου βίντεο) και αναζήτηση μέσω περιήγησης. Έτσι οι χρήστες είναι δυνατόν να αντιμετωπίσουν επιτυχέστερα τον τεράστιο όγκο πληροφοριών που διακινείται στο YouTube (Cha, et al, 2007).

Τα περισσότερα από τα βίντεο που ανεβαίνουν στο YouTube περιέχουν πολύ φτωχή ποσότητα μεταδεδομένων εκτός από τα λάθη τα οποία, δυστυχώς, συνηθίζονται πάρα πολύ και αυτό δυσχεραίνει τη διαδικασία παραγωγής συστάσεων. Αυτό που είναι επιθυμητό από την κοινότητα του YouTube είναι να παραχθούν συστάσεις όσο γίνεται σχετικότερες με τις πρόσφατες αναζητήσεις του χρήστη. Επίσης, ο χρήστης πρέπει να αντιλαμβάνεται για ποιο λόγο προτάθηκε ένα βίντεο σε αυτόν. Η ομάδα των προτεινόμενων βίντεο παράγεται χρησιμοποιώντας την πρόσφατη δραστηριότητα του χρήστη η οποία αφορά βίντεο που παρακολούθησε και αξιολόγησε.

Κατά τη διάρκεια της παραγωγής συστάσεων πρέπει να εξεταστούν δύο ευρείες κατηγορίες δεδομένων:

- δεδομένα που αφορούν το περιεχόμενο όπως η περιγραφή, ο τίτλος, κλπ
- δεδομένα που αφορούν τη δραστηριότητα του χρήστη τα οποία μπορούν να διαιρεθούν σε άμεσες και έμμεσες κατηγορίες για καλύτερη επεξεργασία.

Οι άμεσες κατηγορίες περιλαμβάνουν τη βαθμολογία και τη γενικότερη αξιολόγησης (favoriting/liking). Οι έμμεσες κατηγορίες περιλαμβάνουν τα αποτελέσματα που παράγονται από τον τρόπο που παρακολουθεί ένα βίντεο ο χρήστης (π.χ. παρακολουθεί ένα τμήμα του βίντεο).

3.6.2.2 Σχετικά βίντεο

Ένα πολύ σημαντικό βήμα για την παραγωγή συστάσεων είναι η κατασκευή του πίνακα αντιστοίχισης ενός βίντεο ui σε ένα σύνολο παρόμοιων ή σχετικών βίντεο Ri. Ως παρόμοια βίντεο ορίζονται εκείνα που ο χρήστης είναι πιθανό να παρακολουθήσει μετά την παρακολούθηση ενός αρχικού δοθέντος βίντεο u. Προκειμένου να δημιουργηθεί η αντιστοίχιση γίνεται χρήση μιας τεχνικής γνωστής ως συσχέτιση κανόνων εξόρυξης (Agrawal, et al, 1993) ή μέτρηση της συν-επισκεψιμότητας δύο βίντεο. Για μία δοθείσα περίοδο, συνήθως 24 ώρες, μετριέται για κάθε ζεύγος βίντεο (u_i ,u_j,) πόσο συχνά παρακολουθήθηκαν και τα δύο. Δηλώνοντας ότι η συν-επισκεψιμότητα μετριέται από το c_{ij} ορίζεται ως βαθμός (σκορ) συσχέτισης («συγγένειας») του βίντεο u_j με το βίντεο u_i το εξής:

$$r(u_i, u_j)=\frac{c_{ij}}{f(u_i,u_j)}$$

όπου c_i και c_j είναι συνολικές εμφανίσεις για τα βίντεο u_i και u_j αντίστοιχα ενώ $f(u_i, u_j)$ ένας παράγοντας ομαλοποίησης της συνολικής δημοτικότητας των δύο βίντεο. Έπειτα για ένα δοθέν βίντεο u_i επιλέγεται ένα σύνολο σχετικών βίντεο R_i ως τα πρώτα N υποψήφια βίντεο. Το σύνολο αυτό καθορίζεται από τη βαθμολογία. Συνεπώς κάποια βίντεο με χαμηλή επισκεψιμότητα ή αξιολόγηση δεν συμπεριλαμβάνονται σε αυτά που θα προταθούν. Σημειωτέον ότι η ανωτέρω περιγραφή είναι σχετικώς απλουστευμένη αφού στην πράξη αντιμετωπίζονται επιπρόσθετα προβλήματα όπως άσχετα βίντεο με περιγραφή που, όμως, ταιριάζει σε αυτό που ζητείται.

Τελικώς (Davidson, et al, 2010), για την παραγωγή των υποψήφιων προς σύσταση βίντεο συνδυάζονται τα σχετικά βίντεο με την πρόσφατη δραστηριότητα του

χρήστη στο YouTube. Αυτό μπορεί να περιλαμβάνει όλα τα βίντεο που έχουν παρακολουθηθεί και αυτά που έχουν λάβει οποιασδήποτε μορφής αξιολόγηση. Στη συνέχεια τα υποψήφια βίντεο βαθμολογούνται χρησιμοποιώντας την ποιότητα του βίντεο (δηλαδή πόσο δημοφιλές είναι, κάτι που μετριέται με επισκεψιμότητα, voting, sharing, κλπ) και τις ιδιαιτερότητες του χρήστη (τι επιλέγει να παρακολουθεί). Επειδή πρέπει να προταθεί μόνο ένας σχετικώς μικρός αριθμός βίντεο, γίνεται προσπάθεια να επιτευχθεί η χρυσή τομή μεταξύ του βαθμού συσχέτισης των βίντεο και της ποικιλομορφίας (επιλογή σχετικών από άλλες πηγές-κανάλια). Π.χ. βίντεο που είναι πολύ όμοια (π.χ. από το ίδιο κανάλι) δεν προτείνονται στο σύνολό τους ακόμα κι αν έχουν υψηλό βαθμό συσχέτισης.

Γενικά, το συνολικό σύστημα μπορεί να συνοψιστεί στα εξής τρία βήματα:

α) συλλογή δεδομένων,

β) παραγωγή συστάσεων και

γ) επίδοση των συστάσεων.

3.6.2.3 Αξιολόγηση και αποτελέσματα

Για την αξιολόγηση χρησιμοποιείται ένας συνδυασμός μετρικών. Οι βασικές μετρικές είναι η CTR (Click Through Rate, κλικ μέσω βαθμολόγησης), η μακρά CTR που μετρά μόνο κλικ που οδηγούν σε παρακολούθηση ουσιαστικού τμήματος του βίντεο, η διάρκεια της συνεδρίας (π.χ με την παρακολούθηση ενός μόνο δευτερολέπτου από το βίντεο γίνεται προσπάθεια να μην αξιολογηθεί θετικά), ο χρόνος μέχρι την πρώτη ουσιαστική παρακολούθηση του βίντεο και το ποσοστό των ατόμων που συστήνουν το βίντεο (Zhou, et al 2010).

Το σύστημα των συστάσεων έχει ενσωματωθεί στην αρχική σελίδα και παράγει συστάσεις που αντιπροσωπεύουν το 60% των βίντεο της κεντρικής σελίδας του YouTube. Ελέγχοντας την εκτέλεση των συστάσεων υπάρχει το πρόβλημα της παρουσίασης συστάσεων που τοποθετούνται στην κορυφή από προεπιλογή (presentation bias). Για να διορθωθεί αυτό, εξετάζεται η μετρική CTR από τις σελίδες που έχουν επισκεφθεί και συγκρίνονται οι συστάσεις με άλλα, αλγοριθμικώς παραγόμενα, προτεινόμενα σύνολα βίντεο όπως: α) Βίντεο με τις περισσότερες προβολές σε μία ημέρα, β) Πιο αγαπημένα βίντεο και γ) Βίντεο με την υψηλότερη βαθμολογία.

3.6.3 Αλγόριθμος Google News

Το Google News είναι μια υπηρεσία παροχής προσωποποιημένων πληροφοριών-ειδήσεων μέσα από ένα πλήθος ειδήσεων. Αυτό επιτυγχάνεται με αλγόριθμο συγκέντρωσης και αξιολόγησης αυτών. Η παραγωγή των συστάσεων στηρίζεται σε collaborative αλγόριθμο (Das, et al, 2007) και βασίζεται στο ιστορικό των «χτυπημάτων» (κλικ) των ενεργών χρηστών και στο ιστορικό της ευρύτερης κοινότητας. Υπενθυμίζεται ότι το κλικ προσμετράται ως θετική ψήφος. Ο αλγόριθμος προσπαθεί να δημιουργήσει ένα ειδικό τμήμα που θα έχει μια προσωποποιημένη λίστα με νέα άρθρα.

Οι βασικές προκλήσεις είναι:

α) να παραχθεί αυτή η λίστα σε πραγματικό χρόνο επιτρέποντας το πολύ ένα δευτερόλεπτο στο σύστημα να παράγει τη νέα σύσταση και

β) η λίστα αυτή να αλλάζει πολύ συχνά αφού υπάρχει μια συνεχής ροή νέων άρθρων ενώ την ίδια στιγμή κάποια άρθρα μπορεί να βγουν εκτός ημερομηνίας δηλαδή πάψουν να είναι επίκαιρα.

Επιπροσθέτως ένας ακόμα στόχος είναι να αντιδρά αμέσως στην αλληλεπίδραση του χρήστη και να λάβει υπόψη τα τελευταία άρθρα που διαβάζει. Λόγω της ποικιλίας των άρθρων μία αμιγώς memory-based τεχνική δεν είναι εφαρμόσιμη σε αυτή την περίπτωση. Έτσι χρησιμοποιείται ένας συνδυασμός model based και memory based τεχνικών. Το model based τμήμα βασίζεται σε δύο τεχνικές ομαδοποίησης, την PLSI (Probabilistic latent semantic analysis, Hofmann, 2004) και την MInHash Μέθοδο (Das, et al, 2007).

Η πρώτη (PLSI) είναι μία στατιστική τεχνική για την ανάλυση των δεδομένων και την ευρετηρίασή τους. Χρησιμοποιούνται κρυφές μεταβλητές με ένα πεπερασμένο σύνολο καταστάσεων για κάθε χρήστη. Έτσι, με ένα τέτοιο μοντέλο, μπορεί να αντιμετωπιστεί η πραγματικότητα ότι οι χρήστες έχουν παράλληλα πολλά ενδιαφέροντα σε διάφορα πεδία.

Όμως οι λεπτομέρειες αυτού του αλγορίθμου ξεφεύγουν αρκετά από τα πλαίσια της παρούσας εργασίας (εξάλλου δεν χρησιμοποιείται στην κατασκευασθείσα πλατφόρμα παραγωγής συστάσεων) και, συνεπώς, δεν θα αναλυθεί περαιτέρω.

Πανεπιστήμιο Πειραιώς

3.7 Αξιολόγηση των collaborative filtering αλγορίθμων

3.7.1 Τρία βασικά ερωτήματα κατά την αξιολόγηση

Ύστερα από δύο δεκαετίες ερευνών πάνω σε θέματα των collaborative filtering αλγορίθμων, η αξιολόγησή τους παρουσιάζει ακόμα αρκετά προβλήματα και αρκετές προκλήσεις. Αυτά συνοψίζονται κυρίως σε τρία σημαντικά ερωτήματα που αναφέρονται στη συνέχεια.

3.7.1.1 Τι αξιολογείται σε ένα σύστημα

Το πρώτο ερώτημα αφορά το τι τελικά πρέπει να αξιολογήσουμε σε ένα σύστημα παραγωγής συστάσεων. Ακόμα δεν υπάρχει κοινή γραμμή για το ποια χαρακτηριστικά ενός συστήματος αξιολογούνται. Η συνήθης τάση είναι να εξετάζεται και να αξιολογείται η ακρίβεια του αλγορίθμου. Οι Herlocker et al. (2004) αναγνωρίζουν τρεις τύπους μετρικών για να μετρηθεί η ποιότητα ενός αλγορίθμου:

- Η πρώτη μετρική θεωρείται η ακρίβεια της πρόβλεψης (prediction accuracy). Σύμφωνα με αυτήν μετριέται η διαφορά μεταξύ της βαθμολογίας που προβλέπει το σύστημα ότι ο χρήστης θα ψηφίσει το συγκεκριμένο αντικείμενο με την πραγματική βαθμολογία δηλαδή αυτήν που πραγματικά έδωσε. Αυτή συνήθως μετριέται με το Μέσο Απόλυτο Σφάλμα (MAE) αλλά και με το Μέσο Τετραγωνικό Σφάλμα (RMSE), το οποίο έγινε ιδιαίτερα δημοφιλές τα τελευταία χρόνια αφού χρησιμοποιήθηκε στο διαγωνισμό της Netflix. Άλλες φορές χρησιμοποιείται και το Κανονικοποιημένο Μέσο Απόλυτο Σφάλμα.
- Για την αξιολόγηση της ποιότητας ενός αλγορίθμου, οι Herlocker, et al (2004) θεωρούν ως δεύτερη μετρική την ταξινόμηση της ακρίβειας (classification accuracy). Βάσει αυτής μετριέται το κατά πόσο καλά το σύστημα διαφοροποιεί ένα καλό προϊόν από ένα κακό.
- Τέλος, χρησιμοποιείται η βαθμολογική ακρίβεια (rank accuracy), η οποία μετρά την ικανότητα του συστήματος για ταξινόμηση των προτεινόμενων αντικειμένων, έτσι όπως θα είχε κάνει και ο χρήστης.

Στη βιβλιογραφία μπορούν να βρεθούν μετρικές, οι οποίες δεν σχετίζονται με την ακρίβεια των αλγορίθμων. Μια τέτοια συνήθης μετρική που χρησιμοποιείται είναι ο μέσος όρος, ο οποίος υπολογίζει το ποσοστό των αντικείμενων για τα οποία το

Πανεπιστήμιο Πειραιώς

σύστημα είναι ικανό να πραγματοποιήσει προβλέψεις. Επιπλέον τα τελευταία χρόνια, οι ερευνητές έχουν προσπαθήσει να αξιολογήσουν το βαθμό ικανοποίησης των χρηστών σε σχέση με τα προτεινόμενα αντικείμενα.

3.7.1.2. Τρόπος εκτέλεσης της αξιολόγησης

Το δεύτερο ερώτημα που απασχολεί τη διεθνή ερευνητική κοινότητα αφορά τον τρόπο εκτέλεσης της αξιολόγησης. Στις έρευνες για τα collaborative filtering συστήματα, η πιο κοινή προσέγγιση είναι η αξιολόγηση να γίνεται σε σύστημα εκτός σύνδεσης (offline evaluation). Βασίζεται σε μια βάση δεδομένων που είναι διαιρεμένη σε δύο υποσύνολα: στο training και το evaluation. Στο υποσύνολο training βρίσκονται τα δεδομένα που ο αλγόριθμος 'γνωρίζει', αυτά δηλαδή που χρησιμοποιεί ο αλγόριθμος για να υπολογίσει τις προτάσεις ή την βαθμολογική ακρίβεια. Αυτά μετά συγκρίνονται με τα δεδομένα του υποσυνόλου evaluation (Herlocker et al, 2004).

Παρόλο που η μέθοδος αυτή είναι ευρέως διαδεδομένη, υπάρχουν πολύ μεγάλες διαφορές ανάμεσα σε έρευνες και μελέτες. Για παράδειγμα δεν υπάρχει ξεκάθαρη στρατηγική που θα ακολουθηθεί για να γίνει αυτή η διαίρεση στη βάση των δεδομένων και αυτό δυσχεραίνει την προσπάθεια για οποιαδήποτε σύγκριση των εργασιών αυτών.

3.7.1.3 Επιλογή του καλύτερου αλγορίθμου σε συγκεκριμένο πλαίσιο για βέλτιστο αποτέλεσμα

Το τρίτο και τελευταίο ερώτημα αφορά την επιλογή του καλύτερου αλγορίθμου (ή και αλγορίθμων) για ένα συγκεκριμένο και ειδικό πλαίσιο. Η ερώτηση αυτή είναι πολύ δύσκολο να απαντηθεί βασιζόμενοι στη διεθνή βιβλιογραφία. Αρκετοί αξιολογούν τον προτεινόμενο αλγόριθμο χρησιμοποιώντας μετρικές και μεθοδολογίες οι οποίες προσφέρουν τα καλύτερα αποτελέσματα ή μελετούν τις ευνοϊκότερες συνθήκες για έναν συγκεκριμένο αλγόριθμο. Είναι δύσκολη η αξιολόγηση μιας οικογένειας αλγορίθμων, καθώς και η συμπεριφορά τους σε συγκεκριμένες καταστάσεις ή και η σύγκριση αυτών μεταξύ τους (Herlocker et al, 2004).

Πανεπιστήμιο Πειραιώς

3.7.2 Μετρικές

Ακολουθεί η παράθεση κάποιων σημαντικών μετρικών (Cacheda, et al, 2011) :

1. Μέσος Όρος (Coverage) που, όπως προαναφέρθηκε, χρησιμοποιείται για τον υπολογισμό του ποσοστού των αντικειμένων που το σύστημα είναι ικανό να προτείνει. Δύναται να χρησιμοποιηθεί για τον εντοπισμό αλγορίθμων που προτείνουν μικρό αριθμό αντικειμένων. Αυτά είναι συνήθως πολύ δημοφιλή αντικείμενα που ο χρήστης ήδη γνωρίζει χωρίς τη βοήθεια του συστήματος. Επομένως ένας υψηλός μέσος όρος στη βαθμολογία όχι μόνο είναι επιθυμητός αλλά βοηθά στην καλυτέρευση του βαθμού εμπιστοσύνης των αποτελεσμάτων των μετρικών ακριβείας.

2. Ακρίβεια πρόβλεψης (Prediction Accuracy) που μετράται με διάφορους τρόπους. Αναφέρονται οι πιο δημοφιλείς:

   2.1 Μέσο Απόλυτο Σφάλμα (Mean absolute error, MAE)

   Βάσει αυτού γίνεται η μέτρηση της διαφοράς, σαν απόλυτη τιμή, μεταξύ της πρόβλεψης του αλγορίθμου και της πραγματικής βαθμολογίας του χρήστη. Υπολογίζεται βάσει όλων των διαθέσιμων βαθμολογιών με τον εξής τύπο:

   $$|\bar{E}| = \frac{\sum_i^N |p_i - v_i|}{N}$$

   όπου p_i η πρόβλεψη για το *i* αντικείμενο που προτείνεται και v_i η βαθμολογία του *i* αντικειμένου που έχει προταθεί από τον ενεργό χρήστη.

   2.2 Μέσο τετραγωνικό σφάλμα (Root mean squared error, RMSE)

   Είναι παρόμοια, με την προηγούμενη, μετρική και δίνει μεγαλύτερη έμφαση όταν υπάρχουν μεγαλύτερα σφάλματα. Υπολογίζεται με τον παρακάτω τύπο:

   $$|\bar{E}| = \sqrt{\frac{\sum_i^N (p_i - v_i)^2}{N}}$$

   Όπου οι μεταβλητές ορίζονται όπως στην περίπτωση του MAE.

   2.3 Ακρίβεια και Ανάκληση (Precision and recall)

Πανεπιστήμιο Πειραιώς

Η ακρίβεια, εδώ, ορίζεται ως ο λόγος των σχετικών αντικειμένων ως προς τα προτεινόμενα αντικείμενα, ενώ η ανάκληση αφορά το ποσοστό των σχετικών αντικειμένων που έχουν προταθεί σε σχέση με το συνολικό αριθμό των σχετικών αντικειμένων. Για ένα σύστημα είναι επιθυμητό να έχει υψηλή ακρίβεια και ανάκληση στις τιμές του. Ο υπολογισμός γίνεται από τον τύπο:

$$F_1 = \frac{2PR}{P+R}$$

όπου P είναι η ακρίβεια και R η ανάκληση όπως ορίστηκαν παραπάνω.

