



ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

**ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ**

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Έρμεσοι Μηχανισμοί Φήμης για Κοινωνικά Δίκτυα

Ιωάννης - Άγγελος - Κανελλόπουλος

Επιβλέποντες: **Αφροδίτη Τσαλγατίδου**, Αναπληρώτρια Καθηγήτρια
Ελένη Κουτρούλη, Υποψήφια Διδάκτωρ

ΑΘΗΝΑ

ΙΟΥΛΙΟΣ 2015

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Έμμεσοι Μηχανισμοί Φήμης για Κοινωνικά Δίκτυα

Ιωάννης Α. Κανελλόπουλος

A.M.: 1115201100176

ΕΠΙΒΛΕΠΟΝΤΕΣ: **Αφροδίτη Τσαλαγατίδου**, Αναπληρώτρια Καθηγήτρια
Ελένη Κουτρούλη, Υποψήφια Διδάκτωρ

ΠΕΡΙΛΗΨΗ

Τα κοινωνικά δίκτυα έχουν αλλάξει τον τρόπο που επικοινωνούμε, που μοιράζομαστε ιδέες και πράγματα μεταξύ μας και γενικότερα τον τρόπο με τον οποίο παράγονται τα δεδομένα στο διαδίκτυο αφού πλέον οι χρήστες εισάγουν τα δεδομένα και όχι οι εταιρείες. Στις διάφορες εφαρμογές κοινωνικής δικτύωσης οι χρήστες χρειάζεται να παίρνουν αποφάσεις σχετικά με ποιους χρήστες ή / και πιο περιεχόμενο θα εμπιστευτούν για να διεξάγουν συναλλαγές στα πλαίσια της κάθε εφαρμογής. Κατά συνέπεια, οι εφαρμογές αυτές προκειμένου να είναι αποτελεσματικές χρειάζονται μηχανισμούς που θα βοηθούν τους χρήστες να λαμβάνουν ποιοτικές αποφάσεις εμπιστοσύνης βασισμένες σε διάφορα κριτήρια. Τέτοιοι μηχανισμοί περιλαμβάνονται στα συστήματα φήμης τα οποία υπολογίζουν τιμές φήμης για χρήστες ή περιεχόμενο, με βάση συγκεκριμένες βαθμολογίες (άμεσοι μηχανισμοί φήμης) και κυρίως άλλα στοιχεία όπως οι κοινωνικές σχέσεις, η ομοιότητα μεταξύ χρηστών, οι δραστηριότητα των χρηστών, κ.α.. (χρησιμοποιούμε τον όρο "έμμεσοι μηχανισμοί φήμης" για τους μηχανισμούς που δεν βασίζονται στην άμεση βαθμολόγηση). Η παρούσα εργασία έχει δύο στόχους: 1) την παρουσίαση διαφόρων κατηγοριών συστημάτων φήμης που χρησιμοποιούν έμμεσους μηχανισμούς φήμης, και 2) τη δημιουργία ενός έμμεσου συστήματος φήμης. Συγκεκριμένα, παραθέτουμε τη δική μας υλοποίηση ενός συστήματος φήμης για το Twitter, ένα από τα πιο δημοφιλή κοινωνικά δίκτυα. Στο σύστημα αυτό, το οποίο λέγεται Υπολογιστής Επιροής υπολογίζουμε την επιρροή χρηστών ή του περιεχομένου θεματικών ενοτήτων στο Twitter. Τέλος για την αξιολόγηση του συστήματός μας, το συγκρίνουμε με το με το καθιερωμένο εργαλείο υπολογισμού επιρροής σε κοινωνικά δίκτυα, Klout.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Μηχανισμοί Φήμης

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: Συνεργατικό φιλτράρισμα, Συστήματα συστάσεων, Κοινωνικά δίκτυα, Διαδικτυακή κοινωνική επιρροή, Ανάλυση μέσων ενημέρωσης

ABSTRACT

Social networks have changed the way we communicate, share our ideas and in general the way data are being generated as users input the data instead of companies. In Social Network-based applications, users need to take decisions regarding who other users or / and content they will trust in order to make transactions in the context of each application. Consequently, these applications, in order to be effective, need mechanisms which will help users to take qualitative decisions based on various criteria. Such mechanisms are contained in reputation systems for social networks, which estimate reputation values for users or / and content based on ratings (direct mechanisms), and also on other types of information, such as social relationships, social actions, similarity between users, etc. (we use the term “indirect reputation mechanisms” for the reputation mechanism which are not based on direct ratings). This thesis has two goals: 1) the presentation of various categories of indirect reputation systems, and 2) the design and implementation of an indirect reputation system for Twitter. In our system, which we call Twitter Influence Computer (TIC), we estimate the influence power of a user or of the context of a topic. For the evaluation of TIC, we compare it with Klout, a popular social media analytics application which computes influence scores of social media users.

SUBJECT AREA: Reputation Mechanisms

KEYWORDS: Collaborative filtering, Recommendation systems, Social networks, Online social influence, Media analytics

ΠΕΡΙΕΧΟΜΕΝΑ

1. ΕΙΣΑΓΩΓΗ.....	11
2. ΣΥΣΤΗΜΑΤΑ ΦΗΜΗΣ ΓΙΑ Η-ΚΟΙΝΟΤΗΤΕΣ.....	11
2.1 Η Έννοια της Εμπιστοσύνης στα Συστήματα Φήμης.....	12
2.2 Κατηγορίες των Συστημάτων Φήμης	13
2.2.1 Συστήματα Συστάσεων	13
2.2.2 Συστήματα Φήμης για P2P εφαρμογές.....	14
3. ΤΕΧΝΙΚΕΣ COLLABORATIVE FILTERING.....	15
3.1. Semantic Social Collaborative Filtering	15
3.1.1 Simple Semantic Social CF (Απλό Σημασιολογικό Κοινωνικό ΣΦ).....	16
3.1.2 Secured Semantic Social CF (Ασφαλές Σημασιολογικό Κοινωνικό ΣΦ)	17
3.2. Memory-Based Collaborative Filtering Techniques.....	19
3.2.1 Υπολογισμός της Ομοιότητας.....	19
3.2.2 Πρόβλεψη και Υπολογισμός Σύστασης	20
3.2.3 Παρουσίαση των Κορυφαίων Συστάσεων	21
3.3. Συνεργατικό Φιλτράρισμα με βάση το Μοντέλο	22
3.3.1 Απλός Bayesian Αλγόριθμος	22
3.3.2 Αλγόριθμοι CF με Clustering (Συστάδες).....	22
3.3.3 Regression Algorithms	23
3.4. Υβριδικές τεχνικές για Collaborative Filtering	24
3.4.1 Υβριδικά μοντέλα με τεχνικές CF και CBS.....	24
4. ΕΥΡΕΣΗ ΚΟΜΒΩΝ ΜΕΓΑΛΗΣ ΕΠΙΡΡΟΗΣ ΣΕ ΚΟΙΝΩΝΙΚΑ ΔΙΚΤΥΑ.....	25
4.1. Μοντέλο Υπολογισμού Επιρροής TGT	26
4.1.1 Απλό Μοντέλο Διάχυσης.....	26
4.1.2 Μοντέλο Διάχυσης με Θετικό και Αρνητικό Αντίκτυπο στον Χρήστη	28
4.1.3 Υπολογισμός Πιθανότητας Επιρροής.....	30

4.1.4	Εύρεση Κόμβων με τη Μεγαλύτερη Επιρροή	31
4.2.	Μοντέλο Υπολογισμού της Επιρροής Σχεδιασμένο για Blogs.....	32
4.2.1	Μοντέλο Εύρεσης Κόμβων Επιρροής σε blogs	32
4.2.2	Bloggers με την Μεγαλύτερη Επιρροή	33
4.2.3	Ιδιότητες Εντοπισμού των Κόμβων Επιρροής	34
4.2.4	Υπολογισμός των Κόμβων Επιρροής.....	35
5.	ΠΕΡΙΓΡΑΦΗ ΥΛΟΠΟΙΗΣΗΣ	36
5.1	Υπολογισμός Επιρροής ενός Μηνύματος στο Twitter (Influence of a Tweet)	36
5.2	Υλοποίηση: Υπολογιστής Επιρροής για το Twitter (Twitter Influence Computer - TIC)	37
5.2.1	Υπολογισμός Επιρροής ενός Μηνύματος στο Twitter (Influence of a Tweet).....	38
5.2.2	Υπολογισμός Επιρροής ενός Χρήστη στο Twitter (Influence of a User).....	39
5.2.3	Τεχνολογίες και Εργαλεία που χρησιμοποιήθηκαν για το TIC	40
5.2.4	Πρόσβαση στα Δεδομένα του Twitter	42
6.	ΕΛΕΓΧΟΣ.....	42
6.1	Σύγκριση με το Klout.....	42
6.2	Επιδράσεις και Διαφορετικές Χρήσεις των Βαρών (Weights).....	44
7.	ΣΥΜΠΕΡΑΣΜΑΤΑ	44
	ΠΙΝΑΚΑΣ ΟΡΟΛΟΓΙΑΣ	46
	ΣΥΝΤΜΗΣΕΙΣ – ΑΡΚΤΙΚΟΛΕΞΑ – ΑΚΡΩΝΥΜΙΑ	47
	ΑΝΑΦΟΡΕΣ	48

