Πρόβλεψη Δημοτικότητας Διαδικτυακής Ειδησεογραφίας:
Μελέτη, Πειραματική Αξιολόγηση και Υλοποίηση.

Σπηλιόπουλος Χρήστος

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ Η/Υ
ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ & ΔΙΚΤΥΩΝ

ΜΕΤΑΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΕΠΙΒΛΕΠΟΝΤΕΣ:

1) ΧΟΥΣΤΗ ΑΙΚΑΤΕΡΙΝΗ
2) ΤΣΟΜΠΑΝΟΠΟΥΛΟΥ ΠΑΝΑΓΙΩΤΑ
3) ΜΠΟΖΑΝΗΣ ΠΑΝΑΓΙΩΤΗΣ

Περίληψη

Στόχος του NNY.com είναι να ενισχύσει τον ρόλο που μια κοινότητα αναγνωστών παίζει στην ανάδειξη σημαντικής ειδησεογραφίας. Αυτό θα επιτευχθεί μέσα από την μελέτη, την πειραματική αξιολόγηση και την υλοποίηση ενός αλγορίθμου ανάλυσης περιεχομένου. Στόχος είναι να αναδεικνύονται δυναμικά και άμεσα οι ειδήσεις που σχετίζονται με τα συλλογικά (και μεταβαλλόμενα) ενδιαφέροντα μιας κοινότητας. Με τον τρόπο αυτό δημιουργούνται βρόγχοι ανάδρασης ανάμεσα στο αναγνωστικό κοινό και στους παραγωγούς περιεχομένου, με αποτέλεσμα τόσο την βελτίωση της αναγνωστικής εμπειρίας, αλλά και του τρόπου που οι παραγωγοί δημοσιεύουν ειδήσεις.

Abstract

The goal of NNY.com is to enhance the role that a community of readers plays in recognizing popular news. This will be accomplished via a thorough study, experimental evaluation and implementation of an algorithm used for content analysis. The goal is to dynamically and immediately discover news that are relevant to a community's collective and changing interests. In such a way, feedback loops are created between news consumers (readers) and news producers (publishers), resulting both in the improvement of reading experience and of the way news are published.

Στους γονείς μου, στην Μαριλένα και στον "Μπρετζέμη".

ΠΕΡΙΕΧΟΜΕΝΑ

ΠΕΡΙΛΗΨΗ 2

ABSTRACT 3

ΠΕΡΙΕΧΟΜΕΝΑ 5

1) ΕΙΣΑΓΩΓΗ 7

2) ΑΝΑΛΥΣΗ ΚΑΙ ΠΡΟΒΛΕΨΗ ΤΟΥ DIGG 11

3) ΠΡΟΒΛΕΨΗ ΔΗΜΟΦΙΛΙΑΣ ΜΕ ΒΑΣΗ ΤΟ ΠΕΡΙΕΧΟΜΕΝΟ 14

4) ΠΕΙΡΑΜΑΤΙΚΗ ΕΠΑΛΗΘΕΥΣΗ 17

4.1) ΣΥΛΛΟΓΗ ΔΕΔΟΜΕΝΩΝ ….....17

4.2) ΜΕΘΟΔΟΛΟΓΙΑ ….....19

4.3) ΥΛΟΠΟΙΗΣΗ ΝΟΟΤΡΟΠΙΑΣ ….....23

4.4) ΑΡΧΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ ….....24

4.5) ΒΕΛΤΙΩΣΕΙΣ ….....27

4.6) ΤΕΛΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ ….....28

4.7) ΣΥΜΠΕΡΑΣΜΑΤΑ ….....40

5) ΣΧΕΔΙΑΣΗ ΕΦΑΡΜΟΓΗΣ.....41

5.1) ΠΡΟΓΡΑΜΜΑΤΙΣΤΙΚΑ ΕΡΓΑΛΕΙΑ ….....41

5.2) ΑΡΧΙΤΕΚΤΟΝΙΚΗ ΣΥΣΤΗΜΑΤΟΣ ….....43

5.3) ΑΝΑΛΥΣΗ ΕΦΑΡΜΟΓΗΣ ….....46

5.4) ΥΛΟΠΟΙΗΣΗ ΜΗΧΑΝΗΣ ΑΝΑΖΗΤΗΣΗΣ ….....48

5.5) ΣΥΜΠΕΡΑΣΜΑΤΑ ….....50

6) ΣΥΝΟΨΗ ΚΑΙ ΜΕΛΛΟΝΤΙΚΑ ΣΧΕΔΙΑ 51

7) ΒΙΒΛΙΟΓΡΑΦΙΑ 53

1. ΕΙΣΑΓΩΓΗ

Είναι γεγονός ότι ο Παγκόσμιος Ιστός αποτελεί πλέον αναπόσπαστο κομμάτι της καθημερινότητάς μας και η συνεχής εξέλιξή του το έχει μετατρέψει σε ένα πολυσύνθετο και πολύπλοκο “οργανισμό”. Στην σημερινή μορφή του, στην οποία συχνά αποδίδεται ο όρος Web 2.0, ο Παγκόσμιος Ιστός συνιστά έναν παράδεισο διαμοιρασμού πληροφοριών, κοινωνικοποίησης και συνεργασίας, με επίκεντρο πάντα τον τελικό χρήστη. Οι χρήστες πλεόν αλληλεπιδρούν μεταξύ τους μέσα από δομές όπως ιστότοπους κοινωνικής δικτύωσης, κοινωνικής ειδησεογραφίας, διαμοιρασμού οπτικής, ακουστικής και γραπτής πληροφορίας κ.α. Είναι προφανές ότι η δυναμική ενός τέτοιου συστήματος είναι απρόβλεπτη, συνεχώς μεταβαλλόμενη και διέπεται από προβλήματα όπως η υπερπληροφόρηση (information overload), η ύπαρξη δηλαδή μεγάλου όγκου πληροφορίας που αποτρέπει ή καθιστά δύσκολη την κατανόηση θεμάτων και την λήψη αποφάσεων από τον χρήστη. Για τον λόγο αυτό, δημιουργήθηκαν διάφορες μέθοδοι που προσπαθούν να θεραπεύσουν το παραπάνω πρόβλημα οι οποίες αποτελούν μέρος ενός σχετικά νέου κλάδου της επιστήμης των υπολογιστών, τα λεγόμενα συστήματα προτάσεων (recommender systems).

Οι βασικές μεθοδολογίες / προσεγγίσεις των παραπάνω συστημάτων είναι οι εξής: Συνεργατικό φιλτράρισμα (collaborative filtering) και φιλτράρισμα με βάση το περιεχόμενο (content-based filtering). Και οι δυο μεθοδολογίες βασίζονται στην συγκέντρωση και ανάλυση της συμπεριφοράς των χρηστών και των προτιμήσεών τους, με σκοπό την πρόβλεψη και ανακάλυψη ενδιαφέρουσας πληροφορίας για κάποιον χρήστη. Το συνεργατικό φιλτράρισμα έγκειται στην μοντελοποίηση μιας (κοινωνικής) διαδικασίας, όπως μια συμβουλή ή σύσταση που κάποιος θα αναζητούσε από φίλους ή άτομα με κοινά ενδιαφέροντα. Για παράδειγμα, ο ιστότοπος Last.fm προτείνει μουσική με βάση την σύγκριση των μουσικών προτιμήσεων χρηστών με παρόμοια μουσικά ενδιαφέροντα. Η δεύτερη μεθοδολογία από την άλλη βασίζεται μόνο σε χαρακτηριστικά που προκύπτουν από το ίδιο το πληροφοριακό αντικείμενο και με βάση τις καθαρά προσωπικές προτιμήσεις του χρήστη. Στις περισσότερες περιπτώσεις το σύστημα δημιουργεί ένα προφίλ το οποίο αποτελείται από ένα διάνυσμα βεβαρημένων πληροφοριακών χαρακτηριστικών. Τα βάρη δείχνουν την σημαντικότητα του κάθε χαρακτηριστικού για τον

χρήστη, υπολογίζονται δε χρησιμοποιώντας διάφορες τεχνικές εκμάθησης όπως δέντρα απόφασης, δίκτυα Bayes, νευρωνικά δίκτυα κ.α. Μια διαδικασία η οποία πλέον αποτελεί αναπόσπατο κομμάτι των παραπάνω προσεγγίσεων είναι η συγκέντρωση ανάδρασης από τον χρήστη, η οποία μπορεί να είναι έμμεση ή άμεση. Οι χρήστες παρέχουν θετική ή αρνητική ανάδραση, η οποία χρησιμοποιείται από τα παραπάνω συστήματα για τον υπολογισμό των εξατομικευμένων πληροφοριακών συστάσεων.

Τα προβλήματα που σχετίζονται με τις προαναφερθείσες προσεγγίσεις προφανώς και δεν απουσιάζουν. Στην περίπτωση του collaborative filtering είναι τρία:

1) Είναι βασική προϋπόθεση η ύπαρξη πολλών αρχικών δεδομένων για να κάνουμε ακριβείς προβλέψεις.

2) Στους τομείς στους οποίους εφαρμόζεται παρατηρείται συνήθως πληθώρα χρηστών και προτιμήσεων, κάνοντας την όλη διαδικασία να χρειάζεται μεγάλη υπολογιστική δύναμη για να να περατωθεί.

3) Η πληροφορία, αν και μεγάλη σε όγκο, συχνά είναι διάσπαρτη και αραιή, καθιστώντας δύσκολο τον ακριβή υπολογισμό πληροφοριακών συστάσεων.

Στην περίπτωση του (παραδοσιακού) content-based filtering, υπάρχει το πρόβλημα της έλλειψης ακρίβειας λόγω των πολλών διαστάσεων του προβλήματος, του πλήθους δηλαδή των χαρακτηριστικών που χρησιμοποιούνται για την δημιουργία του προφίλ.

Ιστότοποι κοινωνικής ειδησεογραφίας (social news sites) όπως το Digg και Reddit έγιναν πολύ δημοφιλή στο πλαίσιο του Web 2.0 και της κουλτούρας που το διέπει. Η δημοκρατική τους φύση τα καθιστά πιο ελκυστικά σε σχέση με τα παραδοσιακά μέσα μαζικής ενημέρωσης. Δίνουν την δυνατότητα στους ίδιους τους χρήστες να προβάλλουν τα άρθρα τους, αλλά και να κρίνουν την εγκυρότητα και αξία τους, καθώς και το πόσο ενδιαφέροντα είναι. Αυτό γίνεται μέσω ενός συστήματος ψηφοφορίας, που περιλαμβάνει θετικές και αρνητικές ψήφους. Το σύστημα αυτό λαμβάνει υπόψιν του διάφορες παραμέτρους, όπως τον ρυθμό ψηφοφορίας, την δημοφιλία του ίδιου του ψηφοφόρου, τον χρόνο από την στιγμή δημοσίευσης του άρθρου κ.α., για να εκτιμήσει την δημοφιλία κάθε άρθρου και ανάλογα να το προάγει ή να το υποβαθμίσει. Τα άρθρα που

έχουν προαχθεί εμφανίζονται στην πρώτη σελίδα του ιστότοπου και με αυτόν τον τρόπο τους δίνεται η ευκαιρία να αναδειχτούν ακόμη περισσότερο και να λάβουν επιπλέον ψήφους.

Με εκατομμύρια ενεργούς χρήστες, οι συγκεκριμένοι ιστότοποι αποτελούσαν ένα μεγάλο κανάλι διανομής ειδησεογραφίας, στο πλαίσιο του αποκαλούμενου “real-time” web. Οι παραδοσιακοί αλγόριθμοι collaborative filtering δεν αποδίδουν σε έναν τέτοιο τομέα, εξαιτίας της δυναμικότητας του, τον υψηλό δηλαδή ρυθμό παραγωγής και τον χαμηλό κύκλο ζωής των ειδήσεων-άρθρων. Υπάρχει αδυναμία συγκέντρωσης σε πραγματικό χρόνο των στατιστικών ποσοτήτων που απαιτούνται ώστε να παραχθούν προσωπικές - στοχευμένες προτάσεις. Για αυτό το λόγο οι ιστότοποι κοινωνικής ειδησεογραφίας χρησιμοποιούν, αντί για συνεργατικό φιλτράρισμα, συστήματα ψηφοφορίας ώστε να επιτρέπουν στην κοινότητά τους να αναγνωρίσει τα σημαντικά άρθρα με βάση τα συλλογικά της ενδιαφέροντα.

Παρά την προφανή δημοκρατική τους φύση σε τέτοιες περιπτώσεις τα συστήματα ψηφοφορίας δεν είναι αδιάβλητα. Πάσχουν από φαινόμενα όπως το “Sybil-attack”, που συμβαίνει στην περίπτωση που κάποιος χρήστης ενεργοποιεί περισσότερους από έναν παράλληλους λογαριασμούς. Έτσι μπορεί στοχευμένα να ψηφίζει συγκεκριμένα άρθρα και να κερδίζει όλο και περισσότερο έδαφος στην διαδικασία ανάδειξης δημοφιλίας. Έχουν υπάρξει περιπτώσεις ιδιωτικών εταιριών που αναλάμβαναν την προαγωγή συγκεκριμένων άρθρων έναντι χρηματικής αμοιβής. Επιπλέον, επειδή η δημοτικότητα των άρθρων εξαρτάται από το πλήθος των ψήφων που έχουν αποκομίσει, τα ήδη προαχθέντα θα αποκομίσουν ακόμα περισσότερους. Στο Digg για παράδειγμα, ένας πολύ μικρός πυρήνας πραγματικών χρηστών ελέγχει την πλειονότητα των άρθρων που τελικά καταφέρνουν να πλασαριστούν στην αρχική σελίδα. Επίσης, πρέπει να περάσει κάποιο χρονικό διάστημα μέχρι το άρθρο αυτό να εκδοθεί για πρώτη φορά στο “upcoming section”, και να λάβει τις απαιτούμενες ψήφους για να γίνει δημοφιλές. Στο πλαίσιο του real-time web κάτι τέτοιο αποτελεί σοβαρό μειονέκτημα, καθώς η αξία των ειδήσεων πέφτει κατακόρυφα με το πέρασμα του χρόνου.

Η πρόβλεψη δημοτικότητας (popularity prediction) θα μπορούσε να βελτιώσει τα παραπάνω θέματα. Αν υπήρχε η δυνατότητα να εκτιμηθεί η δημοφιλία του άρθρου με το που δημοσιεύεται, τότε θα άλλαζε όλη η δυναμική του συστήματος ψηφοφορίας, καθώς θα άλλαζαν οι αρχικές συνθήκες. Σε αυτή την εργασία, προτείνουμε μια νέα μεθοδολογία για την πρόβλεψη δημοφιλίας που βασίζεται μόνο στο πραγματικό περιεχόμενο του άρθρου. Χρησιμοποιούμε ένα προφίλ για να αιχμαλωτίσουμε τα συλλογικά ενδιαφέροντα των χρηστών που βασίζεται σε

άρθρα τα οποία οι ίδιοι έχουν ψηφίσει θετικά στο παρελθόν. Το προφίλ αυτό στην συνέχεια εκτιμά το περιεχόμενο των νέων άρθρων και αξιολογεί την σχετικότητα τους με τα συλλογικά ενδιαφέροντα χωρίς να χρειάζεται ψήφους. Θεωρούμε ότι όσο υψηλότερος ο βαθμός σχετικότητας τόσο πιο πιθανό είναι το άρθρο να γίνει δημοφιλές και να ψηφιστεί. Με αυτόν τον τρόπο έναν νέο άρθρο μπορεί να γίνει αντιληπτό με βάση την αντικειμενική σχετικότητα του περιεχομένου του ως προς τα ενδιαφέροντα της κοινότητας, και όχι την υποκειμενική συσσώρευση ψήφων από συγκεκριμένους χρήστες.