Πανεπιστήμιο Πειραιώς

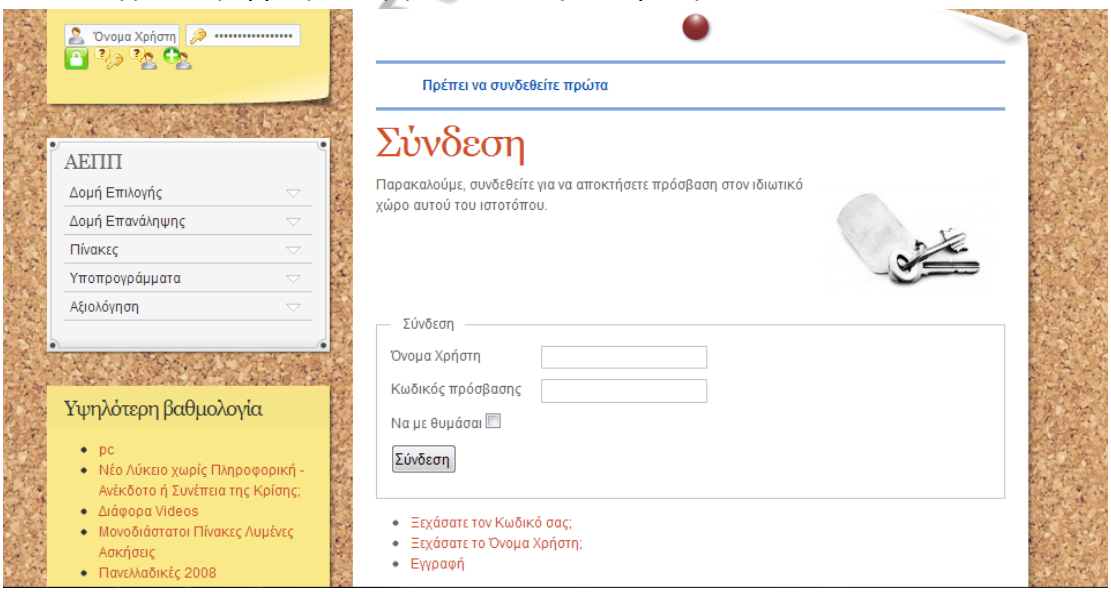
4 Πλατφόρμα και μέθοδος παραγωγής συστάσεων

4.1 Περιγραφή της πλατφόρμας

Θα ακολουθήσει μία σχετικά σύντομη περιγραφή της πλατφόρμας που δημιουργήθηκε στα πλαίσια του μαθήματος «Εφαρμογές και Υπηρεσίες του Παγκόσμιου Ιστού» του 3ου εξαμήνου του Προγράμματος Μεταπτυχιακών Σπουδών «Διδακτική της Τεχνολογίας και Ψηφιακά Συστήματα» στην κατεύθυνση Ηλεκτρονική Μάθηση με τη χρήση του CMS (Content Management System, Συστήματος Διαχείρισης Περιεχομένου) JOOMLA 1.5.22.

Η πλατφόρμα αυτή αφορά, όπως αναφέρθηκε ανωτέρω, το μάθημα «Ανάπτυξη Εφαρμογών σε Προγραμματιστικό Περιβάλλον» (ΑΕΠΠ) και απευθύνεται κυρίως σε μαθητές της Γ' τάξης του Γενικού Λυκείου και της Δ' τάξης του Εσπερινού Λυκείου. Μπορούν όμως ακόμα και καθηγητές Πληροφορικής να χρησιμοποιήσουν το παρεχόμενο εκπαιδευτικό υλικό καθώς και μαθητές του Τομέα Πληροφορικής και Δικτύων των Επαγγελματικών Λυκείων οι οποίοι έχουν καλή γνώση των αλγοριθμικών δομών αφού διδάσκονται σχετικά μαθήματα.

Σημειωτέον ότι η πρόσβαση επιτρέπεται όχι μόνο σε εγγεγραμμένους χρήστες (registered users) αλλά και σε απλούς επισκέπτες. Στην τελευταία περίπτωση υπάρχουν κάποιοι περιορισμοί στη χρήση της πλατφόρμας οι οποίοι θα αναλυθούν στη συνέχεια. Η εγγραφή γίνεται πολύ εύκολα από ειδικό μενού στην αρχική σελίδα της πλατφόρμας όπως φαίνεται στην επόμενη εικόνα.



Εικόνα 13. Φόρμα εγγραφής του χρήστη στην πλατφόρμα

Κατά τη διαδικασία εγγραφής ο υποψήφιος νέος χρήστης θα πρέπει να πληκτρολογήσει έναν ενεργό λογαριασμό ηλεκτρονικού ταχυδρομείου. Για να μπορέσει ο νέος χρήστης να χρησιμοποιήσει το λογαριασμό του είναι απαραίτητο να γίνει έγκριση από το διαχειριστή του συστήματος. Όταν εγκριθεί η εγγραφή, αποστέλλεται στο δοθέντα λογαριασμό ηλεκτρονικού ταχυδρομείου του νέου χρήστη ένας σύνδεσμος ο οποίος επιλέγεται από το χρήστη και ολοκληρώνεται κατ' αυτόν τον τρόπο η διαδικασία εγγραφής. Τονίζεται για μια ακόμα φορά ότι η εγγραφή δεν είναι απαραίτητη αλλά σε αυτήν την περίπτωση δεν θα είναι διαθέσιμες όλες οι δυνατότητες της πλατφόρμας όπως η παραγωγής πλήρως εξατομικευμένων συστάσεων.

Για τις ανάγκες της παρούσας εργασίας κρίθηκε σκόπιμο να αφαιρεθούν κάποια τμήματα της πλατφόρμας διότι δεν συνεισέφεραν προς την εκπλήρωση του σκοπού. Επίσης, προστέθηκαν αρκετές ασκήσεις ούτως ώστε να υπάρχει ικανοποιητικός αριθμός για βαθμολόγηση από τους χρήστες-μαθητές. Επιπροσθέτως, οι ασκήσεις έχουν χωριστεί σε κατηγορίες ανάλογα με το θεματικό περιεχόμενο σύμφωνα με το σχολικό εγχειρίδιο ούτως ώστε να διευκολύνεται ο χρήστης. Τέλος, οι ασκήσεις μπορούν πλέον να προταθούν στους χρήστες μετά την υλοποίηση του κατάλληλου αλγορίθμου παραγωγής συστάσεων, κάτι το οποίο θα αναλυθεί διεξοδικώς στη συνέχεια.

Συγκεκριμένα η πλατφόρμα περιλαμβάνει τα εξής:

1. Κεντρικό μενού με τις επιλογές Αρχική, Blog, Forum, Wiki και Επικοινωνία
2. Μενού «ΑΕΠΠ» με ένα τμήμα των επιλογών του υπομενού Αρχική του κεντρικού μενού (Δομή επιλογής, Δομής Επανάληψης, Πίνακες, Υποπρογράμματα) και μία νέα, την Αξιολόγηση
3. Δυνατότητα σύνδεσης με δημοφιλείς υπηρεσίες κοινωνικής δικτύωσης όπως facebook, flick, myspace, twitter και youtube.
4. Δυνατότητα κλήσης μέσω Skype το οποίο παρέχει όχι μόνο επικοινωνία μέσω φωνής (VOIP) αλλά και chat (ανταλλαγή γραπτών μηνυμάτων on line).
5. Άμεση σύνδεση από την κεντρική σελίδα μας προς το λογισμικό «pseudoglossa» (www.pseudoglossa.gr) το οποίο παρέχει ένα πολύ ενδιαφέρον περιβάλλον εξάσκησης και εκμάθησης των αλγοριθμικών δομών.
6. Άμεση σύνδεση από την κεντρική σελίδα μας προς τον ιστότοπο (http://www.spinet.gr/glossomatheia) που υποστηρίζει το λογισμικό «Γλωσσομάθεια» το οποίο παρέχει στο χρήστη τη δυνατότητα να εκτελέσει προγράμματα και να κατανοήσει καλύτερα το μάθημα.
7. Άμεση σύνδεση από την κεντρική σελίδα μας προς τον ιστότοπο http://users.sch.gr/alkisg/ που υποστηρίζει λογισμικό «ΓΛΩΣΣΑ» όπου ο

Πανεπιστήμιο Πειραιώς

μαθητής μπορεί να το κατεβάσει και να το δουλέψει τοπικά στον υπολογιστή του και που λειτουργεί με παρόμοιο τρόπο με τη «Γλωσσομάθεια» και την «pseudoglossa».

8. Δυνατότητα αναζήτησης οποιουδήποτε λεκτικού μέσα στον ιστότοπο.
9. Σε κάθε δημοσιευμένο άρθρο παρέχεται η δυνατότητα διαμοιρασμού (sharing) στο facebook και ανακοίνωσης στο twitter όπως και η δυνατότητα αξιολόγησης με τη γνωστή επιλογή “like” του facebook.
10. Υπάρχει δυνατότητα αξιολόγησης των άρθρων (rating-voting).
11. Δυνατότητα προσθήκης ετικετών (tag) σε κάθε θέμα

Για την υλοποίηση των ανωτέρω λειτουργιών χρησιμοποιήθηκαν πρόσθετα προγράμματα ή, αλλιώς, επεκτάσεις (extensions). Γενικώς στο JOOMLA η λογική που διέπει τη λειτουργία του είναι ότι ο χρήστης μπορεί να χρησιμοποιήσει έτοιμα προγράμματα που μπορεί να αντλήσει από το Διαδίκτυο και τα οποία απαιτούν εγκατάσταση και, ενίοτε, ειδικές ρυθμίσεις. Τα προγράμματα αυτά χωρίζονται σε τέσσερις βασικές κατηγορίες: πρότυπα (templates), ενθέματα (modules), πρόσθετα (plugins) και εφαρμογές (components). Μέσω λοιπόν αυτών των προγραμμάτων ο διαχειριστής του συστήματος μπορεί να υλοποιήσει πλήθος λειτουργιών και, τελικά, να κατασκευάσει ένα δυναμικό ιστότοπο χωρίς να απαιτούνται ιδιαίτερες προγραμματιστικές γνώσεις. Λόγω του ότι το JOOMLA λειτουργεί με PHP και MySQL ενώ χρησιμοποιεί και HTML είναι ευκόλως αντιληπτό ότι η γνώση των ανωτέρω επιτρέπει στο διαχειριστή να τροποποιήσει, κατά το δοκούν, τις χρησιμοποιηθείσες επεκτάσεις ούτως ώστε να προσθέσει επιπλέον λειτουργικότητα ή, σε τελική ανάλυση, να επιτύχει κάτι διαφορετικό από αυτό που παρέχουν τα έτοιμα προγράμματα. Τέλος, με τη χρήση των επεκτάσεων παράγονται κάποιες συστάσεις, που δεν είναι πλήρως προσωποποιημένες, προς τους χρήστες της πλατφόρμας και αυτό περιγράφεται αναλυτικότερα στη συνέχεια της παρούσας μελέτης.

Πανεπιστήμιο Πειραιώς

4.2 Παραγωγή συστάσεων στο σύστημα

Για να παραχθούν λοιπόν στην παρούσα εργασία συστάσεις σε αρχικό επίπεδο χρειάστηκε να εγκατασταθούν και ρυθμιστούν καταλλήλως κάποιες επεκτάσεις. Αυτές οι αρχικές συστάσεις παράγονται με τους εξής τρόπους:

1. Μέσω σχετικών άρθρων (Most related articles)
2. Μέσω πιο πολυδιαβασμένων άρθρων (most read articles)
3. Μέσω άρθρων με υψηλότερη βαθμολογία (top rated articles)

Ο πρώτος τρόπος επιτυγχάνεται μέσω σχετικών άρθρων που μπορεί να δει ένας χρήστης κάτω από το άρθρο που μελετά. Αυτό γίνεται με τη χρησιμοποίηση των ετικετών (tags) κάθε άρθρου όπως φαίνεται στο επόμενο στιγμιότυπο:

Εικόνα 14: Πρόταση πιο σχετικών άρθρων βάσει ετικετών

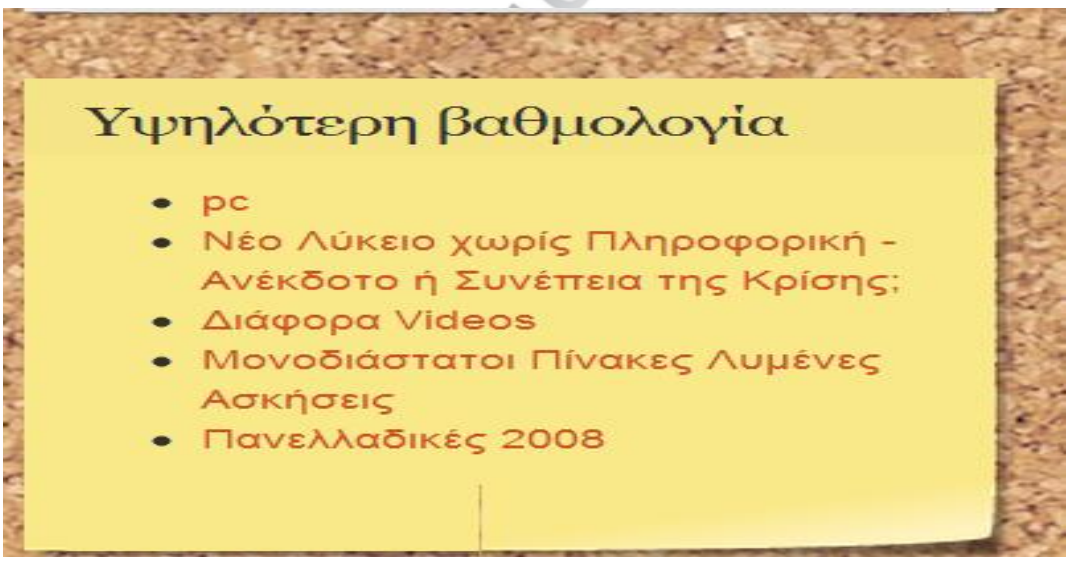
Η παραγωγή αυτών των συστάσεων ανήκει στην κατηγορία των content-based. Θεωρούνται προσωποποιημένες αφού τα άρθρα, που προτείνονται, έχουν να κάνουν με το περιεχόμενο που μελετά τη δεδομένη στιγμή ο χρήστης και δεν είναι δυνατό να παρουσιάσει άσχετα άρθρα με ανόμοιες ετικέτες. Η σύσταση μέσω ετικετών κατέστη δυνατή έπειτα από την εγκατάσταση της εφαρμογής joomla_tag που καταφορτώθηκε (downloaded) από τον ιστότοπο www.joomla.org.

Επίσης, στην πλατφόρμα ο χρήστης έχει τη δυνατότητα να δει τα *πιο πολυδιαβασμένα άρθρα*. Αυτό αποτελεί παραγωγή σύστασης αφού αυτά προκύπτουν από τον αριθμό των κλικ που γίνονται σε κάθε άρθρο. Το κάθε κλικ λαμβάνεται ως θετική ψήφος και κατ' αυτόν τον τρόπο προκύπτει η σύσταση όπως μπορούμε να δούμε στο επόμενο στιγμιότυπο.

Εικόνα 15. Παραγωγή συστάσεων μέσω πιο πολυδιαβασμένων άρθρων

Αυτό επιτυγχάνεται με τη χρήση του ενθέματος most_read το οποίο μεταφορτώθηκε από το Διαδίκτυο (www.joomla.org) και εγκαταστάθηκε μετά από κατάλληλες ρυθμίσεις στην πλατφόρμα από το περιβάλλον του Διαχειριστή.

Επιπλέον η πλατφόρμα παρέχει τη δυνατότητα στους χρήστες να διαβάσουν τα άρθρα τα οποία έχουν τις *υψηλότερες βαθμολογίες* όπως φαίνεται παρακάτω.



Εικόνα 16. Πρόταση των άρθρων με την υψηλότερη βαθμολογία

Αυτή η δυνατότητα παρέχεται από το ένθεμα mod_db8bestratedcontent που καταφορτώθηκε από τον ιστότοπο www.joomla.org. Εν κατακλείδι, οι δύο τελευταίες παραγωγές συστάσεων (πιο πολυδιαβασμένα-most read και υψηλότερη βαθμολογία-top rated) δεν μπορούν να θεωρηθούν προσωποποιημένες, αφού εμφανίζονται στους χρήστες ανεξάρτητα με το τι είδος άρθρα τους απασχολούν.

Τέλος, από τα ανωτέρω συνάγεται εύκολα ότι ο βαθμός εξατομίκευσης των παραγόμενων συστάσεων δεν είναι μεγάλος και γι' αυτό στην παρούσα εργασία επιδιώκεται η επίτευξη αυτού του σκοπού μέσω της χρήσης ενός κατάλληλου αλγορίθμου και της συνεπακόλουθης αναγκαίας περαιτέρω τροποποίησης του συστήματος.

4.3 Επιλογή μεθόδου παραγωγής συστάσεων

4.3.1 Επιλογή αλγορίθμου παραγωγής συστάσεων

Καταρχάς, πρέπει να τονιστεί ότι για την παρούσα εργασία μάς ενδιαφέρουν περισσότερο τα τμήματα που αφορούν την αξιολόγηση των άρθρων διότι, όπως θα εξηγηθεί αναλυτικά στη συνέχεια, βάσει αυτών θα παραχθούν οι πλήρως εξατομικευμένες συστάσεις προς τους εγγεγραμμένους χρήστες του ιστοτόπου.

Σαν μέθοδο παραγωγής συστάσεων θα χρησιμοποιηθεί η collaborative η οποία ελέγχει την ομοιότητα των χρηστών και προτείνει άρθρα σε χρήστες που είναι αρκετά όμοιοι όπως περιγράφηκε στην παράγραφο 2.4.2.

Η μέθοδος content based δεν επιλέγεται διότι είναι λιγότερο ευέλικτη από την collaborative. Όπως αναφέρθηκε στην παράγραφο 2.4.1.4 (περιορισμοί-μειονεκτήματα) τα συστήματα content based δεν μπορούν ξεχωρίσουν ένα καλογραμμένο άρθρο από ένα κακογραμμένο αν αυτά έχουν πολλές κοινές λέξεις-κλειδιά. Επίσης, σε αυτά είναι σημαντικό ότι δίνονται συστάσεις μόνο για παρόμοια αντικείμενα αλλά αποκλείονται εξαρχής τα ανόμοια ενώ πιθανόν να ενδιαφέρουν το χρήστη. Αυτό είναι ίσως και το σπουδαιότερο εγγενές μειονέκτημα της μεθόδου δηλαδή τα ήδη βαθμολογημένα άρθρα θα συνεχίσουν να προτείνονται όπως και τα όμοιά τους ενώ αυτά που δεν είναι τόσο όμοια, θα αγνοούνται ακόμα και αν είναι αξιόλογα. Επιπροσθέτως μπορεί να προτείνουν κάτι που είναι υπερβολικά όμοιο και πρακτικά δεν έχει καμία χρησιμότητα αφού θα ήταν προφανής επιλογή και, τέλος, ο χρήστης πρέπει να έχει αξιολογήσει αρκετά αντικείμενα πριν το σύστημα κατανοήσει τις προτιμήσεις του και του δώσει αξιόπιστες συστάσεις (Shardanand & Maes, 1995).

Από την άλλη πλευρά, τα συστήματα collaborative filtering δεν είναι τόσο ευαίσθητα και επιρρεπή σε αυτά τα προβλήματα αφού η παραγωγή συστάσεων δεν

Πανεπιστήμιο Πειραιώς

βασίζεται στο περιεχόμενο αλλά στις απόψεις των χρηστών (Shardanand & Maes, 1995). Συγκεκριμένα, το σύστημα δεν χρειάζεται να αναλύει περιεχόμενο αλλά μόνο τις προτιμήσεις των χρηστών. Έτσι ακόμα και αν ένα άρθρο έχει ψηφιστεί λίγες φορές, δεν αποκλείεται να προταθεί αν ο ενεργός χρήστης βρεθεί όμοιος με κάποιον άλλο που το έχει ψηφίσει. Με δεδομένο λοιπόν ότι οι συστάσεις σε αυτήν την περίπτωση καθορίζονται από τα ενδιαφέροντα και τις προτιμήσεις των χρηστών, οι οποίες είναι φυσιολογικό να καλύπτουν σχεδόν ολόκληρο το φάσμα των ασκήσεων, προκύπτει ότι θα υπάρχει ένα πιο ευέλικτο και δυναμικό σύστημα. Συνεπώς, αν κάποιοι χρήστες αλλάζουν προτιμήσεις τότε το σύστημα δύναται να προσαρμοστεί αναλόγως.

