

Ανοικτό Πανεπιστήμιο Κύπρου

Σχολή Θετικών και Εφαρμοσμένων Επιστημών

**Μεταπτυχιακό Πρόγραμμα Σπουδών *Πληροφοριακά και
Επικοινωνιακά Συστήματα***

Μεταπτυχιακή Διατριβή



**Υβριδικό σύστημα συστάσεων άρθρων με χρήση
πολλαπλών κριτηρίων και κοινωνικών δικτύων**

Βάϊος Συντσιρμάς

**Επιβλέπων Καθηγητής
Ιωάννης Κατάκης**

Ιούνιος 2016

Ανοικτό Πανεπιστήμιο Κύπρου

Σχολή Θετικών και Εφαρμοσμένων Επιστημών

**Μεταπτυχιακό Πρόγραμμα Σπουδών *Πληροφορικά και
Επικοινωνιακά Συστήματα***

Μεταπτυχιακή Διατριβή

**Υβριδικό σύστημα συστάσεων άρθρων με χρήση πολλαπλών
κριτηρίων και κοινωνικών δικτύων**

Βάϊος Συντσιρμάς

**Επιβλέπων Καθηγητής
Ιωάννης Κατάκης**

Η παρούσα μεταπτυχιακή διατριβή υποβλήθηκε προς μερική εκπλήρωση των
απαιτήσεων για απόκτηση μεταπτυχιακού τίτλου σπουδών
στα Πληροφορικά Συστήματα
από τη Σχολή Θετικών και Εφαρμοσμένων Επιστημών
του Ανοικτού Πανεπιστημίου Κύπρου.

Ιούνιος 2016

ΛΕΥΚΗ ΣΕΛΙΔΑ

Περίληψη

Η διατριβή αυτή ασχολήθηκε με τον σχεδιασμό και την ανάπτυξη ενός υβριδικού συστήματος συστάσεων άρθρων σε χρήστες, χρησιμοποιώντας πολλαπλά κριτήρια για τον υπολογισμό των αποτελεσμάτων.

Το σύστημα σχεδιάστηκε ώστε να είναι σε θέση να λειτουργεί και να προτείνει άρθρα σε πραγματικό χρόνο, με τη χρήση ενός συνόλου κριτηρίων. Γίνεται συνδυασμός τεχνικών online και offline λειτουργίας, ώστε η εφαρμογή να έχει όσο το δυνατόν λιγότερες απαιτήσεις από υλικοτεχνική υποδομή, αλλά οι προτάσεις του συστήματος να γίνονται σε πραγματικό χρόνο.

Η αξιολόγηση της εν λόγω εφαρμογής έγινε στον ιστότοπο athensvoice.gr, ώστε η αποτελεσματικότητα της να είναι πάνω σε πραγματικές συνθήκες λειτουργίας. Η εφαρμογή που περιγράφεται στην διατριβή κάνει χρήση πολλαπλών κριτηρίων - συμπεριλαμβανομένου και του κοινωνικού δικτύου Twitter- για να κάνει προτάσεις στους χρήστες. Όσο για την επιλογή του αριθμού αποτελεσμάτων κάθε κριτηρίου, γίνεται χρήση ενός προσαρμοστικού αλγόριθμου. Η αποτελεσματικότητα της εφαρμογής μετρήθηκε με τη χρήση των δεδομένων από το Google Analytics, συγκρίνοντας περιόδους χρήσης της εφαρμογής με περιόδους μη χρήσης της, αλλά και με δεδομένα του συστήματος, επί των χρηστών που επέλεξαν να διαβάσουν κάποιο άρθρο που τους προτάθηκε. Λόγω των παραπάνω επιλογών, το σύστημα εμφανίζει ορθότερη αντιμετώπιση των προβλημάτων τα οποία παρουσιάζονται στις παραδοσιακές τεχνικές συστάσεων.

Summary

This thesis dealt with the design and development of a hybrid system of articles recommendations to users, using multiple criteria to calculate the results. The system was designed to operate and propose articles in real time, using a set of criteria. With a combination of techniques online and offline mode, the application will have the least possible demands on computational resources, but the system recommendations will be made in real time. The assessment of this application was made on the athensvoice.gr site, so the effectiveness of the implementation is on actual operating conditions. The application that is described in this thesis, uses multiple criteria-including social network Twitter, to make suggestions to users; an adaptive algorithm was used to help choose the number of results of each criterion. The effectiveness of the application was measured with the use of data from Google Analytics, comparing periods of application usage with periods of non-usage, but also from the system data, for users who chose to read an article that was suggested to them. Because of the above options, the system displays a better way of dealing with problems encountered in traditional technical recommendations.

Ευχαριστίες

Ευχαριστώ θερμά τον επιβλέποντα καθηγητή μου Δρ. Ιωάννη Κατάκη για τις πολύτιμες συμβουλές και την καθοδήγηση του κατά την εκπόνηση της παρούσας μεταπτυχιακής διατριβής. Επίσης τον δημοσιογράφο και εκδότη της Athens Voice κ. Φώτη Γεωργελέ για την αμέριστη υποστήριξη του, καθώς και όλους όσοι βοήθησαν με τον τρόπο τους. Ιδιαίτερα τη μητέρα μου Δήμητρα και τον γιό μου Κωνσταντίνο.

Περιεχόμενα

Εισαγωγή	11
Θεωρητικό Υπόβαθρο.....	15
2.1 Φιλτράρισμα βάσει περιεχομένου (Content Based Filtering).....	16
2.1.1 Τα πλεονεκτήματα ενός συστήματος βασισμένο στο περιεχόμενο.....	16
2.1.2 Τα μειονεκτήματα ενός συστήματος βασισμένου στο περιεχόμενο.....	17
2.2 Συνεργατικό φιλτράρισμα (Collaborative Filtering).....	18
2.2.1 Προβλήματα που παρουσιάζονται στα συστήματα συνεργατικού φιλτραρίσματος.....	20
2.3 Υβριδικό φιλτράρισμα (Hybrid recommendations Filtering)	20
2.4 Προκλήσεις και περιορισμοί.....	22
Σχετικές εργασίες	24
3.1 Φιλτράρισμα βασισμένο στο περιεχόμενο	24
3.2 Συνεργατικό φιλτράρισμα	26
3.3 Υβριδικό φιλτράρισμα.....	27
Ανάπτυξη εφαρμογής.....	31
4.1 Κριτήρια επιλογής αποτελεσμάτων	31
4.1.1 Υποσύστημα ανάδρασης χρήστη.....	35
4.2 Αλγόριθμος.....	36
4.3 Δείγμα.....	43
4.4 Αρχιτεκτονική συστήματος.....	45
4.5 Τεχνολογίες υλοποίησης της εφαρμογής.....	46
4.5.1 Υλικοτεχνική υποδομή, λειτουργικό σύστημα, πλατφόρμα υλοποίησης	46
4.5.2 Δημιουργία και καταγραφή δεδομένων Online και Offline	47
4.5.3 Βάση δεδομένων.....	50
Αποτελέσματα χρήσης του συστήματος.....	53
5.1 Σύγκριση με άλλες τεχνικές και απόδοση του συστήματος με χρήση των δεδομένων του συστήματος.....	53

5.2 Στατιστικά στοιχεία συμπεριφοράς των χρηστών και γραφήματα που σχετίζονται με ορισμένα από τα κριτήρια επιλογής αποτελεσμάτων.....	58
5.3 Συνολικά αποτελέσματα με χρήση του google analytics	63
Επίλογος.....	65
6.1 Συμπεράσματα.....	65
6.2 Μελλοντικές εργασίες.....	66
Βιβλιογραφία.....	68
Παραρτήματα.....	
Παράρτημα Α.....	71
Περιγραφή βάσης δεδομένων.....	71
Παράρτημα Β.....	78
Αρχιτεκτονική βάσης δεδομένων Drupal.....	78
Παράρτημα Γ	79
Κώδικας εφαρμογής	79

Κατάλογος Εικόνων

Εικόνα 1:	Σύστημα προτάσεων περιεχομένου.....	14
Εικόνα 2:	Φιλτράρισμα βάσει περιεχομένου.....	15
Εικόνα 3:	Συνεργατικό φιλτράρισμα.....	17
Εικόνα 4:	Υβριδικό φιλτράρισμα.....	20
Εικόνα 5:	Κώδικας καταγραφής των clicks.....	34
Εικόνα 6:	Αλγόριθμος υπολογισμού των προτάσεων του συστήματος.....	38
Εικόνα 7:	Αλγόριθμος διαχείρισης δεδομένων από το σύστημα.....	39
Εικόνα 8:	Αλγόριθμος υπολογισμού αριθμού άρθρων κάθε κριτηρίου.....	40
Εικόνα 9:	Αλγόριθμος ελέγχου ύπαρξης ή μη cookie από το σύστημα.....	41
Εικόνα 10:	Cookie από το σύστημα.....	42
Εικόνα 11:	Καταγραφή των δεδομένων και εισαγωγή στη βάση.....	43
Εικόνα 12:	Εμφάνιση των αποτελεσμάτων στον ιστότοπο athenvoice.gr.....	43

Εικόνα 13:	Αρχιτεκτονική συστήματος.....	44
Εικόνα 14:	Αρχιτεκτονική συνολικού συστήματος drupal και συστάσεων.....	46
Εικόνα 15:	Βάση δεδομένων συστήματος συστάσεων.....	51
Εικόνα 16:	Αριθμός clicks βάσει περιεχομένου σε διάρκεια μίας εβδομάδας.....	53
Εικόνα 17:	Αριθμός clicks με συνεργατικό φιλτράρισμα μίας εβδομάδας.....	54
Εικόνα 18:	Αριθμός clicks ενός μήνα με όλα τα κριτήρια.....	56
Εικόνα 19:	Ποσοστά εμφάνισης κάθε κριτηρίου.....	57
Εικόνα 20:	Συνολικός αριθμός clicks με τις διαφορετικές προσεγγίσεις.....	57
Εικόνα 21:	Ποσοστά ομοιότητας χρηστών.....	58
Εικόνα 22:	Επιλογή άρθρων επόμενης κατηγορίας.....	59
Εικόνα 23:	Επιλογή άρθρων ίδιας κατηγορίας.....	60
Εικόνα 24:	Επιλογή κατηγοριών στη χρονική ζώνη 02:00 - 07:59.....	61
Εικόνα 25:	Επιλογή κατηγοριών στη χρονική ζώνη 08:00 - 12:59.....	61
Εικόνα 26:	Επιλογή κατηγοριών στη χρονική ζώνη 13:00 - 16:59.....	62
Εικόνα 27:	Επιλογή κατηγοριών στη χρονική ζώνη 17:00 - 20:59.....	62
Εικόνα 28:	Επιλογή κατηγοριών στη χρονική ζώνη 21:00 - 01:59.....	62
Εικόνα 29:	Στατιστικά μήνα με το σύστημα και χωρίς.....	63
Εικόνα 30:	Στατιστικά εβδομάδας με και χωρίς το σύστημα.....	64
Εικόνα 31:	Login στο σύστημα.....	67
Εικόνα 32:	Προσωπική σελίδα με λογαριασμό facebook.....	68
Εικόνα 33:	Πίνακας initdata.....	71
Εικόνα 34:	Πίνακας cid_ip.....	72
Εικόνα 35:	Πίνακας click.....	72
Εικόνα 36:	Πίνακας cosine_sim.....	73
Εικόνα 37:	Πίνακας day_profile.....	73
Εικόνα 38:	Πίνακας day_tid.....	73
Εικόνα 39:	Πίνακας nid_profile.....	74
Εικόνα 40:	Πίνακας similarity_up.....	74
Εικόνα 41:	Πίνακας tid_profile.....	74
Εικόνα 42:	Πίνακας time_profile.....	75
Εικόνα 43:	Πίνακας times_tid.....	75
Εικόνα 44:	Πίνακας user_from_nid_to_nid.....	75
Εικόνα 45:	Πίνακας user_from_to.....	76
Εικόνα 46:	Πίνακας user_has_read.....	76

Εικόνα 47:	Πίνακας user_profile.....	77
Εικόνα 48:	Αρχιτεκτονική βάσης δεδομένων συστήματος Drupal.....	78

Κεφάλαιο 1

Εισαγωγή

Η έρευνα πάνω στους τομείς που ασχολούνται με το φιλτράρισμα της πληροφορίας, την αλληλεπίδραση των χρηστών και τη σύσταση περιεχομένου, συνεχώς αυξάνεται. Τα συστήματα αυτά μπορεί να είναι είτε για προϊόντα είτε για υπηρεσίες είτε γενικότερα για άρθρα (περιεχόμενο). Τα συστήματα συστάσεων είναι ιδιαίτερος σημαντικά σε εφαρμογές e-commerce, αφού είναι σε θέση να προβλέψουν αγαθά που ο χρήστης πιθανώς να ενδιαφέρεται γι αυτά, με αποτέλεσμα να προβεί σε κάποια αγορά.

Γνωστοί ιστότοποι που χρησιμοποιούν συστήματα συστάσεων είναι οι εξής:

blinkx http://www.blinkx.com (TV, ταινίες)
IMDB http://imdb.com (TV shows, ταινίες)
Facebook http://www.facebook.com (κοινωνικό δίκτυο)
Last.fm http://www.last.fm (μουσική, ραδιόφωνο)
Google AdSense (διαφημίσεις)
Google News (Ειδήσεις)
Amazon http://www.amazon.com (ηλεκτρονικό εμπόριο)
Ebay http://ebay.com (ηλεκτρονικό εμπόριο)

Το βασικά προβλήματα που αντιμετωπίζουν ιστότοποι που προσφέρουν περιεχόμενο, όχι μόνο σε μια κατηγορία άρθρων αλλά σε πληθώρα κατηγοριών, είναι ότι οι χρήστες - αναγνώστες έχουν να αντιμετωπίσουν ένα περιβάλλον υπερπληροφόρησης (με άρθρα τα οποία ενδέχεται να μην τους ενδιαφέρουν) και συναντούν δυσκολίες στο να εντοπίσουν άρθρα που εντάσσονται στις προτιμήσεις τους. Αποτέλεσμα των παραπάνω είναι η δυσαρέσκεια αλλά και η σπατάλη χρόνου από μεριάς του χρήστη. Η χρησιμοποίηση ενός συστήματος συστάσεων, δείχνει να μπορεί να δώσει λύσεις σε ένα μεγάλο ποσοστό των προβλημάτων αυτών.

Το κύριο χαρακτηριστικό των συστημάτων συστάσεων είναι ότι παρέχουν εξατομικευμένες προτάσεις σε κάθε χρήστη χωριστά, συνήθως υπό τη μορφή ταξινομημένων λιστών. Οι προτιμήσεις των χρηστών προκύπτουν με δύο τρόπους: είτε απευθείας από τους ίδιους τους χρήστες είτε έμμεσα από το σύστημα. Στην πρώτη περίπτωση οι χρήστες δηλώνουν τις προτιμήσεις τους κατά βάση εντός καθορισμένης βαθμολογικής κλίμακας. Στη δεύτερη περίπτωση το ίδιο το σύστημα αποθηκεύει στοιχεία από την αλληλεπίδραση των χρηστών με αυτό, συνήθως με αρχεία cookies, και κατόπιν τα επεξεργάζεται, προσπαθώντας να αντλήσει τις απαραίτητες πληροφορίες για τις προτιμήσεις τους. Ένα γνωστό τέτοιο παράδειγμα είναι η υπηρεσία ειδήσεων του δικτυακού τόπου Google, η Google news, η οποία κάνει στον χρήστη εξατομικευμένες προτάσεις για την ανάγνωση ειδήσεων, βασισμένες σε παλιότερες αναζητήσεις αλλά και στα άρθρα ειδήσεων στα οποία ο χρήστης επιλέγει να διαβάσει.

Σκοπός της παρούσας διατριβής είναι η ανάλυση και ανάπτυξη εφαρμογής για τον ιστότοπο athensvoice.gr. Κάνοντας χρήση των cookies, από το πρόγραμμα περιήγησης του κάθε χρήστη, το σύστημα δημιουργεί ένα προφίλ για τον κάθε χρήστη με δεδομένα όπως: ποια άρθρα έχει διαβάσει, ποιες κατηγορίες άρθρων είναι αυτές που τον ενδιαφέρουν περισσότερο, με ποιους άλλους χρήστες έχει κοινά αναγνωστικά ενδιαφέροντα κ.α.

Η εφαρμογή είναι σε θέση να διαχειρίζεται όχι μόνο άρθρα ειδήσεων, αλλά γενικότερα άρθρα από όλες τις κατηγορίες που υπάρχουν στον ιστότοπο. Είναι βασισμένη σε οχτώ κριτήρια, τα οποία έχουν επιλεγεί από την ανάλυση των απαιτήσεων του ιστότοπου, αλλά και από την μελέτη των επιστημονικών άρθρων που αναγνώστηκαν για αυτή τη διατριβή.

Επειδή η εφαρμογή μπήκε σε λειτουργία σε ιστότοπο ο οποίος είναι σε λειτουργία με πολλά εκατομμύρια σελιδοπροβολών (impressions) το μήνα και για λόγους αξιοπιστίας και ταχύτητας της εφαρμογής και του ιστότοπου, έγινε επιλογή λειτουργίας της εφαρμογής online και offline. Στη κατάσταση online εμφανίζει τα άρθρα τα οποία προτείνει το σύστημα κάνοντας υπολογισμούς βάση των οχτώ κριτηρίων, και σε offline λειτουργία κάνει εργασίες όπως αντίγραφα ασφαλείας, διαγραφή δεδομένων που πλέον δεν είναι χρήσιμα για το σύστημα, αλλά και υπολογισμούς που έχουν μεγάλες απαιτήσεις σε υπολογιστικούς πόρους, όπως η ομοιότητα συνημιτόνου μεταξύ όλων των άρθρων. Η εμφάνιση των προτάσεων του συστήματος γίνεται εντός των σελίδων

των άρθρων σε σημείο που να είναι αντιληπτό από τους χρήστες αλλά και να μην επηρεάσει τη δομή του ιστότοπου.

Τα αναμενόμενα αποτελέσματα της διατριβής, θα δείξουν κατά πόσο η αλληλεπίδραση των χρηστών του ιστότοπου (athensvoice.gr) με το σύστημα θα έχει σαν αποτέλεσμα την αύξηση των σελιδοπροβολών (*impressions*) των άρθρων που επιλέγονται να εμφανιστούν από το σύστημα, αλλά και γενικότερα την συνολική αύξηση των ποιοτικών μετρήσεων, όπως είναι ο χρόνος που παραμένουν στον ιστότοπο, ο αριθμός των σελίδων που βλέπει ένας χρήστης κ.α.

Επιπλέον τα αποτελέσματα της διατριβής δείχνουν τον επηρεασμό του χρήστη από την συλλογικότητα (εάν οι χρήστες διάβασαν άρθρα από κατηγορία που δεν διαβάζουν, επηρεασμένοι από τη προτροπή του συστήματος, βάσει επιλογών άλλων χρηστών).

Στη διατριβή αυτή, το σύστημα που σχεδιάστηκε δεν διαχειρίζεται μόνο άρθρα ειδήσεων, όπως γίνεται με πολλές υλοποιήσεις συστημάτων προτάσεων άρθρων, αλλά άρθρα ειδήσεων και άρθρα ποικίλης ύλης. Βέβαια τα άρθρα των ειδήσεων αντιμετωπίζονται με διαφορετικό τρόπο σε σχέση με τα υπόλοιπα, λόγω της ιδιαιτερότητας που έχουν στη χρονική τους διάρκεια (μία είδηση δεν έχει ενδιαφέρον μετά από μια εβδομάδα, ενώ ένα άρθρο γνώμης μπορεί να αναγνωστεί μετά από ένα μήνα).

Στο τρίτο κεφάλαιο της παρούσας διατριβής, θα αναφερθούμε σε σχετικές εργασίες που έχουν πραγματοποιηθεί στο παρελθόν, σχετικά με το ποια συστήματα συστάσεων έχουν προταθεί στη βιβλιογραφία, ώστε να επιτυγχάνεται η σύσταση και το φιλτράρισμα περιεχομένου και βάσει ποιων παραμέτρων έγινε η επιλογή των κριτηρίων του συστήματος, που σχεδιάστηκε ώστε να υπάρξει ένα σωστό αποτέλεσμα.

Το τέταρτο κεφάλαιο αναφέρεται στην ανάπτυξη του συστήματος που σχεδιάστηκε. Είναι βασισμένο σε οχτώ κριτήρια τα οποία επιλέγηκαν έπειτα από ανάλυση των απαιτήσεων του ιστότοπου (athensvoice.gr) αλλά και των τεχνικών που υπάρχουν και έχουν δημοσιευθεί σε επιστημονικές εργασίες, που μελέτησαν συστήματα συστάσεων. Το σύστημα έχει χαρακτηριστικά λειτουργίας online και offline, παράγει προτάσεις προς τους χρήστες online και τις εμφανίζει σε συγκεκριμένο σημείο στον ιστότοπο, εντός των σελίδων όπου εμφανίζονται τα άρθρα. Σε λειτουργία online, το σύστημα έχει να κάνει με τον υπολογισμό των αποτελεσμάτων για τον κάθε χρήστη ξεχωριστά. Παράδειγμα είναι η αφαίρεση από τα αποτελέσματα του συγκεκριμένου άρθρου που

διαβάζεται τη δεδομένη στιγμή από το χρήστη, αλλά και τους υπολογισμούς που απαιτούνται για τα οκτώ κριτήρια του συστήματος. Στη λειτουργία offline γίνονται προγραμματισμένες διεργασίες (cron jobs) σε ώρες που δεν υπάρχει μεγάλη ανάγκη πόρων από το σύστημα για να προβαίνει σε συστάσεις (μεταξύ 02:00 και 05:00 το πρωί), οι διεργασίες αυτές έχουν μεγάλες απαιτήσεις σε υπολογιστικούς πόρους, επεξεργαστικής ισχύος και μνήμης. Παράδειγμα τέτοιων διεργασιών είναι ο υπολογισμός της ομοιότητας συνημιτόνου μεταξύ των άρθρων που υπολογίζεται η ομοιότητα των νέων άρθρων με όλα τα άρθρα που είναι καταχωρημένα στη βάση δεδομένων του ιστότοπου, την ομοιότητα μεταξύ των χρηστών που αλληλεπιδρούν με το σύστημα, το backup της βάσης και διαγραφή των δεδομένων από τον αρχικό πίνακα που αποθηκεύονται τα δεδομένα, αφού αυτά πρώτα υπολογιστούν και μεταφερθούν στους πίνακες όπου είναι απαραίτητα.

Οι προτάσεις που γίνονται από το σύστημα, όπως προαναφέρθηκε, βασίζονται σε οκτώ κριτήρια. Πέντε από τα κριτήρια είναι βασισμένα στο περιεχόμενο (content based) και τα υπόλοιπα τρία είναι βασισμένα στο συνεργατικό φιλτράρισμα (collaborative filtering). Σε ένα από τα κριτήρια συνεργατικού φιλτραρίσματος γίνεται χρήση του κοινωνικού δικτύου Twitter επιτυγχάνοντας μεγαλύτερη αλληλεπίδραση των χρηστών του ιστότοπου.

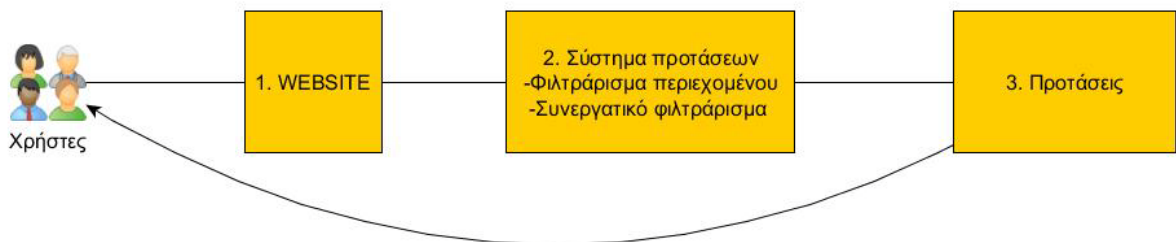
Στο πέμπτο κεφάλαιο παρουσιάζονται τα αποτελέσματα χρήσης του συστήματος. Εάν το σύστημα πρόσφερε θετικά στα στατιστικά του ιστότοπου, αυτό θα γίνει με σύγκριση των στατιστικών προηγούμενης περιόδου με τη περίοδο που θα είναι σε λειτουργία το σύστημα μέσω του Google Analytics, αλλά και με δεδομένα του ίδιου του συστήματος, όπως πόσοι χρήστες επέλεξαν να πατήσουν τις συστάσεις που έγιναν από το σύστημα.

Στο έκτο κεφάλαιο παρουσιάζονται οι μελλοντικές προσθήκες που θα γίνουν στο σύστημα, αλλά και προσθήκες οι οποίες είναι σε φάση ανάπτυξης, όπως είναι η είσοδος στο σύστημα με τη χρήση του κοινωνικού δικτύου Facebook, που πλέον δίνει στο σύστημα τη δυνατότητα να κάνει χρήση πολλών πληροφοριών για τους χρήστες, ώστε να υπολογίζει και να προτείνει στους χρήστες άρθρα κάνοντας χρήση περισσότερων κριτηρίων.

Κεφάλαιο 2

Θεωρητικό Υπόβαθρο

Τα συστήματα συστάσεων (Recommender systems) είναι εφαρμογές οι οποίες αξιοποιούνται σε ιστοσελίδες ειδησεογραφικού περιεχομένου, ηλεκτρονικού εμπορίου και γενικότερα σε ιστοσελίδες οι οποίες, φιλτράροντας το περιεχόμενό τους, διευκολύνουν τους χρήστες να εντοπίζουν ευκολότερα αυτό που τους ενδιαφέρει. Ένα σύστημα συστάσεων μπορεί να χρησιμοποιηθεί για να διευκολύνει τους χρήστες των ιστοσελίδων, που κάνουν χρήση τέτοιων εφαρμογών, στο φιλτράρισμα των περιττών πληροφοριών ή προβλέποντας τα αντικείμενα που ενδιαφέρουν τους χρήστες σύμφωνα με τα ενδιαφέροντά τους. Σε ένα σύστημα συστάσεων οι χρήστες αλληλεπιδρούν με το σύστημα, το σύστημα αποθηκεύει τις απαραίτητες πληροφορίες για τον κάθε ένα χρήστη ξεχωριστά και στο τέλος κάνει τις προτάσεις του προς τους χρήστες. Μια γενική προσέγγιση ενός συστήματος προτάσεων απεικονίζεται στο παρακάτω σχήμα:

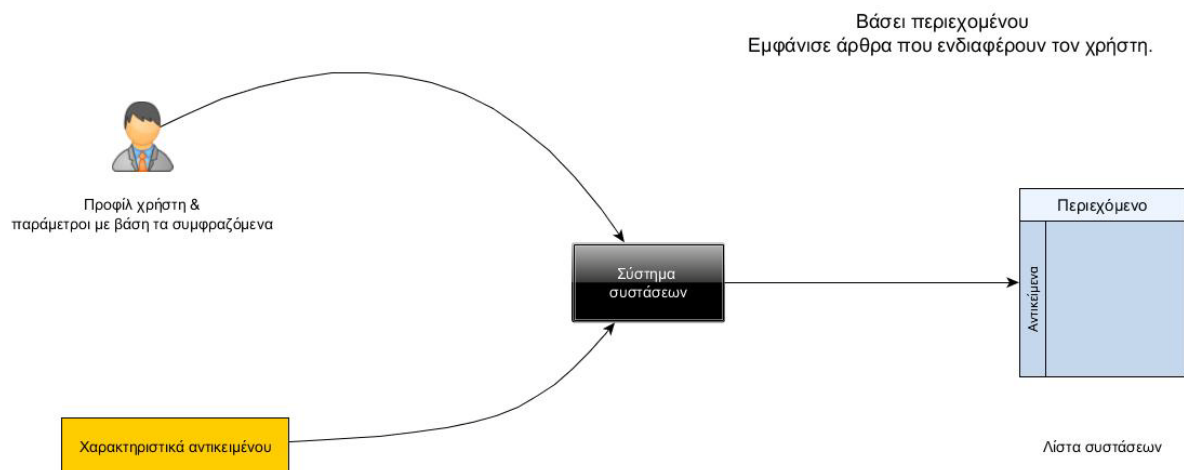


Εικόνα 1. Σύστημα προτάσεων περιεχομένου.