Αν και η πρόβλεψη δημοφιλίας σε άρθρα του Digg έχει ήδη προσελκύσει το ενδιαφέρον των ερευνητών, παραθέτουμε μια μέθοδο βασισμένη σε περιεχόμενο, η οποία μπορεί να παρέχει πρόβλεψη σε πραγματικό χρόνο χωρίς την ανάγκη έστω και λίγων αρχικών ψήφων. Κάτι τέτοιο δεν είναι σε καμμία περίπτωση εύκολο και δεν έχει ξαναγίνει, όπως γίνεται σαφές στην ενότητα 3. Τα συλλογικά ενδιαφέροντα μιας κοινότητας μπορεί να συμπεριλαμβάνουν πολλά θεματικά αντικείμενα που αναπόφευκτα αλλάζουν με την πάροδο του χρόνου. Στην εργασία υιοθετούμε ένα μοντέλο εμπνευσμένο από την βιολογία για να χτίσουμε το προφίλ. Το συγκεκριμένο μοντέλο έχει ήδη εφαρμοστεί επιτυχώς για προσαρμοστικό φιλτράρισμα κειμενικής πληροφορίας, με βάση τα προσωπικά - ξεχωριστά ενδιαφέροντα (personalization), τα οποία μπορεί να είναι διαφορετικά και δυναμικά.

Στο υπόλοιπο του παρόντος εγγράφου, αρχικά εξετάζουμε παρόμοιες εργασίες σχετικά με ανάλυση και πρόβλεψη δημοτικότητας σε άρθρα του Digg, πριν περιγράψουμε την προτεινόμενη με βάση το περιεχόμενο προσέγγιση στο κεφάλαιο 3. Το κεφάλαιο 4 περιγράφει την πειραματική μεθοδολογία, παρουσιάζει τα αποτελέσματα και συζητά τα ευρήματά μας. Στο κεφάλαιο 5 παρουσιάζουμε μια βελτιωμένη έκδοση του αλγορίθμου και επαναλαμβάνουμε την πειραματική διαδικασία. Καταλήγουμε με την ανάλυση και κατασκευή μιας web εφαρμογής η οποία αποτελεί καρπό των όσων προηγήθηκαν.

2. ΑΝΑΛΥΣΗ ΚΑΙ ΠΡΟΒΛΕΨΗ ΤΟΥ DIGG

Το Digg βασίζεται σε ένα πολύπλοκο σύστημα που εκμεταλλεύεται την αναδυόμενη και δυναμική συμπεριφορά από μια μεγάλη κοινότητα ψηφοφόρων. Αυτή εξαρτάται από πολλούς αλληλένδετους παράγοντες σχεδιασμού που αφορούν τη διεπαφή, τους μηχανισμούς κοινωνικής δικτύωσης και τον αλγόριθμο προαγωγής άρθρων. Παράγοντες όπως η επικαιρότητα, σημαντικότητα και ορατότητα ενός άρθρου, το κοινωνικό status του εκδότη μέσα στην κοινότητα κ.α., επιδρούν (ανεξάρτητα) στην προαγωγή της ιστορίας ή στον υποβιβασμό της. Ως εκ τούτου, το Digg έχει ήδη αποτελέσει αντικείμενο μελέτης, πειραμάτων και ανάλυσης. Σε μια σειρά από δημοσιεύσεις [1, 2, 3], η Kristina Lerman μελέτησε και μοντελοποίησε τις παραπάνω διαδικασίες και εν συντομία κατέληξε στο συμπέρασμα ότι:

- Οι ιστότοποι κοινωνικού περιεχομένου (ειδησεογραφίας) επιτρέπουν στους χρήστες να εκμεταλλευτούν άλλους χρήστες σαν ένα είδος μηχανισμού κοινωνικού φιλτραρίσματος ωστε να ανακαλύψουν ενδιαφέρουσες πληροφορίες.

- Η επιτυχία ενός εκδότη ως προς την προαγωγή ενός άρθρου εξαρτάται από το μέγεθος του κοινωνικού του περίγυρου (δικτύου).

- Είναι σύνηθες φαινόμενο οι χρήστες να ψηφίζουν άρθρα που έχουν ψηφιστεί από φίλους.

- Είναι δυνατό κάποιος να εκμεταλευτεί τα μοτίβα κοινωνικής ψηφοφορίας για να προβλέψει επιτυχώς την δημοφιλία ενός άρθρου.

Στο [4], οι συγγραφείς προτείνουν μια μέθοδο κατηγοριοποίησης για να προβλέψουν την δημοφιλία ενός άρθρου. Κάθε κατηγορία αντιστοιχεί σε ένα συγκεκριμένο διάστημα μετρήσεων

ψήφων. Η μέθοδός τους χρησιμοποιεί διάφορα χαρακτηριστικά όπως στατιστικά από τα σχόλια των άρθρων, τον ρυθμό ψηφοφορίας, διαφορές σε θετικές και αρνητικές ψήφους, καθώς και χαρακτηριστικά από την δομή κοινωνικής δικτύωσης των ψηφοφόρων. Διεξήγαγαν διάφορα πειράματα με διάφορους αλγορίθμους κατηγοριοποίησης πάνω σε δεδομένα 37185 διαφορετικών άρθρων, που περιείχαν περισσότερα από 6 εκατομμύρια σχόλια. Στα ευρήματά τους αναφέρουν την ύπαρξη μικρής μείωσης της ακρίβειας της κατηγοριοποίησης όταν χρησιμοποιούνται δεδομένα με χρόνο ζωής μικρότερο των 10 ωρών, σε σχέση με τα τελικά δεδομένα. Έτσι, σύμφωνα με τους συγγραφείς, είναι εφικτή η πρόβλεψη δημοφιλίας ενός άρθρου μόνο με την χρήση δεδομένων που συλλέγονται στις 10 πρώτες ώρες από την δημοσίευσή του.

Μια διαφορετική προσέγγιση περιγράφεται στο [5]. Λογαριθμικοί μετασχηματισμοί χρησιμοποιούνται για την πρόβλεψη δημοφιλίας ενός άρθρου σε μελλοντικό χρόνο t_2, δοθείσης της δημοφιλίας σε χρόνο t_1. Στη συγκεκριμένη περίπτωση τα δεδομένα που χρησιμοποιήθηκαν αγγίζουν τα 1.3 εκατομμύρια άρθρα. Η πειραματική τους ανάλυση οδήγησε σε μια σειρά από ενδιαφέροντα συμπεράσματα:

- Ο αριθμός των μελλοντικών ψήφων είναι σταθερό πολλαπλάσιο των ψήφων στο παρελθόν, αλλά ο πολλαπλασιαστής εξαρτάται από τον χρόνο δειγματοληψίας και πρόβλεψης.

- Άρθρα τα οποία είναι δημοφιλή λίγο μετά την δημοσίευσή τους έχουν την τάση να παραμένουν δημοφιλή.

- Ο συνολικός αριθμός των ψήφων ενός άρθρου φαίνεται να σταθεροποιείται μετά από μια μέρα.

- Ο αριθμός των αρχικών ψήφων και κατά συνέπεια των μελλοντικών εξαρτάται από την χρονική στιγμή δημοσίευσης.

- Το ευρύ κοινωνικό δίκτυο κάποιου εκδότη παίζει σημαντικό ρόλο στον αριθμό των ψήφων που η αρθογραφία του θα λάβει, κυρίως στα αρχικά στάδια του κύκλου ζωής

ενός άρθρου, ενώ φαίνεται να μην επιδρά τόσο δυναμικά στην συνέχεια, όταν πλέον το άρθρο βρίσκεται στην αρχική σελίδα.

Σε ένα τέτοιο πλαίσιο συνθηκών, το αποτέλεσμα τις περισσότερες φορές είναι άδικο, καθώς οι “πλούσιοι γίνονται πλουσιότεροι”, επωφελούμενοι των αδυναμιών του συστήματος. Άρθρα τα οποία αποκομίζουν μεγάλο αριθμό από ψήφους στα αρχικά στάδια, το οποίο είναι ευκολότερο εφόσον ο εκδότης έχει ισχυρή κοινωνική δικτύωση, με μεγάλη πιθανότητα θα προσελκύσουν την προσοχή πολλών αναγνωστών και θα φτάσουν σε σημείο αιχμής μόλις εμφανιστούν στο πρωτοσέλιδο του Digg. Αν επίσης λάβουμε υπόψιν μας το γεγονός ότι ο κύκλος ζωής ενός άρθρου είναι μικρός, με την δημοτικότητά του να έρχεται σε κορεσμό μετά από σύντομο χρονικό διάστημα, μπορούμε να συμπεράνουμε ότι η πρόβλεψη με βάση τις αρχικές ψήφους γίνεται “εκ του ασφαλούς”, ειδικά όταν η περίοδος συγκομιδής των υπό εξέταση δεδομένων είναι παρατεταμένη. Διάφοροι εξωτερικοί παράγοντες, όπως ο ακριβής χρόνος δημοσίευσης επηρεάζουν κατά πολύ την τελική δημοφιλία του άρθρου. Ο αριθμός των αρχικών ψήφων είναι κυρίαρχη παράμετρος ως προς την μετέπειτα πορεία του άρθρου, αλλά αποτελεί παράμετρο που δεν εξαρτάται από την σημασιολογία του άρθρου. Όπως είδαμε, ένα κοινωνικό δίκτυο κάποιου εκδότη ή ακόμα και ο χρόνος δημοσίευσης είναι ικανά να αλλάξουν την τύχη ενός άρθρου. Αυτό σημαίνει ότι μια πραγματικά ενδιαφέρουσα είδηση μπορεί να υποβαθμιστεί σε βάρος κάποιας λιγότερο σημαντικής. Σημειωτέον, καμία από τις προαναφερθείσες προσεγγίσεις δεν λαμβάνει υπόψιν της την πραγματική σημασιολογία ενός άρθρου, αν και οι συγγραφείς στο [5] παραδέχονται ότι “η σημασιολογία ενός άρθρου μπορεί να χρησιμοποιηθεί για να προβλεφθούν οι ρυθμοί επίσκεψης του όταν δεν υπάρχουν άλλα αρχικά δεδομένα”. Αυτή είναι και η οδός που θα ακολουθήσουμε από δω και στο εξής.

3. ΠΡΟΒΛΕΨΗ ΔΗΜΟΦΙΛΙΑΣ ΜΕ ΒΑΣΗ ΤΟ ΠΕΡΙΕΧΟΜΕΝΟ

Η πρόβλεψη δημοφιλίας με βάση το περιεχόμενο βασίζεται στην υπόθεση ότι οι χρήστες ψηφίζουν ανάλογα με τις προτιμήσεις τους και ότι συνολικά, αυτές αντικατοπτρίζουν τα πραγματικά κοινά τους ενδιαφέροντα. Αν μπορούσαμε να ταιριάξουμε το περιεχόμενο των άρθρων με αυτά τα ενδιαφέροντα, θα μπορούσαμε να κάνουμε μια εκτίμηση για την δημοφιλία ενός άρθρου ακριβώς κατά τον χρόνο δημοσίευσης, χωρίς να περιμένουμε να ξεκινήσει η φάση ψηφοφορίας. Επιπλέον, σε αντίθεση με τις ψήφους, το περιεχόμενο δεν μπορεί να αλλάξει με κακόβουλη δραστηριότητα.

Ένας τρόπος για να πραγματοποιήσουμε ένα τέτοιο είδος πρόβλεψης είναι να χρησιμοποιήσουμε ένα προφίλ που αναπαριστά τα συλλογικά ενδιαφέροντα μιας κοινότητας. Αυτό το προφίλ “χτίζεται” πάνω σε άρθρα που έχουν ψηφιστεί στο παρελθόν και χρησιμοποιείται για να αξιολογήσει την σχετικότητα κάθε νέου άρθρου στα συλλογικά ενδιαφέροντα. Το πρόβλημα αυτό θυμίζει το αντίστοιχο του φιλτραρίσματος πληροφορίας με βάση το περιεχόμενο (content-based information filtering), βλέπε [6], αλλά το γεγονός ότι έχουμε να αντιμετωπίσουμε πολλούς χρήστες αλλάζει τα χαρακτηριστικά του. Είναι φυσικό και επόμενο να υποθέσουμε ότι τα ενδιαφέροντα μιας κοινότητας σχετίζονται με πολλές θεματικές ενότητες, ακόμα και μέσα σε βασικές κατηγορίες. Επιπρόσθετα, αυτά τα ενδιαφέροντα είναι δυναμικά και αλλάζουν με την πάροδο του χρόνου εξαιτίας πλήθους παραγόντων, όπως η επικαιρότητα, οι προσωπικές αποκλίσεις - προτιμήσεις κ.α. Οπότε χρειαζόμαστε ένα προφίλ που να μπορεί να “αιχμαλωτίσει” τα γενικότερα ενδιαφέροντα μιας κοινότητας και να μπορεί να προσαρμόζεται σε αυτά με την πάροδο του χρόνου.

Στο πλαίσιο της εργασίας μας υιοθετούμε την Νοοτροπία, ένα μοντέλο δημιουργίας προφίλ που έχει ήδη εφαρμοστεί αποτελεσματικά σε προβλήματα προσαρμοστικού φιλτραρίσματος εγγράφων με βάση το περιεχόμενο και έχει επαληθευτεί πειραματικά ότι πληρεί

τα προαναφερθέντα χαρακτηριστικά. Στην παρούσα μορφή της, η Νοοτροπία έχει πρωτοεμφανιστεί στο [7] και από τότε, έχει αναλυθεί διεξοδικά, επαληθευτεί πειραματικά και έχει εφαρμοστεί στη δημιουργία διάφορων πραγματικών εφαρμογών ([12, 11, 9, 8, 10]). Όπως αναφέρεται στο [7], η Νοοτροπία είναι εμπνευσμένη από το βιολογικό ανοσοποιητικό σύστημα, στο πλαίσιο της θεωρίας Αυτοποίησης. Όπως ένα δίκτυο από αλληλεπιδρώντα αντισώματα σε ένα ανοσοποιητικό σύστημα ορίζει και διαφυλάττει την ταυτότητα ενός οργανισμού, κατ' αναλογία, η Νοοτροπία είναι ένα δίκτυο από αλληλεπιδρώντα χαρακτηριστικά (features) που χρησιμοποιούνται για να ορίσουν και να διατηρήσουν τα ενδιαφέροντα ενός χρήστη ή μιας κοινότητας. Τόσο οι κόμβοι (features) όσο και οι σύνδεσμοι συσχετισμού μεταξύ τους (feature correlations) είναι βεβαρημένοι σε ένα τέτοιο δίκτυο. Το βάρος ενός κόμβου εκφράζει την σημαντικότητά του μέσα στο προφίλ και το βάρος ενός συνδέσμου μετρά το πόσο ισχυρή είναι η συσχέτιση μεταξύ δυο κόμβων - χαρακτηριστικών. Τα χαρακτηριστικά του προφίλ εξάγονται από το περιεχόμενο ενός κειμένου, ή ενός πληροφοριακού αντικειμένου γενικότερα. Δύο χαρακτηριστικά συσχετίζονται μεταξύ τους όταν εμφανίζονται με μεγάλη συχνότητα στο ίδιο πλαίσιο.

Για να αξιολογήσουμε την σχετικότητα ενός κειμένου σε σχέση με τα ενδιαφέροντα ενός χρήστη ή μιας κοινότητας ακολουθείται μια διαδικασία κατευθυνόμενης διάδοσης "ενέργειας". Λέξεις από το προφίλ οι οποίες εμφανίζονται στο κείμενο ενεργοποιούνται και διαδοχικά "διαχέουν" την ενέργειά τους προς άλλες ενεργοποιημένες λέξεις με μεγαλύτερο βάρος από αυτές, εφ' όσον υπάρχει σύνδεσμος μεταξύ τους. Το ποσό της διαχέουσας ενέργειας μεταξύ δύο λέξεων είναι ανάλογο με το αντίστοιχο βάρος του μεταξύ τους συνδέσμου. Όλη αυτή η διαδικασία ξεκινά από την ενεργοποιημένη λέξη με το μικρότερο βάρος και συνεχίζει διαδοχικά, έως ότου εξετάσουμε και την ενεργοποιημένη λέξη με το μεγαλύτερο βάρος. Σε αυτό το σημείο, ένας βαθμός σχετικότητας υπολογίζεται ως το σταθμισμένο άθροισμα των τελικών ενεργειών των λέξεων. Το συνολικό αποτέλεσμα είναι μια μη γραμμική διαδικασία αξιολόγησης σχετικότητας η οποία περιγράφεται με ακρίβεια στο [8], όπου επίσης διαπιστώνεται το εξής γεγονός: λαμβάνοντας υπ' όψιν τους συσχετισμούς ανάμεσα στις λέξεις (χαρακτηριστικά) με τον συγκεκριμένο τρόπο, η Νοοτροπία θεραπεύει πολλά από τα εγγενή προβλήματα διαστάσεων (dimensionality problems) των παραδοσιακών διανυσματικών αναπαραστάσεων προφίλ. Με άλλα λόγια, το προφίλ μπορεί να διαθέτει πληθώρα χαρακτηριστικών χωρίς να γίνεται ασαφές και διφορούμενο. Συγκριτικά πειράματα μεταξύ της Νοοτροπίας και προφίλ διανυσματικού χώρου πάνω στα ίδια βεβαρημένα χαρακτηριστικά δείχνουν ότι όσο ο αριθμός των λέξεων αυξάνεται (άρα και οι διαστάσεις), η Νοοτροπία πετυχαίνει βελτίωση της ακρίβειας μέχρι και

50%. Αυτό αποτελεί συγκριτικό πλεονέκτημα για την διαδικασία της πρόβλεψης δημοφιλίας με βάση το περιεχόμενο, επειδή επιτρέπει στο προφίλ να διατηρεί έναν μεγάλο αριθμό λέξεων που αντιπροσωπεύουν την πληθώρα των θεματικών αντικειμένων που ενδιαφέρουν μια κοινότητα χρηστών.