ΚΑΤΑΛΟΓΟΣ ΣΧΗΜΑΤΩΝ

Σχήμα 1: Απεικόνιση του σημασιολογικού συνεργατικού φιλτραρίσματος.....	16
Σχήμα 2: Ασφαλές σημασιολογικό συνεργατικό φιλτράρισμα με Λίστες Ελέγχου Πρόσβασης	18
Σχήμα 3: Υπολογισμός Επιρροής του TIC για το Δείγμα Χρηστών	43
Σχήμα 4: Υπολογισμός Επιρροής του Klout για το Δείγμα Χρηστών	43

ΚΑΤΑΛΟΓΟΣ ΕΙΚΟΝΩΝ

Εικόνα 1: Βαθμολογικοί δείκτες ενός πωλητή στο Amazon	σελ. 14
Εικόνα 2: Ψευδοκώδικας Υπολογισμού της Επιρροής ενός tweet.....	σελ. 39
Εικόνα 3: Διεπαφή TIC Υπολογισμού της Συνολικής Επιρροής ενός Χρήστη	σελ. 40

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

Πίνακας 1: Απεικόνιση Μοντέλου Αναζήτησης Υλικού.....	σελ. 16
Πίνακας 2: Καθορισμός Μοντέλου	σελ. 17
Πίνακας 3: Αλγόριθμος του Ασφαλούς Μοντέλου SSCF	σελ. 18
Πίνακας 4: Καθορισμός του Ασφαλούς Μοντέλου SSCF.....	σελ. 18
Πίνακας 5: Πίνακας χρηστών με τις βαθμολογίες τους	σελ. 24
Πίνακας 6: συμπληρωμένος με τις βαθμολογίες που προέκυψαν απο τον αλγόριθμο CBS	σελ. 25
Πίνακας 7: Πίνακας χρηστών αναθεωρημένος από τα αποτελέσματα του αλγορίθμου CF	σελ. 25

1. ΕΙΣΑΓΩΓΗ

Τα συστήματα φήμης είναι πολύ σημαντικά εργαλεία τα οποία χρησιμοποιούνται σε διαδικτυακές κοινότητες. Σκοπός τους είναι να βοηθήσουν τους χρήστες να εντοπίσουν πιο γρήγορα ένα αντικείμενο που αναζητούν, μια υπηρεσία ή έναν χρήστη. Παραδείγματα ιστοσελίδων κοινωνικής δικτύωσης που χρησιμοποιούν έντονα τα οφέλη των συστημάτων φήμης είναι ενδεικτικά το YouTube [41], το Facebook [42] και το Twitter [43]. Σε αυτές τις διαδικτυακές εφαρμογές κύριο ζητούμενο είναι να αποσαφηνιστεί ποιός χρήστης έχει μεγάλη επιρροή ή αξίζει προσοχής. Τα έμμεσα συστήματα φήμης για κοινωνικά δίκτυα αποτελούνται από ποικίλες μεθόδους και κατηγορίες από τις οποίες αναλύουμε τις πιο σημαντικές και δημοφιλείς στις ακόλουθες ενότητες. Έπειτα παρουσιάζουμε τη δική μας υλοποίηση ενός συστήματος φήμης για το Twitter, αξιοποιώντας υπάρχουσες έρευνες και μεθοδολογίες τις οποίες βελτιώσαμε και αναβαθμίσαμε για πιο ακριβή αποτελέσματα. Επιπρόσθετα ακολουθεί ο έλεγχος της υλοποίησης μας, την οποία και συγκρίνουμε με το καθιερωμένο εργαλείο υπολογισμού επιρροής σε κοινωνικά δίκτυα, Klout [45] με την παρουσίαση των αντίστοιχων αποτελεσμάτων, και τέλος παρουσιάζουμε τα συμπεράσματα της έρευνας μας.

2. ΣΥΣΤΗΜΑΤΑ ΦΗΜΗΣ ΓΙΑ Η-ΚΟΙΝΟΤΗΤΕΣ

Η έννοια της εμπιστοσύνης και τα συστήματα φήμης αποτελούν μια από τις πιο αποδοτικές λύσεις όταν πρόκειται για τη λήψη αποφάσεων στον κόσμο του διαδικτύου. Η βασική ιδέα είναι ότι ομάδες ή χρήστες, μετά την ολοκλήρωση μιας συναλλαγής έχουν τη δυνατότητα να βαθμολογήσουν εκατέρωθεν την ομάδα αυτή ή τον χρήστη σχετικά με το πόσο αξιόπιστη είναι, δηλαδή πόσο μπορεί να την εμπιστευτεί κάποιος. Με αυτόν τον τρόπο στην επόμενη συναλλαγή οι χρήστες θα μπορούν να εκτιμήσουν τις προηγούμενες κριτικές και να αποφασίσουν πιο αποτελεσματικά αν θα εμπιστευτούν ή όχι τον άλλο χρήστη ή ομάδα. Επομένως η έννοια της εμπιστοσύνης, ενσωματωμένη στα συστήματα φήμης, αποτελεί καθοριστική λειτουργία στις διαδικτυακές συναλλαγές καθώς δίνει τη «δύναμη» στον καταναλωτή να μάθει αρκετές πληροφορίες για τον χρήστη που θα συναλλαχτεί καθώς και τις υπηρεσίες που θα του προσφέρει όπως και στη καθημερινότητα. Στο πραγματικό μη-ηλεκτρονικό κόσμο ο καταναλωτής μπορεί να επισκεφθεί τον πωλητή και να δοκιμάσει τα προϊόντα του ή τις υπηρεσίες του πριν προβεί σε κάποια αγορά. Στις ηλεκτρονικές κοινότητες αυτό όμως δεν είναι εφικτό εκτός αν υλοποιηθούν κατάλληλα συστήματα φήμης, τα οποία ενσωματώνουν έννοιες όπως η εμπιστοσύνη, τα οποία παράγουν ακριβείς συστάσεις (recommendations) για τον εκάστοτε χρήστη.

Η εμπιστοσύνη ενός χρήστη για κάποιον άλλον ενδέχεται να επηρεαστεί από ποικίλους παράγοντες όπως οι εμπειρίες που έχει αποκομίσει από τον στενό του κύκλο, ψυχολογικοί παράγοντες που έχουν διαμορφωθεί στη διάρκεια της ζωής του χρήστη που τις περισσότερες φορές δεν συσχετίζονται άμεσα με τον χρήστη που αποφασίζουμε αν θα εμπιστευτούμε, καθώς επίσης κι από τη φήμη και επιρροή (influence) από τρίτους. Οι ενδεικτικοί παράγοντες που αναφέρθηκαν δεν είναι εύκολο να υλοποιηθούν σε ένα σύστημα. Είναι λοιπόν αναγκαίο να απλοποιηθεί η έννοια της εμπιστοσύνης ώστε να είναι δυνατή η ενσωμάτωση της σε ένα σύστημα φήμης προκειμένου να βαθμολογούνται μεταξύ τους οι χρήστες, και να παράγονται “μέτρα αξιοπιστίας” για τους χρήστες που διευκολύνουν αποφάσεις εμπιστοσύνης.

2.1 Η Έννοια της Εμπιστοσύνης στα Συστήματα Φήμης

Αρκετοί είναι οι ερευνητές που έχουν ορίσει την έννοια της εμπιστοσύνης: Ο Marsh, το 1994 [1] στη διδακτορική του διατριβή στο Πανεπιστήμιο του Sterling, υπολόγιζε στο θεωρητικό του μοντέλο την εμπιστοσύνη με την βοήθεια πρακτόρων που αλληλεπιδρούσαν έχοντας τη δυνατότητα διατήρησης του ιστορικού ενεργειών καθώς και άλλων διάφορων πληροφοριών. Η προσέγγιση αυτή, αν και ιδιαίτερα δημοφιλής, δεν ταιριάζει στο χώρο των κοινωνικών δικτύων, τα οποία αξιοποιούν σύντομες εκφράσεις από το προφίλ του χρήστη, καθώς απαιτεί από τον χρήστη να εισαγάγει μεγάλο όγκο δεδομένων. Το ίδιο πρόβλημα συμβαίνει και με την έρευνα που πραγματοποιήθηκε από τους Castelfranchi & Falcone (1998, 2002) [2, 3]. Το μοντέλο που ανέπτυξαν βασιζόταν κυρίως σε διάφορους ψυχολογικούς παράγοντες αλλά και στο ιστορικό συναλλαγών μεταξύ των χρηστών με αποτέλεσμα και αυτή η μέθοδος να απαιτεί πολλές πληροφορίες.