Ανάλογα με τον τρόπο λειτουργίας τους τα συστήματα συστάσεων διακρίνονται στις παρακάτω τρεις κατηγορίες:

2.1 Φιλτράρισμα βάσει περιεχομένου (Content Based Filtering)

Η μέθοδος δημιουργεί συστάσεις βάσει του περιεχομένου. Για παράδειγμα σε ένα σύστημα συστάσεων βασισμένο στο περιεχόμενο, πρωταρχικό ρόλο παίζει η κατηγορία του άρθρου. Την κατηγορία του άρθρου το σύστημα μπορεί να την ξέρει από τη στιγμή που ένα άρθρο εισέρχεται στο σύστημα ή με ανάλυση περιεχομένου, ώστε το σύστημα να αποφασίσει σε ποια κατηγορία ανήκει το άρθρο. Το σύστημα μαθαίνει τις προτιμήσεις του χρήστη, με βάση τα άρθρα που έχει διαβάσει ή έχει βαθμολογήσει στο παρελθόν, αν φυσικά το σύστημα είναι βασισμένο στη βαθμολόγηση των άρθρων.



Εικόνα 2. Φιλτράρισμα βάσει περιεχομένου.

2.1.1 Τα πλεονεκτήματα ενός συστήματος βασισμένο στο περιεχόμενο.

Ανεξαρτησία του χρήστη. Τα συστήματα φιλτραρίσματος βάσει περιεχομένου αξιοποιούν αποκλειστικά μετρήσεις που προέρχονται από τον ενεργό χρήστη. Αντ' αυτού, το συνεργατικό φιλτράρισμα χρειάζεται αξιολογήσεις από άλλους χρήστες, ώστε να βρει χρήστες που έχουν παρόμοια ενδιαφέροντα.[1]

Διαφάνεια. Μπορεί να δώσει εξηγήσεις για τα συνιστώμενα στοιχεία απαριθμώντας τα χαρακτηριστικά του περιεχομένου που προκάλεσε την σύσταση αυτού του αντικειμένου.

Νέα αντικείμενα (Αντικείμενα τα οποία δεν έχουν βαθμολογηθεί από κανένα χρήστη). Ένα σύστημα φιλτραρίσματος βάσει περιεχομένου είναι σε θέση να συνιστά νέα ή άγνωστα αντικείμενα. Δεν αντιμετωπίζει το πρόβλημα πρώτης βαθμολόγησης.

2.1.2 Τα μειονεκτήματα ενός συστήματος βασισμένου στο περιεχόμενο.

Η ανάλυση περιεχομένου. Είναι σχεδόν ανέφικτο να κατηγοριοποιηθούν άρθρα, τα οποία δεν έχουν κάποιο χαρακτηριστικό ώστε να ανήκουν σε κάποια κατηγορία και περιέχουν ή αποτελούνται από βίντεο ή ήχους (πολυμεσικό υλικό).

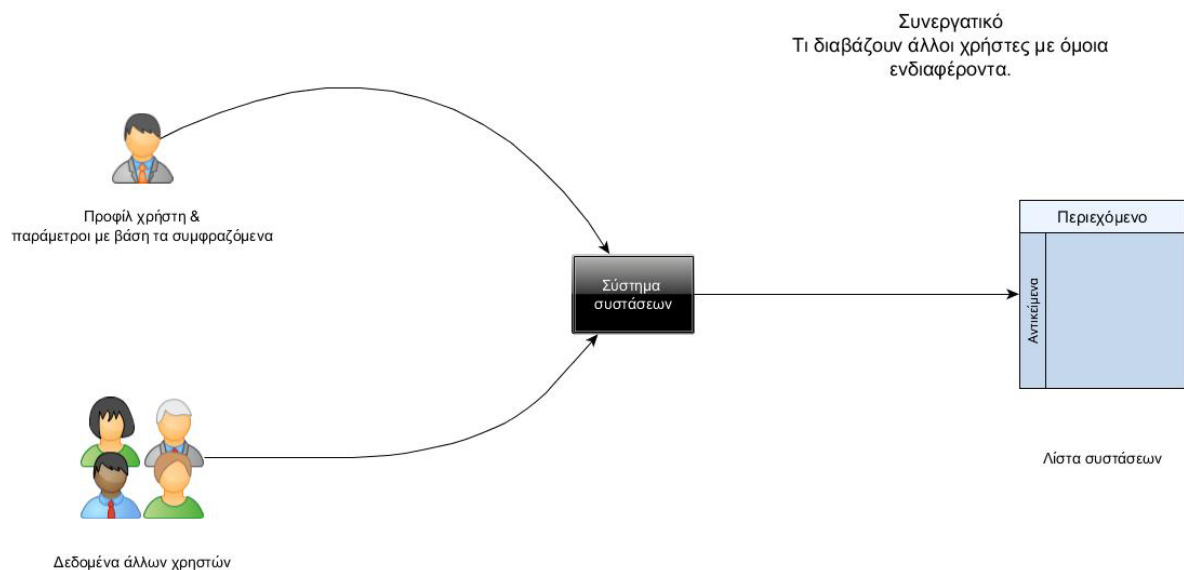
Η μη ποιοτική αξιολόγηση των άρθρων. Για παράδειγμα δυο άρθρα, των οποίων τα χαρακτηριστικά είναι ίδια (ίδια κατηγορία και ίδιος αριθμός εμφανίσεων), το σύστημα τα θεωρεί ίδια, άσχετα αν το ένα είναι ποιοτικότερο από το άλλο.

Το πρόβλημα της υπερεξειδίκευσης. Το σύστημα εμφανίζει άρθρα από κατηγορίες που ενδιαφέρουν τον χρήστη σύμφωνα με το ιστορικό του. Δεν θα εμφανίσει όμως ένα άρθρο το οποίο έχει διαβαστεί από πολλούς χρήστες και έχει πάρει καλές κριτικές, γιατί δεν ανήκει στις κατηγορίες που ενδιαφέρουν τον χρήστη.

Το πρόβλημα Cold-start. Τα συστήματα βασισμένα στο περιεχόμενο δεν μπορούν να αντιμετωπίσουν το πρόβλημα αργής εκκίνησης για κάθε νέο χρήστη. Το σύστημα δεν δύναται να καταλάβει τις προτιμήσεις του χρήστη για να προβλέψει τις κατηγορίες άρθρων που τον ενδιαφέρουν. Θα πρέπει πρώτα να συλλέξει δεδομένα από την συμπεριφορά του ώστε να είναι σε θέση να κάνει σωστές προβλέψεις.

2.2 Συνεργατικό φιλτράρισμα (Collaborative Filtering)

Δημιουργούν συστάσεις με κάποιο μέτρο ομοιότητας μεταξύ των χρηστών. Δηλαδή προϊόντα, άρθρα, υπηρεσίες που διαβάστηκαν ή επιλέχθηκαν από άλλους χρήστες με όμοια ενδιαφέροντα. Σε αυτή την κατηγορία ανήκουν τα User-user Collaborative Filtering και Item-item Collaborative Filtering συστήματα τα οποία χρησιμοποιούν την ομοιότητα μεταξύ των αντικειμένων (και χρηστών αντίστοιχα). Το User-user Collaborative Filtering, που είναι γνωστό και ως k-NN CF, ήταν η πρώτη μέθοδος και παρουσιάστηκε πρώτη φορά στο Group Lens Usenet [23].



Εικόνα 3. Συνεργατικό φιλτράρισμα.

Παράδειγμα της μεθόδου User-user CF είναι η εύρεση της βαθμολογίας που θα βάλει ένας χρήστης σε μία ταινία, σε κλίμακα 0 - 5

	Movie M1	Movie M2	Movie M3	Movie M4
User A	4	?	3	5
User B	?	5	4	?
User C	5	4	2	?

User D	2	4	?	3
User E	3	4	5	?

Πίνακας 1: Βαθμολογίες σε ταινίες από χρήστες.

Στον παραπάνω πίνακα έχουμε 5 χρήστες με τις αξιολογήσεις για 4 ταινίες και θέλουμε να βρούμε το βαθμό που θα δώσει ο χρήστης User C για την ταινία Movie M4.

Θα πρέπει να μετρήσουμε την ομοιότητα των χρηστών χρησιμοποιώντας τη συσχέτιση του Pearson από τον παρακάτω τύπο.

$$sim(a, b) = \frac{\sum_{p \in p} (r_{a,p} - \bar{r}_a)(r_{b,p} - \bar{r}_b)}{\sqrt{\sum_{p \in p} (r_{a,p} - \bar{r}_a)^2} \sqrt{\sum_{p \in p} (r_{b,p} - \bar{r}_b)^2}}$$

Εξίσωση 1. Συσχέτιση Pearson.

Για τον υπολογισμό του συντελεστή Pearson για δύο μεταβλητές r_a και r_b χρησιμοποιείται η έννοια της συνδιακύμανσης. Η συνδιακύμανση ορίζεται ως το γινόμενο των αποκλίσεων της τιμής της μεταβλητής r_a από τον αριθμητικό μέσο της και της τιμής της μεταβλητής r_b από τον αριθμητικό μέσο της για όλες τις παρατηρήσεις. Προκειμένου να οριστεί ένα μέτρο συσχέτισης ανεξάρτητο μονάδων μέτρησης, η συνδιακύμανση διαιρείται με το γινόμενο των τυπικών αποκλίσεων $\sqrt{\sum_{p \in p} (r_{a,p} - \bar{r}_a)^2} \sqrt{\sum_{p \in p} (r_{b,p} - \bar{r}_b)^2}$.

Έπειτα θα πρέπει να βρούμε πόσες γειτονιές θα πρέπει να εξετάσουμε. Δηλαδή πόσοι χρήστες έχουν βαθμολογήσει την συγκεκριμένη ταινία. Και στο τέλος να κάνουμε την πρόβλεψη με τον ακόλουθο τύπο.

$$pred(a, p) = \bar{r}_a + \frac{\sum_{b \in N} sim(a, b)(r_{b,p} - \bar{r}_b)}{\sum_{b \in N} sim(a, b)}$$

Εξίσωση 2. Εύρεση πρόβλεψης.

Στο παράδειγμα για neighborhood = 2 θα έχουμε:

Το μέσο όρο βαθμολογιών σε ταινίες για τον User 3 = 3.667. Την ταινία Movie M4 την έχουν αξιολογήσει 2 χρήστες οι User A και User D. Η ομοιότητα μεταξύ των χρηστών $s(\text{User C}, \text{User A}) = 0.832$ και $s(\text{User C}, \text{User D}) = -0.515$. Η πρόβλεψη για τη βαθμολογία της ταινία θα είναι $pred(C, e) = 4.667$

Στο παράδειγμα, εάν θέλουμε να βρούμε την βαθμολογία της ταινίας Movie M4 από τον χρήστη User C χρησιμοποιώντας τη μέθοδο Item-item CF με την ομοιότητα συνημιτόνου $\cos(d1,d2) = d1*d2/||d1||*||d2||$ θα έχουμε σαν πρόβλεψη βαθμολογίας για τη ταινία $\text{pred}(C,e) = 3.84$

2.2.1 Προβλήματα που παρουσιάζονται στα συστήματα συνεργατικού φιλτραρίσματος.

Το πρόβλημα των νέων χρηστών. Το σύστημα για να προβεί σε προβλέψεις χρειάζεται δεδομένα για τις προτιμήσεις του χρήστη, ώστε να είναι σε θέση να βρει την ομοιότητα μεταξύ άλλων χρηστών του συστήματος.

Το πρόβλημα των αντικειμένων. Ένα αντικείμενο για να συσταθεί θα πρέπει με κάποιο τρόπο να βαθμολογηθεί από τους χρήστες του συστήματος, ώστε να υπάρχουν επαρκή δεδομένα και το αντικείμενο να μπορεί να ενταχθεί στις συστάσεις του συστήματος.

2.3 Υβριδικές συστάσεις(Hybrid recommendations)

Αυτή η μέθοδος συνδυάζει όλες τις μεθόδους, το συνεργατικό φιλτράρισμα και το φιλτράρισμα βασισμένο στο περιεχόμενο, αλλά και άλλες οι οποίες απλώς θα αναφερθούν επιγραμματικά και είναι οι ακόλουθες.

Μη εξατομικευμένη προσέγγιση,

Συνεργατικό φιλτράρισμα με βάση τη μνήμη (memory-based CF),

Συνεργατικό φιλτράρισμα με βάση το μοντέλο (model-based CF),

Φιλτράρισμα βασισμένο στη γνώση (knowledge-based filtering),

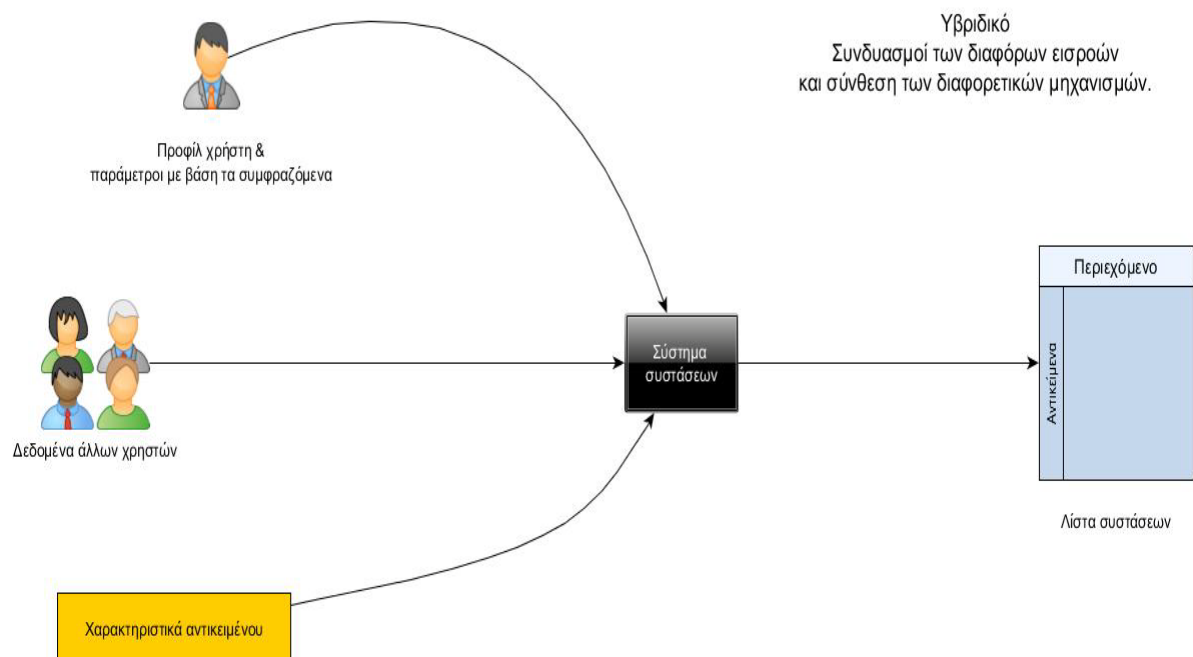
Φιλτράρισμα βασισμένο στα δημογραφικά στοιχεία (demography-based filtering).

Το υβριδικό φιλτράρισμα χωρίζεται σε 7 κατηγορίες. [3] και wikipedia¹

¹ http://en.wikipedia.org/wiki/Recommender_system

1. Σταθμισμένη υβριδοποίηση (Weighted hybrid)
2. Εναλλαγή υβριδοποίησης (Switching hybrid)
3. Μικτή υβριδοποίηση (Mixed hybrid)
4. Συνδυασμός χαρακτηριστικών (Feature combination hybrid)
5. Αύξηση χαρακτηριστικών (Feature augmentation hybrid)
6. Διαδοχική υβριδοποίηση (Cascade hybrid)
7. Υβριδοποίηση μετα-επιπέδου (Meta-level hybrid)

Στα Συστήματα Συστάσεων διακρίνουμε δύο οντότητες, τους χρήστες και τα αντικείμενα. Οι χρήστες είναι αυτοί που ψηφίζουν ή διαβάζουν άρθρα, οπότε εκδηλώνουν το ενδιαφέρον τους για κάποιες κατηγορίες άρθρων, ενώ τα αντικείμενα είναι τα άρθρα τα οποία ανήκουν σε κατηγορίες. Τέτοιες είναι για παράδειγμα η πολιτική, η οικονομία κ.α ή ο συντάκτης του άρθρου, που το σύστημα προσπαθεί να προβλέψει.



Εικόνα 4. Υβριδικό φιλτράρισμα.

2.4 Προκλήσεις και περιορισμοί

Ο επιτυχής σχεδιασμός και η υλοποίηση ενός συστήματος συστάσεων είναι μια διαδικασία που εξαρτάται από πολλούς παράγοντες. Καταρχήν, θα πρέπει να ερευνηθούν μια σειρά από ζητήματα γενικού σκοπού, όπως ερωτήματα σχετικά με τον στόχο του συστήματος, το κοινό στο οποίο θα απευθύνεται, το είδος των προτάσεων που θα κάνει. Έπειτα θα πρέπει να γίνει επιλογή των αλγορίθμων παραγωγής συστάσεων ή δημιουργίας νέων που θα σχετίζονται με τις απαιτήσεις, ο σχεδιασμός της διεπαφής με τον χρήστη, η επιλογή των εργαλείων που θα γίνει η υλοποίηση κ.τ.λ.

Ωστόσο υπάρχουν κάποια ζητήματα τα οποία εμφανίζονται στα περισσότερα συστήματα συστάσεων, ανεξάρτητα από τον τομέα που το κάθε ένα ειδικεύεται. Οι κυριότερες είναι η ανεπάρκεια των αξιολογήσεων, το πρόβλημα της ψυχρής εκκίνησης και η εμπιστοσύνη στις παραγόμενες συστάσεις.

Ανεπάρκεια των αξιολογήσεων: Οι περισσότεροι χρήστες, σε περιβάλλοντα όπου υπάρχει η θετική ή αρνητική ψηφοφορία, πολλές φορές δεν αξιολογούν ένα άρθρο ή ένα αντικείμενο με αποτέλεσμα η μήτρα (matrix) - πίνακας με τις αξιολογήσεις των χρηστών - να είναι αραιή. Αυτό το πρόβλημα μειώνει την πιθανότητα εύρεσης χρηστών με κοινά ενδιαφέροντα. Το ζήτημα αυτό μπορεί να μετριαστεί χρησιμοποιώντας πρόσθετες πληροφορίες ή κάνοντας υποθέσεις βάσει της περιοχής (κατηγορίας) που βρίσκεται ο χρήστης.

Πρόβλημα ψυχρής εκκίνησης (The Cold-start Problem): Όταν νέοι χρήστες ή αντικείμενα εισέλθουν στο σύστημα, δεν υπάρχουν επαρκή δεδομένα, ώστε να υπάρχει σύσταση για τους χρήστες αυτούς ή τα αντικείμενα. Το πρόβλημα αυτό δεν ισχύει μόνο για τα νέα στοιχεία αλλά και για ασαφή στοιχεία. Τέτοιο πρόβλημα μπορεί να είναι μια νέα κατηγορία άρθρων, όπου δεν έχει διαβαστεί από χρήστες ή δεν έχει πολλά άρθρα ώστε να εξαχθούν τα ανάλογα αποτελέσματα. [3]

Εμπιστοσύνη στις παραγόμενες συστάσεις: Το ζήτημα της εμπιστοσύνης στις παραγόμενες συστάσεις εξετάζεται από δύο πλευρές [22]. Η πρώτη αφορά την εμπιστοσύνη του ίδιου του συστήματος στον εαυτό του, δηλαδή στο κατά πόσον οι συστάσεις που παράγει το ίδιο είναι σύμφωνες με τις προτιμήσεις των χρηστών. Η δεύτερη σχετίζεται με τις προσδοκίες των ίδιων των χρηστών από το σύστημα, κατά πόσον δηλαδή οι ίδιοι οι χρήστες θεωρούν ότι έχουν να ωφεληθούν από την αλληλεπίδρασή τους με αυτό. Για παράδειγμα, είναι γνωστό ότι η απόδοση των

συνεργατικών συστημάτων αυξάνει όσο περισσότερα αντικείμενα βαθμολογούν οι χρήστες. Άρα είναι λογικό το σύστημα να έχει περισσότερη εμπιστοσύνη στις συστάσεις που παράγει, όσο αυξάνεται ο όγκος των δεδομένων που επεξεργάζεται. Έχει σημασία το σύστημα να ενημερώνει τον χρήστη για την εμπιστοσύνη των παραγόμενων συστάσεων. Για παράδειγμα, αν το σύστημα προτείνει ένα αντικείμενο με χαμηλό επίπεδο εμπιστοσύνης, τότε το γεγονός αυτό λειτουργεί ως προτροπή προς το χρήστη να ερευνήσει το αντικείμενο περισσότερο προτού λάβει κάποια απόφαση. Έχει παρατηρηθεί ότι η εμπιστοσύνη ενός χρήστη προς το σύστημα αυξάνει αν το τελευταίο δικαιολογεί τις συστάσεις που παράγει. Οι πιο συνηθισμένοι τρόποι περιγραφής της εμπιστοσύνης έχουν τις ρίζες τους στην στατιστική συμπερασματολογία.

Κεφάλαιο 3

Σχετικές εργασίες

Η διατριβή αυτή ασχολείται με την υβριδική προσέγγιση του φιλτραρίσματος περιεχομένου, με χρήση πολλαπλών κριτηρίων επιλογής αποτελεσμάτων. Οι περισσότερες εργασίες που έχουν γίνει σε αυτό το πεδίο έχουν ασχοληθεί με προτάσεις άρθρων τα οποία ανήκουν στην κατηγορία των ειδήσεων. Η συγκεκριμένη διατριβή θα ασχοληθεί όχι μόνο με άρθρα ειδησεογραφικού ενδιαφέροντος αλλά και με άρθρα γενικού ενδιαφέροντος, που οι κατηγορίες ανήκουν στον ιστότοπο athensvoice.gr.

3.1 Φιλτράρισμα βασισμένο στο περιεχόμενο

Σχετικές εργασίες που ασχολήθηκαν με το συνεργατικό φιλτράρισμα είναι βασισμένες στην τεχνική της παραγοντοποίησης πινάκων[24][25]. Οι τεχνικές συνεργατικού φιλτραρίσματος συνήθως υιοθετούνται όταν σπανίζουν τα ιστορικά δεδομένα για την εκπαίδευση του συστήματος. Για την εξατομίκευση των ειδήσεων τα πιο κοινά είναι τα Content-based recommendation έχοντας διαφορετικές προσεγγίσεις. Το Your News [4] είναι ένα σύστημα το οποίο αντλεί το περιεχόμενο του από 8 διαφορετικές πηγές σε μορφή RSS. Το Your News παρατηρεί τη συμπεριφορά των χρηστών του συστήματος και κατασκευάζει για τον κάθε ένα χρήστη ένα μοντέλο, το οποίο αντιπροσωπεύει τα ενδιαφέροντα του χρήστη. Κάνοντας χρήση του μοντέλου αυτού προτείνει στους χρήστες σχετικά άρθρα, έχοντας πρώτα κατηγοριοποιήσει τα άρθρα σε οχτώ θεματικές ενότητες.

Οι θεματικές είναι διαχωρισμένες σε Εθνικά, Κόσμος, Οικονομία κλπ, ώστε να μην συγχέει τα ενδιαφέροντα των χρηστών. Το προφίλ του κάθε χρήστη για την κάθε κατηγορία αντιπροσωπεύεται από ένα βάρος (βαθμολογία) σύμφωνα με το ιστορικό ανάγνωσης του κάθε χρήστη. Η διαδικασία για την εξαγωγή των προτάσεων λαμβάνει υπ όψη το μοντέλο που έχει δημιουργηθεί για τον χρήστη, έτσι ώστε, κάθε ένα νέο άρθρο που εισέρχεται στο σύστημα να είναι γνωστό σε ποιόν θα εμφανιστεί. Η βαθμολογία της ομοιότητας μεταξύ του χρήστη και του άρθρου υπολογίζεται με την χρήση συντελεστών συνημιτόνου σε μια κλίμακα 0 έως 1, έτσι τα άρθρα μπορούν να

ταξινομηθούν σύμφωνα με την ομοιότητα που έχουν με το μοντέλο του χρήστη. Το σύστημα δημιουργεί για τον κάθε χρήστη δύο μοντέλα, ένα βραχυπρόθεσμο και ένα μακροπρόθεσμο. Το βραχυπρόθεσμο προφίλ λαμβάνει υπ όψιν τα τελευταία 20 άρθρα που διαβάστηκαν από τον χρήστη, ενώ το μακροπρόθεσμο λαμβάνει υπ όψιν όλα τα άρθρα που έχουν διαβαστεί από τον χρήστη.

Οι Daniel Billsus and Michael J. Pazzani [5] ανέπτυξαν ένα σύστημα πράκτορα για άρθρα ειδησεογραφικού ενδιαφέροντος, το οποίο μεταβάλεται σύμφωνα με την συμπεριφορά του χρήστη. Κι εδώ όπως και στο [4] δημιουργεί δύο προφίλ (μοντέλα) για τον κάθε χρήστη, ένα βραχυπρόθεσμο και ένα μακροπρόθεσμο. Για την μοντελοποίηση του βραχυπρόθεσμου προφίλ του χρήστη, κάνει χρήση του αλγορίθμου Nearest Neighbor με εύρεση ομοιότητας μεταξύ των κειμένων. Πρώτα μετατρέπει το κείμενο σε μορφή TF-IDF (term-frequency / inverse-document-frequency), και έπειτα, κάνοντας χρήση της ομοιότητα συνημιτόνου, υπολογίζει την συνάφεια των δύο άρθρων. Για την μοντελοποίηση του μακροπρόθεσμου προφίλ των χρηστών γίνεται χρήση του κατηγοριοποιητή Naive Bayes για να διαμορφώσει τις γενικές προτιμήσεις του χρήστη, για τις ειδήσεις που δεν μπορούν να κατηγοριοποιηθούν με το βραχυπρόθεσμο μοντέλο.