Το προφίλ συνεχώς “χτίζεται” και προσαρμόζεται μέσω μιας διαδικασίας αυτο-οργάνωσης που περιγράφεται με λεπτομέρεια στο [7]. Η συγκεκριμένη διαδικασία ρυθμίζει την δομή του δικτύου ανάλογα με “ερεθίσματα” που προέρχονται από τις αποκρίσεις των χρηστών μέσω αλλαγών στα βάρη των λέξεων, αλλά επίσης και μέσω προσάρτησης νέων χαρακτηριστικών, αλλά και διαγραφής αντιστοίχων που έχουν ξεμείνει από βάρος. Για παράδειγμα, αν ένα κείμενο έχει αναγνωριστεί ως ενδιαφέρον και σχετικό για κάποιον χρήστη ή μια κοινότητα, τότε οι λέξεις που υπάρχουν κοινές τόσο στο κείμενο όσο και στο προφίλ θα ενισχυθούν σε βάρος άλλων λέξεων με τα οποία συνδέονται. Αυτοί οι ανταγωνισμοί σε τοπικό επίπεδο παίζουν σημαντικό ρόλο στην δυναμική του δικτύου και στην αναδιανομή των βαρών [8]. Η “μετα-δυναμική” του δικτύου από την άλλη μεριά είναι υπεύθυνη για την προσάρτηση νέων χαρακτηριστικών που δεν προϋπήρχαν στο προφίλ καθώς και για την αφαίρεση αυτών που δεν είναι πλέον ανταγωνιστικά. Κατά συνέπεια, ο ακριβής αριθμός των κόμβων - χαρακτηριστικών δεν είναι προκαθορισμένος, ούτε σταθερός, αλλά αλλάζει δυναμικά με τον χρόνο. Στο [6] έχει γίνει εκτενής συζήτηση για την Νοοτροπία σχετικά με την ύπαρξη βασικών αυτοοργανωτικών και αυτοποιητικών χαρακτηριστικών σε αυτή. Είναι ένα μη γραμμικό σύστημα, ανοιχτό στο περιβάλλον, που λειτουργεί πέρα από σημεία ισορροπίας, συνεχώς μεταβαλλόμενο δομικά με σκοπό την διατήρηση της οργάνωσης και αυτονομίας. Μέσω της δυναμικής του δικτύου, το προφίλ μπορεί να να προσαρμόζεται συνεχώς σε αλλαγές ενδιαφερόντων. Διάφορα πειράματα έχουν αποδείξει ότι η Νοοτροπία μπορεί εξίσου να μάθει αλλά και να ξεχάσει μια θεματική ενότητα [7], όπως επίσης να πετύχει βελτιώσεις στην ακρίβεια της τάξεως του 22% σε σχέση με παραδοσιακούς αλγορίθμους εκμάθησης [10]. Η δυνατότητα της συνεχούς προσαρμοστικότητας σε εναλλασσόμενα ενδιαφέροντα, είναι ζωτικής σημασίας στη περίπτωση των συλλογικών ενδιαφερόντων μιας δυναμικής κοινότητας. Εν κατακλείδι, η Νοοτροπία είναι μια εγγυημένη επιλογή για να προσφέρει λύση στο πρόβλημά της πρόβλεψης δημοφιλίας με βάση το περιεχόμενο, για όλους τους προαναφερθέντες λόγους.

4. ΠΕΙΡΑΜΑΤΙΚΗ ΕΠΑΛΗΘΕΥΣΗ

Για τα πειράματα που διεξήχθησαν, χρησιμοποιήσαμε την διεπαφή (API) του Digg με σκοπό την συγκέντρωση μιας συλλογής άρθρων, σε αναλογία με τις αντίστοιχες συλλογές που χρησιμοποιήθηκαν στα [4, 5] στο κεφάλαιο 2. Η διαφορά έγκειται στην διαφορετική μεθοδολογία επαλήθευσης και στις μετρικές. Όπως προαναφέρουμε, δεν υπήρχε εν γνώση μας κάποια αντίστοιχη μελέτη σχετική με την πρόβλεψη δημοφιλίας με βάση το περιεχόμενο, και ως εκ τούτου αναπτύξαμε μια νέα μεθοδολογία, καθώς δεν προϋπήρχε κάτι ανάλογο προς υϊοθεσία. Στο υπόλοιπο μέρος του κεφαλαίου, πρώτα περιγράφουμε τα δεδομένα της συλλογής μας και την διαδικασία συγκομιδής τους. Έπειτα παραθέτουμε την μεθοδολογία επαλήθευσης, τα πειραματικά αποτελέσματα και συζητούμε τα αποτελέσματα.

4.1 ΣΥΛΛΟΓΗ ΔΕΔΟΜΕΝΩΝ

Η συλλογή μας αποτελείται από 78020 άρθρα που δημοσιεύτηκαν στο Digg από 10964 διαφορετικούς χρήστες την περίοδο 09-01-2011 με 02-04-2011. Για την συγκομιδή των άρθρων χρησιμοποιήσαμε έναν crawler γραμμένο σε python, ο οποίος με την χρήση του Digg API συγκέντρωνε, ανέλυε και αποθήκευε κάθε νέο άρθρο που “κατάφερνε” να εμφανιστεί στη σελίδα των “επερχόμενων” (upcoming) ειδήσεων. Για κάθε άρθρο, αποθηκεύσαμε τα παρακάτω δεδομένα στην βάση δεδομένων:

- **category** : Υπάρχουν συνολικά εννιά κατηγορίες, οι οποίες συνοψίζονται στον πίνακα 1 μαζί με τον αντίστοιχο αριθμό άρθρων που δημοσιεύτηκαν σε κάθε μια από αυτές.

- **content** : Οι λέξεις του περιεχομένου του άρθρου κατόπιν επεξεργασίας (stemming[1] and stopwords removal[2]).

- **timestamp** : Η ημερομηνία δημοσίευσης.

- **initial diggs** : Ο αριθμός των αρχικών ψήφων που έλαβε το άρθρο την στιγμή της δημοσίευσής του (ελάχιστο : 2, μέγιστο : 572, μέσος όρος : 3.15, τυπική απόκλιση : 4.25).

- **boost diggs** : Ο αριθμός των ψήφων που έλαβε το άρθρο μια ώρα μετά την δημοσίευσή του (ελάχιστο : 2, μέγιστο : 1012, μέσος όρος : 7.24, τυπική απόκλιση : 10.84).

- **updated diggs** : Ο αριθμός των ψήφων μετά από μια μέρα, όπου οι τιμές έχουν ουσιαστικά συγκλίνει προς την τελική τους τιμή, όπως αναφέραμε (ελάχιστο : 2, μέγιστο : 1205, μέσος όρος : 31.17, τυπική απόκλιση : 59.54).

category **size**	business 11500	entertainment 12200	lifestyle 10300	offbeat 5300
politics 9100	science 2900	sports 3000	technology 11800	world 8500

Πίνακας 1. Βασικές κατηγορίες του Digg με τον αντίστοιχο αριθμό άρθρων.

1 Stemming: προθεματική ρίζα μιας λέξης.
2 Stopwords removal: αφαίρεση κοινών (συχνών) όρων.

4.2 ΜΕΘΟΔΟΛΟΓΙΑ

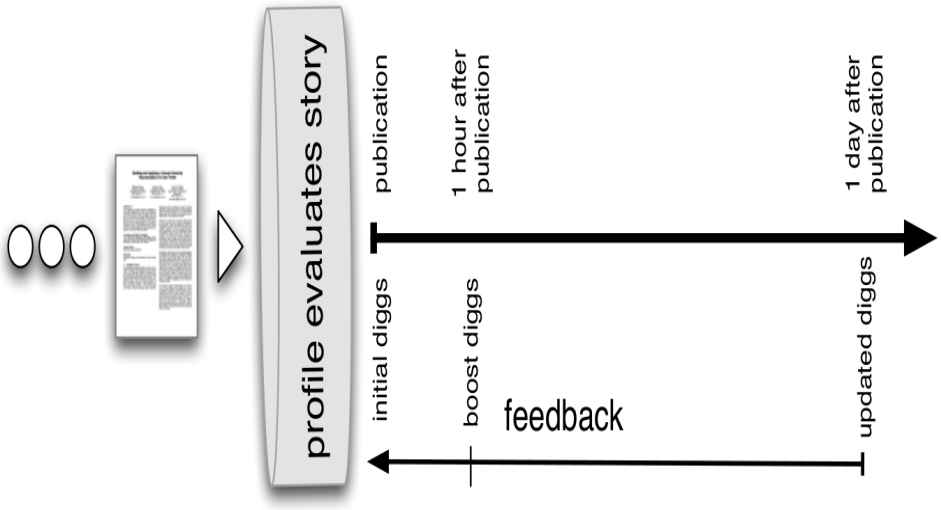
Εκτελέσαμε μια σειρά 10 πειραμάτων, ένα για κάθε θεματική ενότητα και ένα ξεχωριστό για όλα τα άρθρα, με δυο διαφορετικές υλοποιήσεις της Νοοτροπίας. Σε κάθε περίπτωση, η διαδικασία ξεκινά με ένα άδειο προφίλ το οποίο διαπερνά ένα - ένα τα άρθρα κατά χρονολογική σειρά. Κάθε άρθρο αρχικά αξιολογείται με βάση το περιεχόμενό του και του αναθέτουμε ένα βαθμό σημαντικότητας. Σε αυτή την φάση, δεν χρειάζονται ψήφοι, οπότε αυτή η διαδικασία μπορεί να γίνει “επι-τόπου”, ακριβώς δηλαδή κατά την στιγμή της δημοσίευσης. Κατά συνέπεια, οι ψήφοι που κάποιο άρθρο αποκομίζει χρησιμοποιούνται ως ανάδραση (feedback) για την προσαρμογή του προφίλ. Πειραματιστήκαμε με τέσσερις κατηγορίες ανάδρασης:

1) initial diggs
2) boost diggs
3) updated diggs
4) 2 * (updated diggs - initial diggs)

Στην τελευταία περίπτωση, δεν λαμβάνουμε υπόψιν μας τα initial diggs, τα οποία, υπενθυμίζουμε, μπορεί να είναι αποτέλεσμα ενός μεγάλου κοινωνικού δικτύου κάποιου εκδότη ή να εξαρτώνται από τον χρόνο δημοσίευσης. Αντί αυτού χρησιμοποιούμε την διαφορά των τελικών από τους αρχικούς ψήφους ως ένδειξη της “ορμητικότητας” της δημοτικότητας του άρθρου.

Κάθε φορά που αξιολογούμε εκατό άρθρα, μετράμε το πόσο καλά αντιστοιχεί η βαθμολόγησή τους με βάση το περιεχόμενο σε σχέση με την δημοφιλία τους, όπως αυτή υποδηλώνεται από τους τελικούς ψήφους που έχουν αποκομίσει (updated diggs) . Για μια σύγκριση τέτοιου τύπου κάποιος θα μπορούσε να χρησιμοποιήσει ένα (στατιστικό) μέτρο όπως η συσχέτιση Kendall στην ιεράρχηση που προκύπτει. Όμως, τα αρχικά μας πειράματα έδειξαν πως η χρήση ενός μέτρου όπως η απόσταση Kendall είναι ευαίσθητη ειδικά στην περίπτωσή μας όπου έχουμε πολλά άρθρα με τον ίδιο μικρό αριθμό ψήφων και κατά συνέπεια η ιεράρχησή τους είναι τυχαία. Για αυτόν τον λόγο υιοθετήσαμε μια μέθοδο σύγκρισης που περιγράφεται με λεπτομέρεια στο [1]. Σε αυτό το σημείο θα παραθέσουμε μερικά στοιχεία από το συγκεκριμένο

επιστημονικό άρθρο, καθώς και μια υλοποίηση του προτεινόμενου μέτρου σύγκρισης σε γλώσσα Python.



Εικόνα 1. Διαδικασία αξιολόγησης.

Η ιεράρχηση αντικειμένων προϋποθέτει την ύπαρξη κάποιων βαθμολογιών (scores) που έχουν ανατεθεί σε κάθε ένα από αυτά. Πολλά διαφορετικά σχήματα βαθμολόγησης μπορούν να εφαρμοστούν πάνω σε αυτά τα αντικείμενα, λαμβάνοντας υπόψιν τους διάφορες παραμέτρους. Έτσι, δοθέντων δυο διαφορετικών σχημάτων αξιολόγησης - βαθμολόγησης με σκοπό την κατάταξη - ιεράρχηση των αντικειμένων, προκύπτουν δύο εύλογα ερωτήματα:

- Ποιό σχήμα είναι το καλύτερο;
- Πόσο διαφορετικά είναι αυτά τα σχήματα;

Θα μπορούσε κάποιος απλά να συγκρίνει δύο διανύσματα που περιέχουν βαθμολογίες χρησιμοποιώντας έναν συντελεστή συσχέτισης (correlation coefficient) όπως τον αντίστοιχο του Pearson ή του Kendall. Όμως κατ' αυτόν τον τρόπο η ερμηνεία του συγκεκριμένου συντελεστή ως προς την τελική ιεράρχηση έχει χαθεί. Επιπλέον, ο συντελεστής συσχέτισης δεν αποτελεί μέτρο (metric), με συνέπεια να μην μπορούμε να τον ερμηνεύσουμε άμεσα σαν την απόσταση μεταξύ δυο σχημάτων βαθμολόγησης. Θα μπορούσαμε να τον μετασχηματίσουμε σε μέτρο, αλλά ακόμα και τότε δεν θα μπορούσε να αντικατοπτρίσει συγκεκριμένες διαφορές βαθμολογιών που προκύπτουν από την ιεράρχηση. Η εναλλακτική που έχουμε είναι να συγκρίνουμε απ' ευθείας τις βαθμολογίες στην βάση της ιεραρχίας που παράγουν.

Παράδειγμα:

Έστω τρία σχήματα βαθμολόγησης που έχουν αναθέσει διαφορετικές βαθμολογίες σε δυο αντικείμενα i, j.

Σχήμα βαθμολόγησης	**i**	**j**
S1	4	5
S2	5	4
S3	9	1

Τα αντικείμενα i και j βρίσκονται σε ασυμφωνία (discordance) με βάση τα σχήματα βαθμολόγησης S1 και S2, καθώς επίσης και με τα S1, S3. Όμως, μπορούμε να απαντήσουμε στην ερώτηση σε ποια από τις δυο περιπτώσεις υπάρχει μεγαλύτερη ασυμφωνία; Δεν μπορούμε αν δεν δοθεί ορισμός του βαθμού ασυμφωνίας.