Ο ορισμός που δίνεται από τους McKnight & Chervany (1996) [4] είναι πιο ενδιαφέρον, και με τη θεωρητική του προσέγγιση πλησιάζει αρκετά αυτό που απαιτούν οι η – κοινότητες και ειδικά τα κοινωνικά δίκτυα. Αναφέρεται στην εμπιστοσύνη ως τη διάθεση και θέληση μιας ομάδας να εξαρτηθεί από κάτι ή κάποιον σε μια κατάσταση με ένα αίσθημα σχετικής ασφάλειας παρόλο που υπάρχει μια πιθανότητα να προκύψουν αρνητικά αποτελέσματα.

Έτσι λοιπόν η έρευνα που πραγματοποιήθηκε από τον Deutsch, το 1962 [5], τους McKnight & Chervany και ιδιαίτερα του Sztompka (1999) [6], που θα αναλύσουμε¹ παρακάτω, δίνουν έναν πιο απλό, χρήσιμο και αποτελεσματικό ορισμό για την εμπιστοσύνη χωρίς να απαιτούνται πολλές πληροφορίες προκειμένου να μπορεί να χρησιμοποιηθεί για τον υπολογισμό μεταξύ χρηστών. Στη θεωρία του Deutsch η Alice θα εμπιστευτεί τον Bob εάν γνωρίσει ότι ο Bob θα εκτελέσει όλες τις απαραίτητες ενέργειες ώστε το αποτέλεσμα να είναι θετικό. Με παρόμοιο τρόπο ο Sztompka αναφέρεται στην εμπιστοσύνη σαν ένα «στοίχημα» για τις μελλοντικές πράξεις των υπολοίπων ατόμων που αποτελείται από δυο συστατικά. Το ένα είναι η πεποίθηση (belief) και το δεύτερο είναι η δέσμευση (commitment). Η πεποίθηση ότι το άλλο άτομο θα πράξει με έναν συγκεκριμένο τρόπο δεν αρκεί. Εμπιστοσύνη είναι η δέσμευση σε μια ενέργεια που βασίζεται στη πεποίθηση ότι οι μελλοντικές ενέργειες του ατόμου θα οδηγήσουν σε καλά αποτελέσματα. Αυτός ο ορισμός θα υιοθετηθεί και θα χρησιμοποιηθεί στη δημιουργία του συστήματος φήμης που θα αναπτύξουμε.

Πριν αναπτύξουμε ένα σύστημα φήμης οφείλουμε να ανατρέξουμε στον ορισμό της. Σύμφωνα με τον ορισμό του λεξικού της Οξφόρδης¹ (Dictionary of Oxford) ή φήμη ορίζεται ως:

«Φήμη είναι ότι πιστεύεται ή λέγεται γενικά για ένα άτομο ή ομάδα»

Δηλαδή η φήμη είναι μια συνισταμένη απόψεων η οποία προκύπτει από το κοινωνικό δίκτυο ή κοινότητα και είναι ορατή και προσβάσιμη από όλα τα μέλη του δικτύου. Οι έννοιες

¹<http://www.oxforddictionaries.com/definition/english/trust>

της φήμης και της εμπιστοσύνης δεν είναι ταυτόσημες και αυτό φαίνεται με το ακόλουθο χαρακτηριστικό παράδειγμα:

«Σε εμπιστεύομαι λόγω της καλής σου φήμης»

«Σε εμπιστεύομαι παρά τη κακή σου φήμη»

Αυτό το παράδειγμα δείχνει με φανερό τρόπο ότι η εμπιστοσύνη έχει προσωπικό και υποκειμενικό χαρακτήρα και βασίζεται σε ποικίλους παράγοντες. Γενικά η εμπιστοσύνη που προκύπτει από τη προσωπική εμπειρία έχει μεγαλύτερο βάρος από τις γνώμες τρίτων προσώπων αλλά στη περίπτωση που δεν υφίσταται προσωπική εμπειρία, η αξιοπιστία για έναν χρήστη ή ομάδα υπολογίζεται στα λεγόμενα και κριτικές των υπολοίπων μελών της κοινότητας. Μάλιστα στην έρευνα του Tadelis (2001) [7] διαπιστώνεται πως όταν ένα άτομο ενταχθεί σε μια ομάδα κληρονομεί τη φήμη που έχει σχηματιστεί ήδη για αυτή. Αυτό σημαίνει ότι εάν η ομάδα έχει καλή φήμη τότε και τα μέλη την κληρονομούν και το αντίστροφο.

2.2 Κατηγορίες των Συστημάτων Φήμης

Τα συστήματα φήμης γνωρίζουν μεγάλη άνθιση τα τελευταία είκοσι χρόνια, ενσωματώνονται και γίνονται ολοένα και πιο αποδεκτά από τους χρήστες οι οποίοι θέλουν να συναλλάσσονται διαδικτυακά, σε η-κοινότητες, όπως και στη πραγματικότητα. Στον πραγματικό κόσμο κάθε κοινότητα έχει ένα σύστημα φήμης το οποίο συντίθεται από το ιστορικό των ενεργειών των ατόμων που ανήκουν σε αυτήν σχηματίζοντας απόψεις με αποτέλεσμα κάθε απόφαση να λαμβάνεται βάσει αυτών. Το ίδιο λοιπόν απαιτείται στις διαδικτυακές κοινότητες που αποτελούνται από χιλιάδες ξένους μεταξύ τους χρήστες. Έτσι με τα συστήματα φήμης γίνονται ποιοτικότερες συναλλαγές, καλλιεργείται η εμπιστοσύνη και διατηρείται η πίστη (loyalty) των χρηστών στο σύστημα. Υπάρχουν διάφορες κατηγορίες συστημάτων φήμης η-κοινοτήτων, ανάλογα με το είδος της παραγόμενης φήμης, τις χρησιμοποιούμενες πληροφορίες, την επεξεργασία αυτών των πληροφοριών, κ.λπ. Παρακάτω παρουσιάζουμε δύο σημαντικές κατηγορίες, τα συστήματα συστάσεων και τα συστήματα φήμης για συστήματα ομότιμου-προς-ομότιμο (P2P).

2.2.1 Συστήματα Συστάσεων

Τα συστήματα αυτά χρησιμοποιούν διάφορα στοιχεία του προφίλ των χρηστών, π.χ. την ηλικία, το φύλλο, τα ενδιαφέροντα ή ακόμα και τις προηγούμενες αγορές του χρήστη, και στη συνέχεια εφαρμόζουν διάφορες μεθόδους συστάσεων για την παραγωγή συστάσεων προς τους χρήστες. Οι συστάσεις αυτές μπορεί να βασίζονται στο περιεχόμενο (content based recommendation systems) ή σε μεθόδους συνεργατικού φιλτραρίσματος (collaborative filtering) οι οποίες συστήνουν υπηρεσίες και προϊόντα μέσω της εύρεσης παρόμοιων χρηστών. Ενδεικτικά παραδείγματα ηλεκτρονικού εμπορίου (e-commerce) που χρησιμοποιούν συστήματα φήμης είναι η ιστοσελίδα Amazon (amazon.com), Ebay (<http://www.ebay.com>) οι οποίες εξατομικεύουν τα προϊόντα τους για κάθε πελάτη παράγοντας συστάσεις π.χ. βιβλία μαθηματικών σε δασκάλους μαθηματικών και παιχνίδια σε παιδιά αλλά και πιο εξειδικευμένες e-commerce εφαρμογές όπως το Yelp (<http://www.yelp.com>), μια κοινότητα νέων ανθρώπων, όπου χρησιμοποιούνται τα συστήματα φήμης για να ερμηνευτούν υποκειμενικές απόψεις χρηστών για τη σύσταση καλών εστιατορίων, ξενοδοχείων μέχρι καταστημάτων για αγορές ρουχισμού.