Τέλος, δεν πρέπει να παραβλέψουμε το γεγονός ότι στο σύστημα παράγονται ήδη ενός είδους εξατομικευμένες συστάσεις μέσω της επέκτασης που επιστρέφει τα πιο σχετικά άρθρα και, όπως αναφέρθηκε, αυτή η σύσταση είναι βασισμένη στο περιεχόμενο (content based). Επομένως, είναι προτιμότερο να δοκιμαστεί και μια άλλη μέθοδος.

Συνεπώς οι τρεις βασικοί λόγοι που επιλέχθηκε μέθοδος collaborative filtering είναι ότι

- παρέχει μεγαλύτερη ευελιξία αφού δεν αποκλείει απαραιτήτως άρθρα που δεν έχουν λάβει πολλές ψήφους έως τη δεδομένη στιγμή που πρέπει να γίνει η παραγωγή συστάσεων
- δεν αντιμετωπίζει στον ίδιο βαθμό τα εγγενή μειονεκτήματα των μεθόδων content based και
- η μέθοδος content based έχει ήδη εφαρμοστεί μέσω της επέκτασης που αφορά τα πιο σχετικά άρθρα.

4.3.2 Ανάλυση επιλεχθέντος αλγορίθμου κατηγορίας user based

4.3.2.1 Συντελεστής Pearson

Ύστερα από μελέτη των τεσσάρων βασικών κατηγοριών παραγωγής συστάσεων (content, collaborative filtering, knowledge-based και hybrid), οι οποίοι αναλύονται στο θεωρητικό κομμάτι της εργασίας, καταλήξαμε όπως αναφέρθηκε σε αλγόριθμο υποκατηγορίας user-based της κατηγορίας collaborative filtering.

Αυτού του τύπου οι αλγόριθμοι ακολουθούν μια διαδικασία τριών βημάτων που περιλαμβάνει:

Πανεπιστήμιο Πειραιώς

- Υπολογισμό της ομοιότητας μεταξύ του ενεργού χρήστη και των υπολοίπων χρηστών.
- Επιλογή ενός υποσυνόλου των χρηστών (μιας γειτονιάς) βάσει της ομοιότητάς τους με τον ενεργό χρήστη.
- Υπολογισμό της πρόβλεψης χρησιμοποιώντας τις βαθμολογίες της γειτονιάς.

Η αρχική επιλογή μας για το πρώτο βήμα είναι ο περιορισμένος συντελεστής Pearson (Constrained Pearson), ο οποίος χρησιμοποιεί την κεντρική τιμή της βαθμολογίας των χρηστών (στην πλατφόρμα χρησιμοποιείται η κλίμακα 1 έως 5 και γι' αυτό επιλέγεται η τιμή 3) αντικαθιστώντας έτσι το μέσο όρο των βαθμολογιών του κάθε χρήστη. Ο τύπος που χρησιμοποιείται για να βρούμε την ομοιότητα αυτή είναι ο παρακάτω (Shardanand & Maes, 1995):

Τύπος εύρεσης ομοιότητας δύο χρηστών

$$s(a,u) = \frac{\sum_{i \in I_a \cap I_v}(v_{a_i} - 3)(v_{u_i} - 3)}{\sqrt{\sum_{i \in I_a \cap I_v}(v_{a_i} - 3)^2 \sum_{i \in I_a \cap I_v}(v_{u_i} - 3)^2}}$$

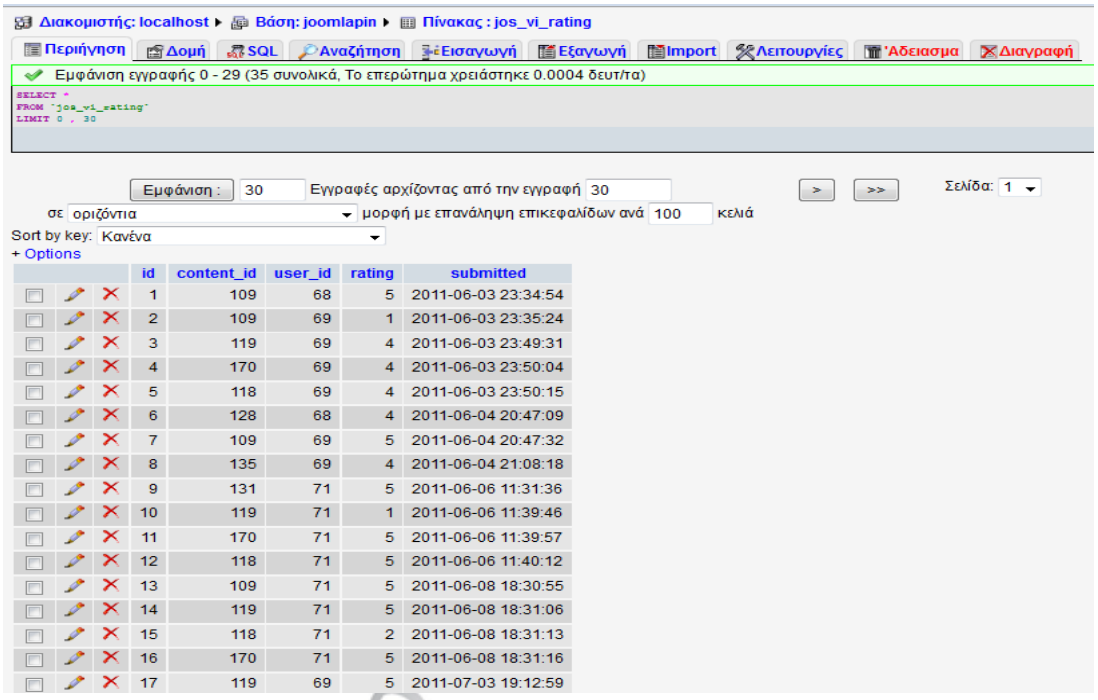
Υπενθυμίζεται ότι, με v_{a_i} συμβολίζουμε τη βαθμολογία του ενεργού χρήστη (αυτού ουσιαστικά που θέλουμε να του παραχθούν συστάσεις) στο i αντικείμενο, ενώ με v_{u_i} τη βαθμολογία του χρήστη u (είναι ο χρήστης για τον οποίο ενδιαφερόμαστε να ελέγξουμε αν είναι όμοιος με τον ενεργό) στο i αντικείμενο. Ο υπολογισμός γίνεται βάσει των βαθμολογιών των αντικειμένων που έχουν βαθμολογηθεί και από τους δύο χρήστες ($I_a \cap I_v$).

Έπειτα η επιλογή της γειτονιές των 'περισσότερο' ομοίων γίνεται με τη βοήθεια ενός προκαθορισμένου ορίου. Αυτό το όριο, λόγω μικρού αριθμού χρηστών και άρθρων καθορίστηκε στο 0.6 (Shardanand & Maes, 1995, Leydesdorff, 2008, Li & Zhang, 2009). Έτσι αν ο δείκτης ομοιότητας ενός χρήστη με τον ενεργό υπερβαίνει ή είναι ίσος με το όριο 0.6 τότε αυτός ο χρήστης εντάσσεται στη γειτονιά.

Πανεπιστήμιο Πειραιώς

4.3.2.2 Επεξεργασία βαθμολογίας κάθε χρήστη

Για να υπολογιστεί ο συντελεστής Pearson μεταξύ δύο χρηστών και, κατά συνέπεια, κάθε ζευγαριού χρηστών που είναι υποψήφιοι όμοιοι, θα χρειαστεί να αξιοποιηθεί η βαθμολογία που έχει δώσει ο κάθε χρήστης. Αυτή η βαθμολογία θα αντληθεί από πίνακα της Βάσης στον οποίο είναι καταχωρισμένες όλες οι βαθμολογίες που έχουν δώσει οι χρήστες της πλατφόρμας. Ο εν λόγω πίνακας είναι ο jos_vi_rating, ένα στιγμιότυπο του οποίου παρατίθεται στη συνέχεια


Διακομιστής: localhost ▸ Βάση: joomlapin ▸ Πίνακας : jos_vi_rating

Περιήγηση | Δομή | SQL | Αναζήτηση | Εισαγωγή | Εξαγωγή | Import | Λειτουργίες | Άδειασμα | Διαγραφή

Εμφάνιση εγγραφής 0 - 29 (35 συνολικά, Το επερώτημα χρειάστηκε 0.0004 δευτ/τα)

```
SELECT *
FROM `jos_vi_rating`
LIMIT 0 , 30
```

Εμφάνιση : 30 Εγγραφές αρχίζοντας από την εγγραφή 30 > >> Σελίδα: 1

σε οριζόντια μορφή με επανάληψη επικεφαλίδων ανά 100 κελιά

Sort by key: Κανένα

+ Options

	id	content_id	user_id	rating	submitted
	1	109	68	5	2011-06-03 23:34:54
	2	109	69	1	2011-06-03 23:35:24
	3	119	69	4	2011-06-03 23:49:31
	4	170	69	4	2011-06-03 23:50:04
	5	118	69	4	2011-06-03 23:50:15
	6	128	68	4	2011-06-04 20:47:09
	7	109	69	5	2011-06-04 20:47:32
	8	135	69	4	2011-06-04 21:08:18
	9	131	71	5	2011-06-06 11:31:36
	10	119	71	1	2011-06-06 11:39:46
	11	170	71	5	2011-06-06 11:39:57
	12	118	71	5	2011-06-06 11:40:12
	13	109	71	5	2011-06-08 18:30:55
	14	119	71	5	2011-06-08 18:31:06
	15	118	71	2	2011-06-08 18:31:13
	16	170	71	5	2011-06-08 18:31:16
	17	119	69	5	2011-07-03 19:12:59

Εικόνα 17. Πίνακας με τις βαθμολογίες που έχουν δώσει οι χρήστες της πλατφόρμας

Ο ανωτέρω πίνακας έχει τρεις βασικές στήλες: η $1^{η}$ είναι η content_id, δηλαδή δίνει έναν αριθμό ο οποίος παίζει το ρόλο του εκπαιδευτικού αντικειμένου-άρθρου που βαθμολογείται, η $2^{η}$ είναι το user_id που συμβολίζει το όνομα-αριθμό του χρήστη και η $3^{η}$ στήλη είναι η βαθμολογία που έδωσε στο άρθρο της $1^{ης}$ στήλης.

Χρειάζεται να κατασκευαστεί για κάθε χρήστη ένας δισδιάστατος πίνακας (τύπου array), ο οποίος θα έχει τα αντικείμενα και τις αντίστοιχες βαθμολογίες. Τα αντικείμενα που δεν έχουν βαθμολογηθεί, πρέπει να εισαχθούν στον πίνακα αυτό και στη θέση της βαθμολογίας τους το μηδέν (0).

Η κατασκευή αυτών των πινάκων (για όλους χρήστες) γίνεται καταμετρώντας όλους τους χρήστες για να γνωστοποιηθεί ο αριθμός των πινάκων που θα κατασκευαστούν (δεν χρειάζεται αυστηρή καταμέτρηση αφού αρκεί κάτι δυναμικό) και όλων των αντικειμένων. Εν συνεχεία στις αντίστοιχες θέσεις θα πρέπει να

Πανεπιστήμιο

τοποθετηθούν οι βαθμολογίες. Όπως προαναφέρθηκε, όποιο αντικείμενο δεν έχει βαθμολογηθεί, πρέπει να τοποθετηθεί σε αυτό μηδέν.

Ακολουθεί ένα παράδειγμα για καλύτερη κατανόηση.

Παράδειγμα δημιουργίας πινάκων με τη βαθμολογία που έχει δώσει κάθε χρήστης

Στον παρακάτω πίνακα αποτυπώνεται η βαθμολογία που έχει δώσει κάθε ένας από τους χρήστες στα αντικείμενα

Content_id (αναγνωριστικό αντικειμένου)	User_id (αναγνωστικό χρήστη)	Rating (Βαθμολογία)
10	100	5
10	200	4
20	100	3
10	300	5
30	400	1

Στο εξεταζόμενο παράδειγμα υπάρχουν λοιπόν 3 αντικείμενα (με ονόματα-αριθμούς 10, 20, 30), 4 χρήστες (100, 200, 300, 400) και οι αντίστοιχες βαθμολογίες.

Συνεπώς, επιθυμούμε να κατασκευάσουμε 4 πίνακες(έστω I ο κάθε πίνακας) για καθέναν από τους 4 χρήστες :

Για το χρήστη 100

10	20	30
5	3	0

Αυτός είναι ο πίνακας των βαθμολογιών του χρήστη με id 100 όπου η πρώτη γραμμή αφορά τα αντικείμενα και η δεύτερη τη βαθμολογία. Στα δύο πρώτα αντικείμενα τοποθετήθηκαν οι βαθμολογίες του δηλαδή το 5 που έδωσε στο αντικείμενο 10 και το 3 που έδωσε στο αντικείμενο 20 , ενώ στο $3^{ο}$ αντικείμενο (με όνομα 30), εισήχθη 0 (μηδέν) αφού δεν το έχει βαθμολογήσει ακόμα (εννοείται ότι δεν αναγκαίο να το ψηφίσει).

Οι υπόλοιποι πίνακες I που χρειάζονται είναι:

Πανεπιστήμιο Πειραιώς

Για το χρήστη 200

10	20	30
4	0	0

Για το χρήστη 300

10	20	30
5	0	0

Για το χρήστη 400

10	20	30
0	0	1

Χρησιμοποίηση μονοδιάστατων πινάκων

Στο σημείο αυτό πρέπει να αναφερθεί ότι δεν χρησιμοποιήθηκαν δισδιάστατοι πίνακες (τύπου array) αφού η PHP δεν έχει ευχέρεια χρήσης σε κάτι τέτοιο και αυτό θα δημιουργούσε πρόσθετες δυσχέρειες στη συνολική διαδικασία. Λόγω αυτού προτιμήθηκε να χρησιμοποιηθούν μονοδιάστατοι πίνακες των οποίων ο χειρισμός είναι πιο εύκολος. Έτσι ο κάθε χρήστης έχει ένα μονοδιάστατο πίνακα (διάνυσμα) με τις βαθμολογίες που έχει δώσει όπου όμως θα υπάρχει η πληροφορία σε ποιο άρθρο αναφέρεται η συγκεκριμένη βαθμολογία. Αυτό φαίνεται καλύτερα στο ακόλουθο παράδειγμα:

I_{100} = [5, 3, 0] βαθμολογία

↓ ↓ ↓

10 20 30 αντικείμενο

όπου ο χρήστης 100 έχει δώσει το βαθμό 5 στο αντικείμενο 10, το βαθμό 3 στο αντικείμενο 20 ενώ δεν έχει ψηφίσει το αντικείμενο 30. Το ανωτέρω επιτυγχάνεται με τη χρήση των συσχετιζόμενων πινάκων (associative arrays) στους οποίους συσχετίζεται ένα συγκεκριμένο κλειδί με μία συγκεκριμένη τιμή. Στην περίπτωσή μας το κλειδί θα είναι το αντικείμενο-άρθρο και η τιμή στη θέση με το κλειδί αυτό θα είναι η βαθμολογία.

Αυτό υλοποιείται επιτυχώς με τον παρακάτω αλγόριθμο που διατυπώνεται σε φυσική γλώσσα κατά βήματα:

Πανεπιστήμιο Πειραιώς

Αλγόριθμος αποθήκευσης βαθμολογίας κάθε χρήστη

1. Για τον τρέχοντα (ενεργό) χρήστη κάνε σύνδεση στη Βάση Δεδομένων του συστήματος χρησιμοποιώντας όλα τα απαραίτητα στοιχεία (όνομα χρήστη, κωδικό, όνομα βάσης). Σε περίπτωση αποτυχίας, τύπωσε μήνυμα λάθους
2. Επίλεξε από τον πίνακα jos_vi_rating, που περιέχει τη βαθμολογία κάθε άρθρου από κάθε χρήστη, τη βαθμολογία που έχει δώσει ο ενεργός χρήστης σε κάθε άρθρο.
3. Αποθήκευσε την κάθε βαθμολογία σε ένα πίνακα (array) όπου το κλειδί κάθε θέσης θα είναι το αντικείμενο και η τιμή κάθε θέσης θα είναι η βαθμολογία

Με τον τρόπο αυτό καθίσταται δυνατή η δημιουργία αυτών των πινάκων.

4.3.2.3 Υπολογισμός ομοιότητας μέσω συντελεστή Pearson

Στη συνέχεια, ξεκινά η δημιουργία του αλγορίθμου υπολογισμού της ομοιότητας των χρηστών μέσω του προαναφερθέντος τύπου της παραγράφου 4.3.2.1 που περιλαμβάνει επαναληπτικές διαδικασίες. Στην περίπτωση που η τιμή του συντελεστή συσχέτισης-ομοιότητας ξεπερνά το όριο, τότε ξεκινά η διαδικασία εύρεσης αντικειμένων που θα μπορούν να προταθούν. Στο σημείο αυτό ουσιαστικά ελέγχονται τα μη κοινά αντικείμενα του ενεργού χρήστη με τον όμοιό του (συνεξεταζόμενο χρήστη) δηλαδή αυτά που έχουν ψηφιστεί από τον όμοιο αλλά όχι από τον ενεργό. Τέλος, τα μη κοινά αντικείμενα που έχουν βαθμολογηθεί με τις υψηλότερες βαθμολογίες (δηλαδή 4 και 5) από τον όμοιο, προτείνονται στον ενεργό χρήστη. Η διαδικασία αυτή περιγράφεται στη συνέχεια με τον εξής αλγόριθμο σε φυσική γλώσσα κατά βήματα. Σημειωτέον ότι για να γίνει υπολογισμός της ομοιότητας και να εξαχθούν συστάσεις θα πρέπει ο επισκέπτης της πλατφόρμας να εισέλθει στο σύστημα ως εγγεγραμμένος χρήστης ούτως ώστε να μπορέσει το σύστημα να κρατήσει τις απαιτούμενες πληροφορίες και να ολοκληρώσει τη διαδικασία.