Για περισσότερες λεπτομέρειες σχετικά με την διατύπωση του βαθμού ασυμφωνίας (degree of discordance) παραπέμπουμε τον αναγνώστη στο [13]. Παραθέτουμε την υλοποίηση του προτεινόμενου μέτρου με χρήση της γλώσσας Python:

```
def IFunction(a, g):

    '''
    helper function taken from
    "Comparing Scores Intended for Ranking" paper
    a = rank fusion parameter
    g = a/b , where b the other rank fusion parameter
    '''

    a, g = float(a), float(g)
    if a!=0 and g >= (1.0/a): return 0.5
    else: return g*a - 0.5*(g**2)*(a**2)


def MinMax(s1i, s1j, s2i, s2j):

    '''
    helper function to calculate aij and bij
    '''

    return min(s1i-s1j,s2i-s2j) , max(s1i-s1j,s2i-s2j)


def rankDistance(S1, S2, g=1):

    '''
    Note that S1, S2 are lists (and they represent score vectors)
    It indicates a measure of discordance between S1, S2
    '''

    suma = 0
    l=len(S1)

    for i in xrange(0,l-1):

        for j in xrange(i+1,l):
```

```
            a, b = MinMax(S1[i],S1[j],S2[i],S2[j]
            d=0
            if b>=a and a>=0: d = IFunction(b,g) - IFunction(a,g)
            elif a<=0 and b>=0: d = IFunction(b,g) + IFunction(-a,g)
            elif a<=0 and b<=0: d = IFunction(-a,g) - IFunction(-b,g)
            suma += d

    return suma
```

Ο συγκεκριμένος βαθμός ασυμφωνίας έχει αποδειχτεί ότι είναι επαρκές μέτρο και παίρνει τιμές στο διάστημα 0 , n(n -1) / 2, όπου n ο αριθμός των υπό σύγκριση βαθμολογιών. Στην περίπτωσή μας (n = 100) η μέγιστη τιμή του μέτρου είναι 4950. Όσο **μικρότερος** αυτός ο αριθμός τόσο καλύτερο το ταίριασμα μεταξύ των δυο βαθμολογιών.

4.3 ΥΛΟΠΟΙΗΣΗ ΝΟΟΤΡΟΠΙΑΣ

Χρησιμοποιήσαμε δυο εκδόσεις της Νοοτροπίας. Στην πρώτη περίπτωση και σε σύγκριση με άλλες εκδόσεις όπως περιγράφεται στα [6, 8], υπάρχουν δυο βασικές διαφορές. Δεν χρησιμοποιήσαμε την τεχνική βεβαρημένων όρων [8] για να αναθέσουμε βάρη σε κάθε λέξη ενός κειμένου, και αντί αυτού θέσαμε το βάρος κάθε λέξης στην ποσότητα 1 / (πλήθος λέξεων κειμένου). Επιπλέον, αν και στο παρελθόν χρησιμοποιήθηκε μια μέθοδο παραθύρου για τη σάρωση των λέξεων του κειμένου και την εύρεση των μεταξύ τους συσχετίσεων [7, 8], εδώ όλες οι λέξεις σε ένα κείμενο θεωρούνται σε συσχέτιση με όλες τις άλλες. Συγκεκριμένα, το βάρος w_{kn} του συνδέσμου μεταξύ δύο λέξεων k και n υπολογίζεται χρησιμοποιώντας την παρακάτω εξίσωση, η οποία υιοθετείται επίσης στο [14].

$$wkn = \frac{fr_{kn^2}}{fr_k \cdot fr_n} \quad (1)$$

Όπου :

fr_{kn} είναι ο αριθμός των φορών όπου το k και το n συνυπάρχουν στο ίδιο κείμενο

fr_k , fr_n είναι ο αριθμός όπου το k και το n εμφανίζονται στο κείμενο.

Χρησιμοποιήσαμε την απλουστευμένη έκδοση της Νοοτροπίας για δύο βασικούς λόγους. Πρώτον αποφεύγουμε να κάνουμε περίπλοκη την διαδικασία αξιολόγησης χωρίς λόγο, με επιπλέον παραμέτρους. Για παράδειγμα, υιοθετώντας ένα απλό σχήμα βεβαρημένων όρων, δεν χρειάζεται να διατηρούμε στατιστικά σχετικά με την συχνότητα εμφάνισης των λέξεων ή να ορίζουμε κατώφλια - όρια για την εξαγωγή όρων. Επιπρόσθετα, χρησιμοποιώντας το πλήρες περιεχόμενο των άρθρων για να βρούμε τις συσχετίσεις των όρων, δεν χρειάζεται να ορίσουμε ένα συγκεκριμένο μέγεθος παραθύρου ολίσθησης (sliding window). Η επιλογή μας αυτή είναι επίσης “ασφαλής” υπό την έννοια ότι το περιεχόμενο των άρθρων του Digg είναι σχετικά μικρό και περιεκτικό. Ο δεύτερος και πιο σημαντικός λόγος έχει να κάνει με το γεγονός ότι αυτή είναι η πρώτη μας προσπάθεια για πρόβλεψη δημοφιλίας με βάση το περιεχόμενο. Αν μπορούμε να αποδείξουμε ότι η πιο απλή έκδοση της Νοοτροπίας μπορεί να ανταπεξέλθει ικανοποιητικά σε αυτή την διαδικασία, τότε μπορούμε να είμαστε πεπεισμένοι ότι μελλοντικές βελτιώσεις είναι πιθανές και ότι η έρευνα σε ένα τέτοιο τομέα μπορεί να αποδώσει καρπούς.

4.4 ΑΡΧΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ

Ο πίνακας 2 παρουσιάζει τον μέσο όρο και την τυπική απόκλιση από τις τιμές ασυμφωνίας για όλα τα άρθρα ανά κατηγορία και ανά τύπο ανάδρασης. Ο πίνακας επίσης περιλαμβάνει και την βάση πάνω στην οποία έγινε η σύγκριση, που υποδηλώνεται με το “by-date”. Πρόκειται για μια σύγκριση που περιλαμβάνει τον βαθμό ασυμφωνίας μεταξύ της ιεράρχησης των άρθρων με βάση την ημερομηνία δημοσίευσης και της ιεράρχησης με βάση τον τελικό αριθμό των ψήφων που αποκόμισαν. Χρησιμοποιήσαμε την ιεράρχηση με βάση τον χρόνο και όχι κάποια άλλη (τυχαία), γιατί ακολουθεί διαισθητικά τον τρόπο παρουσίασης των

άρθρων στην συντριπτική πλειονότητα των ιστότοπων ειδησεογραφίας (τα νεότερα άρθρα εμφανίζονται πρώτα). Τα αποτελέσματα δείχνουν ότι άσχετα με τον τύπο ανάδρασης ή την κατηγορία, οι τιμές ασυμφωνίας της Νοοτροπίας (σε σχέση με τις τελικές ψήφους) είναι σημαντικά καλύτερες (μικρότερες) σε σχέση με την βάση σύγκρισης. Εκτελέσαμε επίσης ένα δίπλευρο t-test μεταξύ των διαφορετικών τύπων ανάδρασης και της βάσης σύγκρισης και σε όλες τις περιπτώσεις η τιμή p που προέκυπτε ήταν πολύ μικρή ($p << 0.00001$).

Όπως ήταν αναμενόμενο, τα αποτελέσματα επίσης δείχνουν ότι όσο πιο πολύ επιμηκύνουμε την περίοδο που μετράμε τις ψήφους, οι βαθμοί ασυμφωνίας μειώνονται. Χρησιμοποιώντας τα boost diggs σαν ανάδραση είναι προφανώς καλύτερο από το να χρησιμοποιήσουμε τα initial diggs, και η απόδοση μεγαλώνει ακόμα περισσότερο στην περίπτωση των updated diggs. Ενδιαφέρον παρουσιάζει το γεγονός ότι η καλύτερη απόδοση πετυχαίνεται όταν δεν λαμβάνουμε υπόψιν μας τα initial diggs (τελευταία περίπτωση στον πίνακα 2, updated - initial). Αυτό αποτελεί άλλη μια ένδειξη ότι οι αρχικοί ψήφοι δεν αντανακλούν πάντα τις προτιμήσεις μιας κοινότητας, όπως έχουμε ήδη συζητήσει.

	first half average	second half average
ALL	933.98	972.46
BUSINESS	750.73	842.22
ENTERTAINMENT	822.29	940.82
LIFESTYLE	1022.79	986.51
OFFBEAT	1044.88	976.04
POLITICS	988.04	937.83
SCIENCE	1130.99	1065.59
SPORTS	1157.65	1034.11
TECHNOLOGY	954.87	982.73
WORLD	720.17	968.71

Πίνακας 3: Μέσες τιμές βαθμών ασυμφωνίας για το πρώτο και δεύτερο μισό της πειραματικής διαδικασίας για κάθε κατηγορία.

CATEGORY		by date	initial	boost	updated	updated-initial
ALL	average	1417.24	984.54	988.71	965.66	953.22
	st.deviation	104.7	202.40	205.45	207.32	208.66
BUSINESS	average	1369.52	865.11	861.78	822.05	796.87
	st.deviation	91.64	184.14	181.76	190.35	187.13
ENTERTAINMENT	average	1411.64	968.95	942.04	899.50	881.77
	st.deviation	142.5	217.39	221.77	219.81	226.64
LIFESTYLE	average	1427.53	993.42	998.13	980.30	1006.21
	st.deviation	143.63	186.80	192.82	186.24	167.64
OFFBEAT	average	1429.85	1047.63	1033.19	1016.97	1009.81
	st.deviation	120.65	186.16	142.30	128.61	119.25
POLITICS	average	1402.54	949.30	960.80	968.88	962.66
	st.deviation	84.30	167.15	168.87	176.68	171.58
SCIENCE	average	1480.54	1088.19	1072.90	1091.45	1097.16
	st.deviation	113.42	197.90	202.88	179.62	169.94
SPORTS	average	1456.94	1224.53	1190.24	1123.01	1098.27
	st.deviation	147.68	239.77	241.63	240.17	210.19
TECHNOLOGY	average	1400.42	1002.52	987.74	970.86	968.80
	st.deviation	112.58	174.24	194.52	185.29	178.58
WORLD	average	1377.71	921.99	915.62	866.20	845.90
	st.deviation	86.06	206.50	210.32	246.11	243.16
	macro average	1417.39	1004.6	995.12	970.49	962.07

Πίνακας 2: Μέσος όρος βαθμού ασυμφωνίας και τυπικές αποκλίσεις για ολόκληρη τη συλλογή άρθρων και ανά κατηγορία.

Αξιοσημείωτο είναι επίσης το γεγονός ότι ο βαθμός ασυμφωνίας για όλα τα άρθρα (η περίπτωση ALL), είναι πολύ καλύτερη από αρκετές κατηγορίες. Αυτό καταδεικνύει την ικανότητα της Νοοτροπίας να διαχειρίζεται έναν μεγάλο πλήθος λέξεων που απαιτείται για να αναπαρασταθούν τα συλλογικά ενδιαφέροντα μιας κοινότητας, με ένα μόνο προφίλ.

4.5 ΒΕΛΤΙΩΣΕΙΣ

Σε αυτό το σημείο επαναλαμβάνουμε τα πειράματα χρησιμοποιώντας μια διαφορετική έκδοση της Νοοτροπίας και λαμβάνοντας υπόψιν κάποια στατιστικά μέτρα που ενυπάρχουν στην συλλογή των άρθρων μας. Ουσιαστικά, υιοθετούμε την μέθοδο ολίσθησης παραθύρου και την τεχνική των βεβαρημένων όρων, αποσκοπώντας σε ακόμα μεγαλύτερη ακρίβεια του προφίλ.

Όπως προαναφέραμε, με την τεχνική ολίσθησης παραθύρου πετυχαίνουμε καλύτερη εξαγωγή συσχετίσεων μεταξύ των όρων ενός άρθρου. Πλέον, όλες οι λέξεις του κειμένου δεν βρίσκονται σε συσχέτιση μεταξύ τους, αλλά θεωρούμε ένα μήκος παραθύρου ίσο με 20 για να εγκλωβίσουμε νοηματικά την ροή του κειμένου και να αντικατοπτρίσουμε καλύτερα την αλληλοεξάρτηση των λέξεων. Επιπλέον, θέσαμε το βάρος των όρων στην αντίστοιχη tf-idf τιμή τους. Για τον σκοπό αυτό, αρχικά εξάγαμε όλες τις συχνότητες των λέξεων από την συλλογή άρθρων μας, και παράγουμε δυναμικά ανά άρθρο το tf-idf κάθε λέξης - όρου. Χρησιμοποιώντας ένα κατώφλι (threshold) τιμών για το φιλτράρισμα όρων με χαμηλό tf-idf (δηλαδή των λιγότερο σημαντικών - περισσότερο κοινών όρων), δίνουμε μεγαλύτερο βάρος σε λέξεις οι οποίες είναι πιο συγκεκριμένες και σημαντικές για την ομαλή λειτουργία της όλης διαδικασίας.

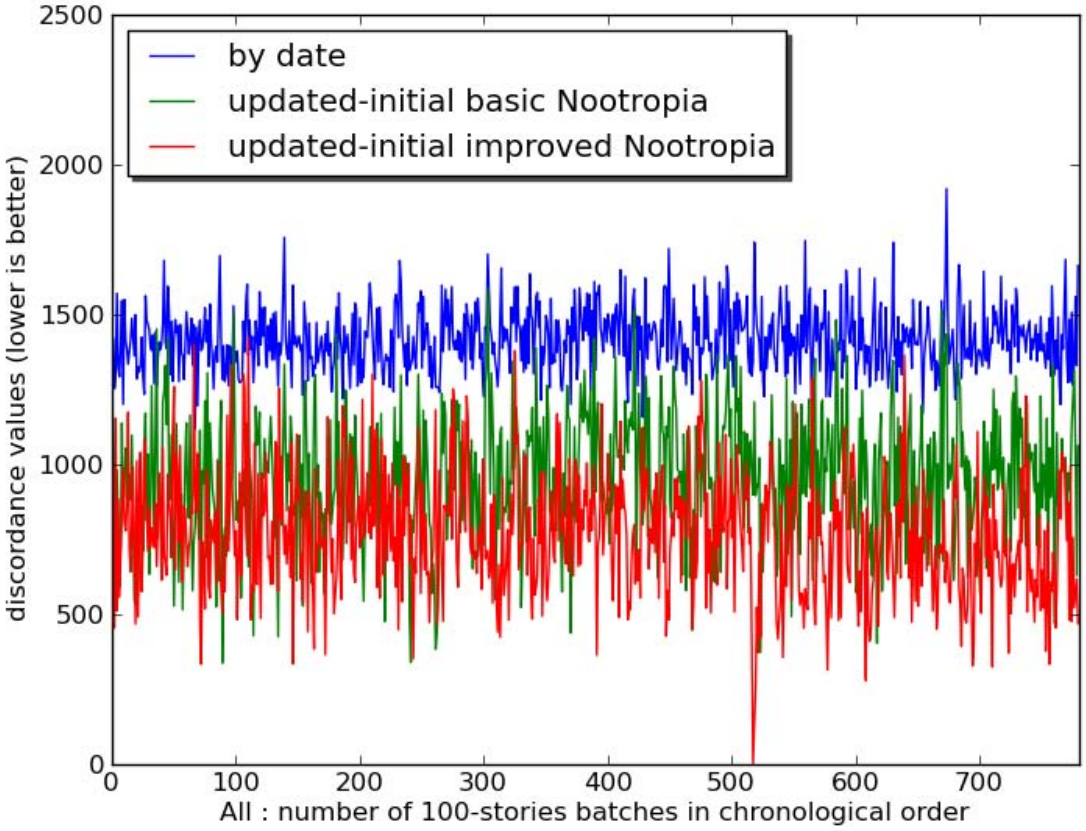
Ορισμός tf-idf (term frequency - inverse document frequency): Μια ποσότητα που χρησιμοποιείτε ευρέως στο πεδίο της ανάκτησης πληροφορίας (information retrieval) και αποτελεί στατιστικό μέτρο που αναπαριστά το πόσο σημαντική είναι μια λέξη μέσα σε κάθε κείμενο μιας συλλογής. Η σημαντικότητα αυξάνεται αναλογικά με το πόσες φορές η λέξη εμφανίζεται στο κείμενο αλλά λαμβάνοντας υπ’ όψιν και την συχνότητα εμφάνισής της στην συλλογή. Σαν μαθηματική αναπαράσταση ορίζεται ως το γινόμενο μεταξύ:

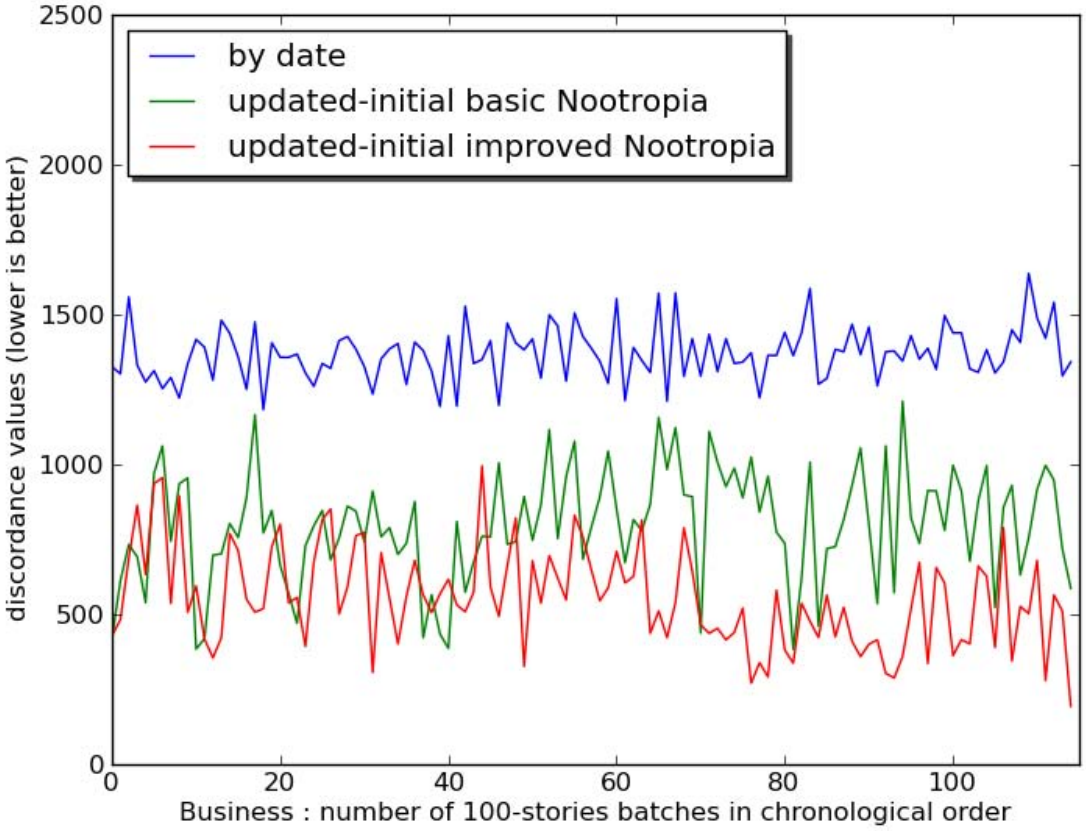
- tf : πόσες φορές εμφανίζεται η λέξη στο κείμενο
- idf : $\frac{|D|}{|\{d \; E \; D{:}t \; E \; d\}|}$ όπου
  - |D| ο συνολικός αριθμός των κειμένων της συλλογής
  - |{d E D : t E d}| ο αριθμός των κειμένων στον οποίο εμφανίζεται η λέξη t.

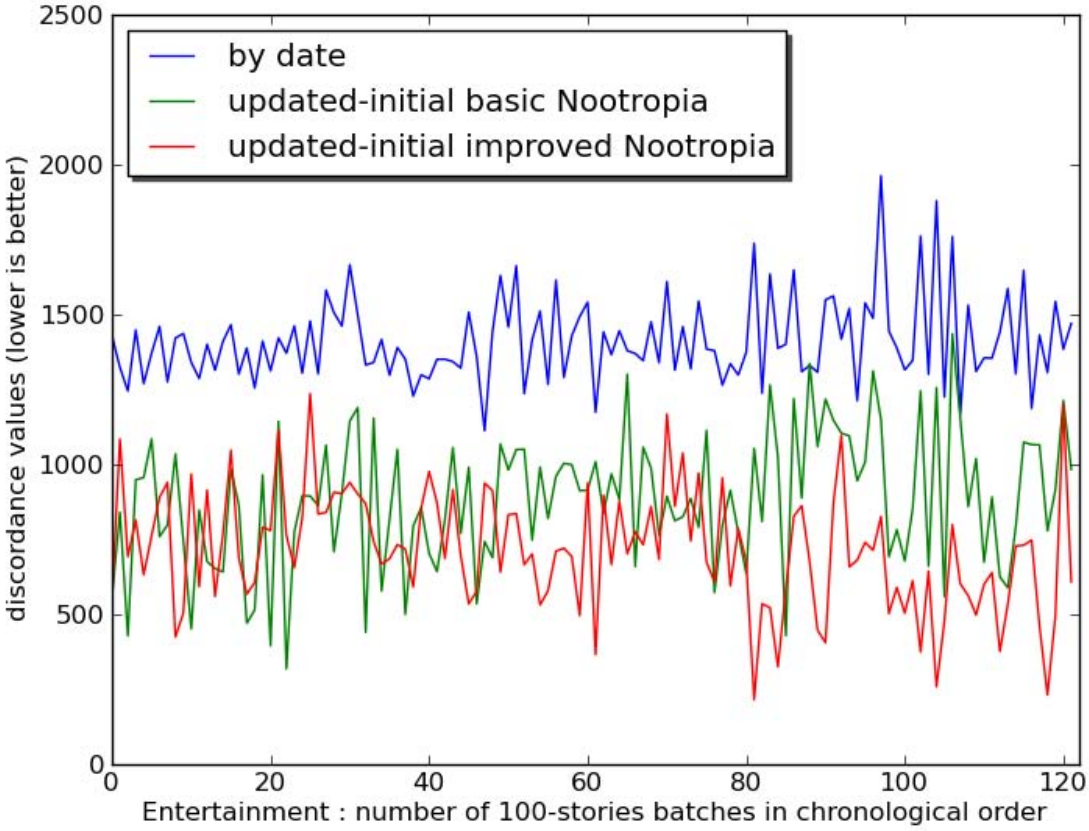
Παρατηρούμε πως η απόδοση του προφίλ βελτιώθηκε κατά μέσο όρο περίπου 20% σε σχέση με τα προηγούμενα αποτελέσματα, και σε αρκετές κατηγορίες βελτιώνεται με την πάροδο του χρόνου. Στον πίνακα 4 συνοψίζουμε τις μέσες τιμές βαθμών ασυμφωνίας για το πρώτο και δεύτερο μισό των πειραμάτων, παρατηρώντας ότι πλέον για 6 από τις 10 κατηγορίες τα αποτελέσματα βελτιώθηκαν.

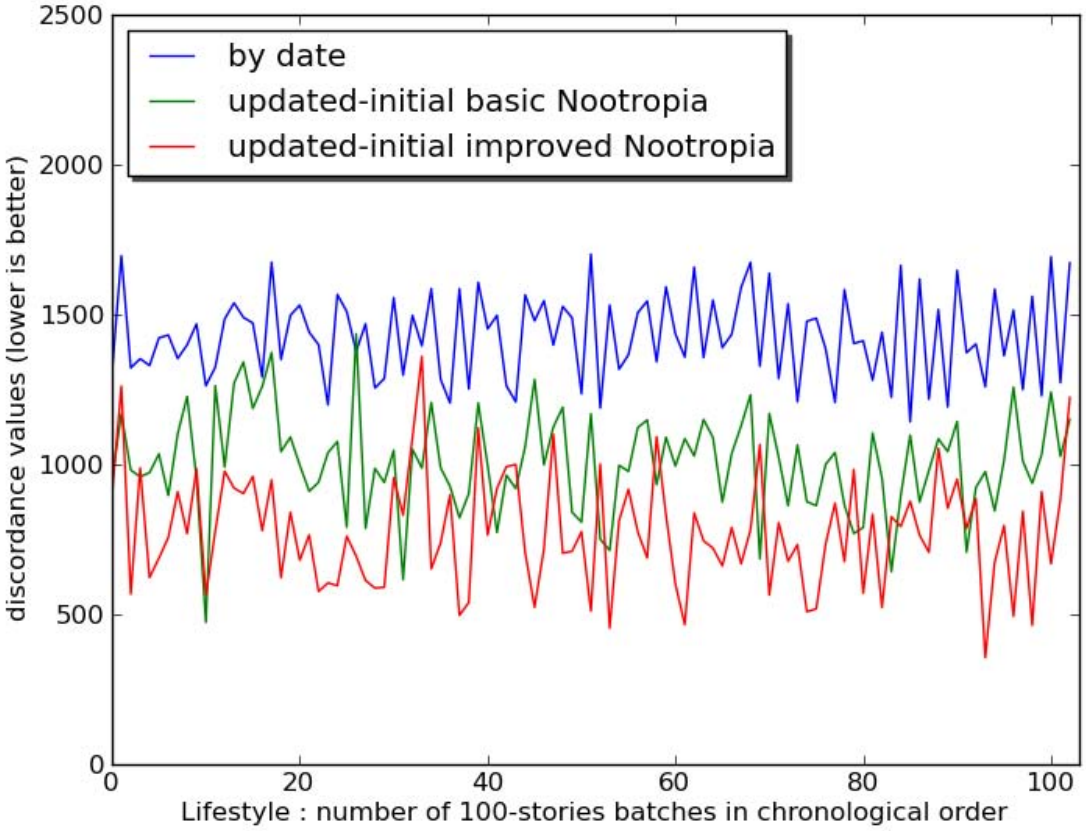
4.6 ΤΕΛΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ

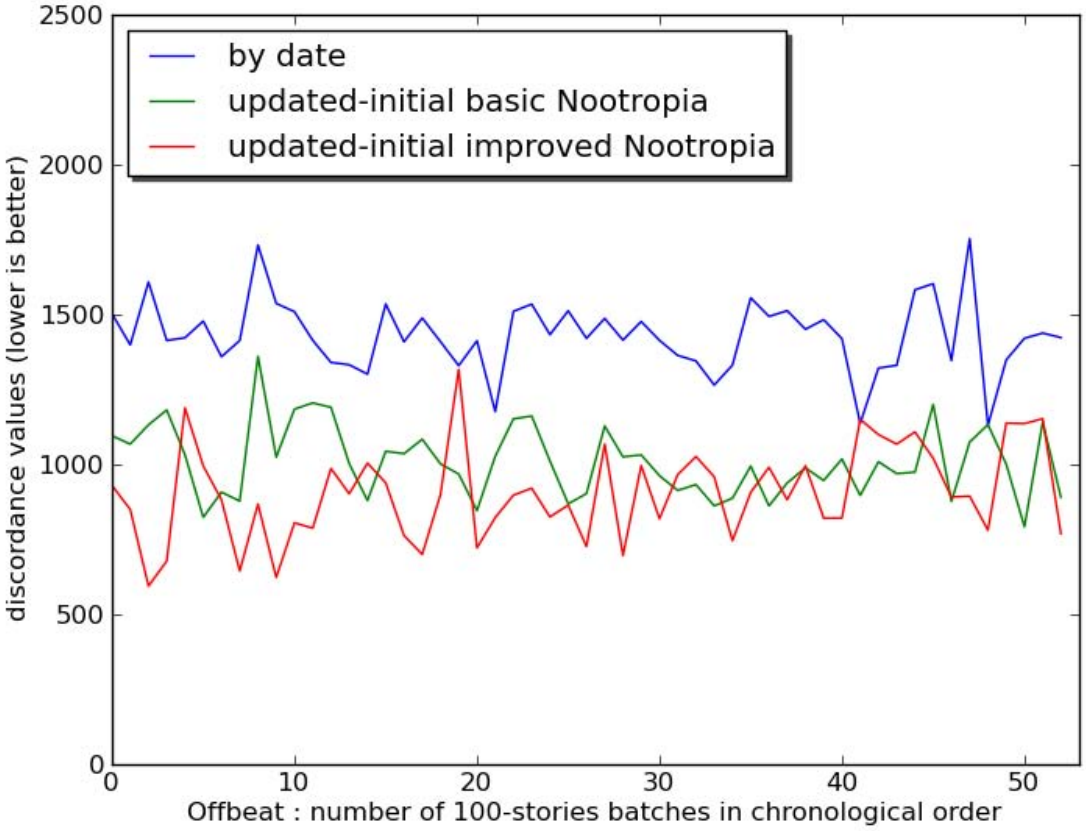
Για να αξιολογήσουμε την απόδοση της Νοοτροπίας στις δυο περιπτώσεις, οι εικόνες 1 - 10 παρουσιάζουν το πώς ο βαθμός ασυμφωνίας αλλάζει με την πάροδο του χρόνου. Ο άξονας των Χ αντιστοιχεί στον αριθμό κάθε τεμαχίου 100 άρθρων, που έχει επεξεργαστεί το προφίλ. Σε αυτό το σημείο πρέπει να υπενθυμίσουμε ότι η πειραματική διαδικασία είναι συνεχής. Το αρχικά άδειο προφίλ αξιολογεί και προσαρμόζεται σε άρθρα μέχρι αυτά να τελειώσουν. Ο άξονας των Υ αντιστοιχεί στον βαθμό ασυμφωνίας μεταξύ των τεμαχίων αυτών. Στην συνέχεια παραθέτουμε μόνο τις τιμές της περίπτωσης updated - initial, όπου και παρατηρήθηκε καλύτερη απόδοση. Με πράσινο φαίνονται τα αποτελέσματα για την πρώτη έκδοση της Νοοτροπίας, ενώ με κόκκινο για την δεύτερη:

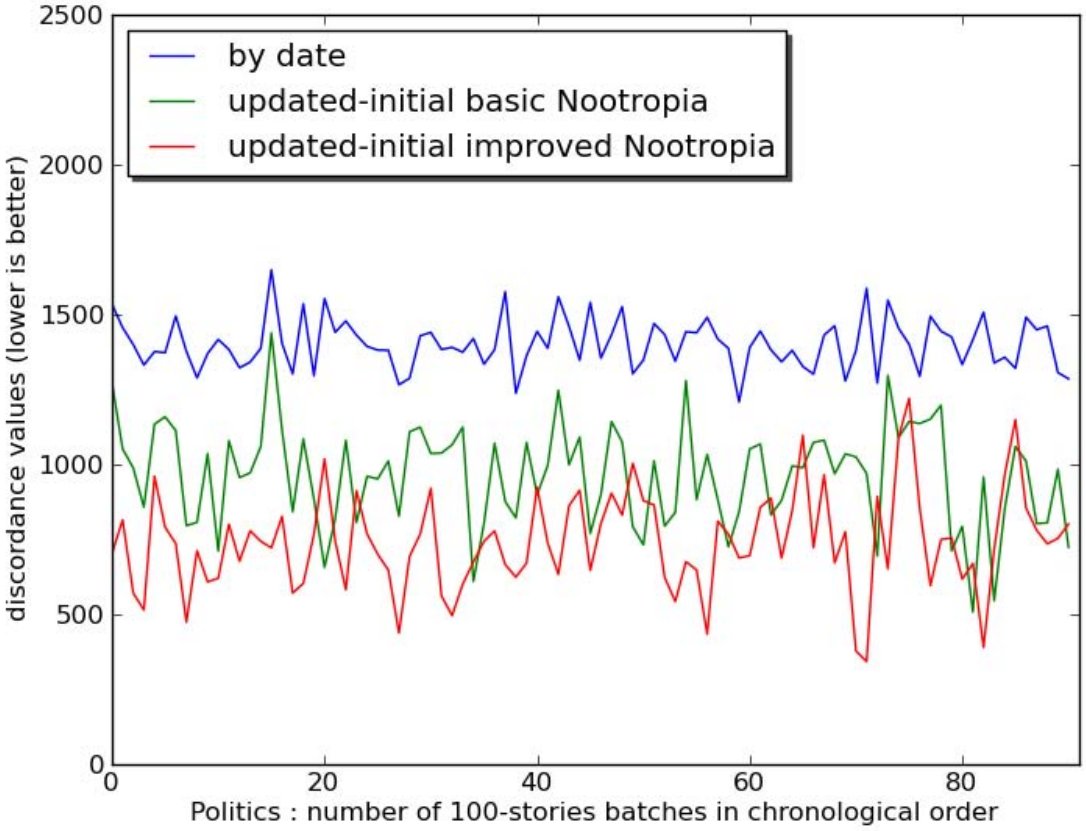


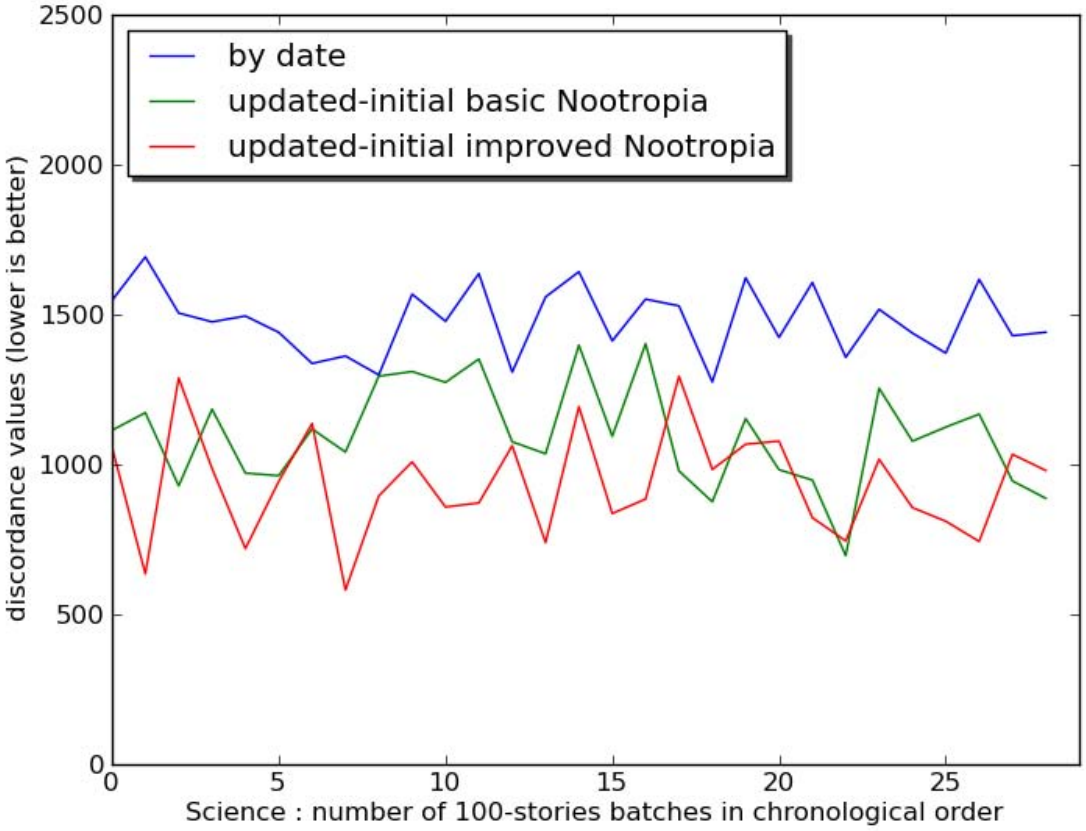


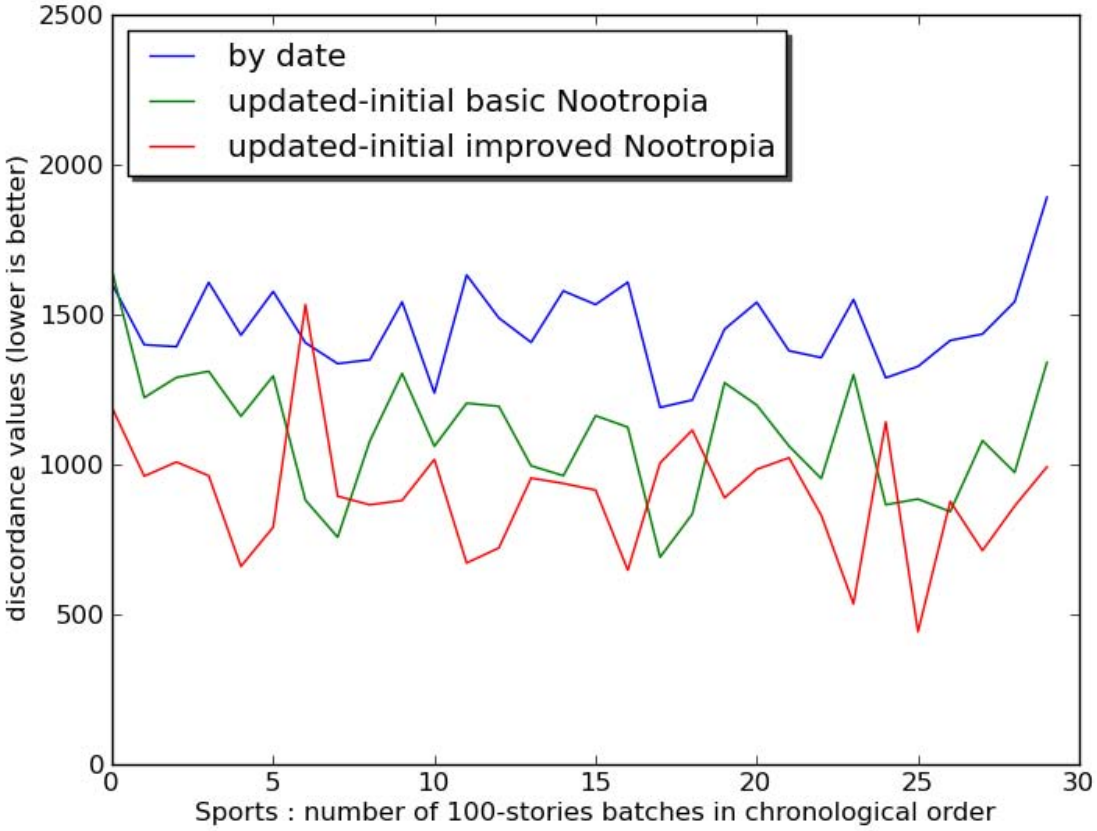


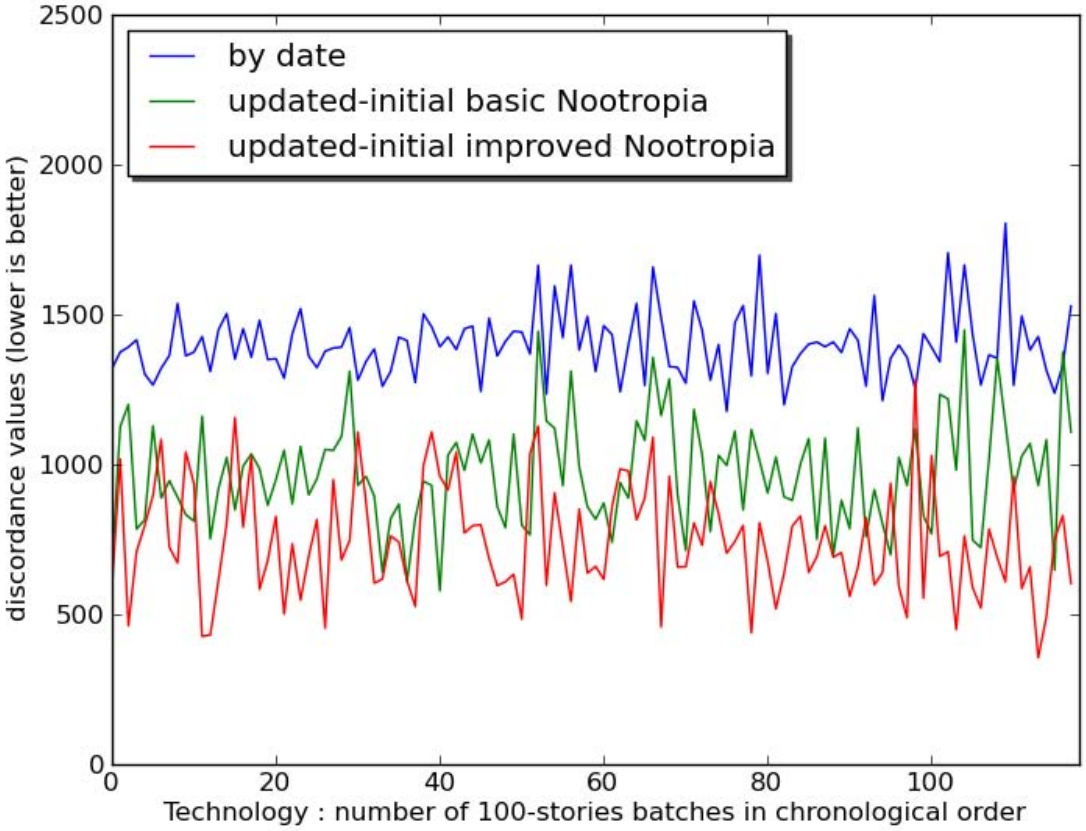


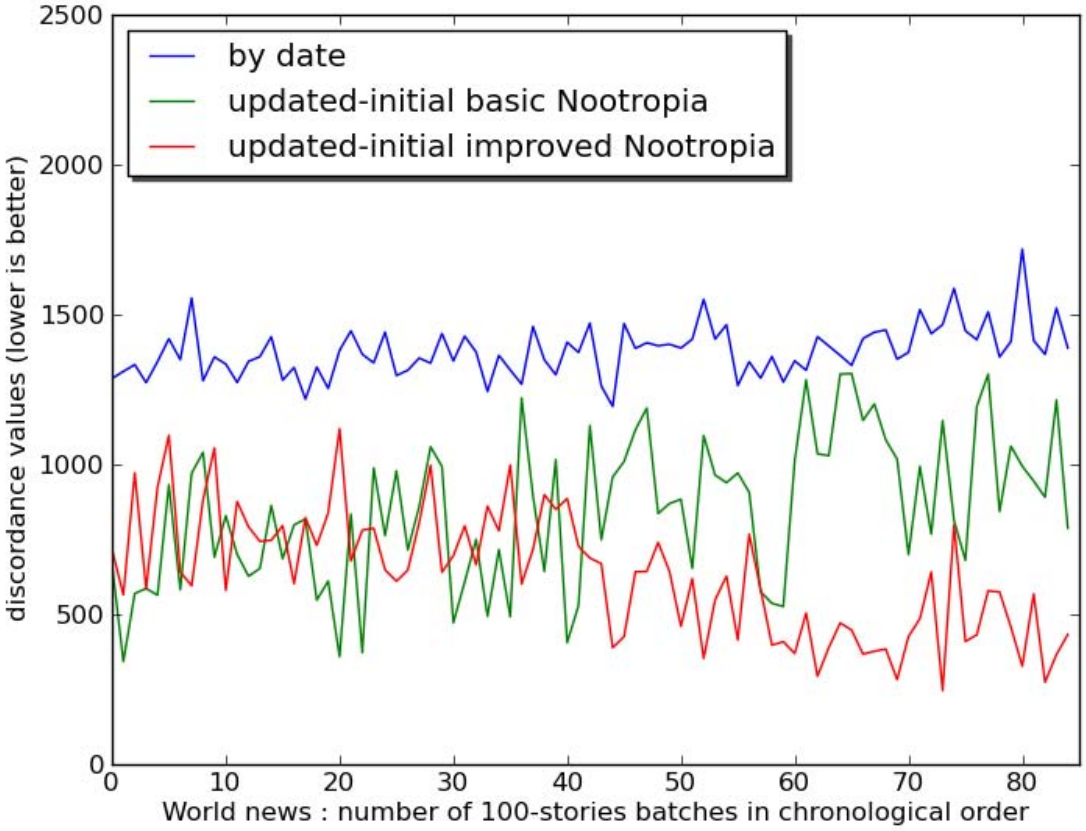












Είναι εμφανές από τα γραφήματα ότι η διαφορά της Νοοτροπίας και της βάσης στην οποία έγινε η σύγκριση (by date) είναι σημαντική και συνεπής για κάθε ξεχωριστό πείραμα που διεξάγαμε. Επιπλέον, παρά τις διακυμάνσεις ο μέσος όρος της απόδοσης σε κάθε περίπτωση επιτυγχάνεται γρήγορα αφού το πρώτο τεμάχιο 100 άρθρων έχει χρησιμοποιηθεί για να προσαρμοστεί το προφίλ. Αυτή η απόδοση διατηρείται μέχρι περατώσεως των πειραμάτων. Αποτελεί δε απόδειξη ότι η Νοοτροπία μπορεί γρήγορα να μάθει τα συλλογικά ενδιαφέροντα μιας κοινότητας και να προσαρμόζεται στις όποιες αλλαγές λαμβάνουν χώρα με το πέρασμα του χρόνου. Σύμφωνα με τον πίνακα 4, όπου συνοψίζονται οι μέσες τιμές βαθμών ασυμφωνίας στο πρώτο και δεύτερο μισό των πειραμάτων, για πέντε από τις δέκα κατηγορίες και ειδικά για αυτές με μικρό αριθμό άρθρων, η απόδοση γίνεται καλύτερη στο δεύτερο μισό. Από την άλλη, για τις υπόλοιπες κατηγορίες, και ειδικά για τις “World” και “Entertainment”, η απόδοση πέφτει με την πάροδο του χρόνου.

	first half average	second half average
ALL	802.57	739.54
BUSINESS	620.93	485.70
ENTERTAINMENT	767.52	667.36
LIFESTYLE	800.25	759.24
OFFBEAT	861.92	957.79
POLITICS	712.87	769.97
SCIENCE	913.13	939.43
SPORTS	936.39	860.97
TECHNOLOGY	763.92	723.30
WORLD	779.84	481.84

Πίνακας 4: Μέσες τιμές βαθμών ασυμφωνίας για το πρώτο και δεύτερο μισό της πειραματικής διαδικασίας για κάθε κατηγορία.

4.7 ΣΥΜΠΕΡΑΣΜΑΤΑ

Συνολικά, μπορούμε να ισχυριστούμε ότι τα πειραματικά μας αποτελέσματα καταδεικνύουν ότι η πρόβλεψη δημοφιλίας είναι εφικτή. Η απόδοση της Νοοτροπίας σε σχέση με την βάση σύγκρισης είναι εμφανώς καλύτερη, και παρά την απλή έκδοση που χρησιμοποιήσαμε, μπορεί να παράγει βαθμολογίες πρόβλεψης δημοφιλίας που θα μπορούσαν να αλλάξουν την δυναμική της παρουσίασης σε ιστότοπους όπως το Digg ή ακόμα να παρέχουν μια διαφορετική ιεράρχηση σε πρόσφατα άρθρα. Αυτό θα είχε ως αποτέλεσμα μια διαφορετική προοπτική παρουσίασης αρθογραφίας, καθώς θα μειώνονταν ο αντίκτυπος που έχει η αρχική ψηφοφορία στην μελλοντική δημοτικότητα ενός άρθρου και θα ενίσχυε τον ρόλο που το περιεχόμενο του παίζει στην όλη διαδικασία. Αν και θεωρητικά, ένα τέλειο και ακριβές σύστημα πρόβλεψης δημοφιλίας με βάση το περιεχόμενο θα μπορούσε να αντικαταστήσει κάποιο ήδη υπάρχων αλγοριθμικό σύστημα ψηφοφορίας, πιστεύουμε ότι υπάρχει μεγαλύτερο όφελος από την υϊοθέτηση ενός υβριδικού συστήματος που να συνδυάζει τις δύο προσεγγίσεις.

5. ΣΧΕΔΙΑΣΗ ΕΦΑΡΜΟΓΗΣ

Στο παρόν κεφάλαιο θα σχεδιάσουμε μια ολοκληρωμένη web εφαρμογή η οποία αποτελεί μια πλατφόρμα παρουσίασης ειδησεογραφίας βασισμένη στην μέθοδο πρόβλεψης δημοφιλίας των άρθρων που έχουμε ήδη περιγράψει. Ουσιαστικά πρόκειται για έναν ιστότοπο όπου οι χρήστες του θα διαμορφώνουν δυναμικά τον τρόπο με τον οποίο τους παρουσιάζονται τα άρθρα, ο οποίος και θα αντανακλά τα συλλογικά τους ενδιαφέροντα. Το κεφάλαιο δομείται ως εξής: Πρώτα θα γίνει μια συνοπτική παρουσίαση των προγραμματιστικών εργαλείων που χρησιμοποιήθηκαν για την κατασκευή της εφαρμογής. Στη συνέχεια θα δοθεί η αρχιτεκτονική της και θα αναλυθούν τα σενάρια χρήσης της. Τέλος, θα κάνουμε μια πιο αναλυτική και επιλεκτική παρουσίαση ορισμένων σημείων με την επίδειξη του αντίστοιχου πηγαίου κώδικα.