Εικόνα 1: Βαθμολογικοί δείκτες φήμης ενός πωλητή στο Amazon

2.2.2 Συστήματα Φήμης για P2P εφαρμογές

Μια άλλη κατηγορία η-κοινοτήτων που χρησιμοποιεί τα συστήματα φήμης είναι τα συστήματα ομότιμων χρηστών (peer-2-peer ή P2P συστήματα). Αυτά τα συστήματα φήμης μπορούν να κατηγοριοποιηθούν σε άλλους δυο τύπους, τα κεντροκοποιημένα και τα κατανεμημένα. Στα κεντροκοποιημένα συστήματα ο βασικός μηχανισμός είναι ότι μετά το πέρας μια συναλλαγής δυο χρηστών, οι βαθμολογίες τους καταχωρούνται σε μια κεντρική αρχή (centralized authority) ενώ στα κατανεμημένα συστήματα κάθε ομάδα η χρήστης συλλέγει τις βαθμολογίες από άλλα μέλη της κοινότητας και όχι από κάποιο κεντρικό σύστημα. Στα P2P συστήματα κάθε κόμβος είναι πελάτης (client) και εξυπηρετητής (server), για αυτό και μερικές φορές αποκαλείται με τον όρο “servant”. Στα P2P συστήματα ένας κόμβος που ενδιαφέρεται να κάνει μια συναλλαγή, π.χ. να κατεβάσει ένα αρχείο, καταρχήν αναζητάει άλλους κόμβους που μπορούν να τον εξυπηρετήσουν, που διαθέτουν δηλαδή τους πόρους που χρειάζεται (φάση αναζήτησης). Στη συνέχεια, αφού έχει βρει υποψήφιους εξυπηρετητές, επιλέγει έναν από αυτούς και προχωράει σε συναλλαγή μαζί του (φάση συναλλαγής). Σε μερικά P2P δίκτυα η φάση της αναζήτησης βασίζεται σε κεντροκοποιημένες λειτουργίες όπως για παράδειγμα στο Napster (<http://www.napster.com>), που χρησιμοποιεί τον διακομιστή πόρων ενώ σε άλλα P2P συστήματα η αναζήτηση βασίζεται σε κατανεμημένες λειτουργίες, όπως στο σύστημα BitTorrent (<http://www.bittorrent.com>), το CSpace (<http://www.cspace.in>). Υπάρχουν και ενδιάμεσες αρχιτεκτονικές όπως το FastTrack όπου υπάρχουν κόμβοι και υπερκόμβοι με τους τελευταίους να κρατάνε πληροφορίες για τους άλλους κόμβους και συνδέονται στο δίκτυο σαν διακομιστές στη φάση της αναζήτησης. Τόσο στη φάση της αναζήτησης όσο και στη φάση της επιλογής χρησιμοποιούνται πληροφορίες σχετικές με την ποιότητα των συναλλαγών των χρηστών ώστε να υπολογιστεί η φήμη των υποψήφιων εξυπηρετητών και να ληφθεί η απόφαση εμπιστοσύνης που αφορά στο χρήστη ο οποίος τελικά θα επιλεγεί ως εξυπηρετητής για την εκάστοτε συναλλαγή.

Λόγω του γεγονότος ότι οι P2P εφαρμογές μπορεί να χρησιμοποιηθούν και για τη μεταφορά παράνομων αρχείων, είναι φανερό ότι τα συστήματα φήμης μπορούν να βελτιώσουν την απόδοση τους κατακόρυφα σε περιπτώσεις όπου τα αρχεία δεν είναι

αξιόπιστα, όπως για παράδειγμα στα αρχεία μουσικής όπου λέγεται ότι διάφορες εταιρίες που κατέχουν τα δικαιώματα μολύνουν το περιεχόμενο πολλών τραγουδιών τους που είναι διαθέσιμα για τους χρήστες (content poisoning). Επομένως τα συστήματα φήμης στα P2P αλλά και γενικά στις η-κοινότητες έχουν ως σκοπό τον καθορισμό των πιο αξιόπιστων χρηστών και το σύνολο εκείνων που παρέχουν τις πιο ποιοτικές πληροφορίες.

Στο επόμενο κεφάλαιο ασχολούμαστε με τα συστήματα συστάσεων, και συγκεκριμένα με τα συστήματα συστάσεων που χρησιμοποιούν τεχνικές συνεργατικού φιλτραρίσματος.

3. ΤΕΧΝΙΚΕΣ COLLABORATIVE FILTERING

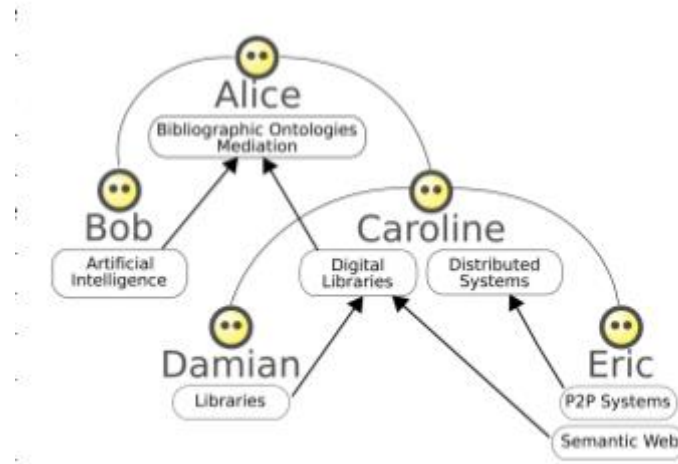
Στον ανοργάνωτο κόσμο του διαδικτύου κάθε πληροφορία που ψάχνουμε φαίνεται ότι βρίσκεται πολύ κοντά αλλά τις περισσότερες φορές εκτός της εμβέλειας και όταν εντέλει τη βρίσκουμε δεν μας είναι χρήσιμη. Οι online κατάλογοι, τα ηλεκτρονικά καταστήματα και οι μηχανές αναζήτησης μας επιστρέφουν αποτελέσματα, τα οποία αρκετές φορές δεν είναι άμεσα συσχετιζόμενα με το αντικείμενο προς αναζήτηση. Έτσι καταλήγουμε στο να ρωτάμε τους γνωστούς μας για ενδιαφέρουσες συστάσεις και προτάσεις. Ακριβώς αυτή είναι η ιδέα του συνεργατικού φιλτραρίσματος ή αλλιώς collaborative filtering (CF), δηλαδή η αυτοματοποίηση της διαδικασίας ερωτήσεων προς τον κύκλο γνωριμιών μας, όταν εμείς αναζητούμε κάποια πληροφορία στο διαδίκτυο. Στη συνέχεια του κεφαλαίου θα αναλύσουμε μερικές από τις πιο δημοφιλείς και αποτελεσματικές τεχνικές που χρησιμοποιούνται για την σχεδίαση αλγορίθμων με CF.

3.1. Semantic Social Collaborative Filtering

Η κατηγορία των μεθόδων Σημασιολογικού Κοινωνικού Φιλτραρίσματος (semantic social collaborative filtering [10]) βασίζεται σε δύο έννοιες: τις κατανεμημένες συλλογές (distributed collections) και τον σχολιασμό των πόρων (annotation of resources). Ο κάθε χρήστης συγκεντρώνει τις πληροφορίες για τις οποίες ενδιαφέρεται σε συλλογές που δημιουργεί [8]. Οι συλλογές που έχει ένας χρήστης στην κατοχή του μπορούν να συνδεθούν στις συλλογές άλλων χρηστών [9]. Αυτή η γνώση μοιράζεται στη συνέχεια στο κοινωνικό δίκτυο. Η διάδοση της γνώσης μέσω της σύνδεσης των συλλογών στο κοινωνικό δίκτυο αντιστοιχεί στο επίπεδο της εξειδίκευσης σε ένα συγκεκριμένο θέμα στο κοινωνικό δίκτυο. Έτσι λοιπόν ο βαθμός της εξειδίκευσης (expertise) που έχει κάθε κόμβος σε μια περιοχή ή θέμα εξαρτάται από την ποιότητα του υλικού που έχει στις συλλογές του, οι οποίες είναι διαθέσιμες στους φίλους του. Επιπρόσθετα ο βαθμός της εξειδίκευσης κάθε χρήστη σε ένα αντικείμενο είναι διαθέσιμος σε όλους εκείνους τους χρήστες που ασχολούνται με αυτό αντικείμενο προκειμένου να γίνεται ποιοτικότερη αναζήτηση πληροφοριών. Επιπλέον εκτός από τη διαχείριση συλλογών η οποία παρέχει την κατηγοριοποίηση των πηγών πληροφοριών, το semantic collaborative filtering αξιοποιεί τα σχόλια που κοινοποιούν οι χρήστες, τα οποία βοηθούν στην γρήγορη εξερεύνηση:

- του περιεχομένου της πηγής
- της σημασίας της πηγής
- της γενικής γνώμης για αυτήν από άλλους χρήστες

Παράδειγμα:



Σχήμα 1: Απεικόνιση του σημασιολογικού συνεργατικού φιλτραρίσματος

Στο παραπάνω παράδειγμα η Αλίκη γράφει μια εργασία στο θέμα “Διαμεσολάβηση στις Βιβλιογραφικές Οντολογίες” (“Mediation in Bibliographic Ontologies”) και προκειμένου να βρει κάποιες πληροφορίες εγγράφεται στη ψηφιακή βιβλιοθήκη του Πανεπιστημίου της. Σύντομα αντιλαμβάνεται ότι δεν είναι αρκετό το υλικό που έχει συγκεντρώσει και οι συλλογές των φίλων της, με τους οποίους συνδέεται, δεν είναι αρκετό. Στη συνέχεια λοιπόν παρατίθενται διάφοροι αλγόριθμοι έτσι ώστε η Αλίκη να βρει τις απαραίτητες πληροφορίες με τη βοήθεια του semantic social collaborative filtering.

3.1.1 Simple Semantic Social CF (Απλό Σημασιολογικό Κοινωνικό ΣΦ)

Το σύστημα που χρησιμοποιεί η Αλίκη για να συλλέξει το υλικό της μπορεί αρχικά να αναπαρασταθεί με τον απλό αλγόριθμο του σημασιολογικού συνεργατικού φιλτραρίσματος [10] ως εξής:

Πίνακας 1: Απεικόνιση Μοντέλου Αναζήτησης Υλικού

Διαδικασία ΛίσταΣυλλογών (p, t)

Για κάθε $p' \in P$

Αν ΑπόστασηΚόμβου(p, p') < ΕμβέλειαΔικτύου

$C' = C' \cup \text{ΣυλλογέςΚόμβου}(p')$

Τέλος για

Ταξινόμηση C' ανάλογα με τη ΤελικήΚατάταξηΣυλλογών

Τέλος Διαδικασίας

Πίνακας 2: Καθορισμός του Μοντέλου

P: Το σύνολο των κόμβων (peers).