Αλγόριθμος παραγωγής συστάσεων

Ακολουθεί ο αλγόριθμος παραγωγής συστάσεων σε φυσική γλώσσα κατά βήματα

1. Κάνε έλεγχο αν έχει γίνει είσοδος ως χρήστης. Σε περίπτωση μη εισόδου με όνομα χρήστη, εμφάνισε μήνυμα λάθους (η πλατφόρμα συνεχίζει να λειτουργεί κανονικά)

Πανεπιστήμιο Πειραιώς

2. Για τον εισελθόντα (ενεργό χρήστη) κάνε έλεγχο αν έχει ψηφίσει κάποιο άρθρο. Αν δεν έχει ψηφίσει, (τότε δεν μπορεί το σύστημα να ¨κατανοήσει¨ τη συμπεριφορά του και να ελέγξει την ομοιότητά του με άλλους χρήστες) εμφάνισε μήνυμα λάθους και τερμάτισε τη διαδικασία (Αυτό σημαίνει ότι δεν θα εξαχθούν προσωποποιημένες συστάσεις αλλά η πλατφόρμα λειτουργεί κανονικά)
3. Εφόσον έχει ψηφίσει κάποιο άρθρο, τότε βρες όλες τις βαθμολογίες που έχει δώσει ο πρώτος από τους υπόλοιπους εγγεγραμμένους χρήστες του συστήματος και καταχώρισέ τες σε πίνακα (ουσιαστικά αντιστοιχίζεται ένας πίνακας ανά χρήστη).
4. Για τον ενεργό χρήστε έλεγξε αν έχει κοινά ψηφίσει κοινά άρθρα με το χρήστη του οποίου η βαθμολογία αναζητήθηκε στο βήμα 3 και καταχωρίστηκε σε πίνακα (Ουσιαστικά θα γίνει σύγκριση του ενεργού χρήστη και του εκάστοτε συνεξεταζομένου για το αν έχουν ψηφίσει κοινά άρθρα και τι βαθμολογία έχουν δώσει σε αυτά).
5. Εφόσον οι δύο ανωτέρω χρήστες έχουν ψηφίσει κοινά άρθρα, τότε υπολόγισε το συντελεστή Pearson (Αναλυτικά ο υπολογισμός με χρήση εντολών PHP παρατίθεται στο παράρτημα). Αν οι δύο χρήστες δεν έχουν ψηφίσει κοινά άρθρα τότε δεν υπάρχει καμία συσχέτιση μεταξύ τους
6. Αν οι δύο χρήστες είναι όμοιοι, τότε έλεγξε ποια άρθρα του όμοιου χρήστη δεν έχει ψηφίσει ο ενεργός. Καταχώρισέ τα (αφού αυτά είναι τα υποψήφια για πρόταση άρθρα) σε έναν πίνακα της Βάσης Δεδομένων ούτως ώστε να γίνει αργότερα λήψη αυτών με την υψηλότερη βαθμολογία. (Τα στοιχεία καταχωρίζονται στον πίνακα (table) jos_similarity ο οποίος περιέχει τα πεδία: ενεργός χρήστης, όμοιος χρήστης(users), συντελεστής pearson, το άρθρο που προτείνεται (item) καθώς και την προτεινόμενη βαθμολογία (prediction) δηλαδή αυτή που έβαλε ο όμοιος χρήστης.)
7. Επανάλαβε τα βήματα 4, 5, 6 για κάθε έναν από τους υπόλοιπους (υποψήφιους όμοιους) χρήστες του συστήματος.
8. Ταξινόμησε τα υποψήφια για πρόταση άρθρα από τον πίνακα jos_similarity πρώτα κατά φθίνουσα σειρά βάσει συντελεστή Pearson και έπειτα κατά φθίνουσα σειρά βάσει προτεινόμενης βαθμολογίας καταχωρίζοντας τα αποτελέσματα σε έναν πίνακα τύπου array. Σε περίπτωση που ένα άρθρο πρόκειται να καταχωριστεί στον array πάνω από μία φορά λόγω υψηλής βαθμολογίας από πολλούς όμοιους χρήστες, τότε καταχώρισέ το μόνο μία φορά (Με τον τρόπο αυτό αποφεύγεται άσκοπη επανάληψη).
9. Από τα τελικώς καταχωρισθέντα άρθρα, εμφάνισε τα πέντε πρώτα (δηλαδή αυτά που συνδυάζουν υψηλότερη τιμή συντελεστή συσχέτισης και υψηλότερη προτεινόμενη βαθμολογία)

Πανεπιστήμιο Πειραιώς

Σημειωτέον ότι αυτό που εμφανίζεται είναι ο τίτλος του άρθρου που λειτουργεί ταυτόχρονα ως σύνδεσμος για να οδηγεί κατευθείαν το χρήστη στο περιεχόμενο του προτεινόμενου άρθρου. Επίσης, από τα ανωτέρω συνάγεται ότι το σύστημα λειτουργεί σε πραγματικό χρόνο (real time) αφού για κάθε χρήστη και κάθε νέα ψήφο υπολογίζει ξανά την ομοιότητα και παράγει διαφορετικές συστάσεις ανάλογα με τις εκάστοτε προτιμήσεις.

4.3.2.4 Συνοπτική περιγραφή σημαντικών τμημάτων της Βάσης Δεδομένων

Στη συνέχεια ακολουθούν μερικά στιγμιότυπα από σημαντικούς πίνακες της βάσης δεδομένων της πλατφόρμας. Τα στιγμιότυπα έχουν ληφθεί με χρήση του λογισμικού phpmyadmin.

Αρχικά παρατίθεται στιγμιότυπο του πίνακα jos_content ο οποίος περιέχει το αναγνωριστικό (id) και τον τίτλο κάθε άρθρου

```
SELECT *
FROM `jos_content`
LIMIT 60 , 30
```

<< < Εμφάνιση : 30 Εγγραφές αρχίζοντας από την εγγραφή 0

σε οριζόντια μορφή με επανάληψη επικεφαλίδων αν

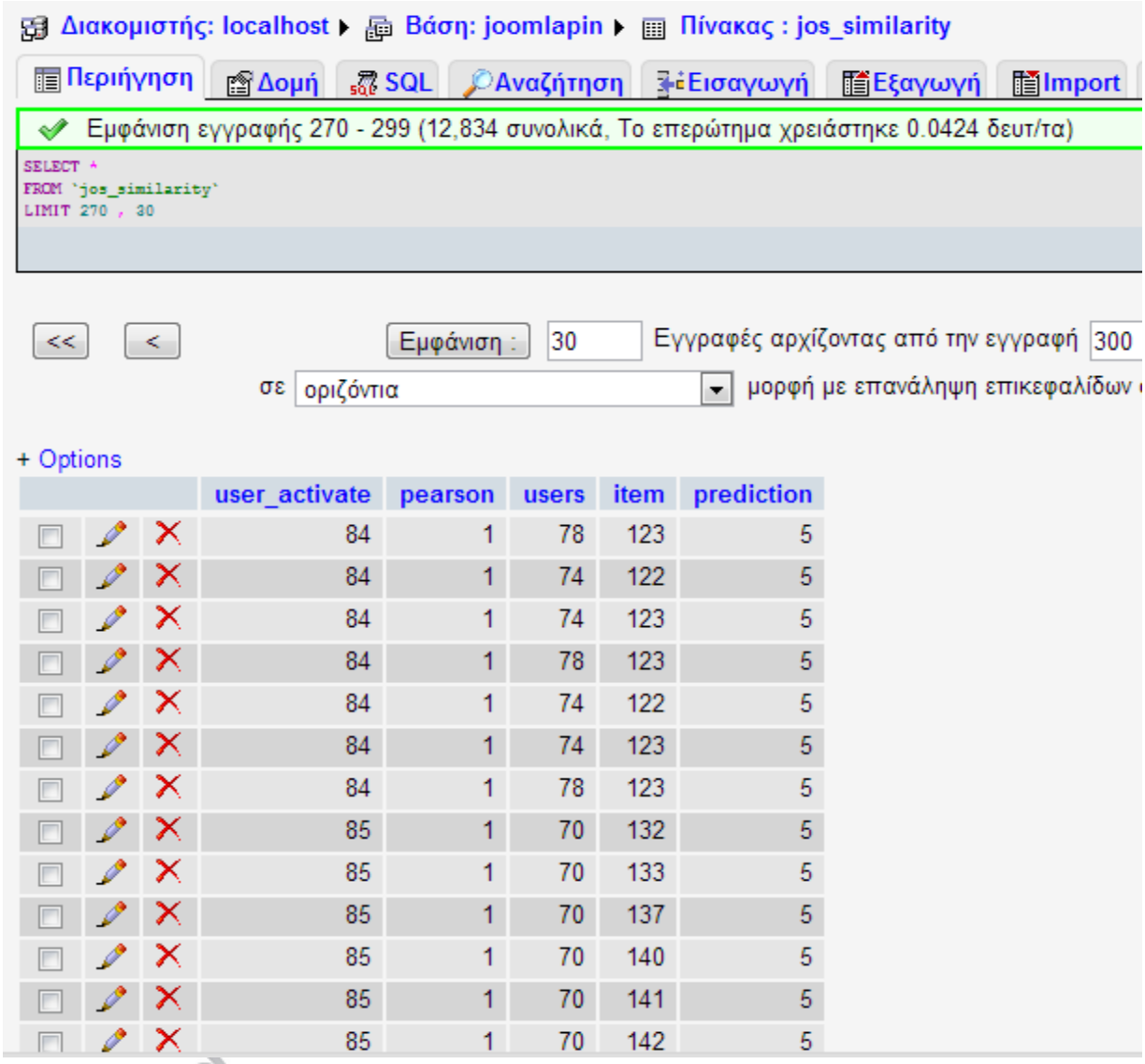
Sort by key: Κανένα

+ Options

			id	title	alias	title_alias	introtext	fulltex
☐	✎	✕	153	Κεφάλαιο 9	-9		<p>{edocs}http://www.pi-schools.gr/download/lesson...	
☐	✎	✕	154	Κεφάλαιο 10	-10		<p>{edocs}http://www.pi-schools.gr/download/lesson...	
☐	✎	✕	155	Κεφάλαιο 11	-11		<p>{edocs}http://www.pi-schools.gr/download/lesson...	
☐	✎	✕	156	Κεφάλαιο 12	-12		<p>{edocs}http://www.pi-schools.gr/download/lesson...	
☐	✎	✕	157	Κεφάλαιο 13	-13		<p>{edocs}http://www.pi-schools.gr/download/lesson	

Εικόνα 18. Πίνακας παράθεσης αναγνωστικού και τίτλου κάθε άρθρου

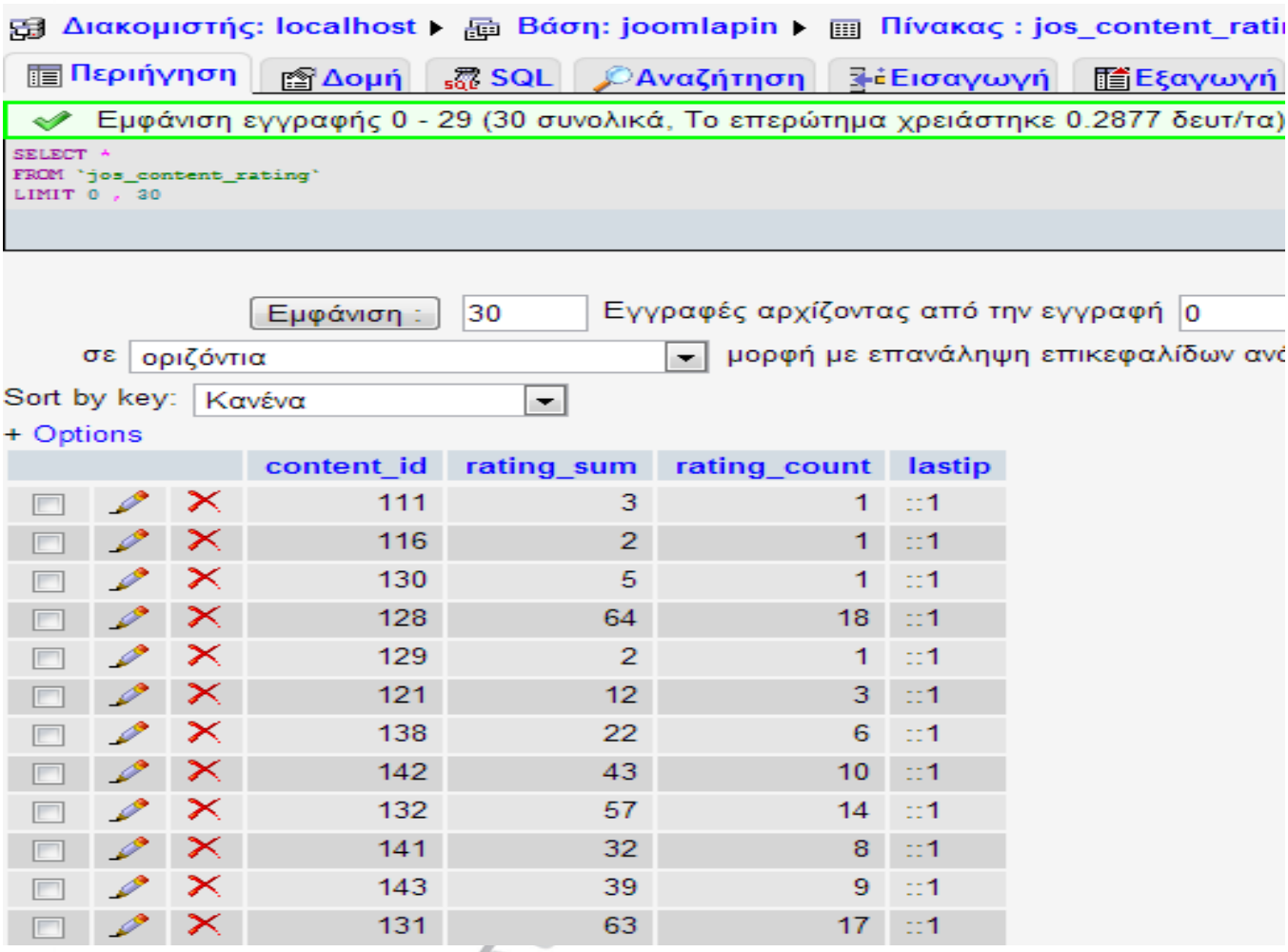
Στη συνέχεια παρατίθεται στιγμιότυπο από τον πίνακα jos_similarity με τη βοήθεια του οποίου παράγονται συστάσεις



user_activate	pearson	users	item	prediction
84	1	78	123	5
84	1	74	122	5
84	1	74	123	5
84	1	78	123	5
84	1	74	122	5
84	1	74	123	5
84	1	78	123	5
85	1	70	132	5
85	1	70	133	5
85	1	70	137	5
85	1	70	140	5
85	1	70	141	5
85	1	70	142	5

Εικόνα 19. Πίνακας υποψηφίων άρθρων, όμοιων χρηστών και πρόβλεψης βαθμολογίας

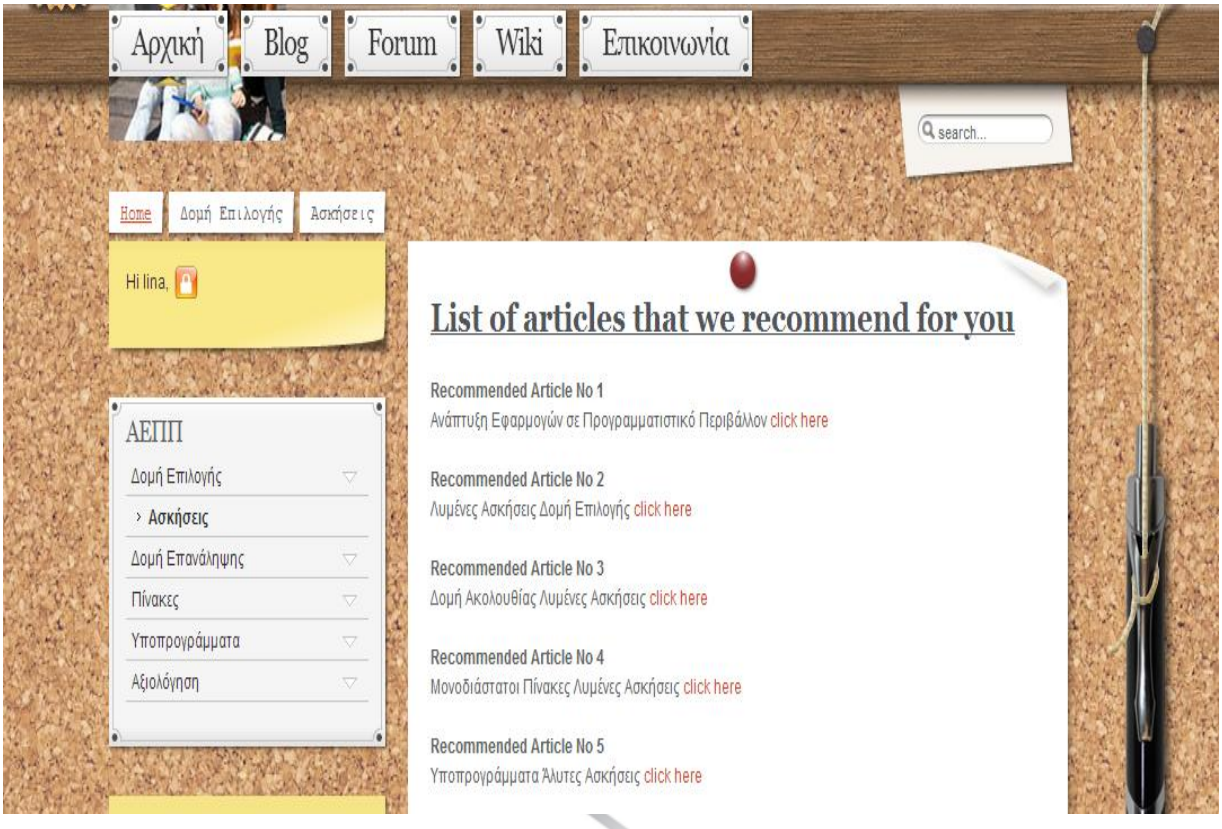
Στη συνέχεια παρατίθεται στιγμιότυπο από τον πίνακα jos_content_rating ο οποίος περιέχει το άθροισμα των ψήφων που έχει λάβει κάθε άρθρο καθώς και πόσοι το ψήφισαν



content_id	rating_sum	rating_count	lastip
111	3	1	::1
116	2	1	::1
130	5	1	::1
128	64	18	::1
129	2	1	::1
121	12	3	::1
138	22	6	::1
142	43	10	::1
132	57	14	::1
141	32	8	::1
143	39	9	::1
131	63	17	::1

Εικόνα 20. Πίνακας που περιέχει το άθροισμα των ψήφων που έχει λάβει κάθε άρθρο.

Όταν ένας χρήστης (π.χ. lina) εισέλθει στο σύστημα, και έχει ψηφίσει κάποιες φορές ούτως ώστε το σύστημα τον «γνωρίσει» και δύναται να ξεκινήσει η διαδικασία εύρεσης όμοιου χρήστη με τον υπολογισμό του συντελεστή Pearson, θα εμφανιστούν οι προτάσεις της πλατφόρμας, όπως φαίνεται καθαρά στο επόμενο στιγμιότυπο.

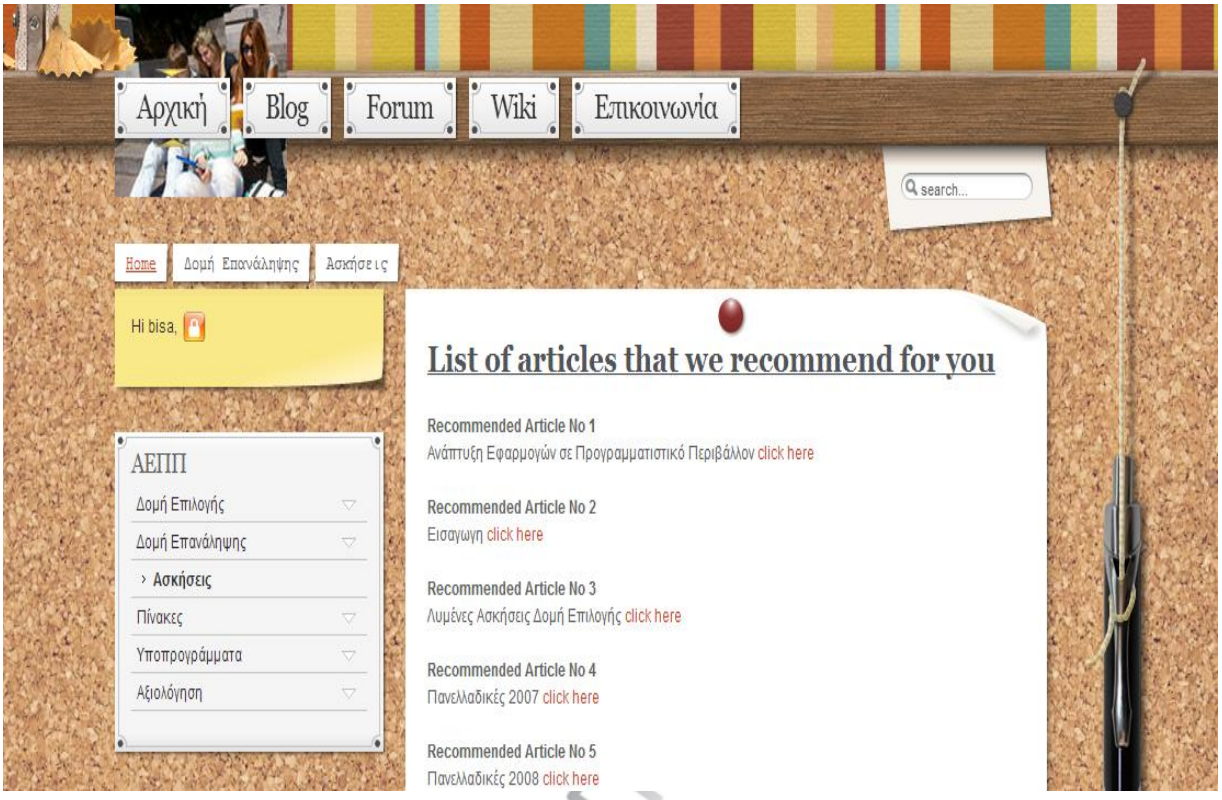


Εικόνα 21. Εξατομικευμένη πρόταση σε συγκεκριμένο χρήστη

Όπως είναι ορατό προτείνονται πέντε άρθρα με τον τίτλο τους και τα οποία μπορεί να τα επιλέξει ο χρήστης πατώντας στην επιλογή «click here». Η σειρά εμφάνισης καθορίζεται, όπως αναφέρεται στην παράγραφό 4.3.2.3, από το συντελεστή Pearson (φθίνουσα σειρά ταξινόμησης) και την προτεινόμενη βαθμολογία (ομοίως φθίνουσα σειρά).