Μια άλλη προσέγγιση φιλτραρίσματος βάσει περιεχομένου, είναι προτάσεις ειδήσεων βασισμένες σε οντολογίες [6]. Το σύστημα Athena είναι μια επέκταση του προηγούμενου συστήματος Hermes, που έκανε χρήση μιας οντολογίας για να αποθηκεύσει έννοιες και τις σχέσεις τους στα άρθρα των ειδήσεων. Παρατηρεί τη συμπεριφορά των χρηστών και δημιουργεί προτάσεις βασισμένες στη συμπεριφορά αυτή. Το σύστημα χρησιμοποιεί ένα κλασικό βάσει λεξιλογίου σύστημα συστάσεων και αλγόριθμους συστάσεων συμασιολογίας για τη σύγκριση των άρθρων που δεν έχουν διαβαστεί από τον χρήστη και το προφίλ του χρήστη που έχει δημιουργήσει το σύστημα.

3.2 Συνεργατικό φιλτράρισμα

Κάποιες εργασίες έχουν ασχοληθεί με τις προτάσεις που μπορούν να γίνουν σε μηχανές αναζήτησης, βάση της ομοιότητας των ερωτημάτων που γίνονται από τους χρήστες [14].

Το MovieLens [21] χρησιμοποιεί τεχνολογία συνεργατικού φιλτραρίσματος [18] και [19] για τη δημιουργία συστάσεων σε ταινίες. Λειτουργεί ομαδοποιώντας τους χρήστες με κοινά ενδιαφέροντα σε ταινίες. Κάθε χρήστης του συστήματος ανήκει σε μια «γειτονιά» μαζί με άλλους χρήστες με παρόμοια ενδιαφέροντα. Οι βαθμολογίες από τους γείτονες χρησιμοποιούνται για να δημιουργήσουν εξατομικευμένες συστάσεις για τον χρήστη [20].

Το MovieLens Unplugged είναι μια εφαρμογή συστάσεων που αφορά ταινίες για φορητές συσκευές και πιο συγκεκριμένα για PDA τα οποία συνδέονται περιστασιακά στο διαδίκτυο. Στόχος της εφαρμογής ήταν, η μελέτη των διεπαφών προκειμένου οι συσκευές να γίνουν όσο το δυνατό πιο φιλικές. Για το πείραμα της εφαρμογής εξετάστηκαν δύο σενάρια, το πρώτο αφορούσε σε ταινίες που παίζονται στους κινηματογράφους και το δεύτερο αφορούσε στις ταινίες που υπάρχουν σε DVD, που ο χρήστης επιθυμούσε να ενοικιάσει.

Τα συμπεράσματα της εν λόγω έρευνας έδειξαν ότι οι χρήστες έμειναν ικανοποιημένοι από το μηχανισμό συστάσεων καθώς είχε θετικά αποτελέσματα.

3.3 Υβριδικό φιλτράρισμα

Προσέγγιση που υλοποιείται με υβριδικό φιλτράρισμα (Content-based και Collaborative) είναι η [7]. Η συγκεκριμένη προσέγγιση κάνει χρήση δεδομένων χρηστών και άρθρων από τον δικτυακό τόπο yahoo. Βασίζεται στην ανταπόκριση των χρηστών (όπως τα clicks που κάνουν στα άρθρα). Κάθε χρήστης αρχικά εκπροσωπείται από ένα διάνυσμα με περισσότερα από 1000 κατηγορηματικά χαρακτηριστικά, τα οποία περιλαμβάνουν δημογραφικά στοιχεία, γεωγραφικά χαρακτηριστικά και κατηγορίες συμπεριφοράς. Σε κάθε επαναφόρτιση της σελίδας ο χρήστης δεν βλέπει τα ίδια άρθρα που του εμφανίστηκαν προηγουμένως. Για την κατηγοριοποίηση των άρθρων χρησιμοποιούνται 100 κατηγορίες οι οποίες περιέχουν τις κατηγορίες των url και τις κατηγορίες των συντακτών.

Η εργασία για τη σύσταση εξατομικευμένων προτάσεων, οι οποίες βασίζονται στη συμπεριφορά των clicks που κάνουν οι χρήστες [8], δίνει βάση στη συμπεριφορά του χρήστη και στο πού επιλέγει να κάνει click, με σκοπό να δημιουργήσει ένα προφίλ για αυτόν. Προτείνει ένα Bayesian μοντέλο για την σύσταση των ειδήσεων στον χρήστη, βασισμένο στα ενδιαφέροντα του χρήστη αλλά και στα trends των άλλων χρηστών. Οι [10] προτείνουν μια προσέγγιση για τη δημιουργία του προφίλ του χρήστη από την συμπεριφορά του στην περιήγηση, λαμβάνοντας υπόψιν το χρόνο που δαπανάει για την ανάγνωση των άρθρων, αλλά και την συμπεριφορά του χρήστη. Το 2012 διεξήχθη μια online αξιολόγηση [11] από διάφορα συστήματα συστάσεων άρθρων ειδήσεων στον ιστότοπο forbes.com με logged-in χρήστες. Αναφέρουν ότι με τη χρήση ενός υβριδικού συστήματος υπήρχε βελτίωση 37% σε σχέση με μεθόδους που βασίζονται στη δημοτικότητα ενός άρθρου.

Η εργασία για την εξατομίκευση περιεχομένου της υπηρεσίας Google news [9] χρησιμοποιεί έναν αλγόριθμο ο οποίος είναι μια μίξη από Memory-based αλγόριθμο και Model-based αλγόριθμο. Ο Memory-based αλγόριθμος, κάνει προβλέψεις χρησιμοποιώντας τις προηγούμενες αξιολογήσεις που έχουν γίνει από τους χρήστες. Τυπικά η πρόβλεψη υπολογίζεται ως ο μέσος όρος των αξιολογήσεων που γίνονται από άλλους χρήστες, όπου το βάρος είναι ανάλογο με την ομοιότητα μεταξύ των χρηστών. Κοινά μέτρα ομοιότητας που περιλαμβάνονται είναι ο συντελεστής Pearson και η ομοιότητα συνημιτόνου. Ο κατά ζεύγη πίνακας ομοιότητας μεταξύ των χρηστών υπολογίζεται offline. Με τον αλγόριθμο Model-based μοντελοποιεί τους χρήστες με

βάση τις προηγούμενες αξιολογήσεις τους. Το μοντέλο αυτό το χρησιμοποιεί για να προβλέψει την αξιολόγηση άρθρων τα οποία δεν έχουν εμφανιστεί. Τα αποτελέσματα των προτάσεων εμφανίζονται σε χρήστες οι οποίοι είναι login με το Google.

Μια άλλη εργασία για το συνεργατικό φιλτράρισμα με βάση τον χρήστη και τα αντικείμενα [12] κάνει χρήση για τον υπολογισμό της τελικής αξιολόγησης, από τρεις πηγές:

- προβλέψεις οι οποίες βασίζονται στις αξιολογήσεις που έχουν γίνει από άλλους χρήστες στο ίδιο αντικείμενο (άρθρο)
- προβλέψεις βασισμένες στις αξιολογήσεις που έχουν γίνει σε διαφορετικά άρθρα, από τον ίδιο χρήστη
- αξιολογήσεις βασισμένες σε δεδομένα από άλλους χρήστες, με μεγάλη ομοιότητα μεταξύ τους, οι οποίοι έχουν αξιολογήσει άλλα άρθρα όμοια μεταξύ τους.

Άλλη εργασία, με αντικείμενο της το συνεργατικό φιλτράρισμα ειδήσεων [13], προτείνει μια προσέγγιση με βάση το φιλτράρισμα περιεχομένου και το συνεργατικό φιλτράρισμα, ώστε να κάνει συστάσεις σε άρθρα ειδήσεων στον ιστότοπο bing.com. Συνδυάζει τις δύο ομοιότητες με ένα μοντέλο γειτονίας. Από το πεδίο των ειδήσεων παίρνει την ομοιότητα μεταξύ των αντικειμένων. Στη φάση της βελτιστοποίησης, συγκεντρώνει τους συνδέσμους των user-items και εξετάζει την ομοιότητα τους μεταξύ των αντικειμένων. Για το φιλτράρισμα βάσει περιεχομένου υπολογίζει την ομοιότητα μεταξύ δύο άρθρων χρησιμοποιώντας το κείμενο των άρθρων και εξετάζει τις σημαντικές λέξεις και την συχνότητα εμφάνισης των λέξεων.

Συνεισφορά διατριβής. Το σύστημα που αναπτύχθηκε, συνδυάζει χαρακτηριστικά από συνεργατικά συστήματα και συστήματα βασισμένα στο περιεχόμενο. Παρουσιάζει ομοιότητες με άλλες προσεγγίσεις που έχουν γίνει [8][9][12][13] και συνδυάζει χαρακτηριστικά και από τις δύο προσεγγίσεις (βάσει περιεχομένου, συνεργατικό) ώστε να δημιουργεί προτάσεις.

Βέβαια υπάρχει διαφορά στην επιλογή των κριτηρίων για την εξαγωγή των αποτελεσμάτων από το σύστημα. Η τελική επιλογή των κριτηρίων έγινε μετά από μελέτη της βιβλιογραφίας, αλλά και ανάλυση των απαιτήσεων του ιστότοπου, μελέτη των δεδομένων από το Google analytics, αλλά και των δεδομένων που συλλέχτηκαν από το σύστημα κατά την διάρκεια της αποθήκευσης των δεδομένων από τους χρήστες.

Μια προσθήκη της συγκεκριμένης διατριβής, που δεν είδαμε να χρησιμοποιείται σε άλλες εργασίες, είναι η χρήση υποσυστήματος ανάδρασης χρήστη για τον υπολογισμό του αριθμού των άρθρων που θα εμφανίζονται σε κάθε κριτήριο. Αυτή η προσθήκη δίνει στο σύστημα μια προσαρμοστικότητα για κάθε χρήστη, η οποία μεταβάλλεται δυναμικά. Επίσης η επιλογή χρησιμοποίησης του κοινωνικού δικτύου Twitter για την εμφάνιση άρθρων, τα οποία υπάρχουν στη βάση δεδομένων του ιστότοπου athensvoice.gr και τα άρθρα αυτά είναι σχετικά με τα twitter trends στην Ελλάδα τη δεδομένη στιγμή, είναι ακόμα μια προσθήκη η οποία δεν χρησιμοποιείται σε συστήματα συστάσεων άρθρων σε ιστότοπους, με τον τρόπο που γίνεται χρήση στην εφαρμογή που υλοποιήθηκε για την παρούσα διατριβή.

Επίσης το σύστημα που αναπτύχθηκε, βασίζεται σε αληθινούς χρήστες και άρθρα τα οποία ανανεώνονται συνεχώς. Τα δεδομένα που συλλέγονται βασίζονται στη συμπεριφορά των χρηστών κατά τη διάρκεια που περιηγούνται στον ιστότοπο και όχι από κάποιο dataset, όπως γίνεται σε άλλες υλοποιήσεις [15] για την εκπαίδευση του συστήματος και την αποτελεσματικότητα της υλοποίησης. Κάποια άλλα συστήματα [17] απαιτούν από τον χρήστη να δημιουργήσει το προφίλ μόνος του. Μια τέτοια προσέγγιση επιβαρύνει τον χρήστη με την συμπλήρωση των απαραίτητων στοιχείων, κάτι που πολύ λίγοι είναι διατεθειμένοι να πράξουν. Σε αντίθεση το σύστημα το οποίο σχεδιάστηκε, δημιουργεί ένα προφίλ για κάθε χρήστη αυτόματα σύμφωνα με τις επιλογές που κάνει.

Από τη μελέτη της βιβλιογραφίας, παρατηρούμε ότι υπάρχουν κοινά χαρακτηριστικά με διάφορες προσεγγίσεις που έχουν υλοποιηθεί, όπως είναι η ομοιότητα κειμένου, η ομοιότητα μεταξύ των χρηστών, τα πιο διαβασμένα, το προφίλ που δημιουργείται για τον κάθε χρήστη, η χρήση των κοινωνικών δικτύων.

Η διαφορετική προσέγγιση της χρήσης των κοινωνικών δικτύων στη διατριβή αυτή είναι μια από τις διαφοροποιήσεις της από άλλες υλοποιήσεις. Οι προτάσεις που γίνονται από το σύστημα, κάνοντας χρήση του Twitter, σχετίζονται με άρθρα τα οποία είναι καταχωρημένα στη βάση δεδομένων του ιστότοπου athensvoice.gr και ο τίτλος περιέχει τις λέξεις κλειδιά των twitter trends στην Ελλάδα, και όχι με εμφάνιση των αποτελεσμάτων του έρχονται από το twitter. Το συνεργατικό φιλτράρισμα, με βάση τους χρήστες, είναι άλλη μια διαφοροποίηση. Το φίλτρο αυτό εμφανίζει άρθρα που ανήκουν σε κατηγορίες που δεν διαβάζει ο χρήστης Α αλλά τις διαβάζουν άλλοι χρήστες, που έχουν μεγάλη ομοιότητα με τον χρήστη Α. Επίσης μία ακόμα διαφοροποίηση είναι

ο προσαρμοστικός αλγόριθμος που αναπτύχθηκε για τη λύση στο πρόβλημα της επιλογής του αριθμού εμφάνισης των άρθρων του κάθε κριτηρίου.

Κεφάλαιο 4

Ανάπτυξη εφαρμογής

Σε αυτό το κεφάλαιο παρουσιάζουμε, τα κριτήρια επιλογής των αποτελεσμάτων, τον αλγόριθμό που σχεδιάστηκε, τη περιγραφή και αρχιτεκτονική του συστήματος, τη βάση δεδομένων που σχεδιάστηκε, αλλά και γενικότερα την ανάπτυξη της εφαρμογής που υλοποιήθηκε με σκοπό να χρησιμοποιηθεί για τη δεδομένη μεταπτυχιακή διατριβή.

4.1 Κριτήρια επιλογής αποτελεσμάτων

Η επιλογή των κριτηρίων για τη λειτουργία του συστήματος έγινε μετά από ανάλυση των δεδομένων από τα google analytics, όπου μας έδειξε ότι πολλοί χρήστες επισκέπτονται τον ιστότοπο συγκεκριμένες ώρες της ημέρας και αναλόγως την ώρα μεταβάλλονται οι επιλογές που κάνουν για την ανάγνωση άρθρων. Η δεύτερη πηγή δεδομένων προς ανάλυση, ήταν από τη βάση δεδομένων του συστήματος, όταν για το σύστημα είχε σχεδιαστεί μόνο η βάση δεδομένων και ήταν σε κατάσταση αποθήκευσης των δεδομένων από τους χρήστες. Για 15 ημέρες το σύστημα αποθήκευε δεδομένα ώστε να είναι εφικτό να εξαχθούν τα απαραίτητα συμπεράσματα για την επιλογή των κριτηρίων, αλλά και να υπάρχει επαρκής πληροφορία για την επιλογή των προσωποποιημένων προτιμήσεων των χρηστών και γενικότερα τι επιλέγουν να διαβάσουν. Τα κριτήρια που επιλέχθηκαν μετά την ανάλυση της συμπεριφοράς των χρηστών στον ιστότοπο είναι τα ακόλουθα:

1. Κριτήριο προφίλ χρηστών. Το 1ο κριτήριο παραγωγής συστάσεων του συστήματος είναι το κριτήριο προφίλ χρήστη (user profile). Το user profile είναι το προφίλ που έχει δημιουργηθεί για κάθε χρήστη κατά την περιήγηση του στον ιστότοπο του athenvoice.gr. Καταγράφεται το cookieID του χρήστη, η κατηγορία των άρθρων που έχει επιλέξει να διαβάσει και πόσες φορές επέλεξε να διαβάσει άρθρα από την κάθε κατηγορία.

Στην αρχή επιλέγονται οι κατηγορίες που διαβάζει ο χρήστης, κατά μέγιστο σε 6 ημέρες, ταξινομημένες με φθίνουσα σειρά (από την κατηγορία με τις περισσότερες εμφανίσεις

προς τη κατηγορία με τις λιγότερες εμφανίσεις). Στα αποτελέσματα του κριτηρίου πραγματοποιείται σύγκριση του πρώτου αποτελέσματος με κάθε ένα από τα υπόλοιπα. Αν η διαφορά μεταξύ τους είναι μεγαλύτερη του 50% τότε το σύστημα αφήνει εκτός αυτά τα αποτελέσματα. Δηλαδή αν το πρώτο αποτέλεσμα έχει τη τιμή 70 και το όγδοο έχει τη τιμή 34, τότε το όγδοο αποτέλεσμα μένει εκτός γιατί το 50% του 70 είναι ο αριθμός 35. Η ενέργεια αυτή γίνεται αν οι συνολικές κατηγορίες είναι από επτά και πλέον. Έπειτα το σύστημα επιλέγει το μέγιστο αριθμό άρθρων (ο μέγιστος αριθμός υπολογίζεται από τον αλγόριθμο ανάδρασης χρήστη, που αναπτύσσεται παρακάτω) σύμφωνα με τα φίλτρα που έχουν αναφερθεί παραπάνω. Στην περίπτωση που κάποιο από τα αποτελέσματα ανήκει σε είδηση τότε η επιλογή γίνεται σύμφωνα με την ημερομηνία, τις τελευταίες 24 ώρες.

2. Κριτήριο επιλογής επόμενης κατηγορίας άρθρων. Το 2ο κριτήριο εξαγωγής των αποτελεσμάτων του συστήματος, είναι η κατά πλειοψηφία εναλλαγή στο προφίλ των κατηγοριών. Δηλαδή η επόμενη κατηγορία άρθρων που επιλέγουν να διαβάσουν οι περισσότεροι χρήστες, όταν προηγουμένως έχουν βρεθεί σε ένα άρθρο άλλης κατηγορίας. Το φιλτράρισμα γίνεται στο σύνολο των χρηστών που αλληλεπιδρούν με το σύστημα και στα αποτελέσματα των κατηγοριών που διαβάζονται από αυτούς. Για παράδειγμα όταν ένας χρήστης διαβάζει ένα άρθρο που ανήκει στην κατηγορία πολιτική, το επόμενο άρθρο που επιλέγει για αυτόν το σύστημα θα ανήκει στην κατηγορία οικονομία, αν αυτό έχει προκύψει από τις συνολικές κινήσεις των χρηστών. Το σύστημα επιλέγει τις είκοσι κατηγορίες, ταξινομημένες με φθίνουσα σειρά (από το ζευγάρι κατηγοριών με τις περισσότερες εμφανίσεις προς το ζευγάρι με τις λιγότερες εμφανίσεις) οι οποίες είναι η επόμενη επιλογή των χρηστών σε σχέση με τη κατηγορία που βρίσκονται. Έπειτα το σύστημα επιλέγει το μέγιστο αριθμό άρθρων, σύμφωνα με τα παραπάνω φίλτρα με τον αλγόριθμο ανάδρασης χρήστη.

3. Κριτήριο επιλογής βάσει χρονικών ζωνών. Το 3ο κριτήριο εξαγωγής των αποτελεσμάτων του συστήματος, είναι το προφίλ που δημιουργείται βάσει της ώρας που οι χρήστες βρίσκονται στον ιστότοπο. Στη διάρκεια της ημέρα υπάρχουν πέντε ζώνες. Η επιλογή των πέντε ζωνών έγινε σύμφωνα με τα δεδομένα από το google analytics, τα οποία δείχνουν την κίνηση των χρηστών στη διάρκεια της ημέρας. Σε κάθε ζώνη καταγράφεται ποιες κατηγορίες έχουν εμφανιστεί αλλά και πόσες φορές εμφανίστηκαν.

Στην αρχή επιλέγονται οι κατηγορίες που ανήκουν στη χρονική ζώνη, όπου ο χρήστης βρίσκεται στον ιστότοπο, ταξινομημένες με φθίνουσα σειρά. Στη συνέχεια γίνεται έλεγχος αν το πρώτο αποτέλεσμα έχει διαφορά μεγαλύτερη από 50% από το καθένα από τα υπόλοιπα αποτελέσματα. Αν η διαφορά είναι μεγαλύτερη, τότε το σύστημα αφήνει έξω τα αποτελέσματα αυτά. Η ενέργεια αυτή γίνεται αν οι συνολικές κατηγορίες είναι περισσότερες από επτά. Έπειτα το σύστημα επιλέγει το μέγιστο αριθμό άρθρων (ο μέγιστος αριθμός υπολογίζεται με χρήση του αλγόριθμου ανάδρασης χρήστη, που αναπτύσσεται παρακάτω) σύμφωνα με τα παραπάνω φίλτρα.

4. Κριτήριο δημοτικότητας. Το 4ο κριτήριο είναι τα πιο διαβασμένα άρθρα. Η επιλογή του συγκεκριμένου κριτηρίου έγινε για να μειωθεί το πρόβλημα της αργής εκκίνησης (cold start problem). Αφού υπάρχουν κάποια άρθρα τα οποία ενδιαφέρουν πολλούς χρήστες, πολύ πιθανό να ενδιαφέρουν και άλλους, οι οποίοι δεν τα έχουν διαβάσει ακόμη. Η επιλογή του αριθμού των άρθρων υπολογίζεται με τον αλγόριθμο ανάδρασης χρήστη.

5. Ομοιότητα κειμένου. Το 5ο κριτήριο είναι το προφίλ των άρθρων και έχει να κάνει με την επόμενη επιλογή ανάγνωσης άρθρου από τους χρήστες. Το σύστημα ελέγχει ποιο είναι το άρθρο που διαβάζει ο χρήστης και επιλέγει να του εμφανίσει (υπολογίζοντας την ομοιότητα συνημιτόνου μεταξύ του άρθρου που διαβάζει τη δεδομένη στιγμή, με όλα τα άρθρα τα οποία έχουν γραφτεί τις τελευταίες 10 ημέρες, ταξινομημένα σύμφωνα με την ομοιότητα τους) εκείνο που πιθανώς θα είναι το επόμενο άρθρο που θα τον ενδιαφέρει να διαβάσει. Ο μέγιστος αριθμός των αποτελεσμάτων υπολογίζεται από με τον αλγόριθμο ανάδρασης χρήστη.

6. Συνεργατικό φιλτράρισμα βάσει αντικειμένων. Το 6ο κριτήριο έχει να κάνει με το συνεργατικό φιλτράρισμα. Γίνεται επιλογή άρθρων που έχουν διαβαστεί από χρήστες που έχουν μεγάλη ομοιότητα μεταξύ τους, αλλά δεν έχουν διαβαστεί από άλλους, που επίσης έχουν ομοιότητα με τους πρώτους. Η επιλογή του αριθμού των άρθρων υπολογίζεται με τον αλγόριθμο ανάδρασης χρήστη. Για την ομοιότητα μεταξύ των χρηστών, ο υπολογισμός γίνεται με το πόσες κοινές κατηγορίες έχουν οι χρήστες μεταξύ τους.

7. Συνεργατικό φιλτράρισμα βάσει χρηστών. Το 7ο κριτήριο υπολογίζει αντικείμενα (άρθρα) από κατηγορίες που δεν διαβάζει ο χρήστης A, αλλά τις διαβάζουν άλλοι χρήστες, οι οποίοι έχουν μεγάλη ομοιότητα με τον χρήστη A. Η ομοιότητα των χρηστών υπολογίζεται από πόσες κοινές κατηγορίες έχουν οι χρήστες μεταξύ τους. Το σύστημα υπολογίζει τον αριθμό των άρθρων που θα εμφανίσει, σύμφωνα με το πόσες διαφορετικές κατηγορίες θα βρει. Η εύρεση γίνεται μέχρι και τρεις ημέρες πίσω και ο μέγιστος αριθμός των άρθρων υπολογίζεται με τον αλγόριθμο ανάδρασης χρήστη.

8. Κριτήριο δημοτικότητας στα κοινωνικά δίκτυα. Το 8ο κριτήριο έχει να κάνει με το κοινωνικό δίκτυο Twitter. Κάθε πέντε λεπτά γίνεται μια κλήση μέσω του API, που προσφέρει το Twitter, και επιστρέφει τα trends του Twitter τη δεδομένη στιγμή στην Ελλάδα. Το σύστημα διαβάζει τις λέξεις κλειδιά από το Twitter και αναζητεί άρθρα στα οποία ο τίτλος ή το κυρίως σώμα του άρθρου περιέχει αυτές τις λέξεις κλειδιά. Η αναζήτηση γίνεται σε άρθρα τα οποία έχουν γραφτεί την τελευταία ημέρα κατά μέγιστο. Λόγω του περιορισμού των κλήσεων προς το Twitter (υπάρχει όριο το οποίο είναι δεκαπέντε κλήσεις ανά δεκαπέντε λεπτά) σε κάθε κλήση που γίνεται, τα δεδομένα που επιστρέφει το Twitter αποθηκεύονται τοπικά στον server σε αρχείο μορφής json για πέντε λεπτά. Με το πέρας του χρόνου των πέντε λεπτών, το σύστημα κάνει εκ νέου μια κλήση, αντικαθιστώντας τα δεδομένα του αρχείου με τα νέα και συνεχίζει την παραπάνω διαδικασία. Το εν λόγω κριτήριο μια καινοτομία που χρησιμοποιείται από το σύστημα, το οποίο έχει υλοποιηθεί.

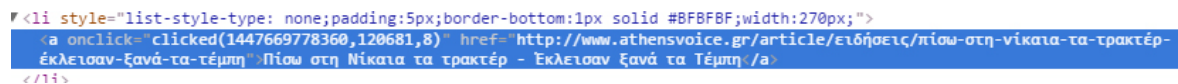
Σε όλα τα κριτήρια λαμβάνονται υπόψιν τα άρθρα που έχουν διαβαστεί από τον κάθε χρήστη, ώστε να μην εμφανίζει στα αποτελέσματα άρθρα τα οποία έχουν ήδη διαβαστεί. Τα τελικά δεκαπέντε άρθρα που επιλέγονται από το σύστημα, εμφανίζονται με τυχαία σειρά με τη χρήση της συνάρτησης της PHP shuffle, ώστε να μην είναι κάποια κατηγορία σε πλεονεκτικότερη θέση από κάποια άλλη (συνήθως τα άρθρα που εμφανίζονται πρώτα στη σειρά επιλέγονται συχνότερα από τους χρήστες σε σχέση με αυτά που εμφανίζονται στις μεσαίες θέσεις). Επίσης γίνεται έλεγχος σε όλα τα κριτήρια, ώστε στο τελικό αποτέλεσμα να μην υπάρχουν διπλογραφίες.

4.1.1 Υποσύστημα ανάδρασης χρήστη

Όπως αναφέρθηκε παραπάνω, το σύστημα είναι βασισμένο σε επτά κριτήρια. Το όγδοο κριτήριο (Twitter trends) επιλέχθηκε να εμφανίζεται αυτόνομο λόγω της ιδιαιτερότητας του. Η επιλογή του αριθμού εμφάνισης των αποτελεσμάτων κάθε ενός από τα οκτώ κριτήρια είναι ανεξάρτητη για κάθε χρήστη και προσαρμόζεται ανάλογα με το τι επιλέγει κάθε χρήστης να διαβάσει.