C: Το σύνολο συλλογών.

ΣυλλογέςΚόμβου: Συνάρτηση που επιστρέφει τις συλλογές του ιδιοκτήτη της (*p*').

ΑπόστασηΚόμβου: Συνάρτηση που υπολογίζει την απόσταση μεταξύ δύο κόμβων.

ΕμβέλειαΔικτύου: Ορίζει τη μέγιστη επιτρεπτή απόσταση μεταξύ δυο κόμβων.

Ταξινόμηση: Συνάρτηση που ταξινομεί με τη βοήθεια του αλγορίθμου του Dijkstra τις συλλογές με κριτήριο την χρησιμότητα της καθεμίας στο θέμα που ψάχνουμε.

ΤελικήΚατάταξηΣυλλογών: Είναι η συνάρτηση που επιστρέφει ταξινομημένα τα αποτελέσματα των συλλογών βασιζόμενα στην απόσταση του κόμβου μας από τον κόμβο ιδιοκτήτη της συλλογής, τον βαθμό ομοιότητας και ποιότητας της συλλογής με το αντικείμενο του θέματος.

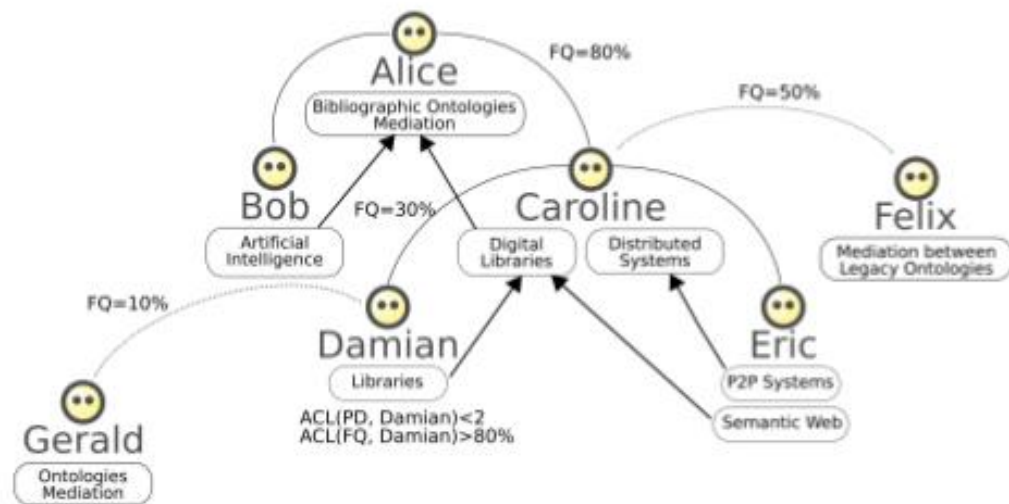
Με τον παραπάνω αλγόριθμο η ψηφιακή βιβλιοθήκη παρουσιάζει στην Αλίκη όλες εκείνες τις συλλογές (collections) που μπορεί να δει ανάλογα το εύρος των φίλων της, οι οποίοι έχουν σχετικές πληροφορίες. Στο τέλος η Αλίκη βρίσκει το ζητούμενο υλικό από την Caroline επειδή ο βαθμός εξειδίκευσής της στον τομέα είναι πολύ υψηλός. Αν και ο φίλος της ενδιαφέρεται στην Τεχνητή Νοημοσύνη επιλέγει να συνδεθεί με τις συλλογές του Eric που έχει υψηλή βαθμολογία στον τομέα “Semantic Web”. Από εδώ και στο εξής η Αλίκη είναι στην πλεονεκτική θέση να λαμβάνει πληροφορίες από τις συλλογές της Caroline και του Eric.

3.1.2 Secured Semantic Social CF (Ασφαλές Σημασιολογικό Κοινωνικό ΣΦ)

Η Αλίκη όμως εξακολουθεί να ψάχνει συλλογές για την εργασία της και έτσι κάνει την εγγραφή της σε μια ανοιχτή ψηφιακή βιβλιοθήκη στην οποία υπάρχει η δυνατότητα να απαγορεύεται η πρόσβαση στις συλλογές κάποιων χρηστών. Η απαγόρευση οφείλεται σε προϋποθέσεις που βασίζονται στην:

- Απόσταση που έχουν που έχουν δυο κόμβοι μεταξύ τους, δηλαδή ένας κόμβος πρέπει να ικανοποιεί τα βήματα που έχει θέσει ο ιδιοκτήτης των συλλογών ως μέγιστη απόσταση για να έχει κανείς πρόσβαση σε αυτές (πχ. ΜέγιστοΌριοΒημάτων = 4).
- Την εμπιστοσύνη που έχει ο ιδιοκτήτης για τον κόμβο που ζητάει πρόσβαση στις συλλογές του, δηλαδή να ικανοποιεί σίγουρα το ελάχιστο όριο εμπιστοσύνης που έχει θέσει (πχ. ΕλάχιστοΌριοΕμπιστοσύνης = 70%).

Για αυτόν τον λόγο η τεχνική αυτή περιλαμβάνει τις Λίστες Πρόσβασης (Access Control Lists ή ACL) [10] οι οποίες επιτρέπουν ή όχι την πρόσβαση στις συλλογές ενός κόμβου, όπως στο ακόλουθο παράδειγμα όπου η Αλίκη επειδή απέχει μόλις δυο βήματα από τον Damien δεν του παρέχεται η πρόσβαση.



Σχήμα 2: Ασφαλές σημασιολογικό συνεργατικό φιλτράρισμα με Λίστες Ελέγχου Πρόσβασης

Πίνακας 3: Αλγόριθμος του Ασφαλούς Μοντέλου SSCF

Διαδικασία Λίστα Συλλογών $ACL(p, t)$
 $C_p = \text{ΣυλλογέςΚόμβου}(p)$
 Για κάθε $p' \in P$
 Αν ΑπόστασηΚόμβου(p, p') < ΕμβέλειαΔικτύου
 Για κάθε $c' \in \text{ΣυλλογέςΚόμβου}(p')$
 Αν ΑπόστασηΚόμβου(p', p) < Απόσταση $ACL(p')$
 Αν ΕμπιστοσύνηΚόμβου(p', p) < ΌριοΕμπιστοσύνης(p')
 $C' = C' \cup \{c'\}$
 Τέλος Αν
 Τέλος Αν
 Τέλος για
 Τέλος για
 Ταξινόμηση C' ανάλογα με τη Τελική Κατάταξη Συλλογών
 Τέλος Διαδικασίας

Πίνακας 4: Καθορισμός του Ασφαλούς Μοντέλου SSCF

Απόσταση ACL : Επιστρέφει τη μέγιστη απόσταση από την οποία μπορεί να απέχει ένας κόμβος προκειμένου να έχει πρόσβαση στις συλλογές του ιδιοκτήτη της.
 ΌριοΕμπιστοσύνης: Επιστρέφει το ελάχιστο όριο εμπιστοσύνης που πρέπει να έχει ο ιδιοκτήτης των συλλογών για έναν κόμβο προκειμένου ο κόμβος αυτός να έχει πρόσβαση στις συλλογές του.
 ΕμπιστοσύνηΚόμβου: Επιστρέφει το ποσοστό εμπιστοσύνης που έχει ένας κόμβος για έναν άλλον.

3.2. Memory-Based Collaborative Filtering Techniques

Οι πρώτες γενιές των συστημάτων συνεργατικού φιλτραρίσματος όπως το GroupLens [11], χρησιμοποιούν τα δεδομένα των βαθμολογιών του χρήστη έτσι ώστε να υπολογίσουν την ομοιότητα (similarity) μεταξύ δυο χρηστών (users) ή αντικειμένων (items) και στη συνέχεια να κάνουν προβλέψεις σύμφωνα με τις υπολογισμένες τιμές (values) ομοιότητας. Τα συστήματα αυτά ονομάζονται συστήματα συνεργατικού φιλτραρίσματος βασιζόμενα στην μνήμη και έχουν υλοποιηθεί επιτυχώς σε μεγάλα ηλεκτρονικά καταστήματα όπως το Amazon και το Ebay. Αυτή η CF τεχνική χρησιμοποιεί ένα δείγμα ή όλη τη βάση δεδομένων, η οποία περιέχει τις βαθμολογίες του χρήστη για κάθε αντικείμενο. Κάθε χρήστης είναι μέλος μιας ομάδας ατόμων με παρόμοια χαρακτηριστικά.