5.1 ΠΡΟΓΡΑΜΜΑΤΙΣΤΙΚΑ ΕΡΓΑΛΕΙΑ

Για την ανάπτυξη της εφαρμογής χρησιμοποιήσαμε στην συντριπτική πλειοψηφία την γλώσσα προγραμματισμού python, ενώ μειοψηφία αποτελούν οι html, css, coffeescript, με τις οποίες κατασκευάστηκε η διεπαφή χρήστη. Ο server μας χτίστηκε με το tornado web framework και η ασύγχρονη εκτέλεση διεργασιών στο παρασκήνιο με το celery. Η βάση δεδομένων μας είναι η mongoDB ενώ για caching και business-logic χρησιμοποιούμε το redis. Ας τα δούμε λίγο πιο αναλυτικά:

Η **python** (http://python.org/) είναι μια διερμηνευόμενη γλώσσα προγραμματισμού υψηλού επιπέδου, με έμφαση στην ευκολία σύνταξης και διαχείρισης του κώδικα, χωρίς να υπολείπεται σε δυνατότητες. Παρέχει στον προγραμματιστή μια πληθώρα βιβλιοθηκών και πολλαπλές μεθοδολογίες δόμησης προγραμμάτων, όπως διαδικαστικό, αντικειμενοστραφή και scripting προγραμματισμό. Το οικοσύστημα το οποίο έχει δημιουργηθεί γύρω της την καθιστά

ευρέως αποδεκτή και σε αυτό έχει βοηθήσει η υϊοθέτησή της από την Google και διάφορα πανεπιστήμια του εξωτερικού (MIT κ.α.).

Το **tornado** (http://www.tornadoweb.org) είναι ένας server και micro web-framework ανοιχτού κώδικα γραμμένο σε python. Η διαφορά του με άλλους servers έγκειται στο ότι είναι non-blocking και εξαιρετικά γρήγορος. Επειδή είναι non-blocking και χρησιμοποιεί epoll ή kqueue[1] , μπορεί να διαχειριστεί χιλιάδες ταυτόχρονες ανοιχτές συνδέσεις, το οποίο το καθιστά ιδανικό για web εφαρμογές πραγματικού χρόνου. Για περισσότερες πληροφορίες σχετικά με την κλιμάκωση των servers ώστε να εξυπηρετούν χιλιάδες χρήστες, ο αναγνώστης παραπέμπεται στο γνωστό C10K πρόβλημα (http://www.kegel.com/c10k.html).

Το **celery** (http://celeryproject.org/) είναι ένα εργαλείο γραμμένο σε python, το οποίο είναι υπεύθυνο για την ασύγχρονη εκτέλεση διάφορων διεργασιών (tasks) στο παρασκήνιο. Πρόκειται για μια ουρά διεργασιών (task (job) queue) που βασίζεται σε κατανεμημένη μετάδοση μηνυμάτων. Είναι ιδανικό για την αποφόρτιση του server από βαριές και χρονοβόρες διεργασίες, οι οποίες μπορούν να εκτελεστούν κάποια στιγμή στο μέλλον.

Για την βάση δεδομένων μας διαλέξαμε να χρησιμοποιήσουμε όχι κάποια κλασσική σχεσιακή (mySQL, postgreSQL), αλλά μια από τις πιο δημοφιλείς NoSQL λύσεις, την **mongoDB** (http://www.mongodb.org/). Αυτό το κάναμε καθαρά για λόγους απόδοσης, καθώς η ειδησεογραφία έχει μικρό χρόνο ζωής, οπότε είναι σαφέστατα πιο καλή η υϊοθέτηση μιας NoSQL λύσης. Η mongoDB είναι μια document-oriented βάση δεδομένων, εξαιρετικά γρήγορη όταν το dataset μπορεί να χωρέσει στην κύρια μνήμη του server. Επιπλέον υπάρχουν αξιοπρεπείς drivers και ORMs για python, όπως οι mongoengine (http://mongoengine.org/) και asyncmongo (https://github.com/bitly/asyncmongo).

Το **redis** (http://redis.io) είναι μια ανοιχτού κώδικα, εξελιγμένη αποθήκη κλειδιών - τιμών. Συχνά αναφέρεται ως ένας server δομών δεδομένων, καθώς μπορεί να περιέχει ως τιμές λίστες, σύνολα, ταξινομημένα σύνολα και δομές κατακερματισμού. Χρησιμοποιείται ώς κρυφή μνήμη καθώς και για οτιδήποτε χρειάζεται ταχύτατη αποθήκευση, επεξεργασία και προσπέλαση.

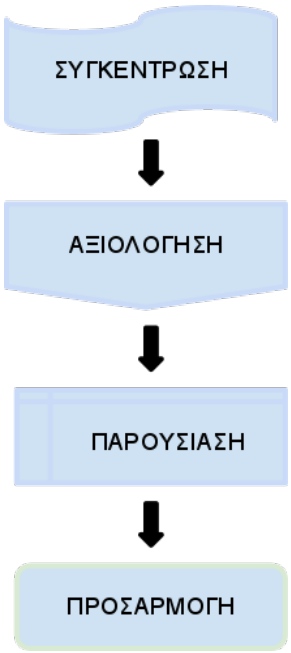
Η **coffeescript** (http://coffeescript.org/) είναι μια νέα scripting γλώσσα προγραμματισμού που μεταγλωττίζεται κατευθείαν σε javascript. Η σύνταξή της παραπέμπει ευθέως σε python /

1 I/O event notification facilities, Linux and BSD respectively.

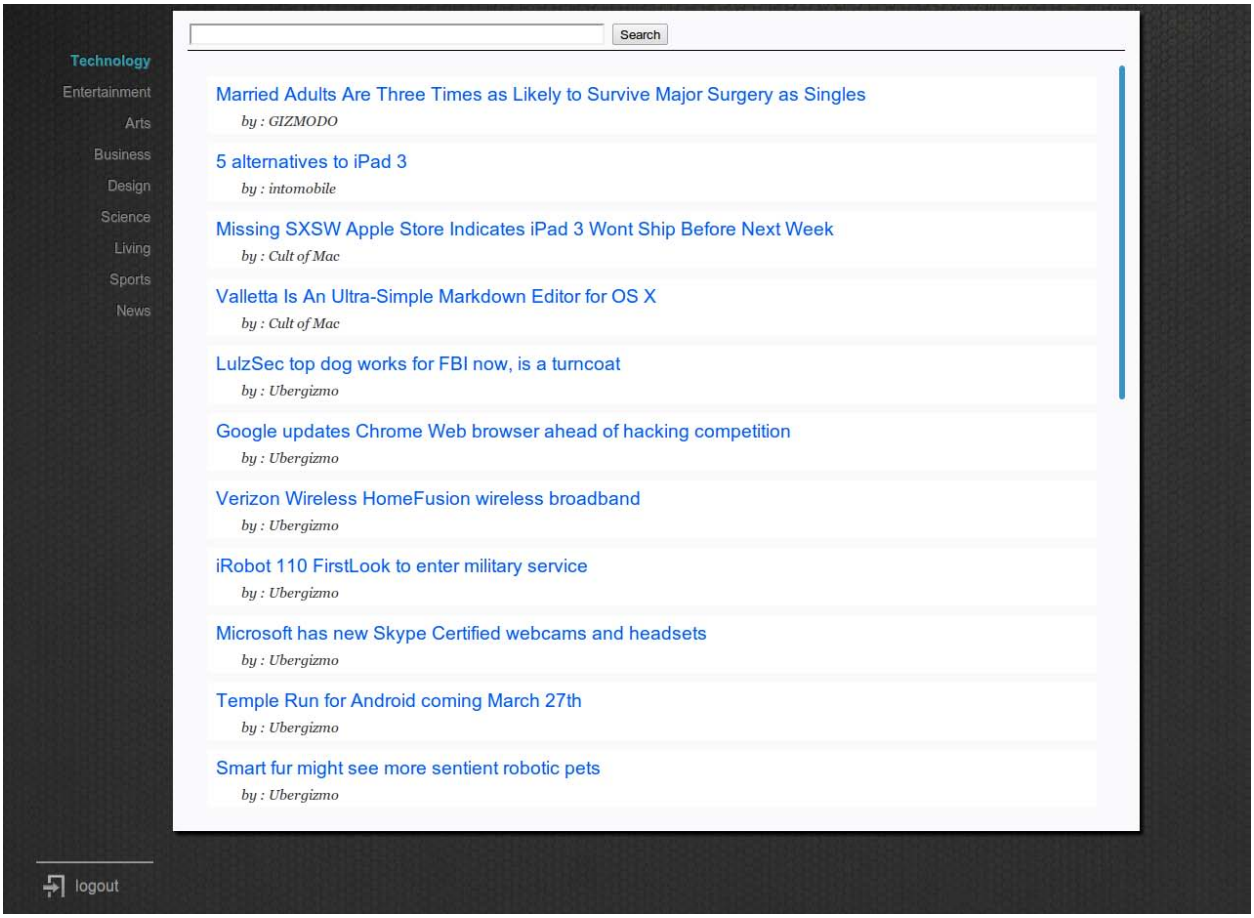
ruby, και ουσιαστικά παρέχει την δυνατότητα στον προγραμματιστή να γράφει βέλτιστο javascript κώδικα χωρίς να περιορίζεται από τις ιδιαιτερότητες της σύνταξης javascript προγραμμάτων.

5.2 ΑΡΧΙΤΕΚΤΟΝΙΚΗ ΣΥΣΤΗΜΑΤΟΣ

Κεντρικό ρόλο στην εφαρμογή παίζουν τα άρθρα, τα οποία πρέπει να συγκεντρωθούν και να αξιολογηθούν από το προφίλ. Για τον σκοπό αυτό είναι απαραίτητη η δημιουργία ενός aggregator, ο οποίος μέσω διάφορων πηγών rss θα παίρνει τα εκάστοτε άρθρα και θα τα προωθεί στο προφίλ. Από εκεί, αφού αξιολογηθούν και τους ανατεθεί κάποιος βαθμός σημαντικότητας - δημοφιλίας, θα αποθηκεύονται στην βάση δεδομένων και θα είναι διαθέσιμα πρός ανάγνωση. Οι χρήστες θα μπορούν να το ψηφίσουν είτε θετικά είτε αρνητικά και αυτό θα αποτελεί την ανάδραση για την προσαρμογή του προφίλ.

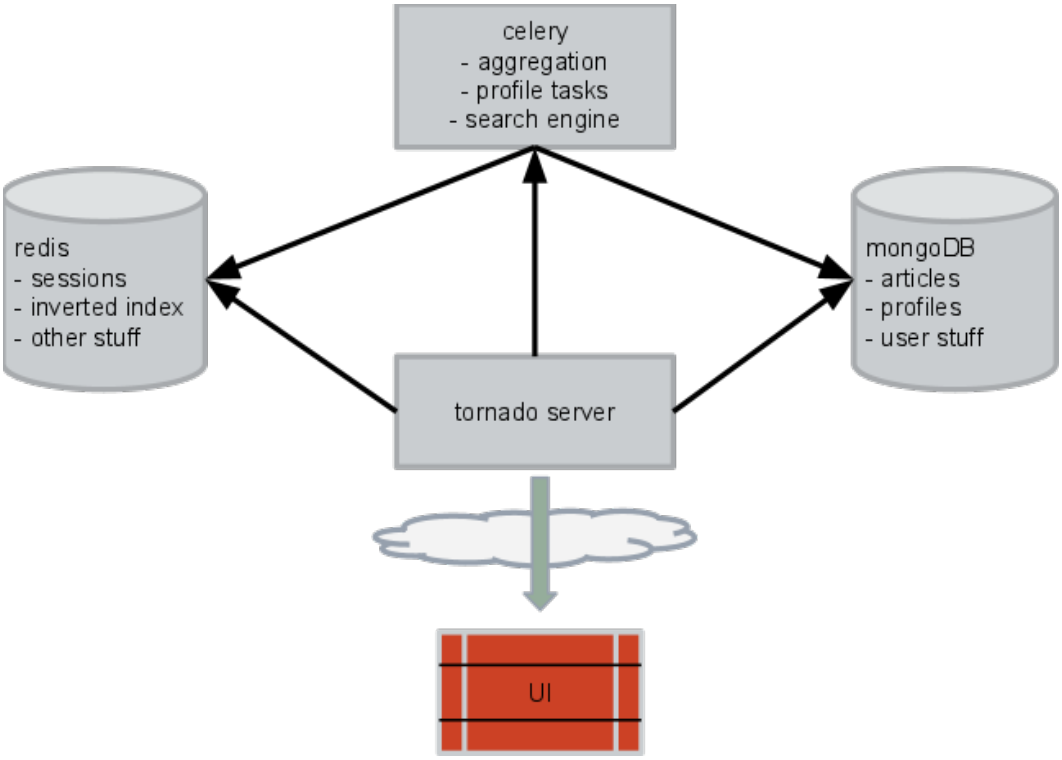


Για μεγαλύτερη ακρίβεια, θα χρησιμοποιήσουμε ένα προφίλ ανά κατηγορία. Επίσης οι ψήφοι, για λόγους απλότητας, θα λαμβάνονται υπόψιν μια φορά για να δοθούν ως ανάδραση (π.χ. μετά από μια μέρα), κάτι που διαισθητικά έρχεται σε συμφωνία με τον μικρό χρόνο ζωής τους και συμφωνεί με τα πειραματικά μας αποτελέσματα. Η παραπάνω διαδικασία μπορεί να εκτελείται ασύγχρονα και το celery είναι ιδανικό για κάτι τέτοιο. Η βάση δεδομένων θα αποθηκεύει ανά πάσα στιγμή αξιολογημένα άρθρα έτοιμα προς παρουσίαση από την διεπαφή χρήστη, με την οποία θα ασχοληθούμε στην συνέχεια. Δίνοντας περισσότερο βάρος στο “engineering” του aggregator και στην βελτιστοποίηση του server, η διεπαφή θα είναι απλή και εύχρηστη, στο πλαίσιο της λογικής kiss (keep it simple stupid). Μια βασική στήλη πλοήγησης με τις κατηγορίες και μια φόρμα μηχανής αναζήτησης, η κυρίως στήλη όπου θα παρουσιάζονται τα άρθρα και μία αρχική σελίδα με δυνατότητες registration / login είναι αρκετές για ένα πλήρως λειτουργικό σύστημα παρουσίασης ειδησεογραφίας.



Εικόνα 1. Κεντρική σελίδα

Ο server θα χειρίζεται κυρίως το σερβίρισμα άρθρων, την αυθεντικοποίηση του χρήστη και τα όποια δευτερεύοντα HTTP calls. Τα sessions των χρηστών θα αποθηκεύονται στην κρυφή μνήμη (redis), οπότε ένα σχήμα που περιέχει όλα τα συστατικά της εφαρμογής καθώς και τους ρόλους τους, φαίνεται στην εικόνα [2]. Η αρχιτεκτονική του συστήματος είναι αρκετά απλή και σε αυτό έχουν βοηθήσει τα εργαλεία υψηλού επιπέδου που χρησιμοποιούνται.



Εικόνα 2. Αρχιτεκτονική εφαρμογής

Επιπρόσθετα, το σύστημα μπορεί να κλιμακωθεί με αρκετή ευκολία, καθώς μπορούμε να κλιμακώσουμε οποιοδήποτε από τα 4 components ανεξάρτητα και εύκολα.