Κατά την επιλογή ενός άρθρου που προτείνει το σύστημα, γίνεται ενημέρωση στο προφίλ του χρήστη για το κριτήριο που επέλεξε. Δηλαδή αν ο χρήστης με το cookieID C επιλέξει ένα άρθρο, το οποίο εμφανίστηκε βάσει ενός από τα οκτώ κριτήρια που υπάρχουν στο σύστημα, τότε το προφίλ του χρήστη ενημερώνεται αυξάνοντας το κριτήριο που επέλεξε κατά ένα.

Κατά την αρχικοποίηση του συστήματος έχει επιλεγθεί να εμφανίζεται τουλάχιστο ένα αποτέλεσμα από το κάθε κριτήριο, ώστε να μην υπάρχει το πρόβλημα της υπερεξειδίκευσης των αποτελεσμάτων. Ακολουθεί εικόνα του κώδικα που καταγράφει το κριτήριο από την επιλογή άρθρου. Πρώτα υπάρχει το cookie ID του χρήστη, ακολουθεί το ID του άρθρου και τέλος ο αριθμός του κριτηρίου, που στην συγκεκριμένη περίπτωση είναι το κριτήριο με τον αριθμό 8.



```
<li style="list-style-type: none;padding:5px;border-bottom:1px solid #BFBFBF;width:270px;">  
<a onclick=" clicked(1447669778360,120681,8)" href=" http://www.athensvoice.gr/article/ειδήσεις/πίσω-στη-νίκαια-τα-τρακτέρ-  
έκλεισαν-ξανά-τα-τέμπη" Πίσω στη Νίκαια τα τρακτέρ - Έκλεισαν ξανά τα Τέμπη /a>  
</li>
```

Εικόνα 5. Κώδικας καταγραφής των clicks.

4.2 Αλγόριθμος

Ο αλγόριθμος που αναπτύχθηκε για τον υπολογισμό των απαραίτητων άρθρων, είναι βασισμένος στην υβριδική προσέγγιση και τη χρήση τεχνικών όπως τη χρονική στιγμή που ο χρήστης βρίσκεται στον ιστότοπο, αλλά και τη χρήση του κοινωνικού δικτύου Twitter (τα twitter trends της Ελλάδας). Η επιλογή των αποτελεσμάτων βασίζεται σε οκτώ κριτήρια τα οποία αναλύονται παραπάνω. Για το πρόβλημα της επιλογής του αριθμού των αποτελεσμάτων κάθε κριτηρίου, επιλέχθηκε η λύση της ανάδρασης χρήστη. Με αυτή τη προσθήκη το σύστημα πλέον είναι προσαρμοστικό για κάθε χρήστη ξεχωριστά σύμφωνα με το ποια κριτήρια επιλέγει.

Αλγόριθμος 1, υπολογισμού των προτάσεων του συστήματος.

1. **procedure** Produce Recommendations
2. uhr <- fn_user_has_read(cookieID)
3. maxNumuser_profile <- call algorithm 3(criteria 1) + 2
4. recommendationUserProfile <- recommendation from
 fn_get_fromuser_profile(cookieID, maxNumuser_profile,uhr)
5. maxNumcategory_profile <- call algorithm 3(criteria 2) + 1
6. recommendationCategoryProfile <- recommendation from
 fn_get_from_tid_profile(cookieID, maxNumcategory_profile, uhr,
 categoryID)
7. maxNumtime_profile <- call algorithm 3(criteria 3) + 1
8. recommendationTimeProfile <- recommendation from
 fn_get_from_time_profile(cookieID, time, maxNumtime_profile, uhr)
9. maxNummost_read <- call algorithm 3(criteria 4) + 1
10. recommendationMostRead <- recommendation from
 fn_get_from_most_read(cookieID, maxNummost_read, uhr)
11. maxNumcategory_profile <- call algorithm 3(criteria 5) + 1
12. recommendationCategoryProfile <- recommendation from
 fn_get_from_tid_profile(cookieID, maxNumcategory_profile, uhr)
13. maxNumarticle_profile <- call algorithm 3(criteria 6) + 1
14. recommendationArticleProfile <- recommendation from
 fn_get_from_nid_profile(cookieID, articleID, maxNumarticle_profile, uhr)
15. maxNumuser_similarity <- call algorithm 3(criteria 7) + 1

16. `recommendationUserSimilarity <- recommendation from
fn_get_from_user_similarity(cookieID, maxNumuser_similarity,
uhr)`
17. `maxNumtwittertrends <- call algorithm 3(criteria 8)`
18. `recommendationTwitterTrends <- recommendation from
fn_get_from_twitter_trends(uhr, maxNumtwittertrends)`
19. `allCriteriaRecommendation <- combine(recommendationUserProfile
,recommendationCategoryProfile, recommendationTimeProfile,
recommendationMostRead, recommendationCategoryProfile,
recommendationArticleProfile, recommendationUserSimilarity,
recommendationTwitterTrends)`
20. `finalRecommendation <- distinct(allCriteriaRecommendation)`
21. `return finalRecommendation`

Αλγόριθμος 1. Αλγόριθμος υπολογισμού των προτάσεων του συστήματος.

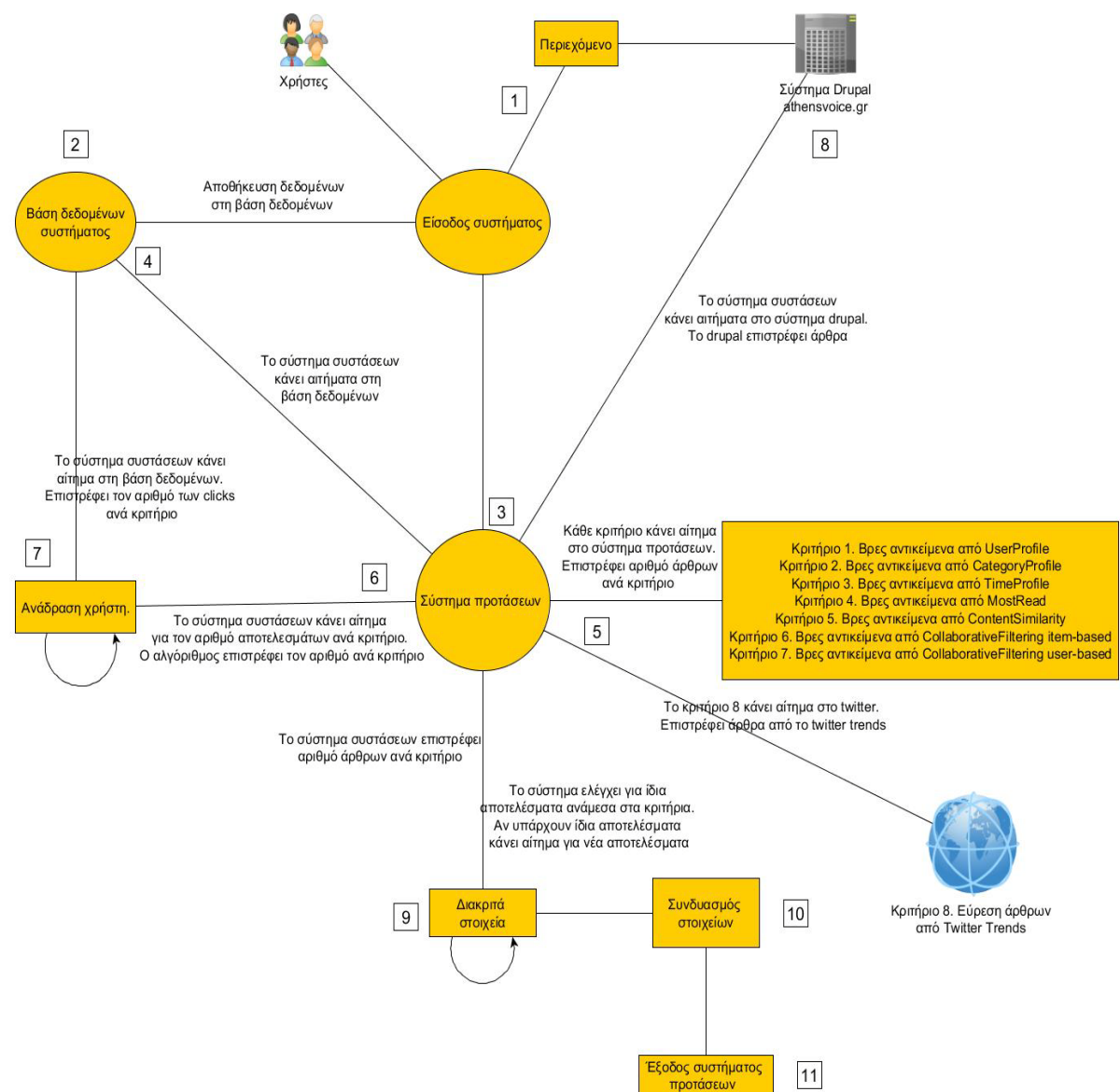
Ο αλγόριθμος για τον υπολογισμό των προτάσεων από το σύστημα, υπολογίζει ποια άρθρα θα εμφανίσει στον κάθε χρήστη αλλά και τον αριθμό των άρθρων αυτών, σε σχέση με τα ενδιαφέροντα του κάθε χρήστη ξεχωριστά. Για την αποφυγή του προβλήματος της υπερεξειδίκευσης των αποτελεσμάτων, επιλέχθηκε για το κριτήριο "προφίλ χρηστών" να υπάρχουν τουλάχιστον δύο (+2) αποτελέσματα, και για τα υπόλοιπα κριτήρια να υπάρχει τουλάχιστον ένα (+1). Ο αλγόριθμος πρώτα υπολογίζει τον αριθμό αποτελεσμάτων κάθε κριτηρίου maxNum, κάνοντας χρήση του αλγόριθμου 3 με παράμετρο τον αριθμό του κριτηρίου. Έπειτα για κάθε ένα κριτήριο υπολογίζει τα άρθρα που θα εμφανίσει στον χρήστη.

- Κριτήριο προφίλ χρηστών `fn_get_fromuser_profile`, με παραμέτρους το `cookieID` του χρήστη, τον αριθμό των αποτελεσμάτων `maxNumuser_profile` και τη `uhr`, δηλαδή τα άρθρα που έχει διαβάσει ο χρήστης ώστε να μην υπάρχουν στα αποτελέσματα.
- Κριτήριο επιλογής επόμενης κατηγορίας `fn_get_from_tid_profile`, με παραμέτρους το `cookieID` το `maxNumcategory_profile` το `uhr` και το `categoryID`, που είναι το `id` της κατηγορίας που επιλέγουν να διαβάσουν οι περισσότεροι χρήστες, όταν βρίσκονται σε κάποια δεδομένη κατηγορία.
- Κριτήριο επιλογής βάσει χρονικών ζωνών `fn_get_from_time_profile`, με παραμέτρους το `cookieID`, τον αριθμό των αποτελεσμάτων `maxNumtime_profile`

τη uhr και την time, που είναι η χρονική στιγμή όπου ο χρήστης βρίσκεται στον ιστότοπο.

- Κριτήριο δημοτικότητας `fn_get_from_most_read`, με παραμέτρους το `cookieID` το `maxNummost_read` και τη `uhr`.
- Κριτήριο ομοιότητας κειμένου `fn_get_from_tid_profile`, με παραμέτρους τις `cookieID`, `maxNumcategory_profile`, `uhr`.
- Κριτήριο συνεργατικού φιλτραρίσματος βάσει αντικειμένων `fn_get_from_nid_profile`, με παραμέτρους το `cookieID`, `maxNumarticle_profile`, `uhr` και `articleID`. Το `id` δηλαδή του άρθρου που διαβάζει ο χρήστης.
- Κριτήριο συνεργατικού φιλτραρίσματος βάσει χρηστών `fn_get_from_user_similarity` με παραμέτρους `cookieID`, `maxNumuser_similarity`, `uhr`.
- Κριτήριο δημοτικότητας στα κοινωνικά δίκτυα `fn_get_from_twitter_trends`, με μεταβλητές τις `maxNumtwittertrends` και `uhr`.

Έπειτα συγχωνεύονται όλα τα άρθρα (`combine`) και γίνεται έλεγχος για διπλοεγγραφές (`distinct`). Στη περίπτωση που υπάρχει διπλοεγγραφή, το σύστημα υπολογίζει κάποιο άλλο άρθρο με τα ίδια χαρακτηριστικά με το άρθρο που αφαιρέθηκε από τον πίνακα των αποτελεσμάτων.

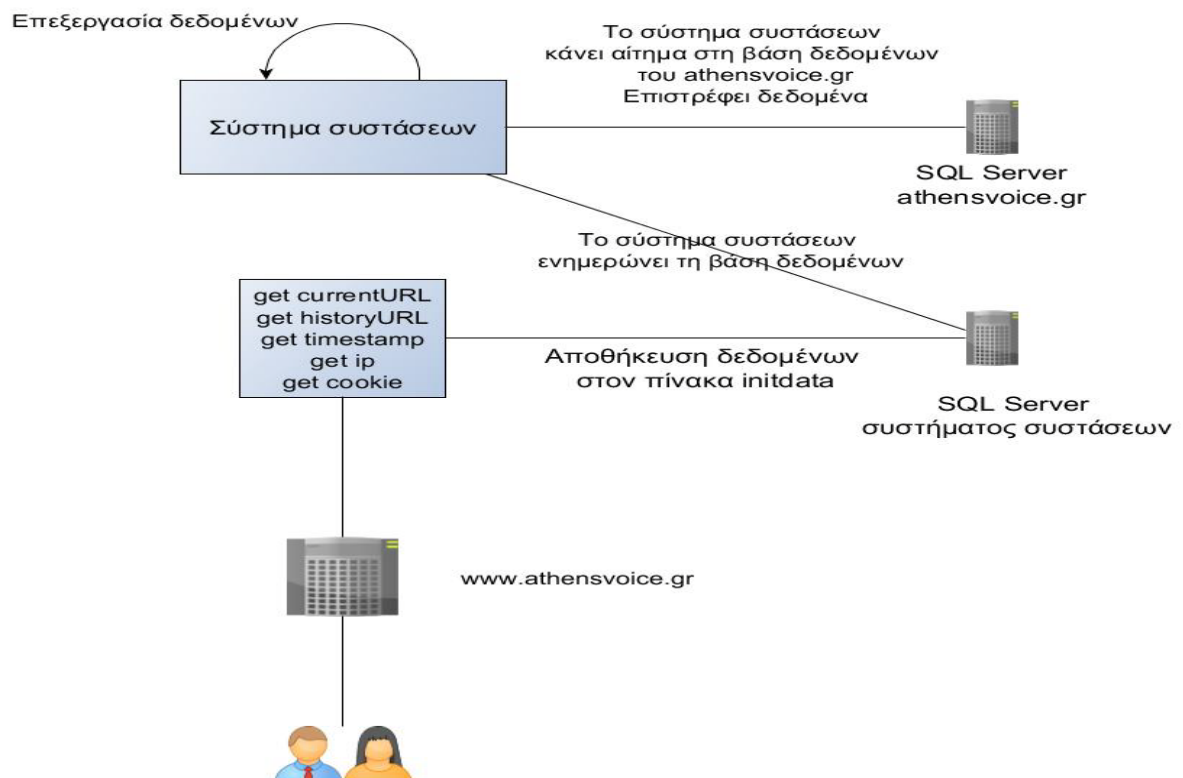


Εικόνα 6. Αλγόριθμος υπολογισμού των προτάσεων του συστήματος.

Αλγόριθμος 2, διαχείριση δεδομένων από το σύστημα.

1. **procedure** User Data
2. get currentURL
3. get historyURL
4. get timestamp
5. get ip
6. call procedure cookieID
7. store to database recommendation system table_name initdata values cookieID, currentURL, historyURL, timestamp
8. recommendation system make request to site database
9. recommendation system make processes data from site database
10. recommendation system update database tables cid_ip, tid_profile, day_profile, day_tid, time_profile, time_tid, nid_profile, user_from_nid_to_nid, user_from_to
11. return recommendation

Αλγόριθμος 2. Αλγόριθμος διαχείρισης δεδομένων από το σύστημα.

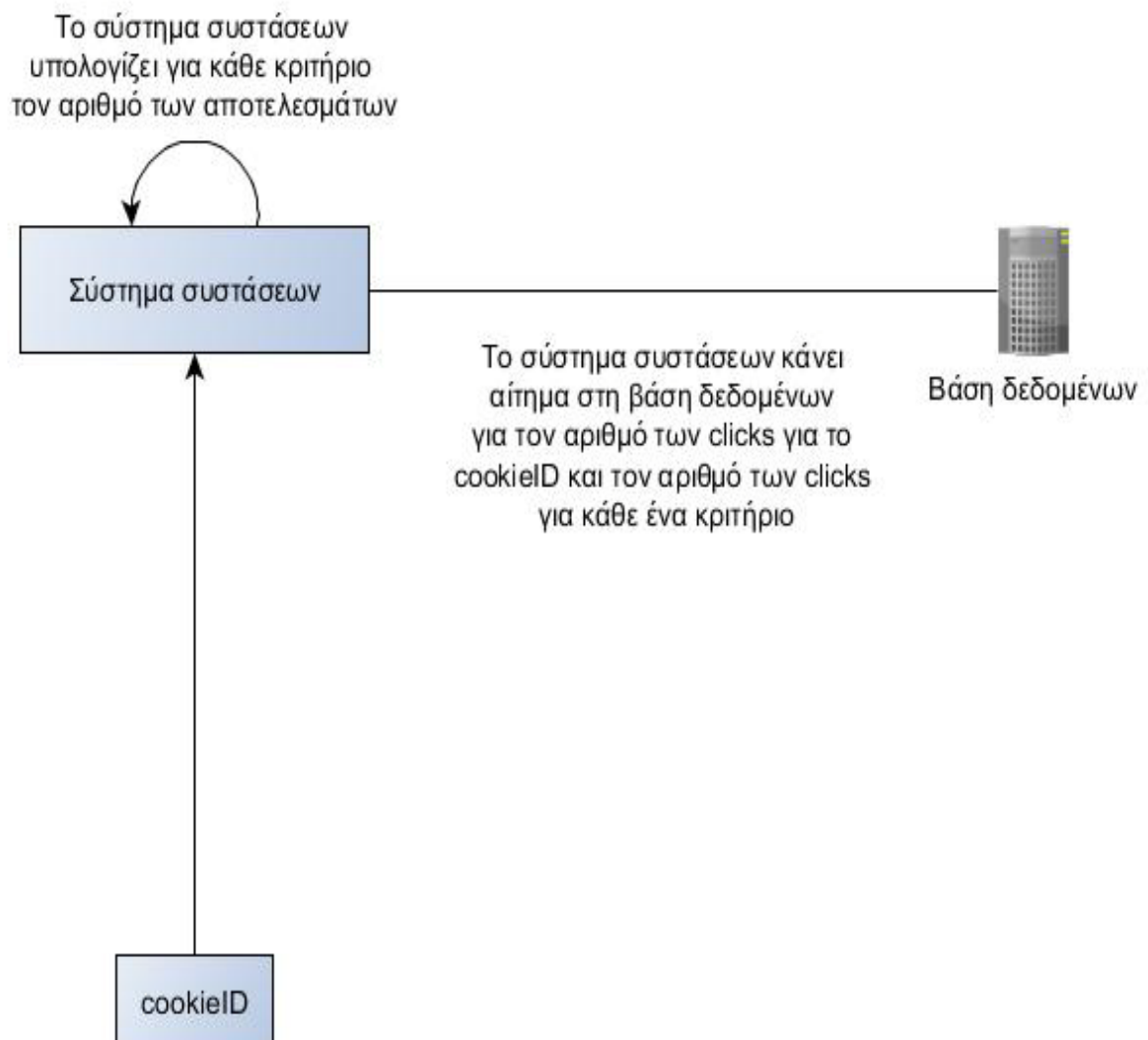


Εικόνα 7. Αλγόριθμος διαχείρισης δεδομένων από το σύστημα.

Αλγόριθμος 3, υπολογισμού αριθμού άρθρων εμφάνισης κάθε κριτηρίου από το σύστημα.

1. **procedure** number of results per criteria(criteria)
2. total_number <- number of clicks from table click for cookieID
3. number_per_criteria <- number of clicks per criteria from table click for cookieID
4. criteria_number <- ((number_per_criteria/total_number)*10)
5. return criteria_number

Αλγόριθμος 3. Αλγόριθμος υπολογισμού αριθμού άρθρων εμφάνισης κάθε κριτηρίου από το σύστημα.

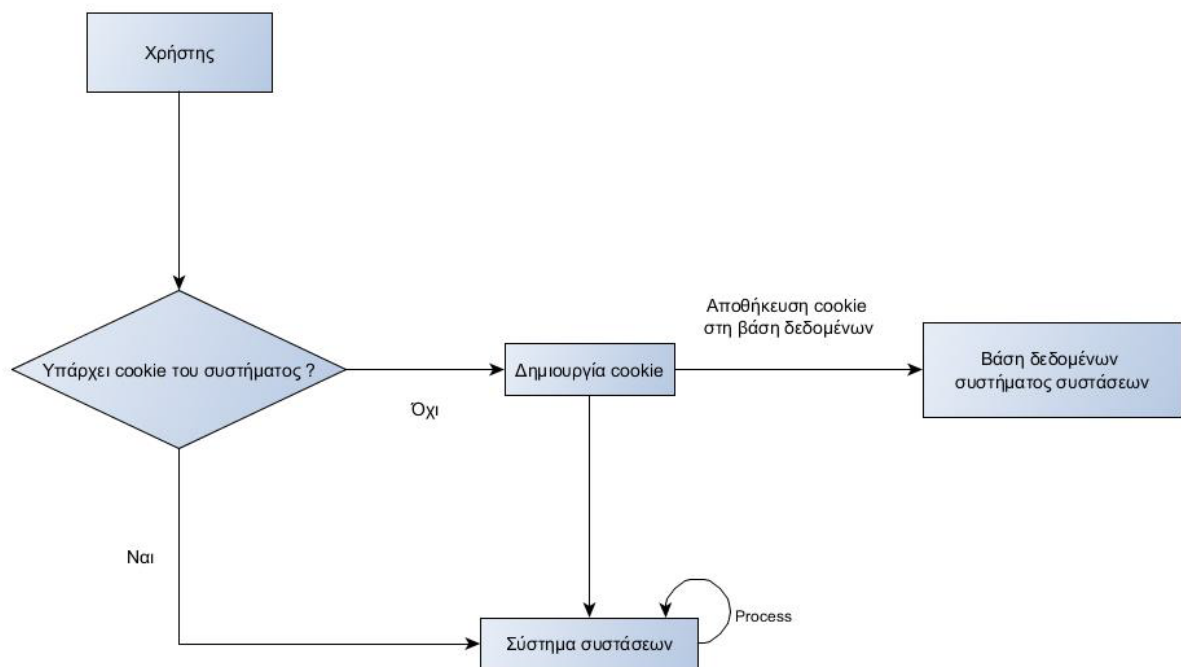


Εικόνα 8. Αλγόριθμος υπολογισμού αριθμού άρθρων κάθε κριτηρίου.

Αλγόριθμος 4, έλεγχος ύπαρξης ή μη cookie από το σύστημα.

1. **procedure** cookieID
2. get cookie from user
3. have_cookie <- connect to database and find if cookie exist
4. if have_cookie = 0 then
5. user <- Server generate cookieID
6. continue
5. else
6. continue
7. return cookieID

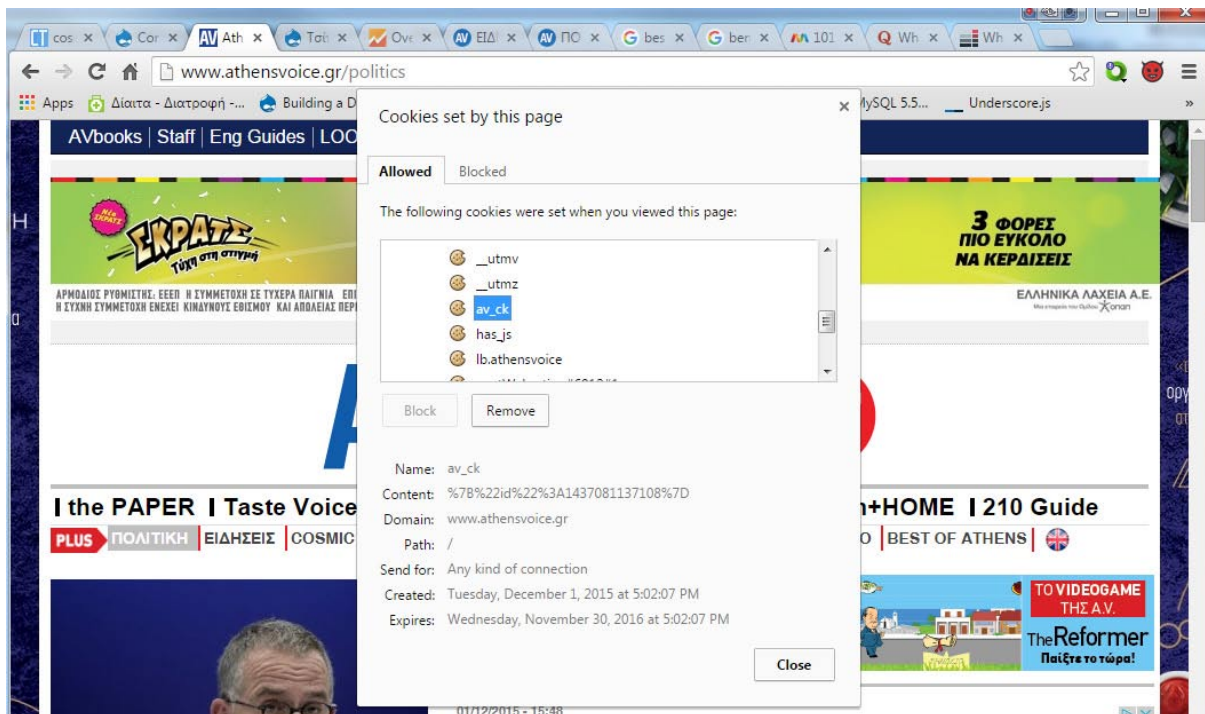
Αλγόριθμος 4. Αλγόριθμος ελέγχου ύπαρξης ή μη cookie από το σύστημα.



Εικόνα 9. Αλγόριθμος ελέγχου ύπαρξης ή μη cookie από το σύστημα.