Ένας από τους πιο γνωστούς memory-based αλγόριθμους είναι αυτός που βασίζεται σε αντικείμενα (item-based) και ακολουθεί τα εξής βήματα [12]:

1. Υπολογίζει την ομοιότητα μεταξύ δύο αντικειμένων λαμβάνοντας υπόψη τις βαθμολογίες που έχουν δοθεί από χρήστες που έχουν βαθμολογήσει και τα δύο.
2. Κάνει πρόβλεψη για το βαθμό που σε έναν χρήστη θα αρέσει ένα συγκεκριμένο αντικείμενο.
3. Παρουσιάζει τις κορυφαίες συστάσεις (top recommendations) αντικειμένων στο χρήστη.

Άλλοι memory-based αλγόριθμοι υπολογίζουν την ομοιότητα μεταξύ δύο χρηστών (user-based) λαμβάνοντας υπόψη τις βαθμολογίες που έχουν δώσει σε αντικείμενα που έχουν βαθμολογήσει από κοινού, π.χ. [13].

3.2.1 Υπολογισμός της Ομοιότητας

Ο υπολογισμός της ομοιότητας μεταξύ δυο αντικειμένων ή χρηστών είναι ένα κρίσιμο βήμα στη τεχνική αυτή. Η βασική ιδέα είναι ότι για να υπολογιστεί η ομοιότητα για δυο αντικείμενα i και j πρώτα πρέπει να βρούμε τους χρήστες εκείνους που έχουν βαθμολογήσει και τα δυο αντικείμενα [13]. Έπειτα υπολογίζουμε την ομοιότητα μεταξύ τους με διάφορες τεχνικές πολλές από τις οποίες παρουσιάζουμε παρακάτω.

3.2.1.1 Ομοιότητα Βασιζόμενη στη Συσχέτιση

Σε αυτή τη μέθοδο η ομοιότητα μεταξύ δυο χρηστών u , v , υπολογίζεται χρησιμοποιώντας τον συσχετισμό Pearson [11], που μετράει τον βαθμό όπου δυο μεταβλητές συσχετίζονται γραμμικά, και προκύπτει με τον παρακάτω τύπο για την περίπτωση του user-based αλγορίθμου:

$$w_{u,v} = \frac{\sum_{i \in I} (r_{u,i} - \bar{r}_u)(r_{v,i} - \bar{r}_v)}{\sqrt{\sum_{i \in I} (r_{u,i} - \bar{r}_u)^2} \sqrt{\sum_{i \in I} (r_{v,i} - \bar{r}_v)^2}}$$

$i \in I$ είναι τα αντικείμενα είναι τα αντικείμενα που και οι δυο χρήστες έχουν βαθμολογήσει και $\bar{r}_u$ ($\bar{r}_v$) μέση βαθμολογία που έχει δώσει ο χρήστης u (v) στα αντικείμενα που έχει βαθμολογήσει από κοινού με τον v (u), ενώ $r_{u,i}$, $r_{v,i}$ είναι οι βαθμολογίες των χρηστών u

και n αντίστοιχα για το αντικείμενο i . Υπάρχουν διάφορες παραλλαγές του Pearson όπως της ομοιότητας Pearson με περιορισμούς (constraints), της Sparkman Correlation, που είναι παρόμοια με του Pearson μόνο που αντί για βαθμολογίες περιέχει βαθμίδες (ranks) αλλά και του Kendall τ correlation όπου σε σχέση με του Sparkman χρησιμοποιεί μόνο τις συσχετιζόμενες κατατάξεις.

3.2.1.2 Vector Cosine – Based Similarity

Η ομοιότητα μεταξύ δυο εγγράφων μπορεί να γίνει παριστάνοντας κάθε έγγραφο σαν έναν πίνακα από λέξεις, όπου κάθε μια έχει έναν μετρητή συχνότητας. Αυτή η τακτική μπορεί να εφαρμοστεί στο collaborative filtering, μόνο που αντί για έγγραφα θα έχουμε αντικείμενα ή χρήστες και αντί για συχνότητες λέξεων θα έχουμε βαθμολογίες (ratings).

Αναλυτικότερα η ομοιότητα μεταξύ δύο οντοτήτων i και j ορίζεται ως το συνημίτονο των n – διανυσμάτων (vectors), που αντιστοιχούν στην i -οστή και j -στή στήλη του πίνακα R , που είναι ο $(m \times n)$ γραμμικός πίνακας χρηστών (m) και αντικειμένων (n) με το i και j να αποτελούν στοιχεία της στήλης που ανήκει στον συνολικό πίνακα R . Η ομοιότητα αυτή υπολογίζεται με τον τύπο [14]:

$$w_{i,j} = \cos(\vec{i}, \vec{j}) = \frac{\vec{i} \cdot \vec{j}}{\|\vec{i}\| * \|\vec{j}\|}$$

Για παράδειγμα αν τα i και j είναι τα διανύσματα $\vec{A} = \{x_1, y_1\}$ και $\vec{B} = \{x_2, y_2\}$ τότε ο τύπος ομοιότητας με βάση το συνημίτονο πινάκων είναι ο εξής:

$$w_{A,B} = \cos(\vec{A}, \vec{B}) = \frac{\vec{A} \cdot \vec{B}}{\|\vec{A}\| * \|\vec{B}\|} = \frac{x_1 x_2 + y_1 y_2}{\sqrt{x_1^2 + y_1^2} \sqrt{x_2^2 + y_2^2}}$$

Στην πράξη οι χρήστες μπορεί να χρησιμοποιούν διαφορετικές κλίμακες βαθμολόγησης, γεγονός το οποίο ο παραπάνω τύπος δεν λαμβάνει υπόψιν. Για αυτόν τον λόγο υπάρχει και η προσαρμοσμένη φόρμουλα που αφαιρεί από κάθε ζευγάρι τον μέσο όρο από όλα τα αντικείμενα που έχουν και οι δυο βαθμολογήσει για μεγαλύτερη ακρίβεια αποτελεσμάτων.

3.2.2 Πρόβλεψη και Υπολογισμός Σύστασης

Η πρόβλεψη για έναν χρήστη είναι το πιο σημαντικό κομμάτι σε ένα σύστημα που εφαρμόζει κάποια τεχνική collaborative filtering. Δεδομένων των πιο κοντινών γειτόνων (nearest neighbors) βάση της ομοιότητας χρησιμοποιούμε ένα σταθμισμένο άθροισμα των βαθμολογιών για να παράγουμε προβλέψεις [15].

3.2.2.1 Σταθμισμένο Άθροισμα από τις Βαθμολογίες

Ένας τρόπος για να κάνουμε πρόβλεψη για έναν χρήστη α σχετικά με ένα αντικείμενο i , δηλαδή πόσο ταιριάζει το αντικείμενο i αυτό για να προταθεί στον χρήστη α , είναι να πάρουμε πρώτα τις βαθμολογίες των γειτόνων u που έχουν ήδη βαθμολογήσει το αντικείμενο i . Έπειτα θα τις αφαιρέσουμε από τον μέσο όρο $\bar{r}_u$, που συμβολίζει τον γενικό μέσο όρο των βαθμολογιών τους για άλλα αντικείμενα που είχαν αγοράσει. Τέλος προσθέτουμε τη μέση βαθμολογία του χρήστη $\bar{r}_\alpha$ και το τελικό αποτέλεσμα αποτελεί την πρόβλεψη [13].

$$P_{\alpha,i} = \bar{r}_\alpha + \frac{\sum_{u \in U} (r_{u,i} - \bar{r}_u) \cdot w_{\alpha,u}}{\sum_{u \in U} |w_{\alpha,u}|}$$

Όπου $\bar{r}_u$, $\bar{r}_\alpha$ είναι ο μέσος όρος των βαθμολογιών των χρηστών u και α , και $w_{\alpha,u}$ το βάρος (ομοιότητα) μεταξύ α και u (β. Τύπο της 3.2.2.1).

3.2.2.2 Απλός Σταθμισμένος Μέσος

Ένας άλλος απλός τρόπος για να κάνουμε μια πρόβλεψη για έναν χρήστη u σχετικά με ένα συγκεκριμένο αντικείμενο i , στους item-based αλγόριθμους [12] είναι να χρησιμοποιήσουμε την μέθοδο του απλού σταθμισμένου μέσου, με τύπο:

$$P_{u,i} = \frac{\sum_{n \in N} r_{u,n} w_{i,n}}{\sum_{n \in N} |w_{i,n}|}$$

όπου το άθροισμα υπολογίζεται για όλα τα αντικείμενα $n \in N$ (δηλαδή τα αντικείμενα που έχει βαθμολογήσει ο u), $r_{u,n}$ η βαθμολογία του χρήστη u για το αντικείμενο n και $w_{i,n}$ η ομοιότητα των αντικειμένων i με το n .