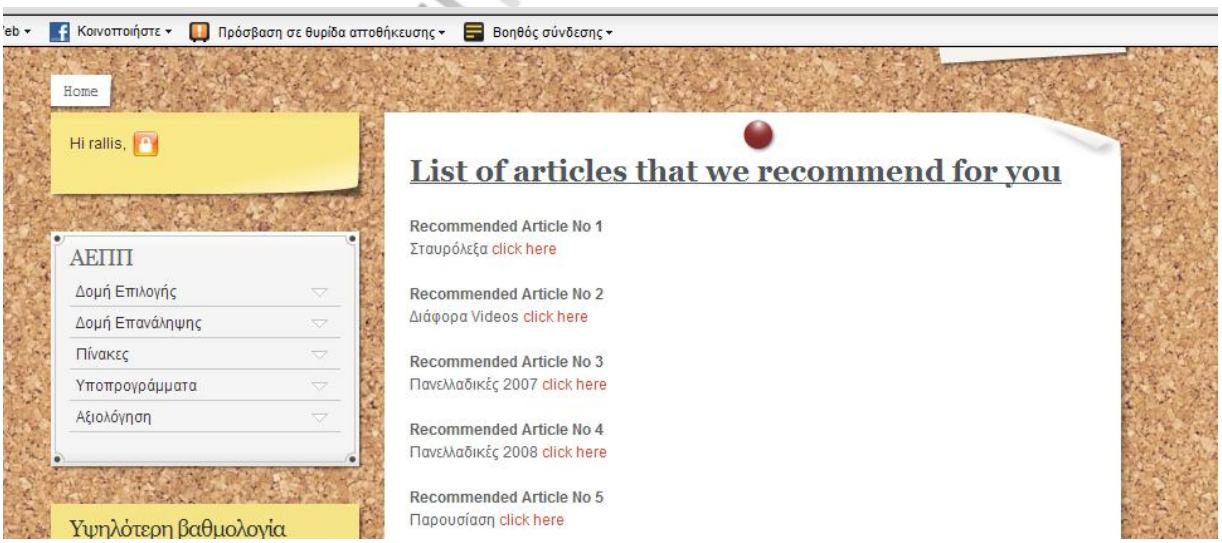
Εννοείται ότι για έναν άλλο χρήστη (π.χ. bisa) θα του προταθούν διαφορετικά άρθρα όπως φαίνεται στο επόμενο στιγμιότυπο

Πανεπιστήμιο



Εικόνα 22. Εξατομικευμένη πρόταση σε συγκεκριμένο χρήστη

Ενώ για έναν ακόμα διαφορετικό χρήστη (με όνομα rallis) θα προκύψουν διαφορετικές συστάσεις όπως φαίνεται στην παρακάτω εικόνα



Εικόνα 23. Εξατομικευμένη πρόταση σε συγκεκριμένο χρήστη

5.Αξιολόγηση του νέου συστήματος

5.1 Περιγραφή δείγματος

Καταρχάς, για την αξιολόγηση επιλέχθηκε ένα δείγμα 20 μαθητών της Β΄ και της Γ' τάξης ΕΠΑΛ (Επαγγελματικού Λυκείου). Οι συγκεκριμένοι μαθητές δεν διδάσκονται το μάθημα Ανάπτυξη Εφαρμογών σε Προγραμματιστικό Περιβάλλον (ΑΕΠΠ), το οποίο διδάσκεται μόνο σε Γενικά Λύκεια αλλά αυτό δεν επηρέασε καθόλου τις δοκιμές διότι έχουν γνώση των αλγοριθμικών δομών αφού φοιτούν στον Τομέα Πληροφορικής και Δικτύων.

Αναλυτικότερα οι μαθητές της Β' τάξης έχουν διδαχθεί αλγοριθμικές δομές με χρήση ψευδογλώσσας και του εγκεκριμένου από το Υπουργείο Παιδείας λογισμικού «Γλωσσομάθεια» καθώς και στοιχεία HTML και JAVA. Από την άλλη πλευρά οι μαθητές της Γ' τάξης έχουν διδαχθεί, εκτός από τα παραπάνω, και Δομημένο Προγραμματισμό σε PASCAL στο οποίο εξετάζονται Πανελλαδικώς.

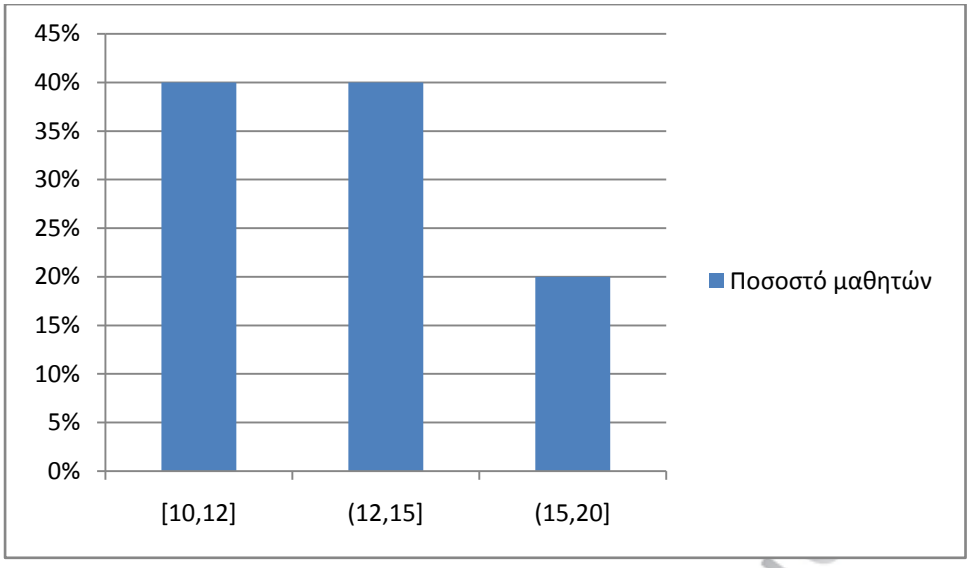
Από αυτά συνάγεται το συμπέρασμα ότι δεν είχαν το παραμικρό πρόβλημα να κατανοήσουν τη λειτουργία του συστήματος. Σημειωτέον ότι οι τελικοί χρήστες του συστήματος είναι μαθητές διαφόρων επιπέδων. Από αυτούς ένα ποσοστό της τάξης του 40% (δηλαδή 8 μαθητές) έχει μέσο όρο βαθμολογίας στα μαθήματα του τομέα Πληροφορικής άνω του 15.

Επίσης, οκτώ μαθητές (40%) έχουν βαθμολογία από 12.1 έως 15 ενώ οι υπόλοιποι τέσσερις (ποσοστό 20%) είχαν βαθμολογία από 10 έως 12. Ακόμα όμως και οι μαθητές με χαμηλότερο βαθμό έχουν αποδεδειγμένα γνώση των αλγοριθμικών δομών που πραγματεύεται η ΑΕΠΠ λόγω της πολύωρης ενασχόλησης με αντικείμενα που άπτονται των γλωσσών προγραμματισμού. Στον επόμενο πίνακα παρατίθενται συγκεντρωτικά οι ανωτέρω πληροφορίες

Αριθμός μαθητών	Ποσοστό	Κλάση βαθμολογίας
8	40%	[10, 12]
8	40%	(12,15]
4	20%	(15, 20]

όπως και στο επόμενο γράφημα

Πανεπιστήμιο Πειραιώς



Εικόνα 24. Διάγραμμα βαθμολογικής κατάταξης του δείγματος

5.2 Σκοπός της αξιολόγησης

Η αξιολόγηση πραγματώνεται σε τρεις φάσεις:

Α. Στην αξιολόγηση του συστήματος με τη χρήση του Μέσου Απόλυτου Σφάλματος

Β. Στην αξιολόγηση μέσω ειδικά διαμορφωμένου ερωτηματολογίου

Γ. Στην εκπλήρωση παιδαγωγικών στόχων

5.2.1 Αξιολόγηση του συστήματος μέσω ΜΑΕ

Στην πρώτη φάση θα χρησιμοποιήσουμε το Μέσο Απόλυτο Σφάλμα (ΜΑΕ) που περιγράφηκε στην παράγραφο 6.2 και δείχνει τη μέση διαφορά της πρόβλεψης του αλγορίθμου από την πραγματική τιμή που έδωσε ο χρήστης. Γενικά το ΜΑΕ είναι ένας πολύ συχνά χρησιμοποιούμενος δείκτης σε τέτοιες περιπτώσεις και θεωρείται αξιόπιστος.

Σκοπός της αξιολόγησης της παρούσας φάσης είναι ο βαθμός ορθότητας της πρόβλεψης. Το σύστημα προτείνει ένα άρθρο θεωρώντας ότι ο χρήστης θα ψηφίσει το εν λόγω άρθρο με την ίδια βαθμολογία. Με την αξιολόγηση αυτή μπορούμε να δούμε κατά πόσο οι προβλέψεις του συστήματος απέχουν ή όχι από την πραγματική βαθμολογία. Εν τέλει, στόχος αυτής της αξιολόγησης είναι η ορθότητα της επιλογής του συγκεκριμένου αλγορίθμου για το συγκεκριμένο πρόβλημα αφού με τη μέτρηση, μέσω του ΜΑΕ, της απόστασης της πρόβλεψης από

Πανεπιστήμιο Πειραι

την πραγματική βαθμολογία ερευνάμε κατά πόσο ο συντελεστής συσχέτισης Pearson λειτουργεί αποτελεσματικά για τη δική μας περίπτωση. Στο σημείο αυτό κρίνεται σκόπιμο να αναφερθεί ότι δεν αμφισβητείται η χρησιμότητα και η αποτελεσματικότητα του δείκτη Pearson συνολικά αλλά το κατά πόσο παράγει επιτυχή αποτελέσματα στο δικό μας σύστημα.

5.2.2 Αξιολόγηση μέσω ειδικά διαμορφωμένου ερωτηματολογίου

Με την επίδοση ειδικά διαμορφωμένου, από τον ερευνητή, ερωτηματολογίου γίνεται η δεύτερη φάση της αξιολόγησης. Σκοπός της αξιολόγησης αυτής είναι η μέτρηση του βαθμού ικανοποίησης του χρήστη από την πλατφόρμα καθώς και η ανατροφοδότηση.

Ειδικότερα μετράται ο βαθμός ικανοποίησης αφού ο χρήστης-μαθητής καλείται να σημειώσει έναν βαθμό στην κλίμακα 1-5 που δηλώνει το επίπεδο αρεσκείας του.

Επιπροσθέτως, ο χρήστης ερωτάται να αποφανθεί ποιος είναι ο καταλληλότερος γι΄ αυτόν από τους χρησιμοποιούμενους στο σύστημα αλγορίθμους παραγωγής συστάσεων. Με την παράμετρο αυτή μελετάμε για μία ακόμα φορά την ορθότητα επιλογής του αλγορίθμου κατηγορίας user-based.

Επιπλέον, ο βαθμός ικανοποίησης μετράται και σε μια τρίτη διάσταση αφού ζητείται η δημιουργία της επιθυμητής λίστας αναδιατάσσοντας την προτεινόμενη. Στο σημείο αυτό, όπως και προηγουμένως, ο σκοπός της αξιολόγησης δεν είναι μόνο η μέτρηση του βαθμού ικανοποίησης αλλά και η διερεύνηση της δυνατότητας εξαγωγής της λίστας προτεινόμενων αντικειμένων με σειρά προτεραιότητας. Τέλος, μέσω αυτών των προτάσεων αναδιάταξης είναι πιθανό να παρουσιαστούν νέες ιδέες οι οποίες μπορούν να δώσουν το έναυσμα για καινούργιες ερευνητικές προσεγγίσεις.

Ως προς το τμήμα της ανατροφοδότησης, αυτή επιτυγχάνεται με τη χρήση του ερωτηματολογίου αφού μέσω αυτού ζητείται η καταγραφή των προτάσεων και των παρατηρήσεων που επιθυμεί να διατυπώσει ο κάθε χρήστης. Ο ρόλος της ανατροφοδότησης είναι καίριος για κάθε επιστημονική έρευνα αφού σκοπός της είναι να καταγραφούν, μελετηθούν και υλοποιηθούν τυχόν διορθωτικές κινήσεις.

Πανεπιστήμιο Πειραιώς

5.2.3 Εκπλήρωση παιδαγωγικών στόχων

Βασικός σκοπός είναι να απαντηθεί το υψίστης σημασίας ερώτημα αν η κατασκευασθείσα πλατφόρμα λειτουργεί ως ένας ακόμα ιστότοπος που παρέχει εκπαιδευτικό υλικό και, συνεπώς, δεν διαχωρίζεται από δεκάδες παρόμοιους ιστοτόπους οι οποίοι απλώς περιέχουν πλούσιο εκπαιδευτικό υλικό ή, και αυτό είναι το μέγα ζητούμενο, λειτουργεί σαν κάτι εντελώς διαφορετικό παρέχοντας εξατομικευμένες συστάσεις στους εγγεγραμμένους χρήστες προσφέροντας, κατ' αυτόν τον τρόπο, καθοδήγηση στους μαθητές. Αυτό απαντάται χρησιμοποιώντας τα δεδομένα του ερωτηματολογίου τα οποία παρατίθενται σε επόμενη παράγραφο.

5.3 Παρουσίαση του νέου συστήματος στους χρήστες

Πριν τη λειτουργία του συστήματος έγινε επίδειξη της χρήσης του από το διδάσκοντα εντός δύο εκπαιδευτικών ωρών. Κατά τη διάρκεια αυτής της επίδειξης εξηγήθηκε στους μαθητές τι είναι ένα σύστημα παραγωγής συστάσεων και παρουσιάστηκαν κάποια παραδείγματα όπως το www.youtube.com και το www.booking.com. Εν συνεχεία οι μαθητές χρησιμοποίησαν την πλατφόρμα χωρίς να κάνουν είσοδο ως χρήστες για να εξοικειωθούν με το περιβάλλον. Κατά το διάστημα αυτό περιηγήθηκαν στις περισσότερες επιλογές της πλατφόρμας ούτως ώστε να έχουν μία καλύτερη συνοπτική εικόνα του συστήματος. Η δοκιμή αυτή διήρκησε περίπου μία εκπαιδευτική ώρα.

Σημειωτέον ότι στο διάστημα αυτό είχαν την ευκαιρία να δουν τρόπους παραγωγής συστάσεων μέσω των πιο σχετικών, πιο πολυδιαβασμένων και υψηλότερα βαθμολογημένων άρθρων. Στο σημείο αυτό ο διδάσκων επέστησε την προσοχή των μαθητών στους ανωτέρω απλοϊκούς τρόπους παραγωγής συστάσεων οι οποίοι λειτουργούν σε κάθε περίπτωση δηλαδή ακόμα κι όταν ο επισκέπτης της πλατφόρμας εισέλθει ως εγγεγραμμένος ή ως μη εγγεγραμμένος χρήστης. Έπειτα, για μία ακόμα εκπαιδευτική ώρα χρησιμοποίησαν το σύστημα με ξεχωριστό όνομα χρήστη ο κάθε μαθητής ούτως ώστε να αποκτήσουν μία καλύτερη εικόνα ενός συστήματος παραγωγής συστάσεων παρατηρώντας πλέον τις συστάσεις από τον υλοποιηθέντα αλγόριθμο της υποκατηγορίας user-based της κατηγορίας collaborative filtering.

Πρέπει να τονιστεί ότι η έκδοση του συστήματος που δοκίμασαν οι μαθητές δεν ήταν η τελική. Για τη δοκιμαστική λειτουργία προστέθηκαν μερικά ακόμα άρθρα σε κάθε κατηγορία των οποίων η παρουσία διαφοροποιεί σε σημαντικό βαθμό τα

Πανεπιστήμιο Πειραιώς

αποτελέσματα. Σημειωτέον ότι το σύστημα, με τον υλοποιηθέντα αλγόριθμο της κατηγορίας collaborative filtering, παράγει συστάσεις μόνο όταν ένας χρήστης εισέλθει με τα στοιχεία του ως εγγεγραμμένος (όνομα χρήστη και κωδικό).

Σε περίπτωση που ένας επισκέπτης του ιστοτόπου δεν είναι εγγεγραμμένος, τότε δεν προτείνεται τίποτα σε αυτόν από τον ανωτέρω αλγόριθμο. Σε αυτή την περίπτωση, όπως αναφέρθηκε, ο επισκέπτης μπορεί να λάβει συστάσεις με τους άλλους τρόπους (υψηλότερης βαθμολογίας, σχετικότερα και πιο πολυδιαβασμένα άρθρα).

5.4 Πιλοτική λειτουργία του νέου συστήματος

Ακολούθως έγινε αρχικοποίηση της βάσης με διαγραφή όλων των ψηφοφοριών ούτως ώστε η πιλοτική λειτουργία του συστήματος να γίνει εκ του μηδενός. Πλέον το σύστημα ήταν έτοιμο προς λειτουργία. Η πιλοτική λειτουργία έλαβε χώρα κατά το διάστημα 20-22 Μαρτίου 2013 σε σχολικό εργαστηριακό χώρο και συγκεκριμένα σε αίθουσα με ηλεκτρονικούς υπολογιστές. Αρχικά οι μαθητές εκλήθησαν να ψηφίσουν πέντε φορές έκαστος ανάλογα με τις προτιμήσεις τους ούτως ώστε η Βάση Δεδομένων να μην είναι κενή. Άλλωστε αυτό είναι ένα από τα εγγενή προβλήματα των συστημάτων παραγωγής συστάσεων δηλαδή η αναγκαιότητα ύπαρξης κάποιων δεδομένων για να ξεκινήσει να λειτουργεί πιο αποτελεσματικά ένας αλγόριθμος.

Εννοείται ότι κατά τη φάση αυτή οι μαθητές θα μπορούσαν να ψηφίσουν περισσότερα εκπαιδευτικά αντικείμενα-άρθρα αλλά προτιμήθηκε τελικά να περιοριστούν στον αριθμό πέντε ούτως ώστε να παραχθούν συστάσεις χωρίς πάρα πολλές ψήφους δηλαδή καταβλήθηκε προσπάθεια να παραχθούν συστάσεις και να είναι αποδοτικό το σύστημα χωρίς πολλά δεδομένα εισαγωγής.