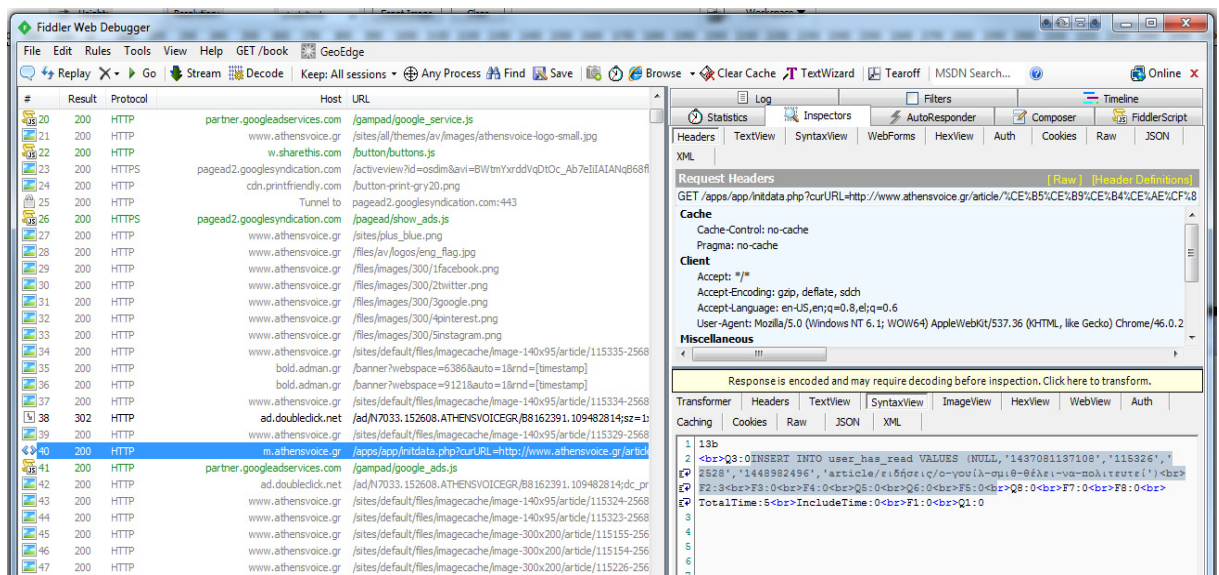
4.3 Δείγμα Χρηστών

Το δείγμα που χρησιμοποιεί το σύστημα πρότασης περιεχομένου, βασίζεται σε αληθινούς χρήστες και προέρχεται από τον ιστότοπο athensvoice.gr. Κατά την φόρτωση οποιασδήποτε σελίδας, υπάρχει κώδικας (JavaScript, Ajax, PHP), ο οποίος ελέγχει αν ο χρήστης έχει το απαραίτητο cookie με όνομα `av_ck`. Αν δεν υπάρχει καταγεγραμμένο στη βάση δεδομένων, τότε του δίνει ένα μοναδικό cookie ID. Στη περίπτωση που υπάρχει cookie, το σύστημα μεταβαίνει στο επόμενο στάδιο.



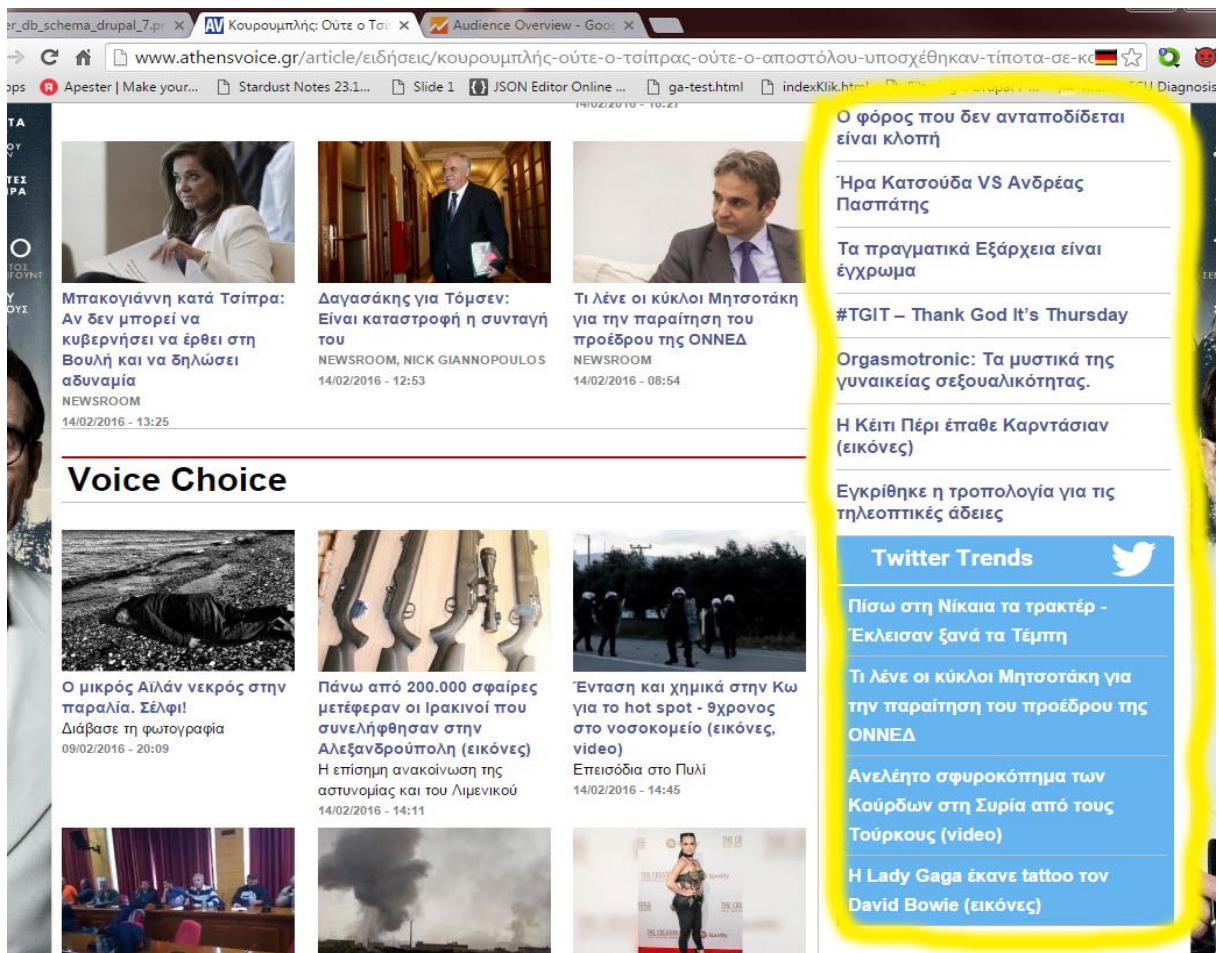
Εικόνα 10. Cookie από το σύστημα.

Επίσης υπάρχει άλλη μια συνάρτηση η οποία καταγράφει τις κινήσεις του χρήστη και τις αποθηκεύει στη βάση δεδομένων. Από τη πληροφορία αυτή, μετέπειτα, μας δίνεται η δυνατότητα να εξάγουμε όλα τα απαραίτητα δεδομένα, ώστε να γίνει ο υπολογισμός των αποτελεσμάτων που θα προταθούν από το σύστημα. Ακολουθεί εικόνα με το αρχείο `initdata.php`, το οποίο καταγράφει τι διαβάζει ο χρήστης. Για την ανάγκη δημιουργίας εικόνας για χρήση στη διατριβή στην εικόνα εμφανίζεται το ερώτημα εισαγωγής των δεδομένων στη βάση, κανονικά στο σύστημα δεν εμφανίζεται το ερώτημα.



Εικόνα 11. Καταγραφή των δεδομένων και εισαγωγή στη βάση.

Η παρακάτω εικόνα εμφανίζει ένα μέρος των προτάσεων που γίνονται από το σύστημα καθώς και το όγδοο κριτήριο που είναι τα Twitter Trends.

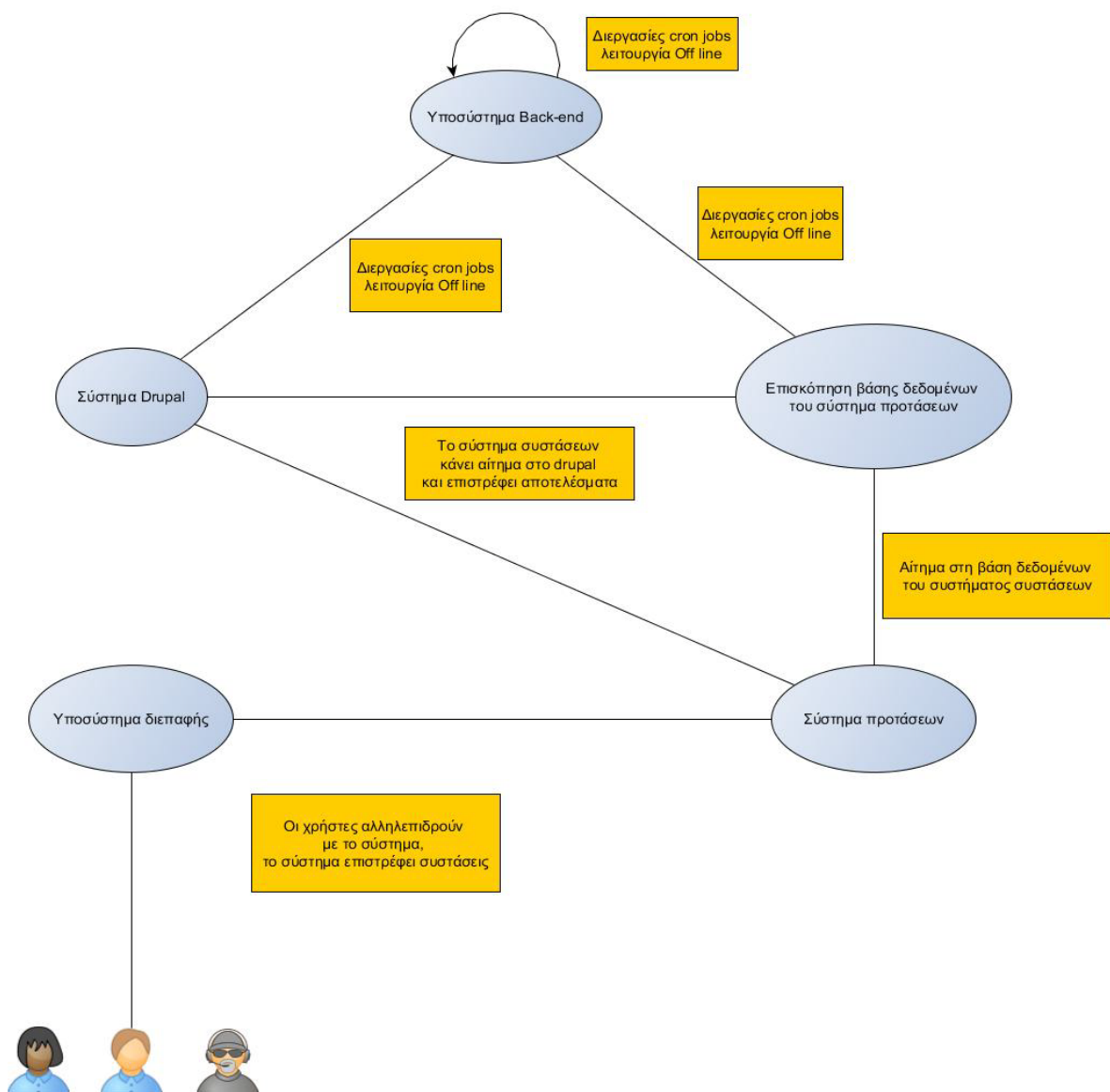


Εικόνα 12. Εμφάνιση των αποτελεσμάτων στον ιστότοπο athensvoice.gr.

4.4 Αρχιτεκτονική συστήματος

Σε αυτή την ενότητα γίνεται περιγραφή του συστήματος που έχει αναπτυχθεί. Για να είναι απολύτως σαφές, έχει γίνει διαχωρισμός του συστήματος στα ακόλουθα:

- Σύστημα Drupal.
- Υποσύστημα διεπαφής.
- Επισκόπηση βάσης δεδομένων του συστήμα προτάσεων.
- Σύστημα προτάσεων
- Υποσύστημα Back-end



Εικόνα 13. Περιγραφή συστήματος.

4.5 Τεχνολογίες υλοποίησης της εφαρμογής

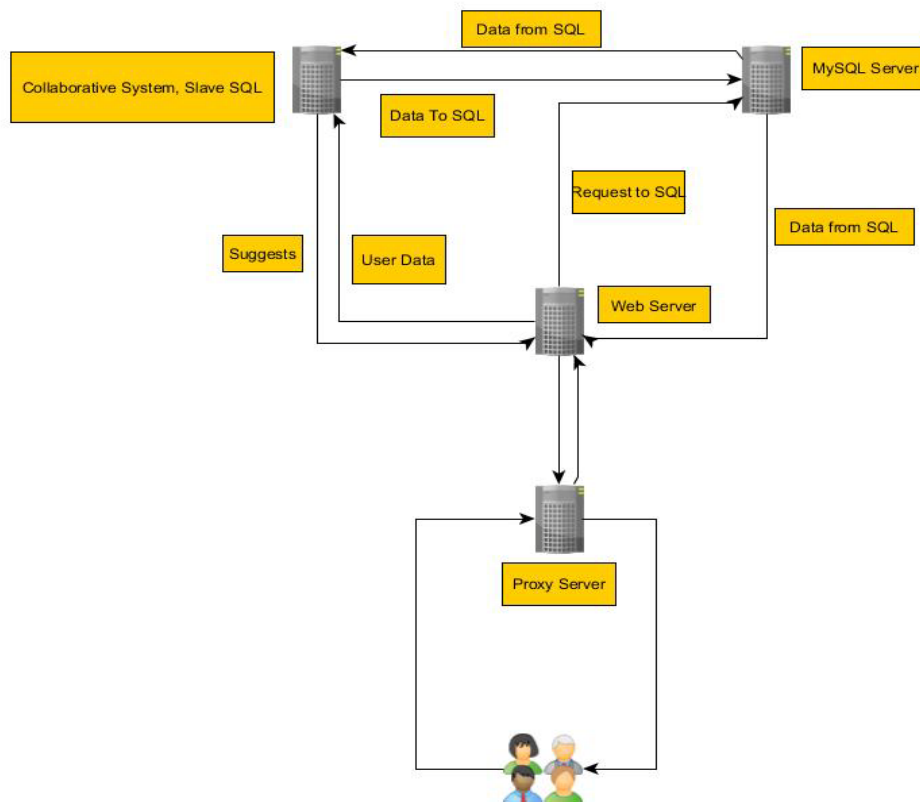
Για την υλοποίηση της εφαρμογής επιλέχθηκε η χρήση γλωσσών προγραμματισμού και υπηρεσιών ανοικτού κώδικα, όπως είναι php, mysql, javascript, Apache server, Mysql Server. Η επιλογή αυτή έγινε γιατί δεν υπάρχει ανάγκη αγοράς αδειών χρήσης για την υλοποίηση της εφαρμογής, αλλά και για λόγους συμβατότητας με το ήδη υπάρχον σύστημα που φιλοξενείται ο ιστότοπος της athenvoice.gr.

4.5.1 Υλικοτεχνική υποδομή, λειτουργικό σύστημα, πλατφόρμα υλοποίησης

Ο ιστότοπος athenvoice.gr είναι υλοποιημένος με την open source πλατφόρμα Drupal. Το site υλοποιείται με τρεις servers και με λειτουργικό σύστημα Linux CentOS 6. Ο πρώτος server εκτελεί χρέη Reverse Proxy και Caching. Για τη λειτουργία αυτή γίνεται χρήση της εφαρμογής Varnish. Ο δεύτερος είναι ο Web Server και "τρέχει" την υπηρεσία Apache, η οποία σερβίρει τις σελίδες στους χρήστες. Ο τρίτος server εκτελεί χρέη SQL Server, κάνοντας χρήση της εφαρμογής MySQL και φιλοξενεί τη βάση δεδομένων του site, η οποία είναι απαραίτητη για τη σωστή λειτουργία του συστήματος συστάσεων άρθρων.

Ο SQL Server που φιλοξενεί τη βάση δεδομένων του site, "τρέχει" σαν Master, ενώ σαν Slave (Replication) "τρέχει" ο Server που φιλοξενεί το σύστημα συστάσεων. Σκοπός είναι να υπάρχει πρόσβαση σε όλα τα απαραίτητα δεδομένα και να μην επιβαρύνεται η λειτουργία του site με τις απαιτητικές διεργασίες, που πρέπει να εκτελούνται για τον υπολογισμό των προτάσεων από το σύστημα.

Στον τελευταίο Server είναι αποθηκευμένα όλα τα απαραίτητα αρχεία για τη λειτουργία του συστήματος συστάσεων. Ακολουθεί σχεδιάγραμμα με την τοπολογία των Servers αλλά και τη ροή των εργασιών.



Εικόνα 14. Αρχιτεκτονική συνολικού συστήματος, drupal και συστάσεων.

4.5.2 Δημιουργία και καταγραφή δεδομένων Online και Offline

Το σύστημα λειτουργεί εγγράφει και επεξεργάζεται τα δεδομένα των χρηστών σε καταστάσεις offline και online. Η επιλογή χρησιμοποίησης και των δύο αυτών τεχνικών, δίνει τη δυνατότητα στην εφαρμογή να έχει όλα εκείνα τα δεδομένα τα οποία είναι απαραίτητα για τη σωστή λειτουργία της και παράλληλα να μπορεί να αντιμετωπίσει τα προβλήματα performance, που προκύπτουν σε ένα τέτοιο σύστημα.

Ένας από τους στόχους της υλοποίησης ήταν το σύστημα να είναι σε θέση να απαντήσει σε χρονικό διάστημα μικρότερο του ενός δευτερολέπτου, έτσι ώστε ο χρήστης να είναι σε θέση να δει τις προτάσεις του συστήματος, ξεκινώντας την ανάγνωση του άρθρου. Στη κατάσταση online το σύστημα ενημερώνει τη βάση δεδομένων για τις αλλαγές στη συμπεριφορά του χρήστη, σε σχέση με τις επιλογές που κάνει κατά τη διάρκεια που περιηγείται στον ιστότοπο. Η επιλογή των δεδομένων που ενημερώνονται online έγινε με κριτήριο τα ατομικά δεδομένα του κάθε χρήστη (ποια άρθρα έχει διαβάσει), ώστε το σύστημα να μην εμφανίζει αποτελέσματα, τα οποία έχουν εμφανιστεί προηγουμένως,

όπως και το προφίλ του κάθε χρήστη, με βάση τις κατηγορίες που επιλέγει να διαβάσει. Οι συναρτήσεις που χρησιμοποιούνται για την online λειτουργία είναι οι ακόλουθες.

- **fn_update_userHasRead()**. Εδώ γίνεται η ενημέρωση της βάσης δεδομένων στον πίνακα (user_has_read). Κατά την εκτέλεση της, εισάγει στη βάση δεδομένων μια εγγραφή που αφορά, το άρθρο (nid) που έχει επιλέξει να διαβάσει ο χρήστης και το (cookieID) του χρήστη, ώστε στην επόμενη εμφάνιση των αποτελεσμάτων, το σύστημα να μην συμπεριλάβει το εν λόγω άρθρο.
- **fn_update_userProfile()**. Η ενημέρωση γίνεται στον πίνακα (user_profile) της βάσης δεδομένων. Η συνάρτηση ενημερώνει μία εγγραφή του πίνακα που αφορά το προφίλ του χρήστη, δηλαδή αν ο χρήστης έχει επιλέξει να διαβάσει ένα άρθρο το οποίο ανήκει σε μια κατηγορία, τότε το σύστημα αυξάνει κατά ένα την δεδομένη κατηγορία. Σε περίπτωση που ο χρήστης διαβάζει πρώτη φορά την κατηγορία αυτή, τότε εισάγει μια νέα εγγραφή με το (cookieID) του χρήστη, το (tid) της κατηγορίας και δίνει τη τιμή 1 στο πεδίο (tidCount).
- **fn_update_time_tid()**. Η ενημέρωση γίνεται στον πίνακα (time_tid) της βάσης δεδομένων. Η συνάρτηση εισάγει μία νέα εγγραφή στον πίνακα που αφορά τον χρήστη με το (cookieID), σε ποιά χρονική ζώνη έγινε η εγγραφή (from_to) και σε ποιά κατηγορία ανήκει το άρθρο (tid) που διάβασε.
- **fn_update_day_tid()**. Η ενημέρωση γίνεται στον πίνακα (day_tid) της βάσης δεδομένων. Η συνάρτηση εισάγει μία νέα εγγραφή στον πίνακα που αφορά τον χρήστη με το (cookieID), ποια μέρα της εβδομάδας έγινε (day) και σε ποια κατηγορία (tid) ανήκει το άρθρο που διάβασε.

Σε λειτουργία offline οι συναρτήσεις που εκτελούνται εισάγουν τις εγγραφές όχι απευθείας στη βάση δεδομένων, αλλά σε αρχείο, το οποίο εκτελείται μέσα στη διάρκεια της ημέρας ανά τακτά χρονικά διαστήματα, ώστε να γίνεται ενημέρωση της βάσης δεδομένων. Ο λόγος χρήσης της τεχνικής αυτής έχει το πλεονέκτημα ότι χρησιμοποιεί μόνο μια σύνδεση με τον sql server και όχι πολλές, που μπορεί να οδηγήσουν σε έλλειψη πόρων από πλευράς του server. Πόρων χρήσιμων για να επεξεργαστεί συνδέσεις οι οποίες είναι απαραίτητες για την σωστή εκτέλεση της εφαρμογής. Οι συναρτήσεις που εισάγουν τα ερωτήματα (sql queries) σε αρχείο, για τη μετέπειτα εκτέλεση τους στη βάση δεδομένων είναι οι ακόλουθες:

- **fn_update_day_profile():** Η ενημέρωση γίνεται στον πίνακα (day_profile) της βάσης δεδομένων, ενώ η συνάρτηση ενημερώνει για το πόσες φορές (tidCount) εμφανίστηκε μία κατηγορία άρθρων (tid) και σε ποια ημέρα της εβδομάδας (day).
- **fn_update_user_from_to():** Η ενημέρωση γίνεται στον πίνακα (user_from_to) της βάσης δεδομένων, ενώ η συνάρτηση ενημερώνει τον χρήστη (cookieID) από ποια κατηγορία άρθρων (tidHistory) διάβαζε, ποια ήταν η επόμενη κατηγορία που επέλεξε να διαβάσει (tidCurrent) και σε ποια χρονική στιγμή έγινε η ενέργεια αυτή (timestamp).
- **fn_update_time_profile():** Η ενημέρωση γίνεται στον πίνακα (time_profile) της βάσης δεδομένων, ενώ η συνάρτηση ενημερώνει τη ζώνη της ημέρας (timezone) και πόσες φορές (tidCount) διαβάστηκε μια κατηγορία άρθρων (tid).
- **fn_update_nidProfile():** Η ενημέρωση γίνεται στον πίνακα (nid_profile) της βάσης δεδομένων, ενώ η συνάρτηση ενημερώνει για το πόσες φορές (times) από ένα άρθρο (nid1) επέλεξαν να διαβάσουν ένα άλλο (nid2) και ποια είναι η ομοιότητα συνημιτόνου μεταξύ των δύο άρθρων (cos_sim). Εδώ πρέπει να διευκρινιστεί ότι η ομοιότητα συνημιτόνου ενημερώνεται από άλλο σημείο της εφαρμογής, οπότε θα γίνει αναφορά στην ίδια ενότητα παρακάτω.
- **fn_update_tidProfile():** Η ενημέρωση γίνεται στον πίνακα (tid_profile) της βάσης δεδομένων, ενώ η συνάρτηση ενημερώνει για το πόσες φορές (times) από μια κατηγορία άρθρων (tidHistory) επέλεξαν κάποια άλλη (tidCurrent).
- **fn_update_user_from_nid_to_nid():** Η ενημέρωση γίνεται στον πίνακα (user_from_nid_to_nid) της βάσης δεδομένων, ενώ η συνάρτηση ενημερώνει τον χρήστη (cookieID) ποια άρθρα διαβάζει (nidCurrent), από ποια άρθρο ήρθε (nidHistory) και σε ποια χρονική στιγμή (timestamp).

Η ενημέρωση της βάσης γίνεται μέσω της εκτέλεσης του αρχείου run_queries.php, που "τρέχει" μέσω cron jobs τέσσερις φορές την ημέρα. Οι ώρες που επιλέγησαν είναι ώρες που ο ιστότοπος έχει μειωμένο traffic.

Η εφαρμογή κάνει χρήση και άλλων πινάκων από τη βάση δεδομένων. Οι πίνακες αυτοί ενημερώνονται μέσω άλλων διεργασιών που εκτελούνται και αυτές με cron jobs. Επίσης υπάρχουν διεργασίες, οι οποίες εκτελούνται δύο φορές το μήνα με σκοπό να διαγράφουν δεδομένα από τη βάση δεδομένων, τα οποία δεν είναι πλέον απαραίτητα για την εφαρμογή. Ο λόγος που επιλέχθηκε η offline λειτουργία είναι θέμα

πολυπλοκότητας, ο υπολογισμός γίνεται ανάμεσα σε μεγάλο όγκο δεδομένων και οι πράξεις που είναι απαραίτητες για τους υπολογισμούς των αλγορίθμων, χρησιμοποιούν μεγάλα ποσοστά από τα resources του server. Τα αρχεία που χρησιμοποιούνται αλλά και μια μικρή ανάλυση για το τι κάνουν, αναφέρονται ακριβώς από κάτω:

- `reset_data.php`: Το αρχείο αυτό εκτελείται μία φορά την εβδομάδα και διαγράφει τον πίνακα `initdata`, όπου εισάγονται όλα τα αρχικά δεδομένα. Έπειτα από την επεξεργασία των δεδομένων, που γίνεται σε άλλα σημεία της εφαρμογής, τα δεδομένα που υπάρχουν στον πίνακα `initdata` δεν είναι πλέον απαραίτητα.
- `insert_cosine_similarity.php`: Το αρχείο εκτελείται 90 φορές την ημέρα και υπολογίζει την ομοιότητα συνημιτόνου, μεταξύ των άρθρων που έχουν εισέλθει πρόσφατα στο σύστημα, με τα ήδη υπάρχοντα. Η πληροφορία που δημιουργείται ενημερώνει από τον πίνακα `nid_profile` το πεδίο `cos_sim`. Η λειτουργία της έχει άμεση σχέση με τη συνάρτηση `fn_update_nidProfile()`.
- `produce_similarity_up.php`: Το αρχείο εκτελείται μία φορά την ημέρα, υπολογίζει την ομοιότητα μεταξύ των χρηστών που είναι στο σύστημα και των νέων που εισάγονται σε αυτό. Τα αποτελέσματα των υπολογισμών εισάγονται στον πίνακα `similarity_up`.
- `run_queries.php`: Το αρχείο εκτελείται μία φορά την ημέρα, εισάγει τις εγγραφές που υπάρχουν στο αρχείο με τις εντολές `sql`, που έχουν εισάγει οι συναρτήσεις `offline`.