3.2.3 Παρουσίαση των Κορυφαίων Συστάσεων

Στο σημείο αυτό έχουμε υπολογίσει τις ομοιότητες μεταξύ των χρηστών ή των αντικειμένων και εμφανίζουμε στον χρήστη προτάσεις όπως για παράδειγμα διάφορα προϊόντα προς αγορά σε περίοπτη θέση στην οθόνη του. Για να δείξουμε τις συστάσεις υπάρχουν δυο τρόποι, οι συστάσεις βασισμένες στον χρήστη ή στα αντικείμενα. Στην πρώτη κατηγορία μπορούμε πρώτα να στραφούμε στα αντικείμενα που έχουν αγοράσει άλλοι χρήστες. Αναλυτικότερα μόλις βρούμε τους ομοιότερους χρήστες, συγκεντρώνουμε τα πιο συχνά (frequent) αντικείμενα που έχουν αγοράσει όλοι και προτείνουμε εκείνα που δεν έχει αγοράσει ο χρήστης στον οποίο γίνονται οι συστάσεις. Μια άλλη τακτική η οποία είναι καλύτερη από την παραπάνω είναι να συγκρίνουμε πρώτα τα αντικείμενα. Σε αυτή τη

μέθοδο εντοπίζουμε πρώτα τα ομοιότερα αντικείμενα ανάλογα με το είδος του αντικειμένου. Στη συνέχεια συγκεντρώνουμε εκείνα που έχει ήδη αγοράσει ο χρήστης και συγκρίνουμε ποια από τα υπόλοιπα μοιάζουν πιο πολύ με αυτά. Τα ομοιότερα αποτελούν τις κορυφαίες επιλογές προς αγορά τις οποίες και παρουσιάζουμε.

3.3. Συνεργατικό Φιλτράρισμα με βάση το Μοντέλο

Η σχεδίαση και η ανάπτυξη των μοντέλων (όπως αλγόριθμοι machine learning και data mining) επιτρέπει σε ένα σύστημα να αναγνωρίζει περίπλοκα πρότυπα βασισμένα σε δεδομένα εκμάθησης (training phase) λαμβάνοντας έτσι ευφυείς προβλέψεις σε ζητήματα collaborative filtering. Παρακάτω παρουσιάζονται μερικές από τις πιο δημοφιλείς μεθόδους που χρησιμοποιούνται με αυτή τη μέθοδο CF.

3.3.1 Απλός Bayesian Αλγόριθμος

Η μεθοδολογία του naive Bayes χρησιμοποιείται ευρέως για ταξινομήσεις, και έχει πολύ καλά αποτελέσματα και σε μεθόδους συνεργατικού φιλτραρίσματος. Συγκεκριμένα, μπορεί να χρησιμοποιηθεί για τον υπολογισμό της πιο πιθανής κλάσης στην οποία ανήκει ένα αντικείμενο, σύμφωνα με τον παρακάτω τύπο. Για μη συμπληρωμένα δεδομένα, ο υπολογισμός της πιθανότητας και ο τύπος υπολογισμού της πιο πιθανής κλάσης για ένα αντικείμενο υπολογίζονται με χρήση δεδομένων παρατήρησης (ο υποδείκτης ο στον παρακάτω τύπο δηλώνει τιμές παρατήρησης):

$$\text{class} = \arg \max_{j \in \text{classSet}} p(\text{class}_j) \prod_o P(X_o = x_o \mid \text{class}_j)$$

Ο απλός αλγόριθμος του Bayes μπορεί να θεωρηθεί ότι δεν ανήκει στη κατηγορία του μοντέλου αλλά της μνήμης καθώς δεν μπορεί να χρησιμοποιηθεί για μη συμπληρωμένα δεδομένα. Για αυτό τον λόγο έχουν αναπτυχθεί επεκτάσεις αυτής της τεχνικής όπως η NB-ELR και TAN-ELR CF [16, 17] όπου έπειτα από μια φάση εκμάθησης μπορούν να προβλέψουν το αποτέλεσμα με υψηλή ακρίβεια. Μια άλλη επέκταση του Simple Bayesian θεωρεί ότι κάθε Bayesian κόμβος περιέχει ένα δέντρο απόφασης (decision tree), όπου κόμβος είναι το αντικείμενο, και οι καταστάσεις του δέντρου οι πιθανές βαθμολογίες για το αντικείμενο.

3.3.2 Αλγόριθμοι CF με Clustering (Συστάδες)

Μια συστάδα (cluster) είναι μια συλλογή από αντικείμενα που περιέχουν όμοια δεδομένα μεταξύ τους και ανόμοια σε σχέση με τα αντικείμενα που βρίσκονται σε άλλες συστάδες [18]. Η μέτρηση της ομοιότητας γίνεται με τη βοήθεια της απόστασης Minkowski, ο οποίος για δυο αντικείμενα $X = x_1, x_2, \dots, x_n$ και $Y = y_1, y_2, \dots, y_n$ ορίζεται από τον τύπο:

$$d(X, Y) = \sqrt[q]{\sum_{i=1}^n |x_i - y_i|^q}$$

Όπου n ο αριθμός των διαστάσεων για τα αντικείμενα X και Y . Για $q = 1$ το αποτέλεσμα μας δίνει την απόσταση Manhattan ενώ για $q = 2$ την Ευκλείδεια απόσταση.

Οι μέθοδοι που χρησιμοποιούνται για τη δημιουργία συστάδων χωρίζονται σε τρεις κατηγορίες:

- Ιεραρχικές (hierarchical)
- Βασισμένες στη πυκνότητα (density based) και
- Διαχωρισμού (partitioning)

Μια γνωστή υλοποίηση clustering με διαχωρισμό είναι η k-means, MacQueen [19]. Από την άλλη οι μέθοδοι πυκνότητας συνήθως ψάχνουν για πυκνές ομάδες αντικειμένων χωριζόμενες από αραιές περιοχές, που αναπαριστούν τον θόρυβο (noise). Γνωστές υλοποιήσεις αυτής της κατηγορίας είναι οι DBSCAN [20] και OPTICS [21]. Τέλος οι ιεραρχικές μέθοδοι χρησιμοποιούν κάποιο κριτήριο για την εξόρυξη του συνόλου των δεδομένων. Οι συστάδες αποτελούν ένα τρόπο παρουσίασης των δεδομένων ο οποίος χρειάζεται περαιτέρω ανάλυση για να προκύψουν προβλέψεις καθώς από μόνος αποτελεί απλώς μια μέθοδο κατηγοριοποίησης των δεδομένων. Στην έρευνα των Sarvar et al. [22], O'Connor και Iterlocker [23] χρησιμοποιείται ένας αλγόριθμος συνεργατικού φιλτραρίσματος με βάση τη μνήμη προκειμένου να κάνουν προβλέψεις για κάθε cluster.

3.3.3 Regression Algorithms

Η κατηγορία αλγορίθμων παλινδρόμησης (regression algorithms) χρησιμοποιεί μια προσέγγιση των βαθμολογιών του χρήστη για να κάνει προβλέψεις. Αν $X=(X_1, X_2, \dots, X_n)$ μια τυχαία μεταβλητή που παριστάνει τις προτιμήσεις ενός χρήστη σε διάφορα αντικείμενα. Τότε το γραμμικό μοντέλο παλινδρόμησης μπορεί να εκφραστεί ως:

$$Y = \Lambda X + N$$

όπου Λ είναι ένας $n \times k$ πίνακας, $N=(N_1, \dots, N_n)$ είναι μια τυχαία μεταβλητή που παριστάνει το θόρυβο στις επιλογές του χρήστη, Y ο $n \times m$ πίνακας με τις προβλέψεις για κάθε αντικείμενο, X είναι ο $k \times m$ πίνακας όπου κάθε στήλη είναι ο υπολογισμός της τιμής της τυχαίας μεταβλητής X για έναν χρήστη. Γνωστή υλοποίηση αυτής της μεθόδου είναι του Vunetic & Obradovic [24][27].

3.4. Υβριδικές τεχνικές για Collaborative Filtering

Αυτή η κατηγορία συνδυάζει αλγορίθμους συνεργατικού φιλτραρίσματος με άλλες τεχνικές συστάσεων (συνήθως με συστήματα βασισμένα στο περιεχόμενο). Τα συστήματα συστάσεων βασισμένα στο περιεχόμενο (content-based recommendation systems) αναλύουν τα δεδομένα του χρήστη, όπως τα μηνύματα του, το ιστορικό πλοήγησης του, το προφίλ του ή ακόμα και τα αρχεία του προκειμένου να παραχθούν ακριβείς προτάσεις [25]. Το αρνητικό με αυτά τα συστήματα είναι ότι δεν μπορούν να δημιουργήσουν αποτελεσματικές προτάσεις στην περίπτωση που ο χρήστης είναι νέος (new user problem) καθώς δεν υπάρχουν αρκετές πληροφορίες ή πολλές φορές είναι δύσκολη η εξόρυξη τους σε αντίθεση με τα συστήματα CF. Τέλος τα CBS πάσχουν από το πρόβλημα της υπερεξειδίκευσης (overspecialization), δηλαδή τη συνεχή σύσταση προϊόντων τα οποία είναι όμοια μεταξύ τους, ενώ για παράδειγμα σε μια ηλεκτρονική κοινότητα για συνταγές μαγειρικής ο χρήστης ενδεχομένως να θέλει να δοκιμάσει και συνταγές από διαφορετικές κουζίνες [26][26].