Επίσης, σε κάθε μαθητή προτείνονταν πέντε άρθρα από τα οποία επέλεγε τρία(3) και τα βαθμολογούσε στην κλίματα 1-5 όπου 1 ήταν η χειρότερη βαθμολογία και 5 η καλύτερη.

Στο σημείο αυτό πρέπει να υπενθυμιστεί ότι όταν ένας χρήστης ψηφίζει, το σύστημα προσπαθεί να βρει τους όμοιούς του χρήστες δηλαδή αυτούς που έχουν δώσει παρόμοια βαθμολογία στα ίδια άρθρα. Η ομοιότητα, όπως αναφέρθηκε, υπολογίζεται με το συντελεστή *Pearson*. Η τιμή του Pearson κυμαίνεται από μείον ένα(-1) που σημαίνει καμία ομοιότητα έως ένα (1) που σημαίνει πλήρη ομοιότητα.

Πανεπιστήμιο Πειραιώς

5.5 Εξαγωγή συστάσεων από το νέο σύστημα

Μετά την ολοκλήρωση της ανωτέρω διαδικασίας το σύστημα μπορεί να προτείνει μια σειρά άρθρων που πληρούν τα παραπάνω κριτήρια. Το πλήθος των άρθρων αυτών μπορεί να είναι μεγάλο, π.χ 20-30 ή και παραπάνω. Για να υπάρχει πράγματι βοήθεια από την πλατφόρμα προς το χρήστη προτείνονται τα πέντε με την καλύτερη-πιο σχετική- βαθμολογία ούτως ώστε να αποφευχθεί η παράθεση πληθώρας άρθρων, κάτι που πιθανότατα θα προκαλέσει σύγχυση στο χρήστη. Έτσι γίνεται αρχικά μια ταξινόμηση κατά φθίνουσα σειρά των άρθρων βάσει του συντελεστή Pearson και έπειτα κατά φθίνουσα σειρά βάσει της βαθμολογίας που έδωσε ο όμοιος χρήστης στα συγκεκριμένα άρθρα, όπως περιγράφεται στην παράγραφο 4.3.2.3.

Πχ. Αν ο ενεργός χρήστης είναι ο 70 και βρεθούν όμοιοί του οι 80,81,82, 83,84,85,86 με συντελεστές Pearson 0.91, 0.91, 0.91, 0.89, 0.88, 0.84 και 0.71 αντίστοιχα τότε θα του προταθούν άρθρα των πέντε πρώτων όμοιων χρηστών, που έχουν υψηλότερο συντελεστή συσχέτισης. Αν στη συνέχεια υπάρχουν πολλά άρθρα με υψηλή βαθμολογία (άνω του τρία) από αυτούς τους χρήστες, κάτι που είναι απόλυτα φυσιολογικό, τότε στην ταξινόμηση θα προηγηθούν αυτά που έχουν λάβει υψηλότερη βαθμολογία από τους χρήστες 80,81,82,83,84.

Σημειωτέον ότι η βαθμολογία των όμοιων χρηστών είναι ουσιαστικά η βαθμολογική πρόβλεψη για το προτεινόμενο αντικείμενο. Επίσης, αν ένα άρθρο προταθεί από πολλούς χρήστες τότε θα εμφανιστεί μόνο μία φορά για αποφυγή άσκοπης επανάληψης. Η εμφάνιση αυτή γίνεται βάσει της ανωτέρω ταξινόμησης δηλαδή κατά φθίνουσα σειρά πρώτα βάσει Pearson και ύστερα βάσει βαθμολογίας που έλαβε το άρθρο αυτό.

Δεν θα πρέπει να παραληφθεί να αναφερθεί για μία ακόμα φορά, προς αποφυγή οποιασδήποτε παρανόησης, ότι κατά την πιλοτική λειτουργία το σύστημα συνέχιζε να παράγει συστάσεις μέσω των πιο πολυδιαβασμένων (δημοφιλών), πιο σχετικών και υψηλότερα βαθμολογημένων άρθρων.

Πανεπιστήμιο Πειραιώς

5.6 Αξιολόγηση με τη χρήση του Μέσου Απόλυτου Σφάλματος

Για την αξιολόγηση του συστήματος χρησιμοποιήθηκε το Μέσο Απόλυτο Σφάλμα (MAE) που δίνει την απόλυτη τιμή της διαφοράς μεταξύ της πρόβλεψης του αλγορίθμου δηλαδή αυτής που προτάθηκε στον τρέχοντα-ενεργό χρήστη και της πραγματικής βαθμολογίας του χρήστη αυτού δηλαδή αυτής που έδωσε τελικά στο προτεινόμενο εκπαιδευτικό αντικείμενο.

Η επιθυμητή τιμή του MAE είναι το μηδέν που δείχνει ότι ο ενεργός χρήστης ψήφισε όπως ακριβώς του προτάθηκαν τα άρθρα ενώ η χειρότερη τιμή είναι το τέσσερα λόγω του ότι η κλίμακα κυμαίνεται από ένα έως πέντε (1-5). Στη συνέχεια γίνεται υπολογισμός του μέσου όρου του MAE κάθε χρήστη. Η συνολική μέση τιμή είναι 1.17 που θεωρείται καλή.

Γενικά όσο πιο σύνθετο είναι ένα σύστημα με περισσότερα άρθρα τόσο δυσκολότερο είναι να επιτευχθεί υψηλή μέση τιμή του MAE αφού λόγω της πληθώρας χρηστών και υλικού μπορούν να βρεθούν καταλληλότερες συστάσεις. Σε ένα μικρότερο σύστημα με μερικές δεκάδες άρθρα η ανωτέρω τιμή σίγουρα είναι πολύ καλή αλλά όχι και εντυπωσιακή. Αναλυτικά τα αποτελέσματα παρατίθενται στον παρακάτω πίνακα

Τιμή MAE	Συχνότητα εμφάνισης	ποσοστό εμφάνισης της τιμής
0.00	1	5.00%
0.33	4	20.00%
0.67	1	5.00%
1.00	4	20.00%
1.33	2	10.00%
1.67	1	5.00%
2.00	2	10.00%
3.00	1	5.00%
4.00	1	5.00%

όπου φαίνεται ότι η τιμή 0.33 παρουσιάστηκε 4 φορές (δηλαδή σε 4 μαθητές), ποσοστό εμφάνισης 20%, κ.ο.κ.

Πανεπιστήμιο Πειραιώς

5.7 Ερωτηματολόγιο

Επίσης, στους χρήστες-μαθητές δόθηκε ένα φύλλο αξιολόγησης με τα θέματα που συνοπτικά παρατίθενται παρακάτω. Ο αναγνώστης μπορεί να δει το ερωτηματολόγιο στην πλήρη έκτασή του στο παράρτημα Β της παρούσας εργασίας

1. Ψηφίστε στην κλίμακα 1-5 (1: καθόλου χρήσιμος, 5:πολύ χρήσιμος) το πώς σας φάνηκε ο αλγόριθμος παραγωγής συστάσεων ανάλογα με το αν σας πρότεινε όντως ενδιαφέροντα άρθρα.
2. Επιλέξτε τον αποδοτικότερο αλγόριθμο παραγωγής συστάσεων από αυτούς που χρησιμοποιεί η πλατφόρμα
   Α. Top rated (βάσει αθροίσματος συνολικής βαθμολογίας)
   Β. Most read (Πιο πολυδιαβασμένο άρθρο βάσει των εμφανίσεων (κλικ) κάθε άρθρου)
   Γ. Most Related (πιο σχετικά άρθρα βάσει ετικετών)
   Δ. Αλγόριθμος κατηγορίας User based (αλγόριθμος με βάση την ομοιότητα των χρηστών).
3. Καταγράψτε τη λίστα προτάσεων του αλγορίθμου κατηγορίας user based. Συμφωνείτε με αυτήν; Αν όχι, προτείνετε τη δική σας δηλαδή να ταξινομήσετε την παραχθείσα, από τον αλγόριθμο, λίστα με βάσει τα δικά σας κριτήρια.
4. Θεωρείτε ότι οι προτάσεις που εξήγαγε ο αλγόριθμος κατηγορίας user-based βοήθησαν κατά τη μελέτη σας;
5. Παραθέστε τις δικές σας προτάσεις για την παραγωγή συστάσεων. Έχετε κάποια ιδέα για το πώς θα μπορούσε να είναι αποδοτικότερη και χρησιμότερη η παραγωγή συστάσεων;

Σημειωτέον ότι όλα φύλλα αξιολόγησης και οι προτάσεις των μαθητών μελετήθηκαν επισταμένως και προέκυψαν από αυτά χρήσιμα συμπεράσματα τα οποία θα παρατεθούν στη συνέχεια.

5.7.1 Βαθμός ικανοποίησης

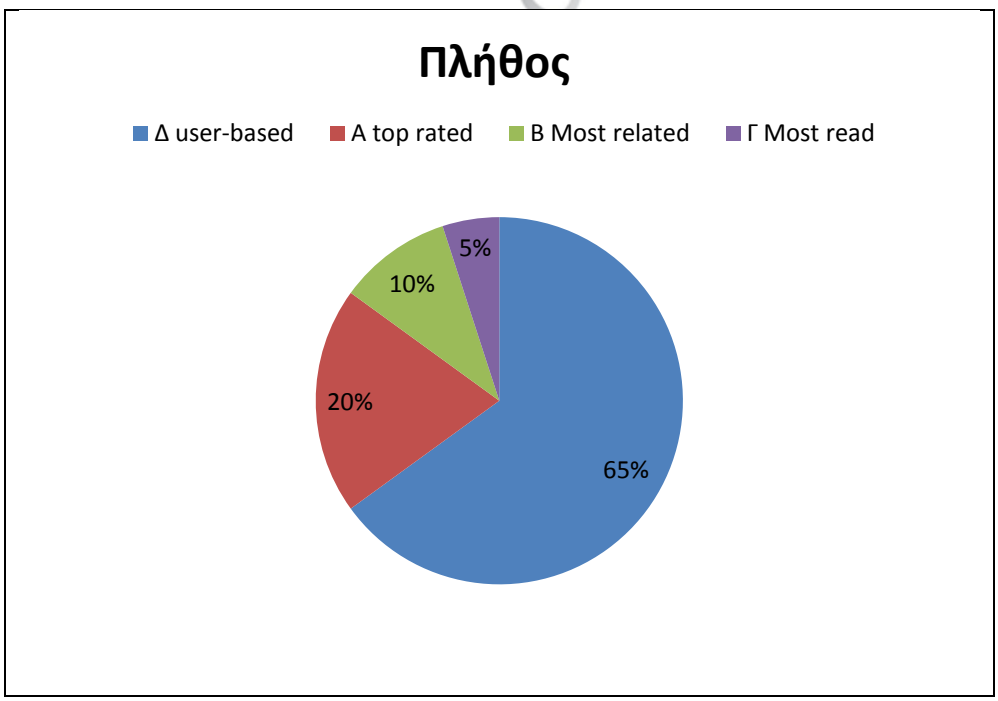
Για το πρώτο ερώτημα η μέση βαθμολογία είναι 3.1 δηλαδή περίπου στο μέσο της κλίμακας 1-5. Συνεπώς ο βαθμός χρησιμότητας δεν ήταν ιδιαίτερα υψηλός αλλά ούτε και χαμηλός. Σε αυτό έπαιξε ρόλο ο σχετικά μικρός αριθμός των άρθρων που ψήφισε κάθε χρήστης. Σε μεγαλύτερο αριθμό ψήφων σίγουρα θα υπήρχε

Πανεπιστήμιο Πειραιώς

μεγαλύτερος δείκτης χρησιμότητας αλλά αυτό είναι πολύ σύνηθες στα συστήματα παραγωγής συστάσεων. Όσο μεγαλώνει το ποσό των διαθέσιμων δεδομένων, τόσο πιο ασφαλείς συστάσεις παράγονται αφού το σύστημα μπορεί να «κατανοήσει» καλύτερα κάθε χρήστη και να προβλέψει καλύτερα τη συμπεριφορά του.

5.7.2 Σύγκριση αλγορίθμων παραγωγής συστάσεων

Όσον αφορά το δεύτερο ερώτημα για τον αποδοτικότερο αλγόριθμο παραγωγής συστάσεων από αυτούς που χρησιμοποιεί η πλατφόρμα, οι μαθητές σε μεγάλο ποσοστό επέλεξαν τον αλγόριθμο κατηγορίας user-based που υλοποιήθηκε κατά τη διάρκεια της παρούσας εργασίας. Συγκεκριμένα, 13 από τους 20, ποσοστό 65%, προτίμησαν τον αλγόριθμο κατηγορίας user-based ενώ δεύτερος σε προτιμήσεις ήταν ο top rated με 4/20 δηλαδή 20%. Ακολούθησαν ο most related με 2/20, ποσοστό 10% και, τέλος, ο most read με μόλις μία ψήφο στις 20, ποσοστό 5%. Τα αποτελέσματα παρατίθενται αναλυτικά στο παρακάτω γράφημα.



Εικόνα 25. Σύγκριση αλγορίθμων παραγωγής συστάσεων

5.7.3 Σύγκριση λιστών

Για το τρίτο ερώτημα που αφορά την καταγραφή της επιθυμητής σειράς της παραχθείσας λίστας, ήταν αναμενόμενο να υπάρξουν πολλές διαφοροποιήσεις αφού υπεισέρχεται το υποκειμενικό στοιχείο το οποίο είναι ιδιαιτέρως έντονο.

Η μέτρηση της απόσταση της επιθυμητής λίστας από την παραγόμενη έγινε με τον Kendall tau (http://en.wikipedia.org/wiki/Kendall_tau_distance) που είναι ένας μετρητής της διαφοράς δύο λιστών. Ο μετρητής κατασκευάστηκε από τον Βρετανό στατιστικολόγο Maurice Kendalll (1907-1983). Σύμφωνα με το μετρητή, όσο ποιο ανόμοιες είναι δύο λίστες τόσο μεγαλύτερη είναι και η απόστασή τους δηλαδή η τιμή του μετρητή Kendall ο οποίος δίνεται από τον τύπο K(T1,T2) = |{(i,j): i<j, (T1(i)<T1(j) ^T2(i) >T2(j)) ⱽ(T1(i) > T1(j)^T2(i) < T2(j))}|

όπου L1 και L2 είναι οι συγκρινόμενες λίστες ενώ T1 και T2 είναι η βαθμολογία στις λίστες των άρθρων αυτών των λιστών. Η τιμή K(T1, T2) θα είναι ίση με μηδέν αν οι δύο λίστες ταυτίζονται ενώ θα είναι ίση με n(n-1)/2, όπου n είναι το μέγεθος της λίστας αν οι δύο λίστες έχουν ακριβώς αντίστροφη σειρά. Πολύ συχνά γίνεται κανονικοποίηση του μετρητή διαιρώντας τον με n(n-1)/2 έτσι ώστε η μέγιστη τιμή που μπορεί να προκύψει να είναι το ένα (1). Στην παρούσα εργασία χρησιμοποιείται ακριβώς αυτός ο κανονικοποιημένος τύπος. Οι μαθητές ψήφισαν ως εξής:

τιμή Kendall tau	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70
Πλήθος	6	1	1	0	5	2	4	1

Όπως ευκόλως παρατηρείται, μερικοί μαθητές και συγκεκριμένα έξι (6/20), ποσοστό 30%, έδωσαν ως επιθυμητή λίστα ακριβώς την παραγόμενη από τον αλγόριθμο. Επιπλέον ένας μαθητής είχε δείκτη 0.10 και ένας άλλος 0.20 δηλαδή σχεδόν ανύπαρκτη ή ελάχιστη διαφοροποίηση ενώ ήταν αρκετοί, συγκεκριμένα 4/20, ποσοστό 20%, είχαν τιμή 0.60 δηλαδή αρκετά μεγάλη διαφοροποίηση από την αρχική λίστα. Επίσης, άλλοι δύο είχαν τιμή Kendall 0.50 που δείχνει μέτρια διαφοροποίηση(ίση απόσταση από την αρχική και την αντίστροφή λίστα) και μόλις ένας είχε δείκτη 0.70 δηλαδή πολύ μεγάλη διαφοροποίηση σε σχέση με την προτεινόμενη λίστα. Η δεύτερη μεγαλύτερη συχνότητα ήταν 0.40, με πέντε μαθητές να έχουν αυτήν τιμή, που δείχνει ότι ήταν περισσότερο κοντά στην προτεινόμενη λίστα παρά στην αντίστροφη. Τέλος, η μέση τιμή του δείκτη Kendall ήταν 0.32 που

Πανεπιστήμιο Πειραιώς

καταδεικνύει ότι οι προτιμήσεις ήταν πιο κοντά στις προτεινόμενες λίστες παρά στις αντίστροφες.

Στο σημείο αυτό πρέπει να τονιστεί για μία ακόμα φορά ότι μετά την αρχική καταγραφή της επιθυμητής λίστας, αρκετοί μαθητές την τροποποίησαν τουλάχιστον δύο φορές. Αξιοσημείωτο ήταν ότι οι μαθητές κατά την κατάρτιση αυτής της λίστας ήταν αρκετά προβληματισμένοι αφού προσπαθούσαν να σκεφτούν τι πραγματικά ήθελαν περισσότερο.

5.7.4 User-based και μελέτη

Στο ερώτημα αυτό οι μαθητές σε υψηλότατο ποσοστό 85% (ήτοι 17 στους 20) απάντησαν καταφατικά στην ερώτηση αν ο αλγόριθμος κατηγορίας user-based βοήθησε κατά τη μελέτη τους στα μαθήματα που περιλαμβάνοντα στο Πρόγραμμα Σπουδών αφού ένιωθαν ότι είχαν κάποιο βοηθό και δεν αισθάνονταν αβοήθητοι κατά την περιήγησή τους στην πλατφόρμα.

5.7.5 Προτάσεις χρηστών

Όσον αφορά το 5^{ο} ερώτημα του φύλλου αξιολόγησης, οι μαθητές παρέθεσαν κάποιες προτάσεις

Α) Αρχικά, παρατίθεται η σημαντικότερη που δεν είναι άλλη από το κλασικό ερώτημα των συστημάτων παραγωγής συστάσεων: να χρησιμοποιηθεί μια μέθοδος που θα παράγει ασυζητητί τις καλύτερες συστάσεις

Β) Αφού κατανόησαν πώς λειτουργεί το σύστημα, θέλησαν να δουν όχι μόνο τα προτεινόμενα αντικείμενα αλλά και τον όμοιο χρήστη ούτως ώστε να μπορούν να ανταλλάξουν απόψεις δια ζώσης ή ηλεκτρονικώς. Σημειωτέον ότι κάτι τέτοιο ήδη γίνεται από το σύστημα. Συγκεκριμένα, κατά την παράθεση των προτεινόμενων αντικειμένων, μπορεί κάλλιστα να εμφανίζεται ο όμοιος χρήστης καθώς και η βαθμολογία-πρόβλεψη. Απλώς κατά την πιλοτική λειτουργία της πλατφόρμας, είχαν αφαιρεθεί αυτά τα μηνύματα για να είναι όσο πιο αμερόληπτοι γίνεται οι χρήστες και όχι να επηρεαστούν από προσωπικές συμπάθειες και φιλίες. Εξάλλου, η εμφάνιση της πρόβλεψης θα οδηγούσε πιθανότατα το χρήστη να δώσει τον ίδιο βαθμό. Αξιοσημείωτο είναι ότι όταν παρουσιάζεται κάτι καινούργιο σε μαθητές,

Πανεπιστήμιο Πειραιώς

συνήθως αρχίζει βομβαρδισμός ερωτήσεων από τον ένα προς τον άλλο του τύπου «τι κάνουμε εδώ;», «τι βαθμό έβαλες εσύ;», κλπ.