4.5.3 Βάση δεδομένων

Η υλοποίηση της εφαρμογής έγινε με χρήση της εφαρμογής MySQL 5.1.37 και οι πίνακες είναι σε μορφή `InnoDB` με χρήση `index`. Επίσης έγιναν οι απαραίτητες ρυθμίσεις στο `configuration` του `mysql`, ώστε να εκτελεί όλα τα ερωτήματα χωρίς τη χρήση αντιγραφής των αποτελεσμάτων σε προσωρινό φάκελο. Επίσης έγινε αλλαγή του `file system` των σκληρών δίσκων συστοιχίας RAID-10 (τεχνολογίας SSD), που φιλοξενούνται τα δεδομένα της βάσης δεδομένων σε μορφή XFS. Κάποια από τα πεδία σβήνονται μετά από κάποιες ημέρες. Όταν γίνει το απαραίτητο φιλτράρισμα της πληροφορίας που περιέχουν, η πληροφορία αυτή μεταφέρεται σε άλλους πίνακες που χρησιμοποιούνται από το σύστημα συστάσεων. Επίσης, λόγω των πολλών εγγραφών

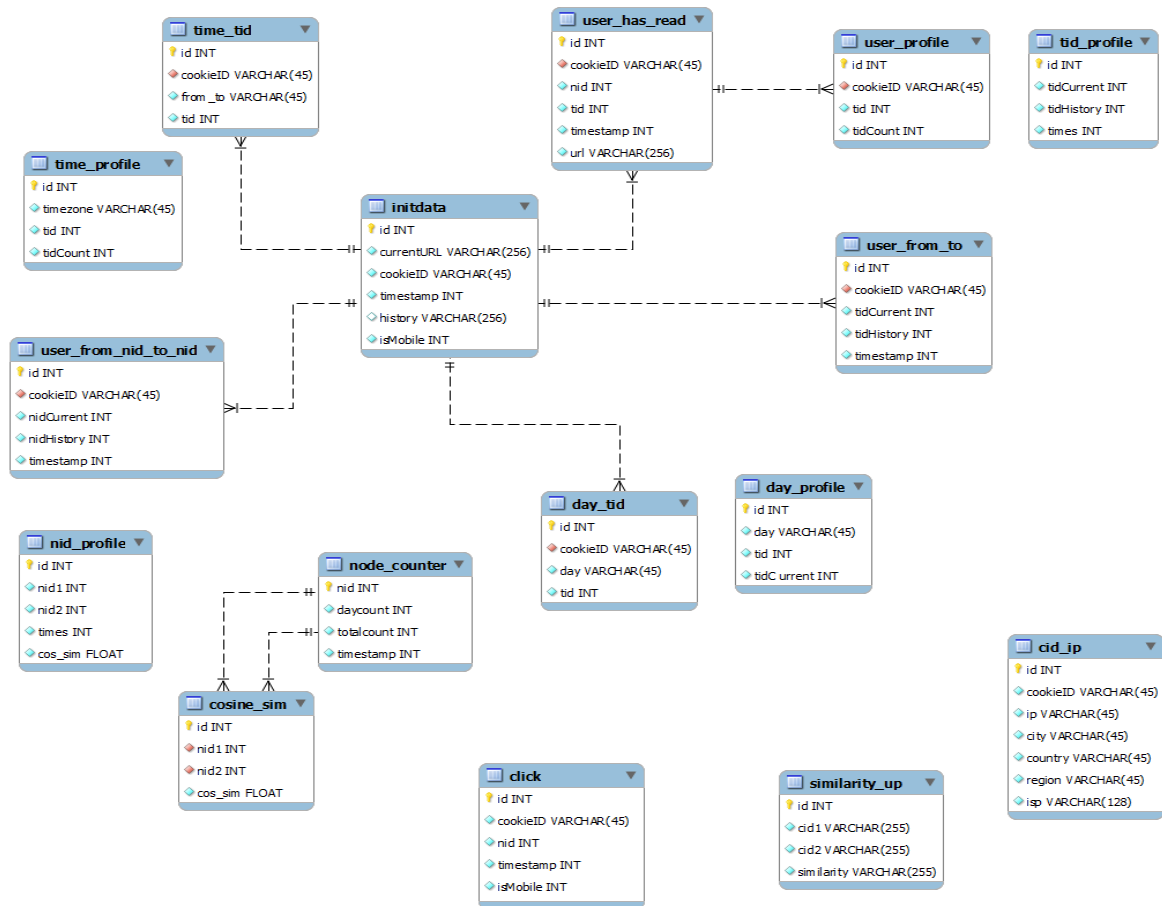
και διαγραφών, εκτελείται, μια φορά την ημέρα ένα script, το οποίο κάνει optimize στη βάση.

Έχει γίνει χρήση πινάκων οι οποίοι ανήκουν στο CMS Drupal, με το οποίο έχει γίνει η υλοποίηση του ιστότοπου athenstvoice.gr. Οι πίνακες αυτοί μας παρέχουν πληροφορίες όπως η κατηγορία των άρθρων, τα ίδια τα άρθρα, πόσες φορές έχει διαβαστεί ένα άρθρο, το URL του άρθρου (που είναι απαραίτητο για την εξαγωγή της απαραίτητης πληροφορίας), τον τίτλο του και τη φωτογραφία του.

Οι πίνακες που έχουν δημιουργηθεί για τη λειτουργία της εφαρμογής είναι οι ακόλουθοι:

- **initdata:** Είναι ο πίνακας στον οποίο αποθηκεύεται η αρχική πληροφορία.
- **cid_ip:** Καταγράφεται το ip address του χρήστη. Από το ip address μετέπειτα μπορούμε να βρούμε τη γεωγραφική περιοχή του χρήστη.
- **cosine_sim:** Είναι ο πίνακας όπου υπάρχει η ομοιότητα συνημιτόνου μεταξύ των άρθρων.
- **day_profile:** Είναι ο πίνακας στον οποίο καταγράφεται η ημέρα, το ID της κατηγορίας και πόσες φορές συνολικά διαβάστηκε η κατηγορία αυτή τη δεδομένη ημέρα.
- **nid_profile:** Πόσες φορές επιλέχθηκε από ένα άρθρο κάποιο άλλο, ποια άρθρα είναι αυτά και ποιά είναι η ομοιότητα μεταξύ τους.
- **similarity_up:** Ποιά είναι η ομοιότητα επί της εκατό μεταξύ δύο χρηστών.
- **tid_profile:** Ποια είναι η επόμενη κατηγορία που επιλέγουν να διαβάσουν οι χρήστες και πόσες φορές έχει γίνει.
- **time_profile:** Στις ζώνες που έχουν επιλεγεί, ποια κατηγορία διαβάζεται και πόσες φορές.
- **time_tid:** Ο κάθε χρήστης ποια κατηγορία διάβασε και σε ποια ζώνη.
- **user_from_nid_to_nid:** Για τον κάθε χρήστη, ποιο άρθρο ήταν το επόμενο που επέλεξε να διαβάσει.
- **user_from_to:** Ο κάθε χρήστης, ποια ήταν η επόμενη κατηγορία που επέλεξε να διαβάσει.
- **user_has_read:** Ποια είναι τα άρθρα που διάβασε ο χρήστης και ποια χρονική στιγμή.
- **user_profile:** Ο κάθε χρήστης πόσες φορές έχει διαβάσει μια κατηγορία άρθρων.

Η αρχιτεκτονική της βάσης δεδομένων του συστήματος είναι η ακόλουθη:



Εικόνα 15. Βάση δεδομένων συστήματος συστάσεων

Κεφάλαιο 5

Αποτελέσματα χρήσης του συστήματος

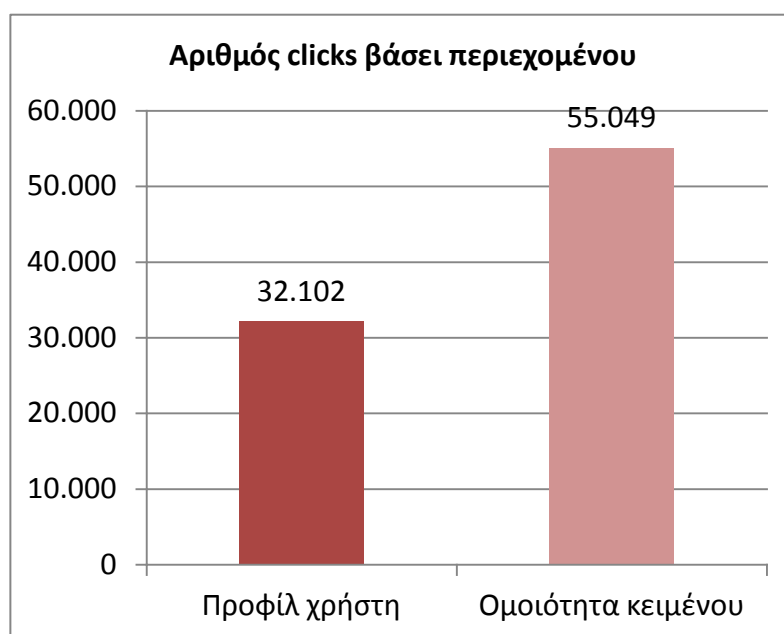
Στο κεφάλαιο αυτό παρουσιάζονται τα αποτελέσματα της χρήσης του συστήματος που σχεδιάστηκε και χρησιμοποιήθηκε για τον ιστότοπο athensvoice.gr και διαχειρίστηκε πραγματικά άρθρα και πραγματικούς χρήστες. Για τον υπολογισμό των αποτελεσμάτων έγινε χρήση στατιστικών (τα οποία αποθηκεύει το ίδιο το σύστημα) όπως είναι ο αριθμός των clicks που έγιναν από τους χρήστες στις προτάσεις του συστήματος, η συμπεριφορά και η ομοιότητα των χρηστών, αλλά και στατιστικών από το Google Analytics. Γίνεται σύγκριση αποτελεσμάτων μόνο με χρήση φιλτραρίσματος βάσει περιεχομένου, συνεργατικού φιλτραρίσματος και υβριδικού φιλτραρίσματος. Παρουσιάζονται γραφήματα τα οποία έχουν σχέση με ορισμένα από τα κριτήρια που επιλέχθηκαν για την εξαγωγή των αποτελεσμάτων. Τέλος πραγματοποιείται σύγκριση των αποτελεσμάτων με προηγούμενες περιόδους, όπου δεν υπήρχε το εν λόγω σύστημα με τη χρήση των στατιστικών από το google analytics.

5.1 Σύγκριση με άλλες τεχνικές και απόδοση του συστήματος με χρήση των δεδομένων του συστήματος

Για την αποτελεσματικότητα του συστήματος, έγινε σύγκριση με άλλες τεχνικές, επιλέχθηκε για μία εβδομάδα το σύστημα να κάνει μόνο προτάσεις βάσει περιεχομένου και για άλλη μία εβδομάδα να γίνονται μόνο προτάσεις με συνεργατικό φιλτράρισμα. Τα δεδομένα που χρησιμοποιήθηκαν είναι από τους χρήστες οι οποίοι αλληλεπιδράσαν με το σύστημα (επέλεξαν κάποιο άρθρο που τους πρότεινε το σύστημα). Βάσει περιεχομένου είναι το 1ο κριτήριο, το οποίο είναι το προφίλ του κάθε χρήστη και το 5ο κριτήριο, που είναι η ομοιότητα συνημιτόνου μεταξύ των άρθρων που υπάρχουν στον ιστότοπο.

Στο συνεργατικό φιλτράρισμα ανήκουν το 2ο κριτήριο (που είναι το προφίλ των κατηγοριών που επιλέγουν οι περισσότεροι χρήστες), το 3ο κριτήριο (που είναι η επιλογή άρθρων σύμφωνα με την ώρα που ο χρήστης αλληλεπιδρά με το σύστημα), το 4ο κριτήριο (που είναι τα πιο διαβασμένα άρθρα), το 6ο κριτήριο (που είναι η ομοιότητα μεταξύ των χρηστών), το 7ο κριτήριο (που είναι οι διαφορετικές κατηγορίες μεταξύ χρηστών με μεγάλη ομοιότητα μεταξύ τους) και το 8ο κριτήριο (που είναι τι διαβάζουν οι Έλληνες χρήστες του Twitter). Τα αποτελέσματα των μετρήσεων εμφανίζονται παρακάτω.

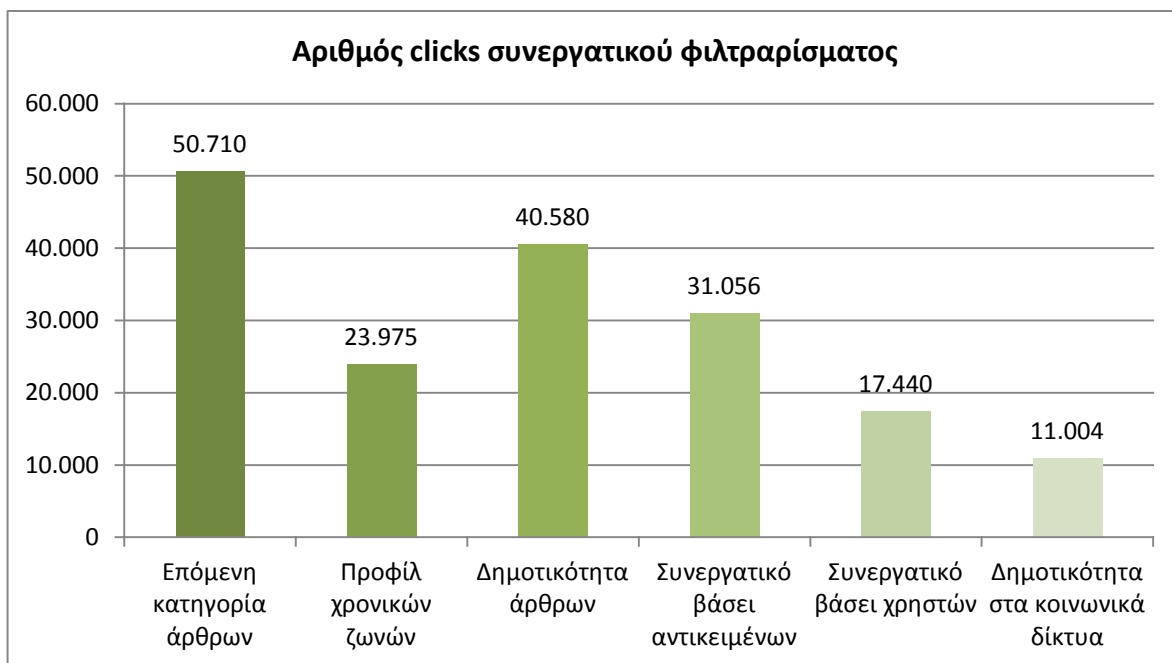
Στατιστικά προτάσεων βάσει περιεχομένου σε διάρκεια μιας εβδομάδας.



Εικόνα 16. Αριθμός clicks βάσει περιεχομένου σε διάρκεια μίας εβδομάδας.

Από τα αποτελέσματα λειτουργίας του συστήματος σε διάρκεια μίας εβδομάδας με φιλτράρισμα βάσει περιεχομένου, ο αριθμός των clicks που έγιναν στα άρθρα που προτάθηκαν από το σύστημα ήταν 87.151. Αν τον αριθμό αυτό τον αναγάγουμε σε μήνα τότε γίνεται 373.504. Παρατηρούμε ότι υπάρχει μια αύξηση της τάξης του 69% σε σχέση με το ποσοστό που κατέχει το φιλτράρισμα βάσει περιεχομένου με την χρήση και των οχτώ κριτηρίων, όπως παρουσιάζεται στην εικόνα 20. Η αύξηση αυτή είναι λογική γιατί τα δύο αυτά κριτήρια μοιράζονται 15 θέσεις αποτελεσμάτων.

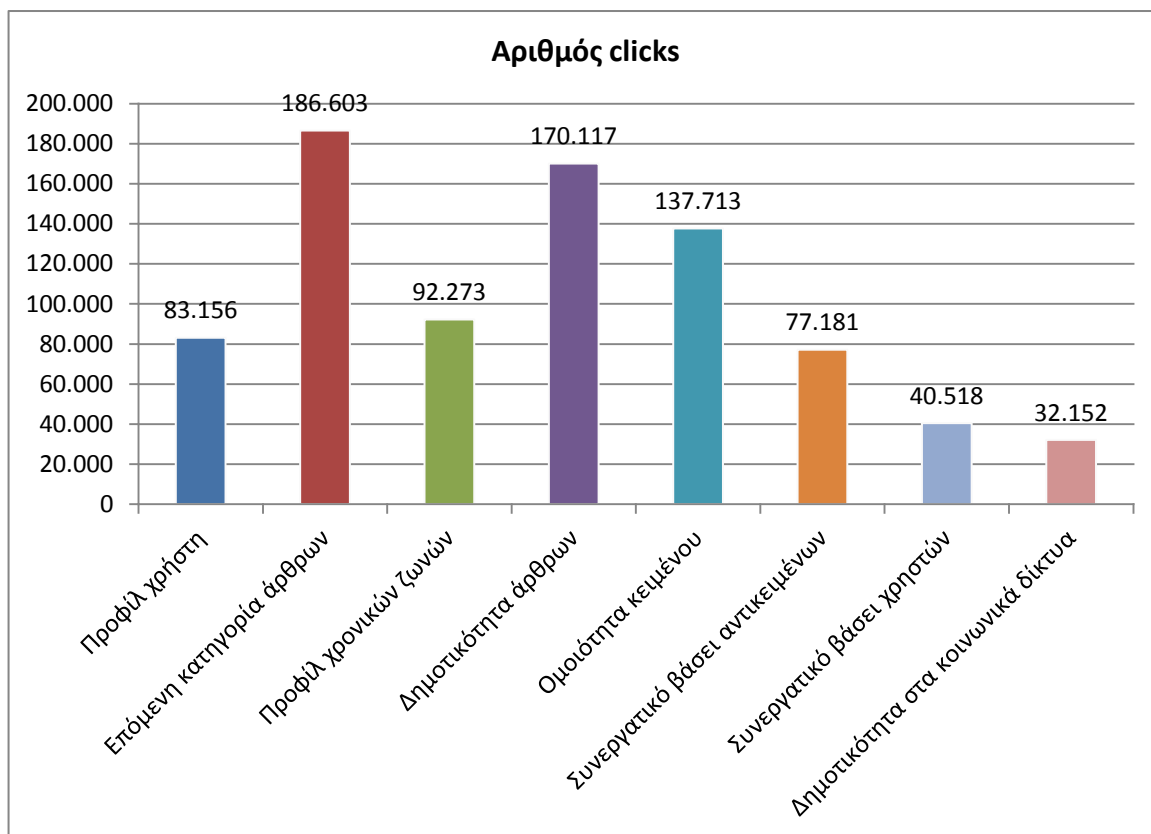
Στατιστικά προτάσεων συνεργατικού φιλτραρίσματος σε διάρκεια μιας εβδομάδας.



Εικόνα 17. Αριθμός clicks με συνεργατικό φιλτράρισμα μίας εβδομάδας.

Από τα αποτελέσματα λειτουργίας του συστήματος σε διάρκεια μίας εβδομάδας με χρήση συνεργατικού φιλτραρίσματος, ο αριθμός των clicks που έγιναν στα άρθρα που προτάθηκαν ήταν 174.765. Αν τον αριθμό αυτόν τον αναγάγουμε σε μήνα τότε γίνεται 748.993. Παρατηρούμε ότι υπάρχει μια αύξηση της τάξης του 25% σε σχέση με το ποσοστό που κατέχει το συνεργατικό φιλτράρισμα με την χρήση και των οχτώ κριτηρίων, όπως παρουσιάζεται στην εικόνα 20. Η αύξηση αυτή είναι λογική γιατί τα έξι αυτά κριτήρια μοιράζονται 15 θέσεις αποτελεσμάτων. Συγκρίνοντας με τα συνολικά αποτελέσματα (εικόνα 20), αν το σύστημα έτρεχε μόνο με συνεργατικό φιλτράρισμα θα είχαμε μια πτώση 70.720 clicks (ποσοστό 8,63%) σε σχέση με τη λειτουργία και των οχτώ κριτηρίων.

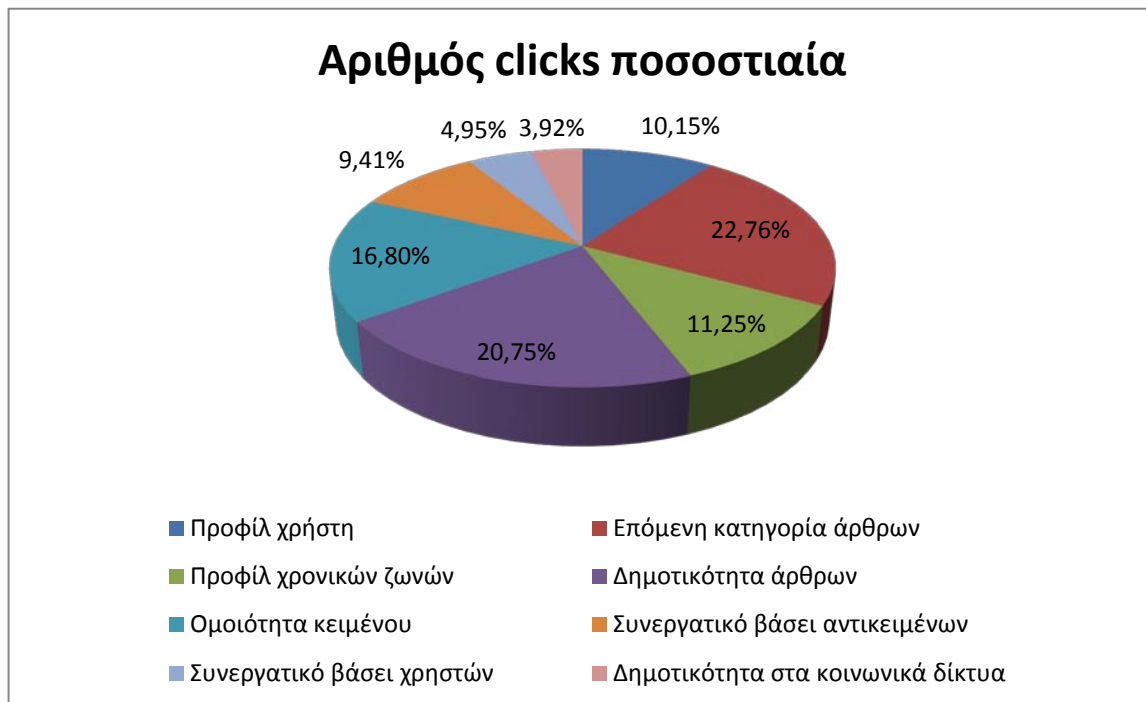
Στατιστικά προτάσεων με χρήση και των οχτώ κριτηρίων σε διάρκεια ενός μήνα.



Εικόνα 18. Αριθμός clicks ενός μήνα με όλα τα κριτήρια.

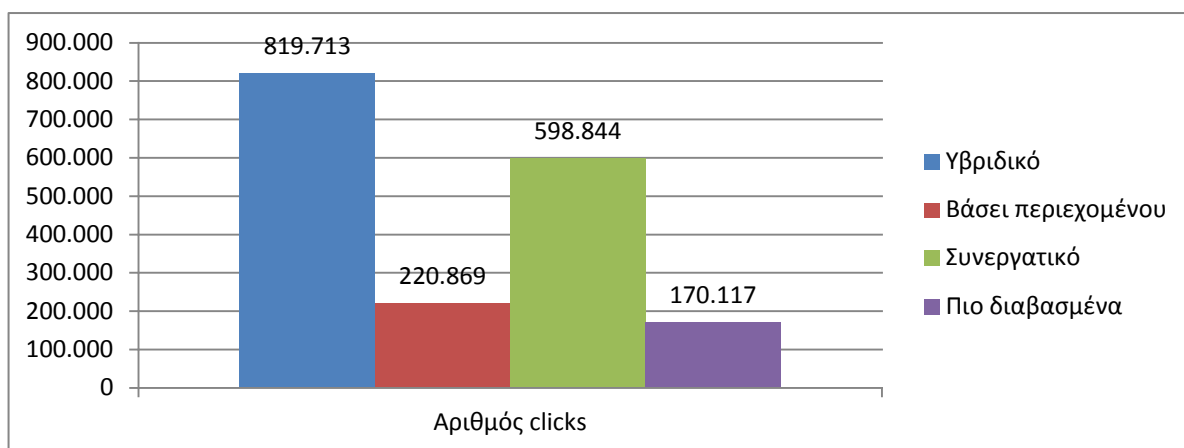
Εντύπωση προκαλεί η τελευταία θέση του κριτηρίου 8, το οποίο εμφανίζει τα twitter trends και είναι σε πλεονεκτικότερη θέση σε σχέση με τα υπόλοιπα, λόγω της διαφοροποίησης στο τρόπο εμφάνισης προς τον χρήστη.

Τα ποσοστά των clicks που έγιναν για το κάθε ένα κριτήριο σε διάρκεια ενός μήνα.



Εικόνα 19. Ποσοστά εμφάνισης κάθε κριτηρίου.

Παρακάτω παρουσιάζονται τα συνολικά αποτελέσματα της κάθε προσέγγισης σε διάρκεια ενός μήνα.



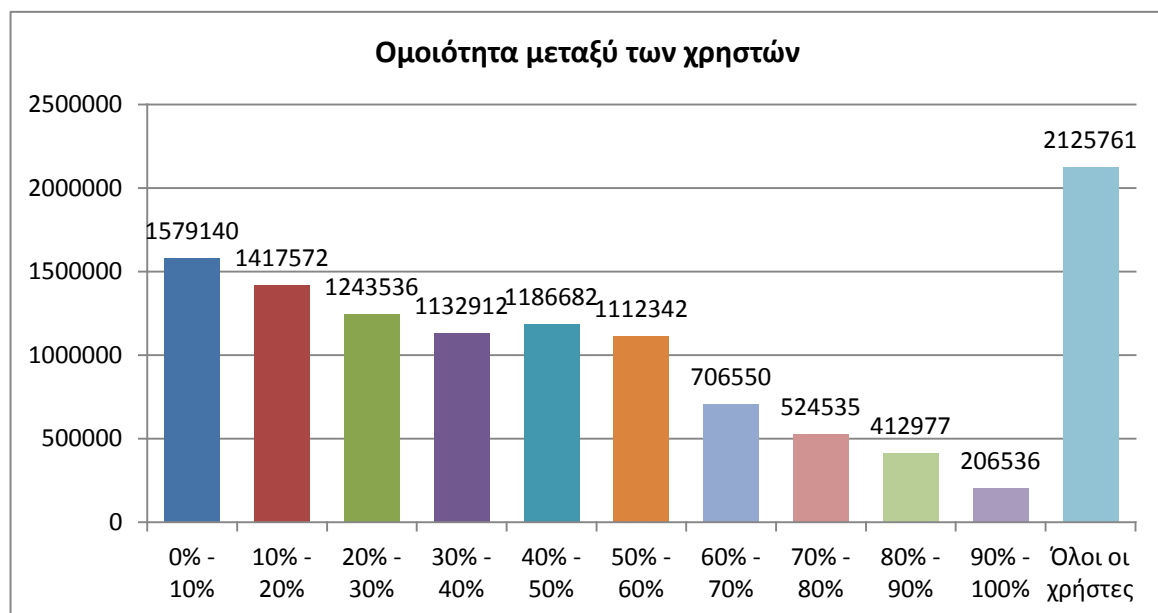
Εικόνα 20. Συνολικός αριθμός clicks με τις διαφορετικές προσεγγίσεις.