5.3 ΑΝΑΛΥΣΗ ΕΦΑΡΜΟΓΗΣ

Η συγκέντρωση (aggregation) των άρθρων γίνεται περιοδικά και ανα τακτά χρονικά διαστήματα, και μοντελοποιείται ως ένα celery task. Για κάθε γνωστή πηγή rss του συστήματος, εξετάζουμε τα urls των άρθρων που αυτό φέρει. Στη συνέχεια, εαν αυτά δεν βρίσκονται στην cache (redis), το οποίο σημαίνει ότι τα έχουμε επεξεργαστεί ήδη, τα προωθούμε σε μια δεύτερη υπηρεσία, η οποία και “καθαρίζει” το περιεχόμενο της σελίδας στην οποία δείχνει το url από περιττά στοιχεία όπως διαφημίσεις κτλ. Για τον σκοπό αυτό διαλέξαμε το diffbot (www.diffbot.com). Αφού λάβουμε το επεξεργασμένο κείμενο, το αρχειοθετούμε στην συλλογή άρθρων μας, μετά από την εξής διαδικασία:

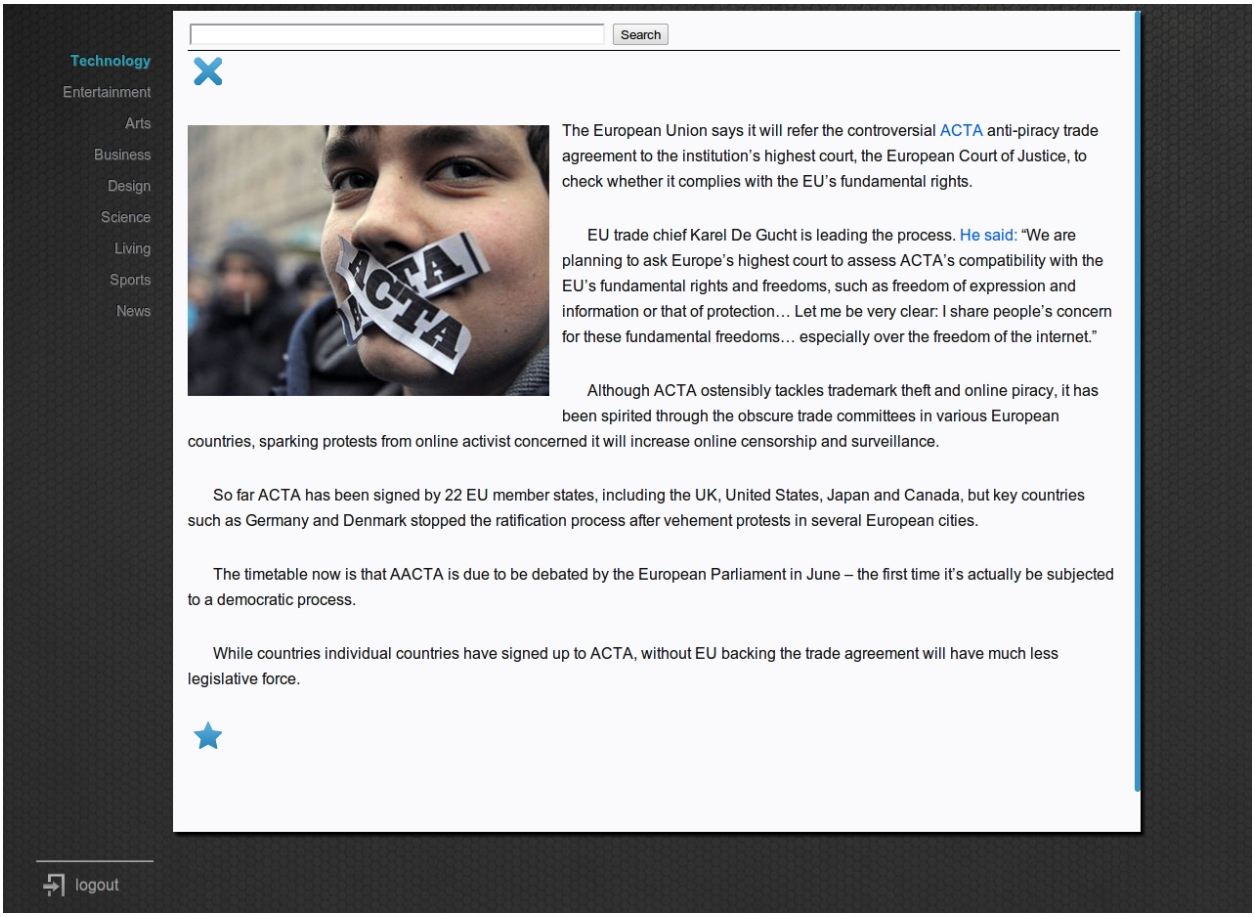
- Εξαγωγή των λέξεων (tokenization)
- Αφαίρεση κοινών λέξεων (stop words removal)
- Εξαγωγή προθεματικής ρίζας (stemming)
- Υπολογισμός tf (term frequencies calculation)
- Ανανέωση df (document frequencies)
- Υπολογισμός tf-idf
- Εξαγωγή χαρακτηριστικών προς προσαρμογή / αξιολόγηση

Η εξαγωγή χαρακτηριστικών γίνεται με βάση κάποιο κατώφλι (tfidf threshold). Κατόπιν τα στέλνουμε για αξιολόγηση στην υπηρεσία που παρέχει το προφίλ της Νοοτροπίας, παίρνουμε σαν απάντηση τον βαθμό (σκορ) αξιολόγησης και αποθηκεύουμε το άρθρο με όλα τα στοιχεία που χρειάζεται να φέρει. Το πλήρες σχήμα (ή καλύτερα collection) ενός άρθρου είναι:

```
class Article(Document):

    ''' Represents an article ready for presentation '''

    title = StringField(required=True, unique=True)
    content = StringField()
    adapt_features = StringField()
```

```
evaluate_features = StringField()
url = URLField()
date_added = DateTimeField(default=datetime.datetime.now)
category = StringField()
votes = IntField()
score = FloatField()
adapted = BooleanField()

meta = {'indexes': ['-date_added', 'category']}
```

Οι χρήστες μπορούν να ψηφίσουν θετικά ένα άρθρο, πατώντας το κουμπί που βρίσκεται στο τέλος του άρθρου (αστέρι), και αυτό μεταφέρεται αυτόματα στην βάση δεδομένων αυξάνοντας το πεδίο votes κατά ένα.



Εικόνα 3. Σελίδα ανάγνωσης άρθρου

Μετά το πέρας 24 ωρών, ξεκινάει η φάση προσαρμογής του προφίλ. Συγκεντρώνουμε όλα τα μη προσαρμοσμένα άρθρα με χρόνο ζωής μεγαλύτερο των 24 ωρών ανα κατηγορία, και τα προωθούμε στο API της Νοοτροπίας για προσαρμογή. Η συγκεκριμένη διαδικασία μοντελοποιείται ως ένα περιοδικό celery task.

Ο πλήρης πηγαίος κώδικας της εφαρμογής είναι διαθέσιμος στην διεύθυνση https://github.com/hymloth/NNYdissertation.

5.4 ΥΛΟΠΟΙΗΣΗ ΜΗΧΑΝΗΣ ΑΝΑΖΗΤΗΣΗΣ

Αναπόσπαστο κομμάτι πολλών ιστότοπων αποτελεί η δυνατότητα για εύρεση πληροφορίας σε πραγματικό χρόνο. Ειδικά στην ειδησεογραφία, η εύρεση άρθρων με βάση λέξεις κλειδιά είναι κάτι παραπάνω από αναγκαίο. Για τον σκοπό αυτό, στο πλαίσιο της εφαρμογής μας, αναπτύξαμε μια απλή μηχανή αναζήτησης, η οποία αποτελεί ξεχωριστό κομμάτι από τα υπόλοιπα (μπορεί να χρησιμοποιηθεί δηλαδή και ανεξάρτητα). Παραπέμπουμε τους ενδιαφερόμενους στο https://github.com/hymloth/pyredise, όπου είναι διαθέσιμος ο πηγαίος κώδικας.

Συνοπτικά, χρησιμοποιούμε το redis για να κρατάμε τον inverted index στην κύρια μνήμη του συστήματος, ενώ με την χρήση python επεξεργαζόμαστε τα queries και κάνουμε την καταχώρηση των εγγράφων (indexing). Ο inverted index είναι της μορφής:

term :
doc-id1 : tf1
doc-id2 : tf2
….

και μοντελοποιείται ώς ένα ταξινομημένο σύνολο (sorted set), με κλειδιά τα αναγνωριστικά των εγγράφων που περιέχουν τον εκάστοτε όρο και τιμές την συχνότητα του όρου στο έγγραφο.

Η μηχανή αναζήτησης παρέχει δυο βασικές μεθόδους επερωτήσεων:

- tfidf ranking
- tfidf ranking + proximity ranking

Στην πρώτη, χρησιμοποιούμε το πασίγνωστο tfidf ranking σχήμα για να ιεραρχήσουμε τα αποτελέσματα με περισσότερη ακρίβεια. Στην τελευταία περίπτωση, κάνουμε χρήση των θέσεων των λέξεων στο κείμενο για να κάνουμε ακριβέστερη την ιεράρχηση, με ένα υβριδικό σχήμα που χρησιμοποιεί το tfidf ranking και μια βαθμολογία που υπολογίζεται από την απόσταση των όρων της επερώτησης στα έγγραφα. Για τον σκοπό αυτό, αποθηκεύουμε τις θέσεις των όρων (posting list) σε μια δομή λεξικού (dictionary):

term :
doc-id1 : [posting list 1]
doc-id2 : [posting list 2]
….

Προφανώς, η διαδικασία εύρεσης κοστίζει σε χρόνο, και σε καμμία περίπτωση δεν πρέπει να εκτελεστεί από κώδικα του server (tornado), ο οποίος πρέπει να “μπλοκάρει” όσο το δυνατόν λιγότερο. Για αυτό το λόγο, όταν ο server δέχεται κάποιο request αναζήτησης, το προωθεί στην ουρά του celery. Εκεί, αφού εκτελεστεί η αναζήτηση, το αποτέλεσμα δημοσιεύεται σε ένα κανάλι publish / subscribe που μας παρέχει το redis και συλλέγεται ασύγχρονα για να προωθηθεί στον χρήστη.

5.5 ΣΥΜΠΕΡΑΣΜΑΤΑ

Η εφαρμογή που παρουσιάστηκε σε αυτό το κεφάλαιο συνιστά αφ' ενός μια προσπάθεια που βασίζεται στα αποτελέσματα των πειραμάτων που προηγήθηκαν και αφ' εταίρου μια επίδειξη ανάπτυξης μιας πιλοτικής υλοποίησης με χρήση τεχνολογίας αιχμής. Γνωρίζουμε βέβαια ότι η συγκεκριμένη υλοποίηση δεν είναι σε καμμία περίπτωση ένα πλήρες προϊόν, αλλά αποτελεί το πρωτόλειο μιας προσπάθειας που βάζει τις βάσεις για την δημιουργία μιας ολοκληρωμένης πλατφόρμας παρουσίασης ειδησεογραφίας.

6. ΣΥΝΟΨΗ ΚΑΙ ΜΕΛΛΟΝΤΙΚΑ ΣΧΕΔΙΑ

Στο πλαίσιο της εργασίας μας ασχοληθήκαμε με την πρόβλεψη δημοφιλίας με βάση το περιεχόμενο στον τομέα της ειδησεογραφίας. Παρουσιάσαμε συνοπτικά διάφορες μεθοδολογίες που έχουν εφαρμοστεί στο πεδίο και επισημάναμε τις αδυναμίες και τα προβλήματά τους. Μελετήσαμε την περίπτωση του ιστότοπου κοινωνικής ειδησεογραφίας, Digg.com, παραθέτοντας προηγούμενες απόπειρες πρόβλεψης της δημοτικότητας των ειδήσεων που παρουσιάζονται σε αυτό. Αναφέραμε τα συγκριτικά πλεονεκτήματα της περίπτωσης όπου η δημοφιλία ενός άρθρου θα υπαγορευόταν καθαρά από το περιεχόμενό του. Αναπτύξαμε μια πρωτοποριακή μεθοδολογία για την επαλήθευση της παραπάνω πρότασης, εκτελώντας μια σειρά πειραμάτων πάνω σε δεδομένα του Digg με την χρήση της Νοοτροπίας. Το κύριο συμπέρασμά μας είναι ότι η πρόβλεψη δημοφιλίας με βάση το περιεχόμενο είναι εφικτή και θεραπεύει πολλά από τα προβλήματα που παρουσιάζονται με την χρήση παραδοσιακών αλγορίθμων. Αυτό δημιουργεί μια εντελώς διαφορετική αντιμετώπιση και προοπτική για την παρουσίαση αρθογραφίας.

Η ανάγκη για δίκαιη και άμεση παρουσίαση ειδησεογραφίας σε μια κοινότητα χρηστών με κοινά ενδιαφέροντα αποτελεί αφορμή για ανάπτυξη εφαρμογών που θα πληρούν αυτά τα χαρακτηριστικά. Αν και τα αποτελέσματά μας είναι κάτι παραπάνω απο θετικά, η δυνατότητα περαιτέρω βελτίωσης της συνολικής ακρίβειας αφήνει αρκετά περιθώρια για έρευνα στο πλαίσιο των recommender systems για την πρόβλεψη και ανάδειξη δημοφιλούς περιεχομένου. Όπως αναφέραμε στο κεφάλαιο 4, ένα υβριδικό μοντέλο που συνδυάζει τους αποδοτικότερους collaborative filtering και content-based αλγορίθμους θα αποτελούσε την ιδανική λύση. Η υλοποίηση ενός τέτοιου συστήματος αποτελεί πρόκληση σε επίπεδο σχεδίασης, αρχιτεκτονικής και κλιμάκωσης.

Επιπλέον, η δυνατότητα προβολής ακριβούς εξατομικευμένης (personalized) πληροφορίας θα συντελούσε σε μια πιο ολοκληρωμένη λύση παρουσίασης ειδησεογραφίας. Ο χρήστης θα είχε την δυνατότητα ενημέρωσης τόσο με βάση τα καθαρά προσωπικά του

ενδιαφέροντα όσο και με αυτά που καθορίζονται από το “ρεύμα των πολλών”. Τα δοχεία αυτά είναι συγκοινωνούντα. Η δημοφιλία έχει να κάνει με πληροφοριακά χαρακτηριστικά που προκύπτουν από την ζύμωση του ατομικού με το συλλογικό. Πιστεύουμε ότι μια ολιστική προσέγγιση στο ζήτημα με δίκαιη κατανομή των “βαρών” σε αλγοριθμικό επίπεδο θα συντελούσε στην δημιουργία ενός τέλειου συστήματος προβολής και λήψης ειδησεογραφίας. Εργαζόμαστε σκληρά προς αυτή την κατεύθυνση.

7. ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] K. Lerman. Social information processing in social news aggregation. IEEE Internet Computing: special issue on Social Search, 11(6):16–28, 2007.

[2] K. Lerman. Social networks and social information filtering on digg. In Proceedings of 1st International Conference on Weblogs and Social Media (ICWSM-07), 2007.

[3] K. Lerman. User participation in social media: Digg study. In Proceedings of the 2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Workshops, WI-IATW '07, pages 255–258, Washington, DC, USA, 2007. IEEE Computer Society.

[4] S. Jamali and H. Rangwala. Digging digg: Comment mining, popularity prediction, and social network analysis. Web Information Systems and Mining, International Conference on, 0:32–38, 2009.

[5] G. Szabo and B. A. Huberman. Predicting the popularity of online content. Commun. ACM, 53:80–88, August 2010.

[6] D. W. Oard. The state of the art in text filtering. User Modeling and User-Adapted Interaction: An Internation Journal, 7(3):141–178, 1997.

[7] N. Nanas and A. De Roeck. Autopoiesis, the immune system and adaptive information filtering. Natural Computing, 8(2):387–427, 2009.

[8] N. Nanas, V. Uren, and A. De Roeck. Nootropia: a user profiling model based on a self-organising term network. In G. Nicosia, V. Cutello, P. J. Bentley, and J. Timmis, editors, Artificial Immune Systems, Third International Conference (ICARIS 2004), volume LNCS 3239 of Lecture Notes in Computer Science, pages 146–160. Springer, Heidelberg, Germany, 2004.

[9] N. Nanas, M. Vavalis, and A. De Roeck. A network-based model for high-dimensional information filtering. In Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval, SIGIR ’10, pages 202–209, New York, NY, USA, 2010. ACM.

[10] N. Nanas, M. Vavalis, and A. De Roeck. Words, antibodies and their interactions. Swarm Intelligence, 4(4):275–300, 2010.

[11] N. Nanas, M. Vavalis, and E. Houstis. Personalised news and scientific literature aggregation. Information Processing and Management, 46:268–283, 2010.

[12] N. Nanas, M. Vavalis, and L. Kellis. Immune learning in a dynamic information environment. In Artificial Immune Systems, 8th International Conference (ICARIS 2009), volume LNCS 5666 of Lecture Notes in Computer Science, pages 192–205. Springer, Heidelberg, Germany, 2009.

[13] N. L. Bhamidipati and S. K. Pal. Comparing scores intended for ranking. IEEE Trans. on Knowledge and Data Engineering, 21:21–34, January 2009

[14] Y. C. Park and K.-S. Choi. Automatic thesaurus construction using bayesian networks. Information Processing and Management., 32(5):543–553, 1996.