Γ) Δυνατότητα δημιουργίας της επιθυμητής λίστας μέσω της πλατφόρμας και όχι μέσω φύλλων αξιολόγησης.

Δ) Επίσης, θα ήθελαν η επιθυμητή λίστα να προτείνεται και στους όμοιους χρήστες. Το θέματα Γ και Δ εμπεριέχουν μεγάλο βαθμό δυσκολίας κατά την υλοποίησή τους και σίγουρα ξεπερνούν αρκετά τα όρια της παρούσας εργασίας.

Ε) Δυνατότητα παράθεσης των πέντε (γενικότερα ν) πιο ανόμοιων αντικειμένων. Γενικώς αυτό μπορεί να γίνει πολύ εύκολα από το σύστημα με την προσθήκη ελάχιστων γραμμών κώδικα PHP και MySQL απλώς είναι πιθανότερο να λειτουργήσει αποπροσανατολιστικά παρά δημιουργικά και γόνιμα.

5.8 Ανάλυση αποτελεσμάτων αξιολόγησης-συμπεράσματα

Στις ακόλουθες παραγράφους αναλύεται το κατά πόσον επιτεύχθηκαν οι τεθέντες στόχοι της παραγράφου 7.1

5.8.1 Στόχοι του Μέσου Απόλυτου Σφάλματος

Όπως προαναφέρθηκε ο μέσος όρος του MAE ήταν 1.17. Αυτό σημαίνει ότι η πλατφόρμα παρήγαγε συστάσεις με πρόβλεψη πού ήταν πολύ κοντά στην πραγματική βαθμολογία που έδωσαν οι χρήστες αφού απείχε περίπου μία μονάδα. Μία ανάλογη τιμή του MAE σε ένα σύστημα με πολλούς χρήστες και πληθώρα αντικειμένων και δεδομένων, ενδεχομένως να μην ήταν αρκετά ικανοποιητική. Εν κατακλείδι φαίνεται ότι ο βαθμός ορθότητας της πρόβλεψης ήταν λίαν ικανοποιητικός και σωστά επελέγη ο εν λόγω αλγόριθμος για το πρόβλημά μας και, συνεπώς, πραγματώνονται οι στόχοι της συγκεκριμένης αξιολόγησης.

Πανεπιστήμιο Πειραιώς

5.8.2 Στόχοι του ερωτηματολογίου

Στο πρώτο ερώτημα που αφορά τη βαθμολόγηση της ικανοποίησης του χρήστη από τον αλγόριθμο, η μέση τιμή ήταν 3.1 με άριστα το πέντε. Εδώ παρατηρείται ότι ο βαθμός ικανοποίησης του χρήστη δεν ήταν ιδιαιτέρως υψηλός και φαινομενικά οι χρήστες δεν έμειναν πολύ ευχαριστημένοι. Αυτό, όμως, όπως αναφέρθηκε, οφείλεται στο σχετικά περιορισμένο αριθμό των άρθρων.

Στο δεύτερο ερώτημα, αυτό της σύγκρισης των αλγορίθμων, το ποσοστό ανήλθε στο 65% για τον user-based και οι υπόλοιποι ακολούθησαν με πολύ σημαντική διαφορά. Άρα ήταν αρκετά ικανοποιητικός ο βαθμός καταλληλότητας του αλγορίθμου και, κατά συνέπεια, της ορθότητας της επιλογής του.

Αρχικά φαίνονται κάπως ασύμβατα και αταίριαστα τα αποτελέσματα του πρώτου και του δεύτερου ερωτήματος αλλά δικαιολογείται πλήρως από το γεγονός ότι όλοι οι μαθητές πρώτοι φορά χρησιμοποίησαν ένα εξατομικευμένο σύστημα παραγωγής συστάσεων, κάτι που είναι λίαν εντυπωσιακό γι' αυτούς. Τους υπόλοιπους τρεις αλγορίθμους, λίγο ή πολύ, τους είχαν συναντήσει σε πολλούς ιστοτόπους αφού είχαν δει αρκετές φορές λίστες με τα πιο δημοφιλή (αυτά με την υψηλότερη βαθμολογία), τα πιο πολυδιαβασμένα και τα πιο σχετικά, βάσει ετικετών, άρθρα. Ήταν όμως η πρώτη φορά που χρησιμοποιούσαν σύστημα με προσωποποιημένες συστάσεις και αυτό προκάλεσε το θαυμασμό πολλών από αυτούς, ειδικότερα όταν έβλεπαν να προτείνονται άρθρα πολύ σχετικά με τις δικές τους προτιμήσεις.

Στο τρίτο ερώτημα, όπως προαναφέρθηκε, ζητήθηκε η αναδιάταξη της προτεινόμενης λίστας. Η μέτρηση της απόστασης-διαφοράς των λιστών μέσω της μετρικής Kendall tau ήταν 0.32 που σημαίνει αρκετά κοντά ως προς την επιθυμητή σειρά προτεραιότητας και, επομένως, υπήρξε σημαντικός βαθμός ικανοποίησης.

Στα ερωτήματα των παρατηρήσεων-προτάσεων υπήρξε τροφή για διορθωτικές παρεμβάσεις στην πλατφόρμα και μελλοντική έρευνα.

5.8.3 Παιδαγωγικοί στόχοι

Στο τέταρτο ερώτημα του ερωτηματολογίου, που περιλάμβανε την ερώτηση κατά πόσο ο ιστότοπος κατάφερε να λειτουργήσει καθοδηγητικά και υποστηρικτικά ως προς τη μελέτη των μαθητών, ένα συντριπτικό ποσοστό της τάξης του 85% (ήτοι 17 στους 20) απάντησε καταφατικά. Αυτό το συμπέρασμα μάς επιτρέπει να δηλώσουμε με βεβαιότητα ότι ο αλγόριθμος επιτέλεσε τους παιδαγωγικούς στόχους.

Πανεπιστήμιο Πειραιώς

6.Σύνοψη και μελλοντική έρευνα

6.1 Σύνοψη της εργασίας

Κατά την παρούσα εργασία έγιναν συνοπτικά τα εξής:

- Παρουσίαση του προβλήματος της αναζήτησης πιο εξειδικευμένων πληροφοριών σε Διαδικτυακούς τόπους.
- Δικαιολόγηση της αναγκαιότητας της δημιουργίας συστημάτων παραγωγής προσωποποιημένων συστάσεων.
- Παρουσίαση ιστοτόπων που χρησιμοποιούν συστήματα παραγωγής συστάσεων
- Παρουσίαση και ανάλυση κατηγοριών αλγορίθμων παραγωγής συστάσεων όπως:
    1. βάσει περιεχομένου (content based),
    2. συνεργατικής διήθησης (collaborative filtering),
    3. βασισμένα στη γνώση (knowledge based) και
    4. υβριδικές
- Περαιτέρω ανάλυση της κατηγορίας collaborative filtering αφού από αυτή την κατηγορία επελέγη τελικά ο αλγόριθμος που χρησιμοποιήθηκε στο νέο σύστημα.
- Αξιολόγηση αλγορίθμων της κατηγορίας collaborative filtering.
- Ανάλυση του τρόπου λειτουργίας ευρέως γνωστών ιστοτόπων και των αλγορίθμων της κατηγορίας collaborative filtering που χρησιμοποιούνται για την παραγωγή συστάσεων. Συγκεκριμένα, αναλύθηκαν οι αλγόριθμοι του amazon, του youtube και του Google news.
- Περιγραφή της πλατφόρμας που χρησιμοποιήθηκε για την παραγωγή συστάσεων. Πολύ συνοπτικά, η πλατφόρμα αφορά το μάθημα «Ανάπτυξη Εφαρμογών σε Προγραμματιστικό Περιβάλλον» που εξετάζεται Πανελλαδικώς στην Τεχνολογική κατεύθυνση του Γενικού Λυκείου και περιλαμβάνει εκπαιδευτικό υλικό όπως θεωρία, λυμένες και άλυτες ασκήσεις. Σκοπός της πλατφόρμας είναι η λειτουργία αυτής ως ιστότοπος που παρέχει εξατομικευμένη βοήθεια στους μαθητές κατά την αναζήτηση του κατάλληλου εκπαιδευτικού υλικού.
- Παρουσίαση των μεθόδων παραγωγής συστάσεων της πλατφόρμα με τη χρήση πρόσθετων προγραμμάτων (extensions) που μεταφορτώθηκαν από το Διαδίκτυο. Συγκεκριμένα, στην πλατφόρμα χρησιμοποιούνται επεκτάσεις, δηλαδή ειδικά προγράμματα που υπάρχουν σε πολλούς ιστοτόπους, μέσω των οποίων προτείνονται στον επισκέπτη της πλατφόρμας τα πιο δημοφιλή, τα πιο πολυδιαβασμένα και τα πιο σχετικά, βάσει ετικετών, άρθρα. Οι δύο πρώτοι τρόποι δεν παράγουν εξατομικευμένες συστάσεις αφού

Πανεπιστήμιο Πειραιώς

προτείνονται στον επισκέπτη μια δεδομένη χρονική στιγμή ανεξάρτητα από τις επιλογές του. Αντιθέτως, ο τελευταίος τρόπος, που αφορά τα πιο σχετικά άρθρα, έχει στοιχεία εξατομίκευσης αφού προτείνονται άρθρα που έχουν κάποιες κοινές ετικέτες με άλλα άρθρα που έχει επισκεφθεί στο παρελθόν ο εν λόγω επισκέπτης της πλατφόρμας. Φυσικά παρουσιάζονται και τα προβλήματα που έχει η συγκεκριμένη μέθοδος βάσει ετικετών. Από τα ανωτέρω συνάγεται ότι δεν ικανοποιούνται οι τεθέντες στόχοι για πλήρως εξατομικευμένες συστάσεις.

- Στη συνέχεια αναφέρεται η κατηγορία αλγορίθμων από την οποία θα αντληθούν τα απαραίτητα στοιχεία για να παράγει το σύστημα προσωποποιημένες συστάσεις υψηλής ποιότητας. Ακολούθως επεξηγούνται οι λόγοι που επιλέχθηκε να χρησιμοποιηθεί η κατηγορία collaborative filtering και η υποκατηγορία user-based.
- Σύμφωνα με την κατηγορία αυτή, θα πρέπει να υπολογιστεί η ομοιότητα των χρηστών το οποίο σημαίνει ότι θα πρέπει να βρεθούν οι όμοιοι χρήστες δηλαδή αυτοί που έχουν δώσει παρόμοια βαθμολογία σε κοινά αντικείμενα. Εν συνεχεία θα προταθούν σε κάθε χρήστη, που βρέθηκε όμοιος με έναν άλλον, τα αντικείμενα που έχει ψηφίσει ο όμοιός του και όχι ο ίδιος.
- Έπειτα παρατίθεται ο μαθηματικός τύπος από τον οποίο προκύπτει ο δημιουργηθείς αλγόριθμος που υπολογίζει την ομοιότητα των χρηστών. Παράλληλα, και αυτό είναι το σημαντικότερο ίσως τμήμα της υλοποίησης του συστήματος, αναλύονται διεξοδικώς τα βήματα ενσωμάτωσης του αλγορίθμου κατηγορίας collaborative filtering ούτως ώστε να λειτουργήσει σωστά στην πλατφόρμα. Αυτή η διαδικασία περιλαμβάνει συνοπτικά την αξιοποίηση της βαθμολογίας κάθε εγγεγραμμένου χρήστη, τον υπολογισμό της ομοιότητας δύο χρηστών βάσει της παρόμοιας βαθμολογίας που έχουν δώσει σε κοινά αντικείμενα-άρθρα, τις απαραίτητες τροποποιήσεις της βάσης δεδομένων του συστήματος και την πρόταση συγκεκριμένου αριθμού, που έχει οριστεί σε πέντε, άρθρων στον όμοιο χρήστη. Σημειωτέον ότι ο αλγόριθμος ενσωμάτωσης των βημάτων της μεθοδολογίας collaborative filtering παρατίθεται σε φυσική γλώσσα κατά βήματα ενώ τα σημαντικότερα τμήματα κώδικα PHP, MySQL και HTML περιλαμβάνονται στο παράρτημα.
- Τέλος, περιλαμβάνεται η αξιολόγηση του συστήματος με τη χρήση γνωστών μετρικών όπως είναι το Μέσο Απόλυτο Σφάλμα και ο δείκτης απόστασης Kendall tau. Ουσιαστικά οι μαθητές (χρήστες του συστήματος) εκλήθησαν να βαθμολογήσουν όλους τους χρησιμοποιούμενους αλγορίθμους παραγωγής συστάσεων της πλατφόρμας. Τα αποτελέσματα έδειξαν ότι οι χρήστες, αν και κάποιοι δεν ικανοποιήθηκαν αρκετά από τις παραγόμενες συστάσεις, προτίμησαν σε πολύ μεγάλο ποσοστό ως βέλτιστο τρόπο παραγωγής

Πανεπιστήμιο Πειραιώς

συστάσεων τον αλγόριθμο της κατηγορίας collaborative filtering. Εν κατακλείδι φαίνεται οι τεθέντες επιδιωκόμενοι στόχοι ικανοποιήθηκαν.

6.2 Μελλοντική έρευνα

Μετά την επιτυχημένη, βάσει της αξιολόγησης, χρήση ενός συστήματος παραγωγής συστήματος που περιλάμβανε όλα τα ανωτέρω περιγραφόμενα στοιχεία, θα παρατεθούν μία σειρά από προτάσεις για μελλοντική έρευνα και περαιτέρω επέκταση του συστήματος:

- Αύξηση του δείγματος έτσι ώστε να εξαχθούν ασφαλέστερα συμπεράσματα. Στο υπάρχον σύστημα χρησιμοποιήθηκαν 20 μαθητές, με την αύξηση αυτού του αριθμού να μπορεί εύκολα να πραγματοποιηθεί.
- Θα είναι σημαντικό να ελεγχθεί η πλατφόρμα όχι μόνο σε μία σχολική μονάδα αλλά και σε περισσότερες με μεγαλύτερη διαβάθμιση του επιπέδου επίδοσης των μαθητών. Σκοπός θα είναι να ελεγχθεί αν η χρήση μιας τέτοιας μεθόδου παραγωγής συστάσεων θα βελτιώσει όντως την επίδοση των μαθητών. Για το λόγο αυτό θα πρέπει οι μαθητές να εξεταστούν σε θέματα του μαθήματος πριν χρησιμοποιήσουν την πλατφόρμα και έπειτα από την πιλοτική λειτουργία να εξεταστούν σε θέματα παραπλήσιας δυσκολίας για να ελεγχθεί το κατά πόσο βελτιώθηκαν ή όχι.
- Ο εμπλουτισμός του περιεχομένου της πλατφόρμας είναι οπωσδήποτε αναγκαίος. Σίγουρα στις υπάρχουσες θεματικές κατηγορίες η προσθήκη νέου εκπαιδευτικού υλικού θα προσδώσει μεγαλύτερη δυναμική στην πλατφόρμα αλλά και η προσθήκη νέων κατηγοριών θα είναι ενδιαφέρουσα (π.χ η παράθεση λυμένων επαναληπτικών διαγωνισμάτων ανάλογης δυσκολίας με αυτά των Πανελλαδικών εξετάσεων).
- Ενσωμάτωση της δυνατότητας αξιολόγησης του συστήματος μέσω κώδικα PHP και MySQL. Η εν λόγω κίνηση σίγουρα θα ανεβάσει το επίπεδο της πλατφόρμας αλλά υπάρχουν σοβαρά τεχνικά θέματα που θα πρέπει να αντιμετωπιστούν σε αυτή την περίπτωση. Γι' αυτό άλλωστε και συνιστά μια πολυσύνθετη πρόκληση. Σε αυτήν την περίπτωση θα είναι ασυγκρίτως ευκολότερη η αξιολόγηση πολύ μεγαλύτερου δείγματος και, συνεπώς, θα μπορούν να εξαχθούν ασφαλέστερα συμπεράσματα.
- Υλοποίηση κάποιου αλγορίθμου της κατηγορίας content based για να γίνει στη συνέχεια έλεγχος κατά πόσον υπερτερεί στη συγκεκριμένη πλατφόρμα η μέθοδος αυτή ή ο ήδη υλοποιημένος αλγόριθμος της κατηγορίας collaborative filtering (υποκατηγορίας user-based).

Πανεπιστήμιο Πειραιώς

- Υλοποίηση υβριδικού αλγορίθμου. Όπως αναφέρθηκε στην παράγραφο 2.4.4 οι υβριδικές μέθοδοι υπερβαίνουν τα μειονεκτήματα των content based και των collaborative filtering κατηγοριών. Απλώς οι τεχνικές δυσχέρειες και ο συνολικός σχεδιασμός ενός τέτοιου εγχειρήματος εμπεριέχουν σημαντικό βαθμό δυσκολίας.
- Μεγαλύτερη εξατομίκευση. Στο υπάρχον σύστημα το προφίλ του χρήστη, βάσει του οποίου υπολογίζεται η ομοιότητα με άλλους χρήστες, καθορίζεται αποκλειστικά από τις βαθμολογίες που έχει δώσει. Συνεπώς θα είναι εξαιρετικά ενδιαφέρον να συμπεριληφθούν κι άλλοι παράγοντες όπως φύλο, ηλικία, επίδοση. Για να γίνει αυτό θα πρέπει να ενσωματωθούν οι παράγοντες αυτοί στο συντελεστή Pearson με συγκεκριμένη βαρύτητα ο καθένας. Η εισαγωγή των τιμών αυτών θα μπορεί να γίνει με ξεχωριστή φόρμα στην αρχική σελίδα της πλατφόρμας.