Στην πρώτη στήλη εμφανίζεται ο συνολικός αριθμός των clicks που έγιναν στο σύστημα σε πλήρη λειτουργία, δηλαδή με τη χρήση και των οχτώ κριτηρίων. Από αυτά το 26,49%

αφορά αποτελέσματα φιλτραρίσματος βάσει περιεχομένου και απεικονίζονται στη δεύτερη στήλη, ενώ σε ποσοστό 73,05% είναι το συνεργατικό φιλτράρισμα που εμφανίζεται στη τρίτη στήλη. Τέλος στη τέταρτη στήλη, για λόγους σύγκρισης (ο αριθμός των clicks που θα γινότανε αν δεν υπήρχε το σύστημα), εμφανίζεται ο αριθμός των clicks που έγινε στα πιο διαβασμένα άρθρα, δηλαδή το κριτήριο 4. Συμπεραίνουμε ότι με τη χρήση του συστήματος συστάσεων που σχεδιάστηκε και υλοποιήθηκε, υπάρχει μια αύξηση 480% στον αριθμό των clicks που γίνεται στα άρθρα που εμφανίζονται στη δεδομένη θέση.

5.2 Στατιστικά στοιχεία συμπεριφοράς των χρηστών και γραφήματα που σχετίζονται με ορισμένα από τα κριτήρια επιλογής αποτελεσμάτων.

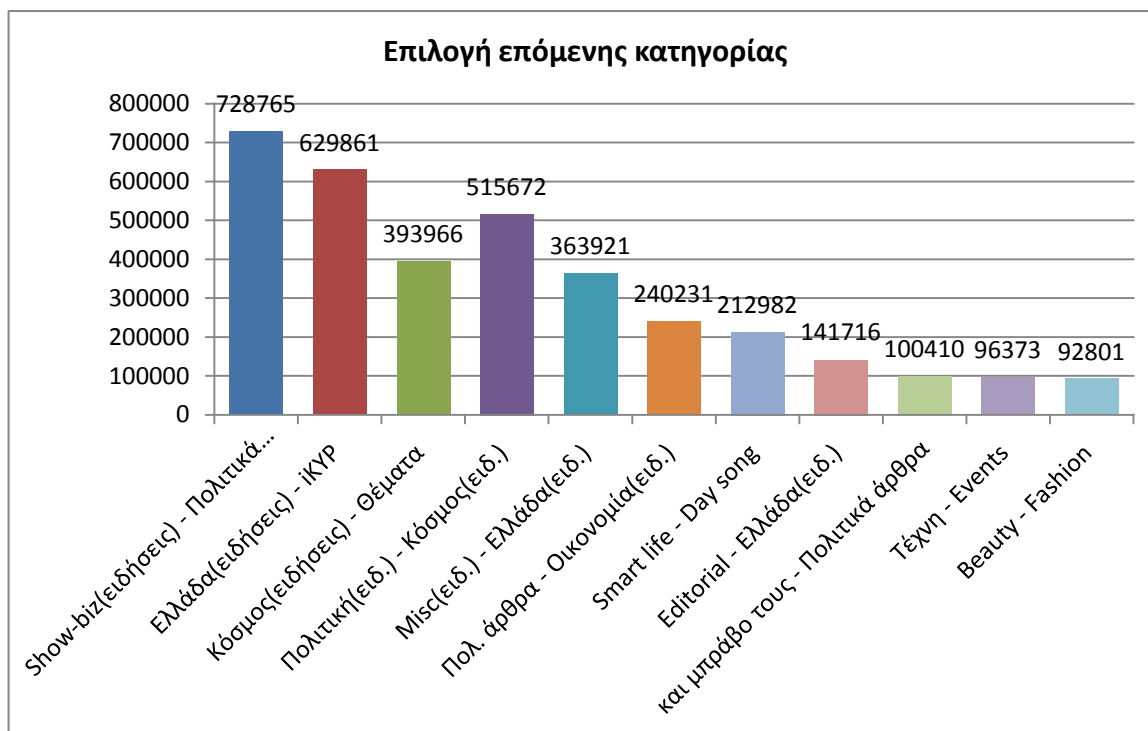
Από τη λειτουργία του συστήματος συστάσεων άρθρων, τα δεδομένα που συλλέχτηκαν μας παρέχουν την δυνατότητα να εξάγουμε συμπεράσματα για το κατά πόσο στα κριτήρια που επιλέχτηκαν υπάρχουν πρότυπα, για την συμπεριφορά των χρηστών και κατά πόσο οι χρήστες ενός ιστότοπου έχουν κοινά αναγνωστικά ενδιαφέροντα.



Εικόνα 21. Ποσοστά ομοιότητας χρηστών.

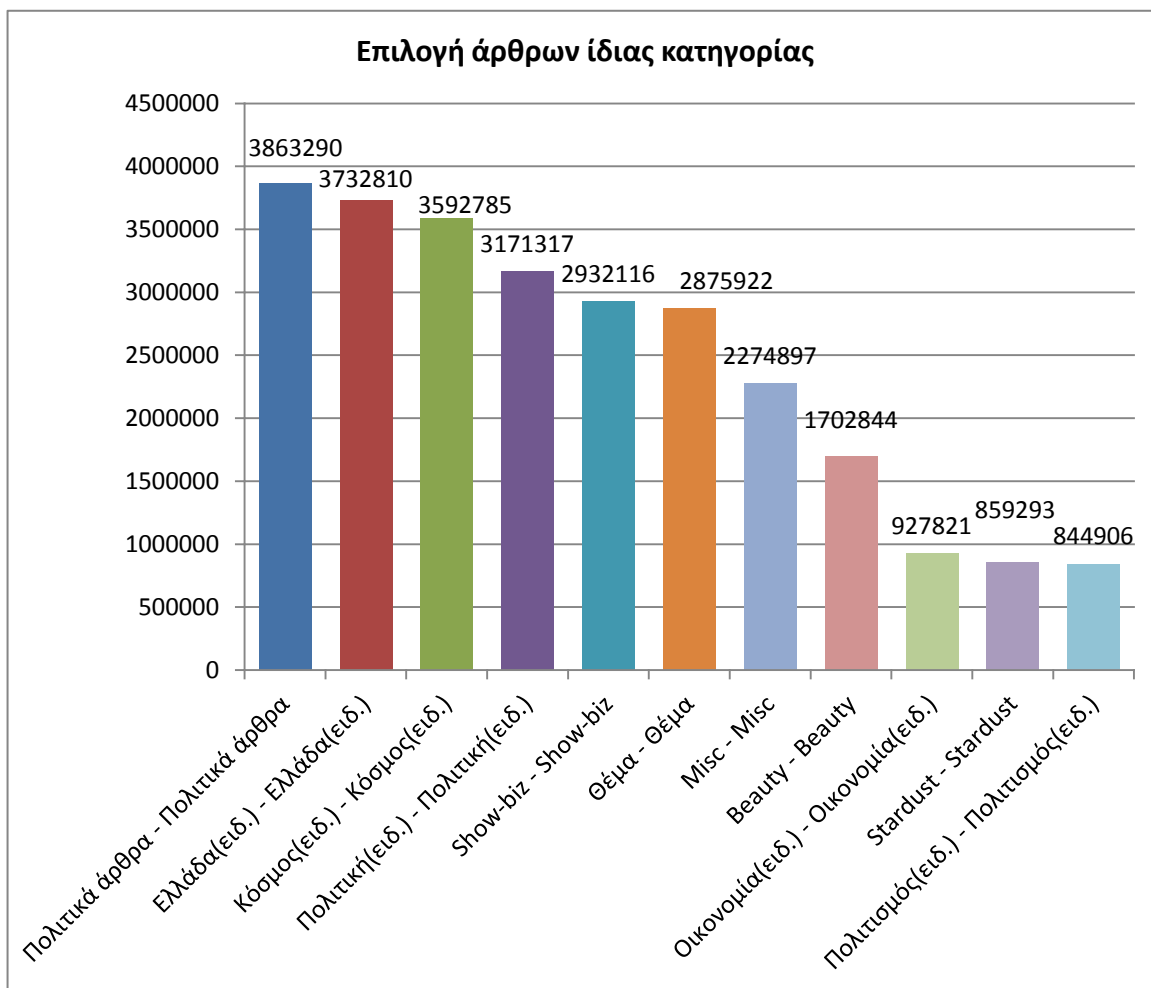
Από την παραπάνω εικόνα που σχετίζεται με τα κριτήρια 6 και 7, παρατηρούμε ότι, σε ποσοστό μεγαλύτερο του 50%, υπάρχουν χρήστες με κοινά αναγνωστικά ενδιαφέροντα με ποσοστά ομοιότητας από 30% - 60%. Επίσης υπάρχει ένα ποσοστό περίπου 10% που η ομοιότητα του κυμαίνεται μεταξύ του 90% και 100%. Και τα δύο κριτήρια μαζί

αντιπροσωπεύουν το 14,37% του συνόλου των χρηστών, που επέλεξαν να διαβάσουν άρθρα, τα οποία τους πρότεινε το σύστημα συστάσεων.



Εικόνα 22. Επιλογή άρθρων επόμενης κατηγορίας.

Στο παραπάνω γράφημα (εικόνα 22) εμφανίζονται τα αποτελέσματα του κριτηρίου 2, παρουσιάζονται οι επόμενες κατηγορίες που επιλέγουν να διαβάσουν οι χρήστες όταν είναι σε διαφορετική κατηγορία, δηλαδή από την κατηγορία Α στη κατηγορία Β. Τα δεδομένα αφορούν την συμπεριφορά των χρηστών στον ιστότοπο γενικότερα και όχι με την επιλογή των άρθρων που προτείνονται από το σύστημα. Το κριτήριο 2 "επιλογής επόμενης κατηγορίας" είναι το κριτήριο το οποίο είναι το πρώτο σε επιλογές ανάγνωσης άρθρων από τους χρήστες, σύμφωνα με το γράφημα της εικόνας 18. Παρατηρούμε ότι υπάρχει λογική για την επιλογή της επόμενης κατηγορίας άρθρου, εκτός ίσως την πρώτη στήλη, όπου από ειδήσεις show-biz, η επόμενη επιλογή είναι πολιτικά άρθρα.

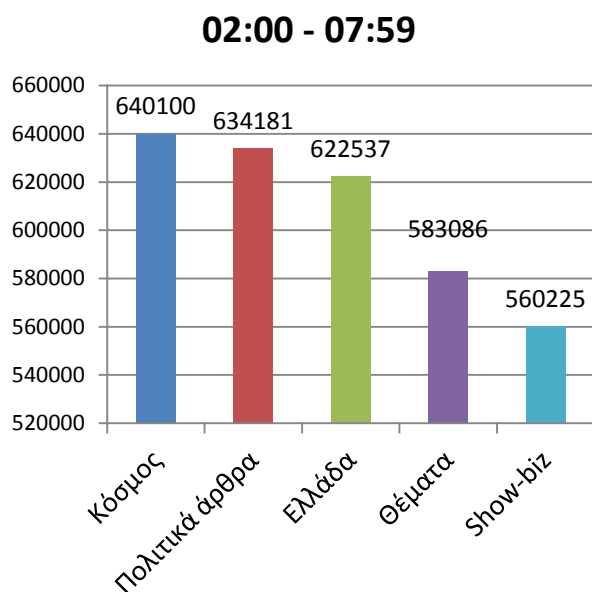


Εικόνα 23. Επιλογή άρθρων ίδιας κατηγορίας.

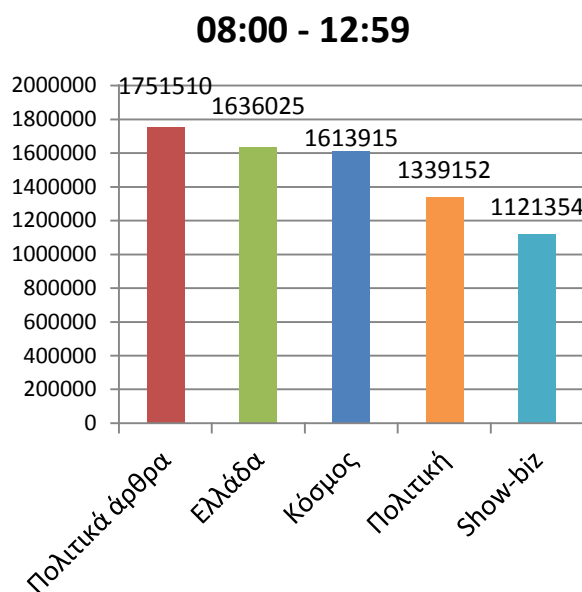
Στο παραπάνω γράφημα (εικόνα 23) παρουσιάζονται οι επόμενες κατηγορίες που επιλέγουν να διαβάσουν οι χρήστες, χωρίς τον περιορισμό οι κατηγορίες να είναι διαφορετικές. Τα δεδομένα αφορούν την συμπεριφορά των χρηστών στον ιστότοπο γενικότερα και όχι με την επιλογή των άρθρων που προτείνονται από το σύστημα. Παρατηρούμε ότι ο αριθμός της επιλογής άρθρων ίδιας κατηγορίας είναι αισθητά μεγαλύτερος από τον αριθμό επιλογής άρθρων διαφορετικής κατηγορίας. Οι λόγοι που δεν επιλέχθηκε η επιλογή άρθρων ίδιας κατηγορίας είναι αφενός γιατί στον ιστότοπο υπάρχει η συγκεκριμένη λειτουργικότητα και αφετέρου γιατί τα άρθρα που ανήκουν στην ίδια κατηγορία εμφανίζονται από τα άλλα κριτήρια, αν συντρέχουν οι προϋποθέσεις που ισχύουν στα υπόλοιπα κριτήρια.

Στις παρακάτω εικόνες (24,25,26,27,28) παρουσιάζονται οι επιλογές κατηγοριών από τους χρήστες, ανάλογα την χρονική ζώνη που επισκέπτονται τον ιστότοπο που αφορά το κριτήριο 3. Τα γραφήματα μας δείχνουν τί διαβάζουν οι χρήστες κατά την διάρκεια της ημέρας. Παρατηρούμε ότι ανάλογα τη χρονική ζώνη υπάρχει μια διαφοροποίηση

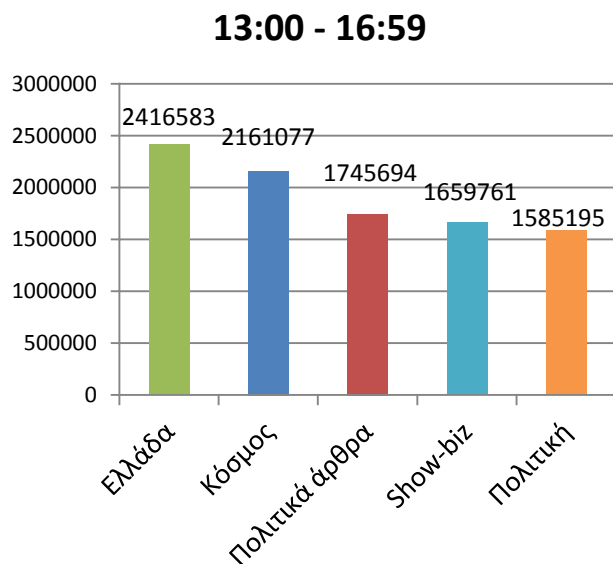
στις επιλογές των χρηστών. Στο γράφημα της εικόνα 24 βλέπουμε ότι στη πεντάδα υπάρχει η στήλη "Θέματα", επίσης θα πρέπει να αναφερθεί ότι στη συγκεκριμένη χρονική ζώνη οι περισσότεροι χρήστες του ιστότοπου είναι από το εξωτερικό. Στο γράφημα της εικόνας 25 την πρώτη θέση την έχει η στήλη "Πολιτικά άρθρα" και οι υπόλοιπες στήλες αφορούν κατηγορίες των ειδήσεων, δηλαδή είναι η χρονική ζώνη που οι χρήστες θέλουν να ενημερωθούν για το τι γίνεται στον κόσμο, στην Ελλάδα και στη πολιτική σκηνή. Στα επόμενα δύο γραφήματα (εικόνες 26,27) επικρατεί η ίδια εικόνα με το προηγούμενο γράφημα (εικόνα 25). Στο τελευταίο γράφημα (εικόνα 28) παρατηρούμε ότι υπάρχει η στήλη "Θέματα", όπως και μια πιο χαλαρή διάθεση στην επιλογή ανάγνωσης άρθρων. Υπάρχει η στήλη "Θέματα" δεν υπάρχει η στήλη "Πολιτικά άρθρα" και οι ειδήσεις της πολιτικής είναι στην πέμπτη θέση.



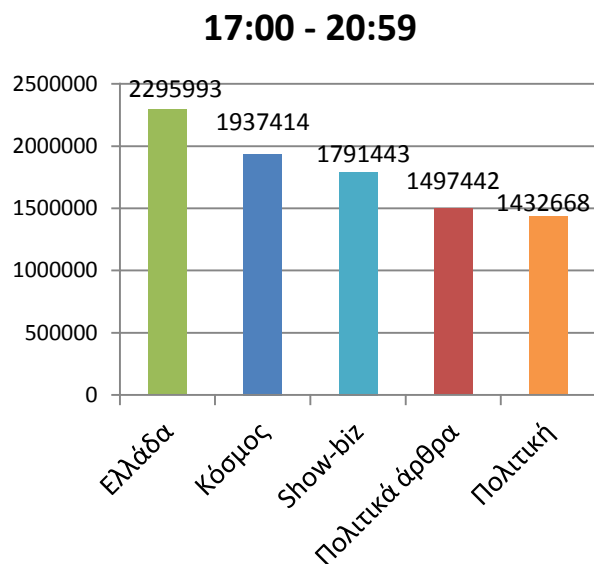
Εικόνα 24. Επιλογή κατηγοριών στη χρονική ζώνη 02:00 - 07:59.



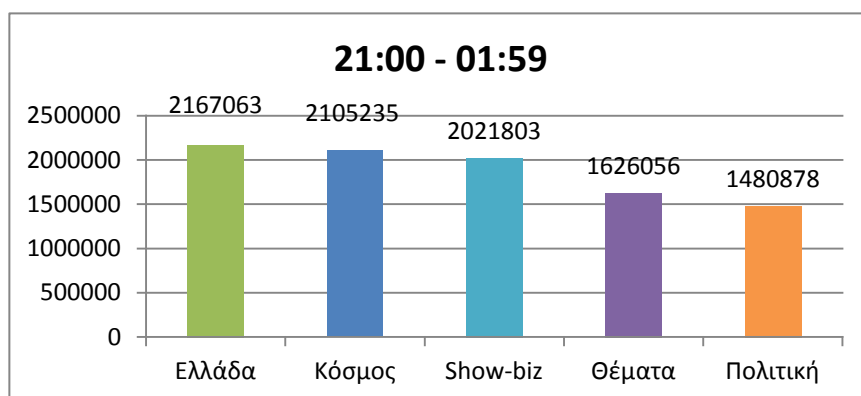
Εικόνα 25. Επιλογή κατηγοριών στη χρονική ζώνη 08:00 - 12:59.



Εικόνα 26. Επιλογή κατηγοριών στη χρονική ζώνη 13:00 - 13:59.



Εικόνα 27. Επιλογή κατηγοριών στη χρονική ζώνη 17:00 - 20:59.

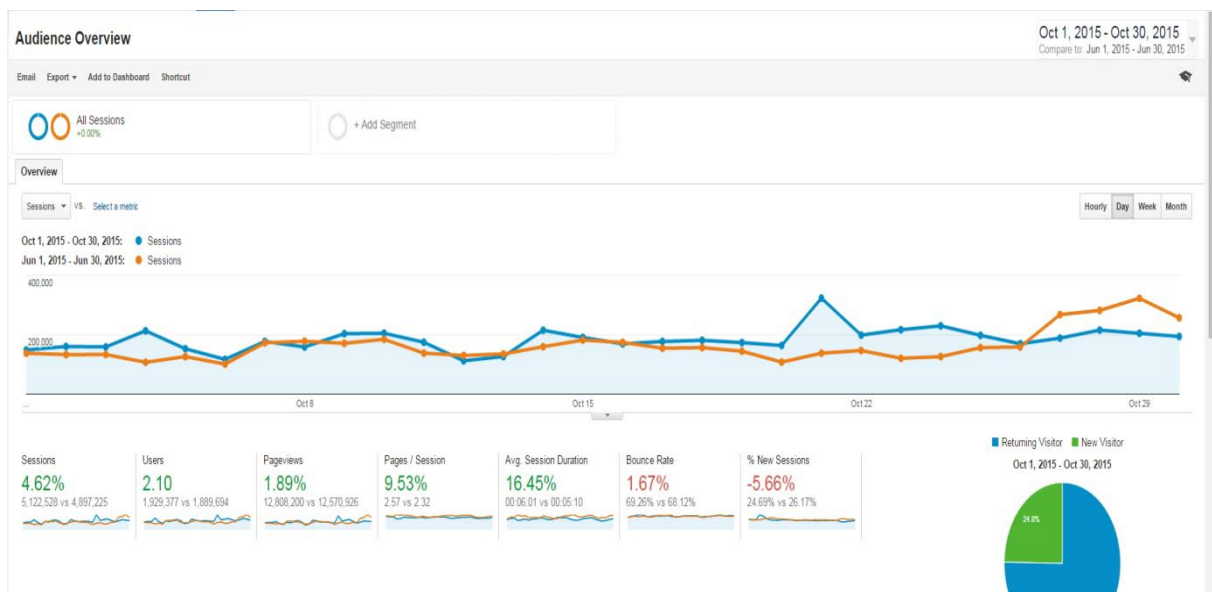


Εικόνα 28. Επιλογή κατηγοριών στη χρονική ζώνη 21:00 - 01:59.

5.3 Συνολικά αποτελέσματα με χρήση του google analytics

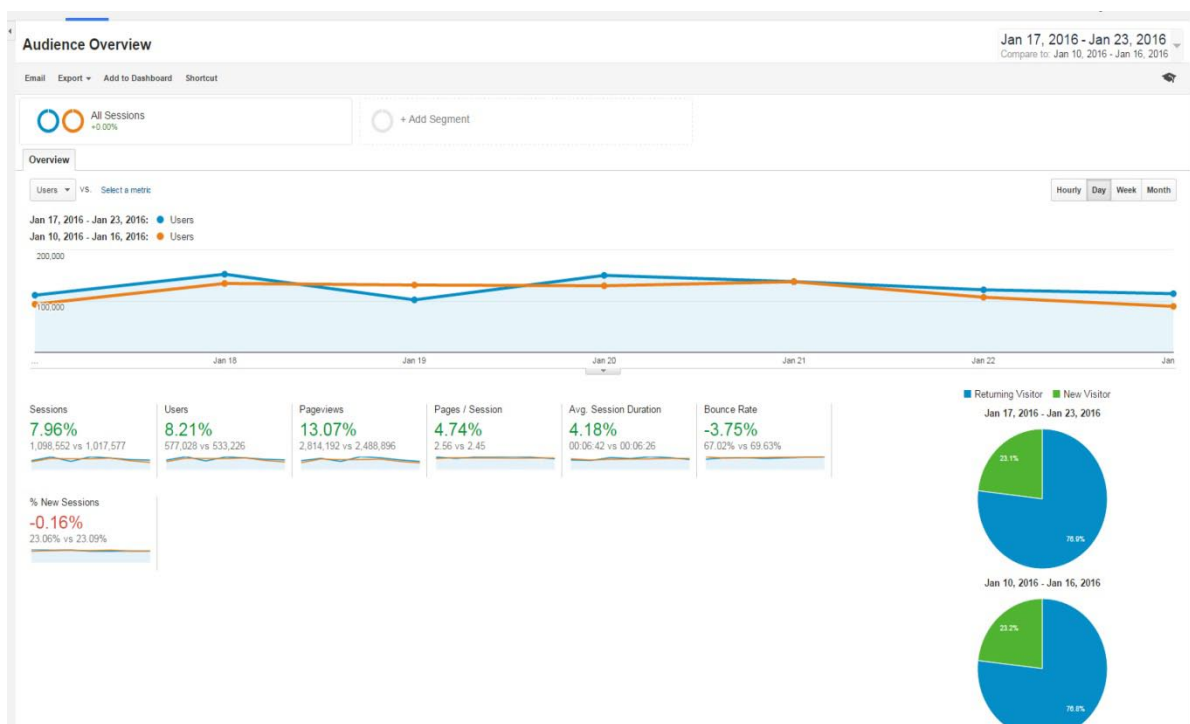
Τα οφέλη του συστήματος στον ιστότοπο υπολογίζονται με την χρήση του Google Analytics, συγκρίνοντας το χρονικό διάστημα λειτουργίας τους (ένα μήνα) σε σχέση με κάποιο άλλο μήνα, στον οποίο το σύστημα συστάσεων δεν ήταν σε λειτουργία. Επίσης για διάρκεια μίας εβδομάδας το σύστημα βγήκε εκτός λειτουργίας, ώστε η σύγκριση να γίνει στον ίδιο μήνα.

Ακολουθεί εικόνα με τα αποτελέσματα του google analytics για διάρκεια ενός μήνα λειτουργίας του συστήματος, σε σύγκριση με κάποιον μήνα που δεν ήταν σε λειτουργία το σύστημα.



Εικόνα 29. Στατιστικά μήνα με το σύστημα και χωρίς.

Ακολουθεί εικόνα με τα αποτελέσματα του google analytics για διάρκεια μίας εβδομάδας λειτουργίας του συστήματος, σε σύγκριση με την προηγούμενη εβδομάδα που δεν ήταν σε λειτουργία.



Εικόνα 30. Στατιστικά εβδομάδας με και χωρίς το σύστημα.

Τις ημερομηνίες 17 - 23 Ιανουαρίου το σύστημα ήταν σε λειτουργία (μπλε γράφημα) και τις ημερομηνίες 10 - 16 Ιανουαρίου το σύστημα δεν ήταν σε λειτουργία (πορτοκαλί γράφημα). Παρατηρούμε ότι τις ημερομηνίες που ήταν σε λειτουργία, υπήρχε μια αύξηση 4.18% στο χρόνο παραμονής και μια αύξηση 4.74% στο μέσο όρο των σελίδων που διαβάζονται σε κάθε επίσκεψη.

Κεφάλαιο 6

Επίλογος

Σε αυτή την διατριβή αναπτύχθηκε ένα υβριδικό σύστημα συστάσεων περιεχομένου, βασισμένο σε οκτώ κριτήρια, για τον υπολογισμό των προτάσεων περιεχομένου στους χρήστες. Αναπτύχθηκε ένα υποσύστημα ανάδρασης χρήστη, ώστε το σύστημα να είναι προσαρμοστικό για τον κάθε χρήστη ξεχωριστά και να υπολογίζει τον αριθμό των άρθρων για κάθε κριτήριο που θα προτείνει στον κάθε χρήστη.

Το σύστημα που αναπτύχθηκε για τη διατριβή, μπήκε σε λειτουργία στον ιστότοπο athensvoice.gr και έκανε προτάσεις από άρθρα τα οποία υπάρχουν στον ιστότοπο και ανανεώνονται συνεχώς, σε πραγματικό χρόνο στους χρήστες του ιστότοπου.

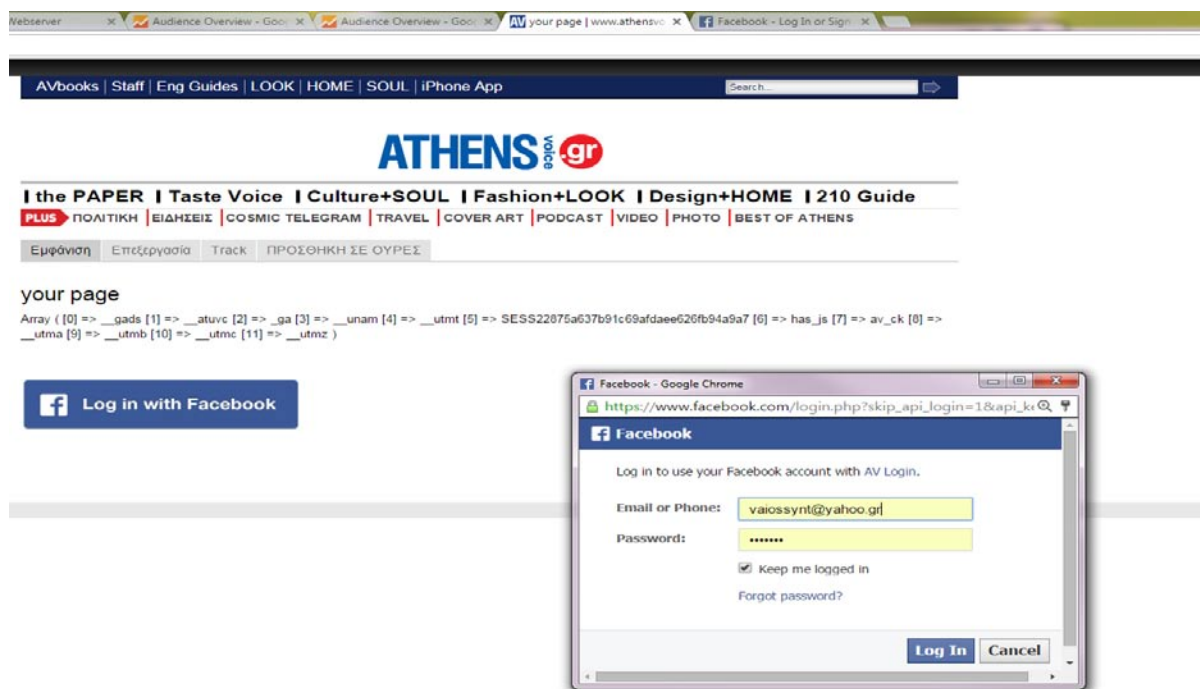
6.1 Συμπεράσματα

Από τα αποτελέσματα που παρουσιάστηκαν στο Κεφάλαιο 5, συμπεραίνουμε ότι το σύστημα που σχεδιάστηκε, αναπτύχθηκε και μπήκε σε λειτουργία στον ιστότοπο athensvoice.gr, βελτίωσε τα νούμερα των ποιοτικών χαρακτηριστικών του ιστότοπου. Με τη χρήση του υβριδικού συστήματος συστάσεων άρθρων βασισμένου στα κριτήρια που επιλέχτηκαν, υπήρχε βελτίωση σε σχέση με την χρήση άλλων τεχνικών. Τεχνικών όπως συστήματα βασισμένα στο περιεχόμενο, συστήματα βασισμένα στο συνεργατικό φιλτράρισμα ή άρθρων βασισμένα στη δημοτικότητα (των πιο διαβασμένων άρθρων).

6.2 Μελλοντικές εργασίες

Μελλοντικά το σύστημα συστάσεων θα προσφέρει στους χρήστες που το χρησιμοποιούν μια προσωποποιημένη σελίδα, η οποία θα τους εμφανίζει μόνο τα άρθρα που τους ενδιαφέρουν. Στη σελίδα αυτή θα πρέπει να κάνουν login με τον λογαριασμό που διαθέτουν στο facebook, ώστε να είναι σε θέση να δουν τις προτάσεις. Από τη στιγμή που ο χρήστης κάνει login με τον λογαριασμό του, αυτομάτως το σύστημα αποκτάει πρόσβαση σε πολύτιμα δεδομένα για τη λειτουργία του συστήματος, όπως φίλους, ηλικία και όποια άλλη πληροφορία στην οποία επιτρέπει πρόσβαση το facebook.

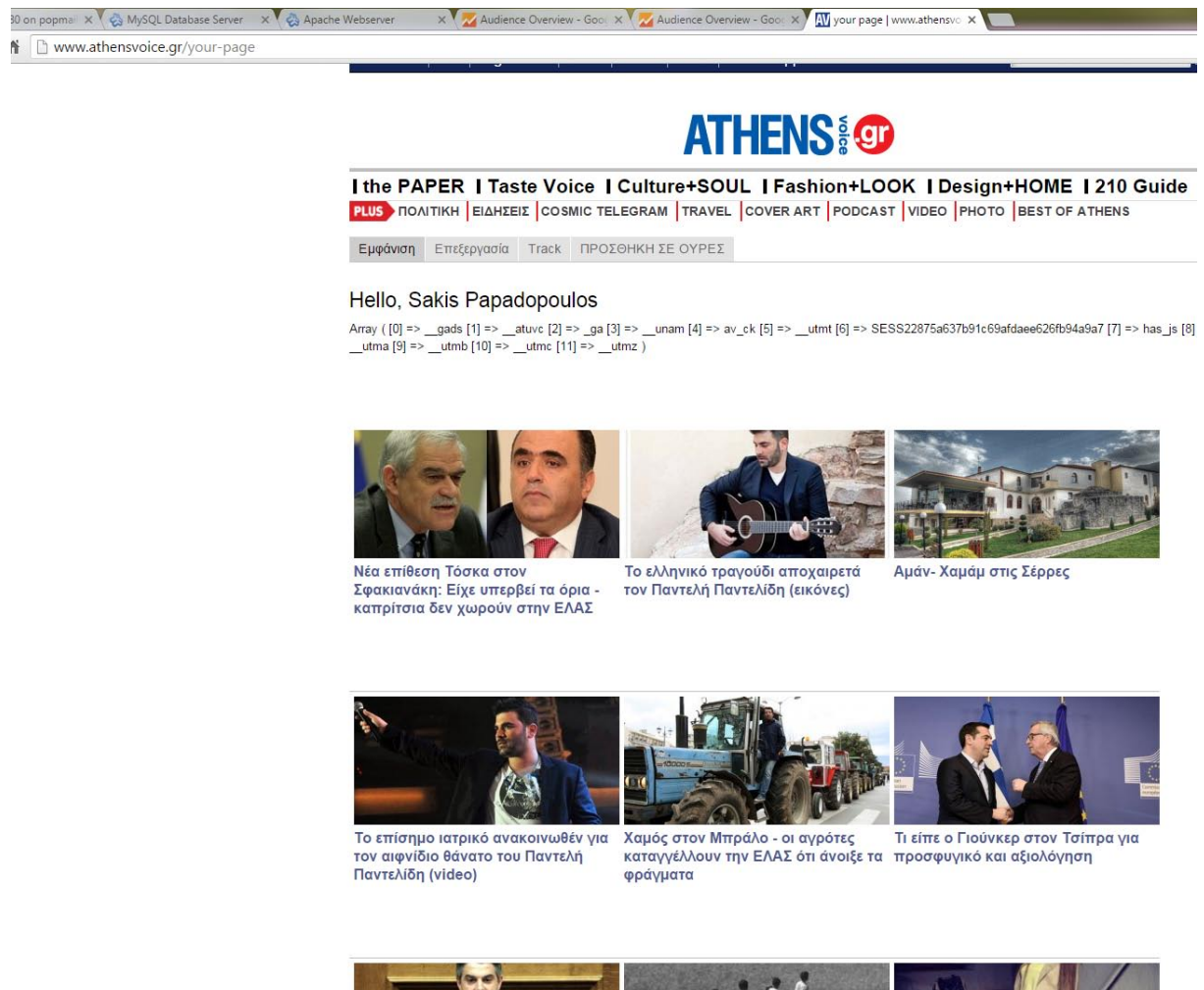
Με αυτό τον τρόπο, το σύστημα θα είναι σε θέση να εμφανίζει προτάσεις σύμφωνα με το τι διαβάζουν οι φίλοι του χρήστη στο athenstvoice.gr και δεν έχει διαβαστεί από τον ίδιο, ποια άρθρα διαβάζουν άτομα που ανήκουν στο ίδιο ηλικιακό group και, το κυριότερο, από τη στιγμή που ο χρήστης έχει κάνει login, το σύστημα θα είναι σε θέση να μην του εμφανίζει ξανά τα άρθρα τα οποία διάβασε από το κινητό του τηλέφωνο ή από κάποιον άλλο υπολογιστή. Ακολουθούν εικόνες με την υπό κατασκευή προσθήκη.



Εικόνα 31. Login στο σύστημα.

Η εικόνα (31) απεικονίζει την προτροπή του συστήματος ώστε ο χρήστης να κάνει login με το σύστημα.

Η παρακάτω εικόνα απεικονίζει όταν ο χρήστης έχει κάνει login στο σύστημα.



Εικόνα 32. Προσωπική σελίδα με λογαριασμό facebook.

Επίσης υπάρχει η δυνατότητα προσθήκης του δημογραφικού κριτηρίου. Στη βάση δεδομένων υπάρχει το ip address του κάθε χρήστη, οπότε το σύστημα θα είναι σε θέση να ξέρει σε ποια περιοχή είναι ο χρήστης και να υπολογίζει εάν υπάρχουν άρθρα που αφορούν την περιοχή του ή τι διαβάζουν άλλοι χρήστες που κατοικούν σε αυτή την περιοχή, και να τους τα προτείνει.

Βιβλιογραφία

- [1] Content-based Recommender Systems: State of the Art and Trends - Pasquale Lops, Marco de Gemmis and Giovanni Semeraro.
- [2] Tutorial: Recommender Systems, International Joint Conference on Artificial Intelligence Deijing. August 4, 2013.
- [3] Collaborative Filtering Recommender Systems By Michael D. Ekstrand, John T. Riedl and Joseph A. Konstan, University of Minnesota.
- [4] J. Ahn, P. Brusilovsky, J. Grady, and D. He. Open user profiles for adaptive news systems: help or harm? In Conf. on WWW, pages 11-20, 2007.
- [5] D. Billsus and M. Pazzani. A hybrid user model for news story classification. In Conf. on User Modeling, pages 99-108, 1999.
- [6] W. IJntema, F. Goossen, F. Frasincar, and F. Hogenboom. Ontology-based news.
- [7] L. Li, W. Chu, J. Langford, and R. Schapire. A contextual-bandit approach to personalized news article recommendation. In Conf. on WWW, 2010.
- [8] J. Liu, P. Dolan, and E. Pedersen. Personalized news recommendation based on click behavior. In Proc. of IUI, pages 31-40, 2010.
- [9] A. Das, M. Datar, A. Garg. Google news personalization: Scalable online collaborative filtering, pages 271-280, May 2007.
- [10] Liang, T.-P. and Lai, H.-J. Discovering User Interests from Web Browsing Behavior: An Application to Internet News Services, IEEE Computer Society, Los Alamitos, CA, USA, 2002.
- [11] E. Kirshenbaum, G. Forman, and M. Dugan. A live comparison of methods for personalized article recommendation at forbes.com. In ECML - PKDD, pages 51-66, 2012.
- [12] Jun Wang, Arjen P. de Vries, and Marcel J. T. Reinders. 2006. Unifying user-based and item-based collaborative filtering approaches by similarity fusion.

- [13] Zhongqi Lu and Zhicheng Dou and Jianxun Lian and Xing Xie and Qiang Yang. 2015 . Content-Based Collaborative Filtering for News Topic Recommendation. AAAI Conference on Artificial Intelligence.
- [14] Richard McCreadie, Craig Macdonald, and Iadh Ounis. 2013. News vertical search: when and what to display to users. 36th international ACM SIGIR conference on Research and development in information retrieval (SIGIR '13). ACM, New York, NY, USA, 253-262.
- [15] Liang Tang, Yexi Jiang, Lei Li, Chunqiu Zeng, and Tao Li. 2015. Personalized Recommendation via Parameter-Free Contextual Bandits. 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '15). ACM, New York, NY, USA, 323 - 332.
- [16] Billsus, D., & Pazzani, M. A hybrid user model for news story classification. In Proceedings of the Seventh International Conference on User Modeling. 1999.
- [17] Tan, A. and Tee, C. "Learning User Profiles for Personalized Information Dissemination," Proceedings of 1998 IEEE International Joint conference on Neural Networks, pp. 183- 188, May 1998
- [18] Joseph Konstan, Brad Miller, David Maltz, Jon Herlocker, Lee Gordon, and John Riedl. Grouplens: Collaborative filtering for use net news. Communications of the ACM, March 1997.
- [19] J. Ben Schafer, Joseph Konstan, and John Riedl. Recommender systems in ecommerce In Proceedings of the ACM Conference on Electronic Commerce (EC '99), 1999.
- [20] Kirsten Swearingen and Sinha Rashmi. Interaction design for recommender systems. In Designing Interactive Systems 2002. ACM, 2002.
- [21] Bradley N. Miller, Istvan Albert, Shyong K. Lam, Joseph A. Konstan, John Riedl GroupLens Research Group: MovieLens Unplug.
- [22] G. Shani and A. Gunawardana. Evaluating recommendation systems. In F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, editors, Recommender Systems Handbook, pages 257–297.

- [23] Joseph A. Konstan, Bradley N. Miller, David Maltz, Jonathan L. Herlocker, Lee R. Gordon, and John Riedl. 1997. GroupLens: Applying Collaborative Filtering to Usenet News. *Commun. ACM* 40, 3: 77–87.
- [24] Ajit P. Singh and Geoffrey J. Gordon. A Unified View of Matrix Factorization Models 2008.
- [25] Matrix factorization techniques for recommender systems. Yehuda Koren, Robert Bell, Chris Volinsky. IEEE Computer Society 2009, pages 42 - 49.

Παράρτημα Α

Περιγραφή βάσης δεδομένων

Στο παράρτημα Α παρουσιάζεται οι πίνακες που απαρτίζουν τη βάση δεδομένων που υλοποιήθηκε για την εφαρμογή.

Column Name	Datatype
 id	 INT(10)
 currentURL	 VARCHAR(256)
 cookieID	 VARCHAR(20)
 timestamp	 VARCHAR(45)
 history	 VARCHAR(256)
 isMobile	 INT(10)

Εικόνα 33. Πίνακας initdata.

Το initdata είναι ο πίνακας στον οποίο αποθηκεύεται όλη η αρχική πληροφορία. Τα πεδία που τον απαρτίζουν είναι τα ακόλουθα:

- id: Πρωτεύων κλειδί του πεδίου.
- currentURL: Το URL (άρθρο) στο οποίο βρίσκεται ο χρήστης.
- cookieID: Το cookie που έχει δώσει το σύστημα στον χρήστη.
- timestamp: Η χρονική στιγμή που έγινε η επίσκεψη του χρήστη σε μορφή δευτερολέπτων.
- history: Από που ήρθε ο χρήστης (άλλο άρθρο ή από άλλο ιστότοπο), σε μορφή URL.
- isMobile: Παίρνει τιμές 1 ή 0, ανάλογα αν η επίσκεψη έγινε από κινητή συσκευή ή όχι.

Column Name	Datatype
id	INT(10)
cookieID	VARCHAR(128)
ip	VARCHAR(45)
city	VARCHAR(45)
country	VARCHAR(45)
region	VARCHAR(45)
isp	VARCHAR(128)

Εικόνα 34. Πίνακας cid_ip.

Στον πίνακα cid_ip καταγράφεται το ip address του χρήστη. Από το ip address μπορούμε να βρούμε τη περιοχή του χρήστη, όπως και τον πάροχο του. Τα πεδία που τον απαρτίζουν είναι τα ακόλουθα:

- id: Πρωτεύων κλειδί.
- cookieID: Το cookie του χρήστη.
- ip: Το ip address του χρήστη.
- city: Η πόλη.
- country: Η χώρα.
- region: Η περιοχή.
- isp: Ποιός είναι ο πάροχος του.

Column Name	Datatype
id	INT(10)
cookieID	VARCHAR(20)
nid	INT(10)
timestamp	VARCHAR(45)
criteria	INT(10)

Εικόνα 35. Πίνακας click.

Ο πίνακας click καταγράφει ποιός χρήστης έκανε click πάνω σε ποίο άρθρο, από το σύνολο των άρθρων που έχει επιλεγεί από το σύστημα.

- id: Το πρωτεύων κλειδί.
- cookieID: Το cookie του χρήστη.
- nid: Το id του άρθρου που έγινε click.

- timestamp: Η χρονική στιγμή που έγινε η πράξη.
- criteria: Ποιό ήταν το κριτήριο εμφάνισης του άρθρου.

Column Name	Datatype
id	INT(10)
nid1	INT(10)
nid2	INT(10)
cos_sim	DECIMAL(6,5)

Εικόνα 36. Πίνακας cosine_sim.

Ο πίνακας cosine_sim αποθηκεύει την ομοιότητα συνημιτόνου μεταξύ δύο άρθρων. Τα πεδία που τον απαρτίζουν είναι τα ακόλουθα.

- id: Το πρωτεύων κλειδί.
- nid1: Το id του ενός άρθρου προς εξέταση.
- nid2: Το id του άλλου άρθρου προς εξέταση.
- cos_sim: Η ομοιότητα συνημιτόνου των συγκεκριμένων άρθρων.

Column Name	Datatype
id	INT(10)
day	VARCHAR(45)
tid	INT(10)
tidCount	INT(10)

Εικόνα 37. Πίνακας day_profile.

- id: Το πρωτεύων κλειδί.
- day: Η μέρα (mon,tue, wed,thu, fri,sat,sun).
- tid: Το id της κατηγορίας του άρθρου.
- tidCount: Πόσες συνολικά φορές διαβάστηκε η κατηγορία (tid) τη δεδομένη ημέρα (day).

Column Name	Datatype
id	INT(10)
cookieID	VARCHAR(20)
day	VARCHAR(45)
tid	INT(10)

Εικόνα 38. Πίνακας day_tid.

- id: Το πρωτεύων κλειδί.

- cookieID: Το cookie του χρήστη.
- day: Ποιά μέρα γίνεται η αναφορά.
- tid: Η κατηγορία άρθρων που διάβασε ο χρήστης τη δεδομένη ημέρα.

Column Name	Datatype
id	INT(10)
nid1	INT(10)
nid2	INT(10)
times	INT(10)
cos_sim	DECIMAL(6,5)

Εικόνα 39. Πίνακας nid_profile.

- id: Το πρωτεύων κλειδί.
- nid1: Το id του άρθρου στο οποίο αναφερόμαστε.
- nid2: Το id του δεύτερου άρθρου που συγκρίνουμε.
- times: Πόσες φορές από το άρθρο με nid1, επιλέγει το άρθρα με nid2 σαν επόμενο.
- cos_sim: Η ομοιότητα συνημιτόνου μεταξύ των δύο άρθρων.

Column Name	Datatype
id	INT(10)
cid1	VARCHAR(255)
cid2	VARCHAR(255)
similarity	DECIMAL(8,5)

Εικόνα 40. Πίνακας similarity_up.

- id: Το πρωτεύων κλειδί.
- cid1: Το cookieID του πρώτου χρήστη προς σύγκριση.
- cid2: Το cookieID του δεύτερου χρήστη προς σύγκριση.
- similarity: Η ομοιότητα μεταξύ ενός ζεύγους χρηστών.

Column Name	Datatype
id	INT(10)
tidCurrent	INT(10)
tidHistory	INT(11)
times	INT(10)

Εικόνα 41. Πίνακας tid_profile.

- id: Το πρωτεύων κλειδί.
- tidCurrent: Το id της κατηγορίας του άρθρου όπου διαβάζει ο χρήστης.

- tidHistory: Το id της κατηγορίας του άρθρου από όπου ήρθε ο χρήστης.
- times: Πόσες συνολικά φορές έγινε η δεδομένη επιλογή.

Column Name	Datatype
id	INT(10)
timezone	VARCHAR(45)
tid	INT(10)
tidCount	INT(10)

Εικόνα 42. Πίνακας time_profile.

- id: Το πρωτεύων κλειδί.
- timezone: Η ημέρα χωρισμένη σε πέντε ζώνες (02:00-07:59,08:00-12:59,13:00-16:59,17:00-20:59,21:00-01:59).
- tid: Τα id των κατηγοριών που ανήκουν τα άρθρα.
- tidCount: Πόσες φορές διαβάστηκε μια κατηγορία σε συγκεκριμένη ζώνη.

Column Name	Datatype
id	INT(10)
cookieID	VARCHAR(20)
from_to	VARCHAR(20)
tid	INT(10)

Εικόνα 43. Πίνακας times_tid.

- id: Το πρωτεύων κλειδί.
- cookieID: Το cookie του χρήστη.
- from_to: Η ζώνη στην οποία έγινε η ενέργεια.
- tid: Το id της κατηγορίας του άρθρου που διάβασε ο χρήστης στις συγκεκριμένες ζώνες της ημέρας.

Column Name	Datatype
id	INT(10)
cookieID	VARCHAR(20)
nidCurrent	INT(10)
nidHistory	INT(10)
timestamp	VARCHAR(45)

Εικόνα 44. Πίνακας user_from_nid_to_nid.

- id: Το πρωτεύων κλειδί.
- cookieID: Το cookie του χρήστη.

- nidCurrent: Το id του άρθρου το οποίο διαβάζει ο χρήστης τη δεδομένη χρονική στιγμή.
- nidHistory: Το id του άρθρου από το οποίο ήρθε ο χρήστης.
- timestamp: Η χρονική στιγμή όπου έγινε η παραπάνω ενέργεια.

Column Name	Datatype
id	INT(10)
cookieID	VARCHAR(20)
tidCurrent	INT(10)
tidHistory	INT(10)
timestamp	VARCHAR(45)

Εικόνα 45. Πίνακας user_from_to.

- id: Το πρωτεύων κλειδί.
- cookieID: Το cookie του χρήστη.
- tidCurrent: Το id της κατηγορίας του άρθρου όπου διαβάζει ο χρήστης τη δεδομένη χρονική στιγμή.
- tidHistory: Το id της κατηγορίας του άρθρου από όπου ήρθε ο χρήστης.
- timestamp: Η χρονική στιγμή που έγινε η ενέργεια.

Column Name	Datatype
id	INT(10)
cookieID	VARCHAR(20)
nid	INT(10)
tid	INT(10)
timestamp	VARCHAR(45)
url	VARCHAR(255)

Εικόνα 46. Πίνακας user_has_read.

- id: Το πρωτεύων κλειδί.
- cookieID: Το cookie του χρήστη.
- nid: Το id του άρθρου που διαβάζει ο χρήστης τη δεδομένη χρονική στιγμή.
- tid: Το id της κατηγορίας του άρθρου.
- timestamp: Η χρονική στιγμή που έγινε η ενέργεια.
- url: Ο υπερσύνδεσμος (url) του άρθρου.

Column Name	Datatype
id	INT(10)
cookieID	VARCHAR(20)
tid	INT(10)
tidCount	INT(10)

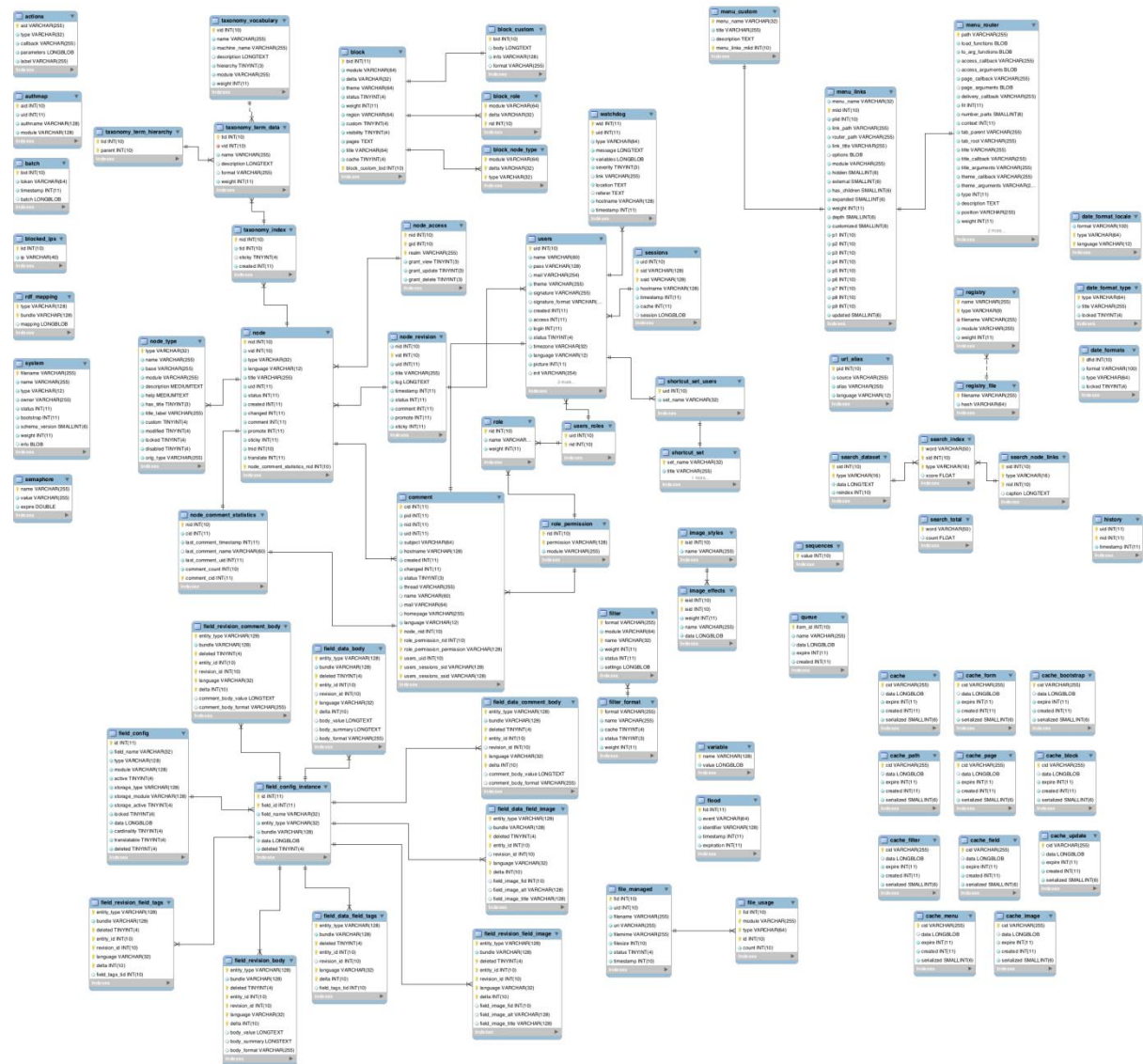
Εικόνα 47. Πίνακας user_profile.

- id: Το πρωτεύων κλειδί.
- cookieID: Το cookie του χρήστη.
- tid: Το id της κατηγορίας.
- tidCount: Ο αριθμός εμφάνισης άρθρων της παραπάνω κατηγορίας

Παράρτημα Β

Αρχιτεκτονική βάσης δεδομένων Drupal

Η δομή της βάσης δεδομένων της πλατφόρμας Drupal είναι η ακόλουθη.



Εικόνα 40. Αρχιτεκτονική βάσης δεδομένων συστήματος Drupal.