Πανεπιστήμιο Πειραιώς

Παράρτημα

I. Καταχώριση της βαθμολογία κάθε χρήστη σε πίνακα

Αυτό υλοποιείται με τον παρακάτω αλγόριθμο:

```
function getUserRatings($user) {
$db_host = 'localhost';
$db_user = 'root';
$db_pwd = '';
$database = 'joomlapin';
if (!mysql_connect($db_host, $db_user, $db_pwd))
   die("Can't connect to database");
```

Στο σημείο αυτό γίνεται σύνδεση στη βάση. Αν δεν καταστεί αυτό εφικτό, τότε εμφανίζεται μήνυμα λάθους

```
if (!mysql_select_db($database))
   die("Can't select database");
$sql = "SELECT a.id, b.rating\n"
   . "FROM jos_content a\n"
   . "LEFT JOIN (SELECT * FROM jos_vi_rating WHERE jos_vi_rating.user_id ="
.$user.") b ON (a.id = b.content_id)\n"
   . "ORDER BY a.id ";
```

Με το ανωτέρω ερώτημα γίνεται λήψη όλων των βαθμολογιών που έχει καταχωρίσει ο ενεργός χρήστης

```
$result = mysql_query($sql);
```

Εδώ εκτελείται το ερώτημα

```
if (!$result) {
   die("Query to show fields from table failed");
}
```

Μήνυμα λάθους σε περίπτωση αποτυχίας

```
$j = 0;
$arr = array();
```

Στη συνέχεια με το βρόχο της εντολής while κατασκευάζεται ένας προσωρινός πίνακας (array) για να αποθηκευτούν τα αποτελέσματα του παραπάνω ερωτήματος

Πανεπιστήμιο Πειραιώς

```
while($row = mysql_fetch_row($result))
{
   //$arr[] = $row[1];
   $arr[$row[0]] = $row[1];
   $j = $j + 1;
}
mysql_free_result($result);
return $arr;
```

Τελικά επιστρέφεται ο πίνακας με τις βαθμολογίες που έχει δώσει κάθε χρήστης

```
}
$rat = getUserRatings(activate_user);
}
```

II. Υλοποίηση και τεκμηρίωση του αλγορίθμου παραγωγής συστάσεων

Στη συνέχεια θα παρατεθούν τα κυριότερα τμήματα κώδικα σε PHP που περιλαμβάνουν εντολές της MySQL και της HTML.

```
if ($user_id ==0)
 γίνεται έλεγχος αν ο επισκέπτης έχει εισέλθει ως εγγεγραμμένος χρήστης
 echo "πρέπει να εισέλθεις ως εγγεγραμμένος χρήστης για εξατομικευμένες
συστάσεις <br />";
  else // exei ginei eisodos ws xristis

    foreach ($rat as $key=>$val)
    {
      if ($rat[$key]==NULL)
         $metr=$metr + 1;
    }
```

Πραγματοποιείται έλεγχος για το αν ο χρήστης έχει ψηφίσει κάποιο άρθρο. Αν δεν το έχει κάνει αυτό, τότε δεν μπορεί το σύστημα να ¨κατανοήσει¨ τη συμπεριφορά του και να ελέγξει την ομοιότητά του με άλλους χρήστες

Πανεπιστήμιο Πειραιώς

```
 if ($num_items ==$metr)
 echo ("current user den psifises pote <br />");
```

εμφανίζεται ένα σχετικό μήνυμα σε περίπτωση που δεν έχει ψηφίσει κανένα άρθρο

Η διαδικασία που ακολουθεί εφαρμόζεται για κάθε χρήστη.

```
  $ratus = getUserRatings($j);
```

Ο πίνακας (array) ratus περιέχει τη βαθμολογία που έδωσε ο χρήστης που συνεξετάζεται με τον ενεργό δηλαδή ο χρήστης που είναι υποψήφιος όμοιος. Ουσιαστικά θα γίνει σύγκριση του ενεργού χρήστη και του εκάστοτε συνεξεταζομένου για το αν έχουν ψηφίσει κοινά άρθρα και τι βαθμολογία έχουν δώσει σε αυτά.

```
       $metr=0;
        $metr1=0;
        Η μεταβλητή metr1 ελέγχει αν υπάρχουν κοινά ψηφισθέντα αντικείμενα.
        $sum1=0;
        $sum2=0;
        $sum3=0;
        $pearson= -2;

  foreach ($rat as $key=>$val)
 {
  if (($rat[$key] != NULL) AND ($ratus[$key] != NULL))
```

Αν ένα αντικείμενο έχει ψηφιστεί από τον ενεργό και τον εξεταζόμενο χρήστη υπολογίζονται όλα τα μεγέθη που είναι απαραίτητα για το συντελεστή Pearson

```
    {
      $a = $rat[$key]-3;
      $b = $ratus[$key]-3;
      $gin = $a * $b;
      $sum1 = $gin + $sum1;
      $sum2 = $sum2 + $a * $a;
      $sum3 = $sum3 + $b * $b;
      $metr1 = 1;
    }
 }
$pearson= array();
```

δημιουργείται ο πίνακας pearson για την αποθήκευση του ομώνυμου συντελεστή

```
if ($metr1 == 1)
```

δηλαδή υπάρχουν αντικείμενα που έχουν ψηφιστεί και από τους δύο χρήστες

Πανεπιστήμιο Πειραιώς

```
 {
  if ($sum1==0)
```
αυτό σημαίνει ότι έχουν δώσει τρία σε όλα τα άρθρα και μηδενίζεται το άθροισμα που είναι ο αριθμητής

```
    $pearson[$j] = 0.6;
```
Δίνεται μια προκαθορισμένη τιμή στο συντελεστή αφού υπάρχει ομοιότητα

```
  else
```
υπολογίζεται πλέον ο συντελεστής pearson
```
    $pearson[$j] = $sum1 / (sqrt ($sum2 * $sum3));
  }
else
```
δηλαδή δεν έχουν ψηφίσει κοινά αντικείμενα και, συνεπώς, δεν υπάρχει ομοιότητα
```
 $pearson[$j] = -1;
```

Οι παρακάτω τρεις εντολές δεν χρησιμοποιούνται στη δοκιμασμένη έκδοση της πλατφόρμας. Δείχνουν τους χρήστες που συνεξετάζονται και την τιμή που έχει ο συντελεστής Pearson για το κάθε ζευγάρι χρηστών. Ουσιαστικά παρέχει τα απαραίτητα στιγμιότυπα με τα οποία ελέγχεται η ορθότητα προσωρινών αποτελεσμάτων

```
echo " energos xristis = $user_id <br /> ";
echo " pearson= $pearson[$j] <br /> ";
echo " xristis poy eksetazetai me ton energo = $j <br /> ";
```

Στη συνέχεια γίνεται έλεγχος για το αν ο συντελεστής Pearson έχει τιμή πάνω από 0.5 και ο ενεργός χρήστης είναι διαφορετικός από τον συνεξεταζόμενο

```
if (($pearson[$j] >0.5) && ($j != $user_id))
```
οι χρήστες user_id kai j είναι όμοιοι

Η παρακάτω διαδικασία εκτελείται για κάθε βαθμολογία του ενεργού χρήστη
```
   if (($rat[$key] == NULL) AND ($ratus[$key]> 3))
```

Αν ο ενεργός χρήστης δεν έχει ψηφίσει το εξεταζόμενο άρθρο και η προτεινόμενη βαθμολογία από τον όμοιο χρήστη είναι πάνω από τρία τότε

```
mysql_query("INSERT INTO jos_similarity (user_activate,pearson, users, item,
prediction)
VALUES ('$user_id', '$pearson[$j]' ,'$j', '$key','$ratus[$key]' )");
```

Πανεπιστήμιο Πειραιώς

καταχωρίζονται τα στοιχεία στον πίνακα (table) jos_similarity ο οποίος περιέχει τα πεδία: ενεργός χρήστης, όμοιος χρήστης(users), συντελεστής pearson, το άρθρο που προτείνεται (item) καθώς και την προτεινόμενη βαθμολογία (prediction) δηλαδή αυτή που έβαλε ο όμοιος χρήστης.

Οι επόμενες δύο εντολές δεν χρησιμοποιούνται στη δοκιμασμένη έκδοση αλλά, απλώς, παρέχουν κάποια στιγμιότυπα που βοηθούν στον έλεγχο της ορθότητας.

```
 echo " item poy psifistike = $key <br /> ";
  echo " vathmologia pou dothike ratus= $ratus[$key] <br /> ";
```

Το παρακάτω τμήμα κώδικα είναι υψίστης σημασίας διότι αποθηκεύει στον array results τα προτεινόμενα άρθρα κατά φθίνουσα σειρά βάσει συντελεστή Pearson. Αν η τιμή του Pearson είναι ίδια, ταξινομεί κατά φθίνουσα σειρά βάσει προτεινόμενης βαθμολογίας (πεδίο prediction). Τέλος, αν ένα άρθρο πρόκειται να προταθεί στον ενεργό χρήστη πάνω από μία φορά (δηλαδή από αρκετούς όμοιους χρήστες) τότε θα καταχωριστεί στον array results μόνο μία φορά και, συγκεκριμένα, από την πλειάδα του jos_similarity που προηγείται στην ανωτέρω περιγραφόμενη ταξινόμηση. Με τον τρόπο αυτό αποφεύγεται άσκοπη επανάληψη

```
$results = mysql_query("SELECT distinct item, pearson, prediction FROM
jos_similarity where jos_similarity.user_activate=$user_id order by pearson desc,
prediction desc ");

$counter=0;
```

Ο μετρητής counter μετρά τα προτεινόμενα άρθρα που εμφανίζονται. Τονίζεται για μια ακόμα φορά ότι ο αριθμός τους έχει οριστεί σε πέντε(5).

```
while ($row = mysql_fetch_array( $results ))
 {
  if ((($rat[$row[item]]) ==NULL) && $counter<5)
    {
     $counter=$counter +1;
      echo " proteinomeno article No $counter <br /> ";
     echo " proteinomena item $row[item] pearson $row[pearson] provlepsi
$row[prediction]";
```

στη συνέχεια λαμβάνεται ο τίτλος του προτεινόμενου άρθρου και στο χρήστη θα εμφανιστεί ένας σύνδεσμος που οδηγεί σε αυτό

Πανεπιστήμιο Πειραιώς

```
$key= $row[item];
 $article =& JTable::getInstance("content");
      $article->load($key);
      echo $article->get("title");
?>
<a href=
"http://localhost/joomlapin/index.php?option=com_content&view=article&id=<?ph
p echo $key; ?>">  click here   </a> <br /> <br />
<?php

          } //if (($rat[$row[item]]) ==NULL)

    }   // while ($row = mysql_fetch_array
mysql_free_result($results);
}
```

Πανεπιστήμιο Πειραιώς

III. Ερωτηματολόγιο που δόθηκε στους χρήστες του συστήματος

1. Ψηφίστε στην κλίμακα 1-5 (1: καθόλου χρήσιμος, 5:πολύ χρήσιμος) το πώς σας φάνηκε ο αλγόριθμος παραγωγής συστάσεων ανάλογα με το αν σας πρότεινε όντως ενδιαφέροντα άρθρα.
2. Επιλέξτε τον αποδοτικότερο αλγόριθμο παραγωγής συστάσεων από αυτούς που χρησιμοποιεί η πλατφόρμα
   Α. Top rated (βάσει αθροίσματος συνολικής βαθμολογίας)
   Β. Most read (Πιο πολυδιαβασμένο άρθρο βάσει των εμφανίσεων (κλικ) κάθε άρθρου)
   Γ. Most Related (πιο σχετικά άρθρα βάσει ετικετών)
   Δ. Αλγόριθμος κατηγορίας User based (αλγόριθμος με βάση την ομοιότητα των χρηστών) ο οποίος προτείνει τα πέντε πιο σχετικά άρθρα σε κάθε χρήστη.

3. Καταγράψτε τη λίστα προτάσεων του αλγορίθμου κατηγορίας user based. Συμφωνείτε με αυτήν; Αν όχι, προτείνετε τη δική σας δηλαδή να ταξινομήσετε την παραχθείσα, από τον αλγόριθμο, λίστα με βάσει τα δικά σας κριτήρια τα οποία δεν είστε υποχρεωμένοι να καταγράψετε. Η δική σας λίστα δεν έχει κανένα περιορισμό δηλαδή μπορεί να διαφέρει εξ' ολοκλήρου από την αρχική και να έχει την ακριβώς αντίστροφη σειρά.

4. Θεωρείτε ότι οι προτάσεις που εξήγαγε ο αλγόριθμος user-based βοήθησαν κατά τη μελέτη σας; Νιώθατε ότι κατά την αναζήτησή σας είχατε έναν πολύτιμο σύμβουλο που μπορούσε να προτείνει χρήσιμα αντικείμενα ή αισθανθήκατε πως δεν υπήρξε η παραμικρή καθοδήγηση από την πλατφόρμα και χρησιμοποιούσατε έναν απλό ιστότοπο;

5. Παραθέστε τις δικές σας προτάσεις για την παραγωγή συστάσεων. Έχετε κάποια ιδέα για το πώς θα μπορούσε να είναι αποδοτικότερη και χρησιμότερη η παραγωγή συστάσεων;

Πανεπιστήμιο Πειραιώς

Βιβλιογραφία

Adomavicius G. & Tuzhilin A., "Context-aware recommender systems," in Recommender Systems Handbook: A Complete Guide for Research Scientists and Practitioners. Springer, 2010.

Adomavicius G. & Tuzhilin A., Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE Trans. on Knowledge and Data Engineering, 17(6):734–749, 2005.

Aggarwal C.C., Wolf J.L., Wu K-L. & Yu P.S., "Horting Hatches an Egg: A New Graph-Theoretic Approach to Collaborative Filtering," Proc. Fifth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, Aug. 1999.

Agrawal R., Imielinski T. & Swami A., Mining association rules between sets of items in large databases. SIGMOD Rec., 22(2):207–216, 1993.

Ansari A., Essegaier S. & Kohli R., "Internet Recommendations Systems," J. Marketing Research, pp. 363-375, Aug. 2000.

Baeza-Yates R. & Ribeiro-Neto B., Modern Information Retrieval. Addison-Wesley, 1999.

Balabanovic M. & Shoham Y., "Fab: Content-Based, Collaborative Recommendation," Comm. ACM, vol. 40, no. 3, pp. 66-72, 1997.

Bercovitz B., Kaliszan F., Koutrika G., Liou H., Mohammadi Zadeh Z. & Garcia-Molina H., CourseRank: A Social System for Course Planning, California, 2009

Bogers T., "Movie recommendation using random walks over the contextual graph," 2010.

Breese J.S., Heckerman D. & Kadie C., "Empirical Analysis of Predictive Algorithms for Collaborative Filtering," Proc. 14th Conf. Uncertainty in Artificial Intelligence, July 1998.

Bridge D., G¨oker M., McGinty L. & Smyth B., Case-based recommender systems, Knowledge Engineering Review 20 (2005), no. 3, 315–320

Burke R., Knowledge-based recommender systems, Encyclopedia of Library and Information Science 69 (2000), no. 32, 180–200.

Cacheda F., Carneiro V., Fernadez D. & Formoso V., Comparison of collaborative filtering algorithms: Limitations of current techniques and proposals for scalable, high-performance recommender systems, Journal, ACM Transactions on the Web (TWEB)

Πανεπιστήμιο Πειραιώς

TWEB Homepage archive, Volume 5 Issue 1, February 2011, Article No. 2, ACM New York, NY, USA

Cha M., Kwak H., Rodriguez P., Ahn Y-Y. & Moon S., "I Tube, YouTube, Everyday Tubes: Analyzing the World's Largest User Generated Content Video System," in ACM Internet Measurement Conference, October 2007

Chakrabarti S, Mining the web: Discovering knowledge from hypertext data, Science and Technology Books, 2002.

Cheng X., Dale C. & Liu J.. " Statistics and Social Network of YouTube Videos", In Proceedings of the 16th International Workshop on Quality of Service, 2008

Creel J., Maslov A., Mikeal C. & Speight, The Orellana Project: An Adaptive Recommendation System for Amazon.com, 2010

Das A., Datar M., Garg A. & Rajaram S., Google News Personalization: Scalable Online Collaborative Filtering, WWW Conference 2007, May 8–12, 2007, Banff, Alberta, Canada

Davidson J., Liebald B., Liu J., Nandy P. & van Vleet T., The YouTube Video Recommendation, RecSys2010, September 26–30, 2010, Barcelona, Spain.

Delgado J. & Ishii N., "Memory-Based Weighted-Majority Prediction for Recommender Systems," Proc. ACM SIGIR '99 Workshop Recommender Systems: Algorithms and Evaluation, 1999.

Felfernig A. & Burke R., Constraint-based recommender systems: technologies and research issues, Proceedings of the 10th International Conference on Electronic Commerce (ICEC '08) (Innsbruck, Austria), ACM, 2008, pp. 1–10.

Felfernig A., Friedrich G., Jannach D. & Zanker M., An integrated environment for the development of knowledge-based recommender applications, International Journal of Electronic Commerce 11 (2006–07), no. 2, 11–34.

Getoor L. & Sahami M., "Using Probabilistic Relational Models for Collaborative Filtering," Proc. Workshop Web Usage Analysis and User Profiling (WEBKDD '99), Aug. 1999.

Goldberg D., Nichols D., Oki B. M. & Terry D.. "Using collaborative filtering to weave an information tapestry, 1992

Gyongyi, Z.; Koutrika, G.; Pedersen, J. & Garcia-Molina, H. Questioning Yahoo! Answers. In 1st Workshop on Question Answering on the Web, 2008.

Herlocker J.L., Konstan J.A., Terveen L.G., & Riedl J.T., "Evaluating Collaborative Filtering Recommender Systems," ACM Trans. Information Systems, vol. 22, no. 1, pp.

Πανεπιστήμιο Πειραιώς

5-53, 2004.

Hofmann T., “Collaborative Filtering via Gaussian Probabilistic Latent Semantic Analysis,” Proc. 26th Ann. Int’l ACM SIGIR Conf., 2003.

Jannach D., Zanker M, Felfernig A. & Friedrich G., Recommender Systems: An Introduction, Cambridge University Press, 2011

Koutrika G., Bercovitz B. & Garcia-Molina H., FlexRecs: Expressing and Combining Flexible Recommendations, 2009 (a)

Koutrika G., Bercovitz B., Kaliszan F., Liou H. & Garcia-Molina H., “CourseRank: A Closed-Community Social System Through the Magnifying Glass”, 2009 (b)

Lampe, C.; Ellison, N. & Steinfield, C., Changes in use and perception of Facebook. In Proc. of the Int’l CSCW Conf. 2008

Leydesdorff L., The relation between Pearson’s correlation coefficient r and Salton’s cosine measure Journal of the American Society for Information Science & Technology, 2008

Li Q. & Zhang Y., An Efficient Mining Algorithm for Top-k Strongly Correlated Item Pairs, Fourth International Conference on Internet Computing for Science and Engineering, 2009

Linden G., Smith B. & York J., Amazon.com Recommendations Item-to-Item Collaborative Filtering, industry report, 2003

Lorenzi F. & Ricci F., Case-based recommender systems: A unifying view, Intelligent Techniques for Web Personalisation, Lecture Notes in Artificial Intelligence, vol. 3169, Springer, 2005, pp. 89–113.

Pazzani M., “A Framework for Collaborative, Content-Based, and Demographic Filtering, Artificial Intelligence Rev., pp. 393-408, Dec. 1999.

Pennock D. M., Horvitz E., Lawrence S. & Lee Giles C. , Collaborative Filtering by Personality Diagnosis: A Hybrid Memory- and, Model-Based Approach, 2000

Prekopcsák Z., “Content organization and discovery:state-of-the-art and new ideas for P2P-Fusion”, May 2007

Resnick P., Iacovou N., Suchak M., Bergstrom P. & Riedl J. “GroupLens: An Open Architecture for Collaborative Filtering of Netnews. In Proceedings of CSCW ’94, Chapel Hill, NC., 1994

Rich E., “User Modeling via Stereotypes,” Cognitive Science, vol. 3, no. 4, pp. 329-354, 1979.

Πανεπιστήμιο Πειραιώς

Salton G., Automatic Text Processing. Addison-Wesley, 1989.

Salton G., Wong A. & Yang C.S., A vector space model for information retrieval, Journal of the American Society for Information Science 18 (1975), no. 11, 613–620.

Sarwar B., Karypis G., Konstan J., & Riedl J., Item-Based Collaborative Filtering Recommendation Algorithms, WWW10, May 1-5, 2001, Hong Kong. ACM 1-58113-348-0/01/0005

Shardanand U. & Maes P., "Social Information Filtering: Algorithms for Automating 'Word of Mouth'," Proc. Conf. Human Factors in Computing Systems, 1995.

Tran T. & Cohen R., "Hybrid Recommender Systems for Electronic Commerce," Proc. Knowledge-Based Electronic Markets, Papers from the AAAI Workshop, Technical Report WS-00-04, AAAI Press, 2000.

Zanker M., Jessenitschnig M. & Schmid W., Preference Reasoning with Soft Constraints in Constraint-Based Recommender Systems Constraints, Springer, 15 (2010), no. 4, 574–595

Zhou R., Khemmarat S. & Gao L., The impact of YouTube recommendation system on video views, IMC '10 Proceedings of the 10th ACM SIGCOMM conference on Internet measurement, Pages 404-410 ACM New York, NY, USA ©2010

Πανεπιστήμιο Πειραιώς

Σύνδεσμοι

1. http://encyc.org/wiki/Wandex).
2. http://www.google.com/intl/el/about/company/history/
3. www.joomla.org
4. www.amazon.com
5. http://en.wikipedia.org/wiki/Amazon.com
6. www.booking.com
7. www.imdb.com
8. https://en.wikipedia.org/wiki/Google_Search
9. http://en.wikipedia.org/wiki/Netflix_Prize
10. www.pseudoglossa.gr
11. http://www.spinet.gr/glossomatheia
12. http://users.sch.gr/alkisg/
13. http://en.wikipedia.org/wiki/Kendall_tau_distance
14. http://www.Amazon.com/gp/aws/landing.html Amazon Web Services

Πανεπιστήμιο Πειραιώς