3.4.1 Υβριδικά μοντέλα με τεχνικές CF και CBS

Ο ενισχυμένος CF αλγόριθμος με CBS χρησιμοποιεί τον απλό Bayesian αλγόριθμο για την κατηγοριοποίηση του περιεχομένου και συμπληρώνει τις βαθμολογίες που δεν έχει αγοράσει ο χρήστης. Έπειτα εφαρμόζει τη σταθμισμένη μέθοδο του Pearson CF επανεκτιμώντας τα αντικείμενα εκείνα τα οποία έχουν αγοραστεί και βαθμολογηθεί από πολλούς όμοιους ενεργούς χρήστες [28]. Παράδειγμα:

Πίνακας 5: Πίνακας χρηστών με τις βαθμολογίες τους

	Age	Sex	Career	zip	I_1	I_2	I_3	I_4	I_5
U_1	32	F	writer	22904			4		
U_2	27	M	student	10022	2		4	3	
U_3	24	M	engineer	60402		1			
U_4	50	F	other	60804		3	3	3	3
U_5	28	M	educator	85251	1				

Πίνακας 6: Πίνακας βαθμολογιών συμπληρωμένος με τις βαθμολογίες που προέκυψαν από τον αλγόριθμο CBS.

I_1	I_2	I_3	I_4	I_5
2	3	4	3	2
2	2	4	3	2
3	1	3	4	3
3	3	3	3	3
1	2	4	1	2

Πίνακας 7: Πίνακας χρηστών αναθεωρημένος από τα αποτελέσματα του αλγορίθμου CF.

I_1	I_2	I_3	I_4	I_5
2	3	4	2	3
3	4	2	2	3
3	3	2	3	3
3	3	3	3	3
1	3	1	2	2

Τα αποτελέσματα του Πίνακα 7 είναι πιο ακριβή γιατί ο αλγόριθμος CF λαμβάνει υπόψιν τις βαθμολογίες των όλων των χρηστών που έχουν αγοράσει τα συγκεκριμένα αντικείμενα παράγοντας μια βαθμολογία προερχόμενη από πολλούς χρήστες.

Μια υλοποίηση αυτής της μεθόδου είναι Ansari et al. [29] όπου προτείνει ένα μοντέλο με Bayesian προτίμηση, όπου στατιστικά ενσωματώνει πολλούς τύπους πληροφοριών, όπως οι προτιμήσεις των χρηστών και οι γνώμες των ειδικών (experts). Τα αποτελέσματα τους έδειξαν ότι η μίξη των δυο τεχνικών πετυχαίνουν καλύτερα αποτελέσματα από το απλό συνεργατικό φιλτράρισμα.

Τέλος υπάρχουν και οι υβριδικές υλοποιήσεις συνδυάζοντας αποκλειστικά μεθόδους CF. Αυτή η κατηγορία συνδυάζει αλγορίθμους μοντέλου και μνήμης. Τα αποτελέσματα έχουν δείξει ότι παράγουν πιο ακριβείς συστάσεις από ότι κάθε μέθοδος ξεχωριστά [30][31]. Σχετικά παραδείγματα τέτοιων ερευνών είναι το Σύστημα Διάγνωσης της Προσωπικότητας (Personality Diagnosis) [31]. Το PD διαλέγει για κάθε χρήστη τυχαία έναν άλλον και υπολογίζει την πιθανότητα ομοιότητας των δυο χρηστών και αν είναι υψηλή τότε συστήνονται στον πρώτο χρήστη τα προϊόντα στα οποία ο δεύτερος έχει δώσει υψηλή βαθμολογία.

4. ΕΥΡΕΣΗ ΚΟΜΒΩΝ ΜΕΓΑΛΗΣ ΕΠΙΡΡΟΗΣ ΣΕ ΚΟΙΝΩΝΙΚΑ ΔΙΚΤΥΑ

Στο κεφάλαιο αυτό εξετάζουμε την λειτουργία της εύρεσης των πιο σημαντικών κόμβων οι οποίοι ασκούν επιρροή σε άλλους κόμβους. Η επιρροή αυτή έχει ως αποτέλεσμα τη λήψη ή την αποφυγή παρόμοιων ενεργειών από τους χρήστες που είναι δέκτες της. Βλέπουμε πως ένας χρήστης επηρεάζεται όχι μόνο θετικά από τις ενέργειες των φίλων του σε ένα

online κοινωνικό δίκτυο άλλα και αρνητικά μειώνοντας τις πιθανότητες του να δράσει με παρόμοιο τρόπο όπως και ο φίλος του. Επιπρόσθετα έχουν γίνει εκτενείς έρευνες στον τομέα αυτόν, τα αποτελέσματα των οποίων χρησιμοποιούνται σήμερα, για τη χρησιμοποίηση μιας μικρής ομάδας χρηστών με μεγάλη επιρροή προκειμένου να διαφημιστούν προϊόντα σε μεγάλο πλήθος ατόμων χωρίς μεγάλο κόστος καθώς η διαφήμιση είναι στοιχούμενη και οι χρησιμοποιούνται οι κόμβοι του κοινωνικού δικτύου με τη μεγαλύτερη εμβέλεια. Παρακάτω θα αναλύσουμε διάφορους τρόπους διαλογής και επιλογής τέτοιων κόμβων χρησιμοποιώντας τη τεχνική του μοντέλου διάχυσης (diffusion model) και μιας τεχνικής που χρησιμοποιείται για ηλεκτρονικές κοινότητες αναρτήσεων (blogs), που έχουν δομή παρόμοια με τα ηλεκτρονικά κοινωνικά δίκτυα που εξετάζουμε, η οποία εντοπίζει με μεγάλη ακρίβεια τους κόμβους με τη μεγαλύτερη επιρροή βάσει ποικίλων στατιστικών στοιχείων που προκύπτουν μέσα από την δραστηριότητα των χρηστών μέσα σε αυτές.

4.1. Μοντέλο Υπολογισμού Επιρροής TGT

Σε αυτήν την υποενότητα θα αναλύσουμε το μοντέλο υπολογισμού επιρροής Trust General Threshold (TGT) [31], το οποίο λαμβάνει υπόψη του θετικές και αρνητικές επιρροές στα κοινωνικά δίκτυα εμπιστοσύνης. Το μοντέλο αυτό βασίζεται στο απλό μοντέλο διάχυσης που περιγράφεται στην ενότητα 4.1.1.

4.1.1 Απλό Μοντέλο Διάχυσης

Ο αλγόριθμος αρχίζει με το μοντέλο διάχυσης, που ορίζει ποιοί χρήστες θα ενεργήσουν αφού κάποιος άλλος γείτονας τους (φίλος τους) έχει ήδη ενεργήσει [31]. Σε έναν γράφο κοινωνικού δικτύου $G(V,E)$, όπου V το σύνολο των κόμβων και E οι ακμές προς τους γείτονες τους (οι σχέσεις των κόμβων με άλλους), το μοντέλο διάχυσης M ορίζεται ως $\sigma M(S)$ όπου S ο αριθμός των ενεργών χρηστών με $S \subseteq V$ και συμβολίζει την εξάπλωση της επιρροής της ομάδας S . Δημοφιλείς υλοποιήσεις αυτής της κατηγορίας συστημάτων είναι το Independent Cascade (IC) και το Linear Threshold (LT).

Το Independent Cascade (Model) είναι ένα μοντέλο διάχυσης όπου η πληροφορία μεταδίδεται μέσω των cascade κόμβων που μπορούν να έχουν δυο καταστάσεις είτε ενεργή είτε ανενεργή. Όταν είναι στην πρώτη σημαίνει ότι ο χρήστης έχει ήδη υιοθετήσει ή επηρεαστεί από την πληροφορία ενώ στην δεύτερη το αντίθετο. Μια εκδοχή του αλγορίθμου που χρησιμοποιείται για το IC είναι ο ακόλουθος:

Διαδικασία Διάχυσης (Γράφος g , Λίστα seeds)

Πίνακας ενεργοί Χρήστες, Αποτελέσματα

Στοίβα κόμβοι

```

Για κάθε  $s \in seeds$ 
  κόμβοι.εισαγωγή( $s$ )
  Όσο (κόμβοι.μέγεθος > 0)
    κόμβος  $\kappa$  = κόμβοι.εξαγωγή()
    ενεργοίΚόμβοι.εισαγωγή( $\kappa$ )
    Για κάθε γείτονα  $\gamma$  του κόμβου  $\kappa$ 
       $\alpha$  = τυχαίοςΑριθμός()
      Αν  $\alpha < \kappa.πιθανότηταΕπιρροής$ 
        Αν !ενεργοίΚόμβοι.μέλος( $\gamma$ )
          κόμβοι.εισαγωγή( $\gamma$ )
      Τέλος αν
    Τέλος αν
  Τέλος για
Τέλος όσο
Τέλος για
Τέλος Διαδικασίας

```

Επιπλέον υπάρχει και το γραμμικό μοντέλο κατωφλιού (linear threshold model) που έχει παρόμοια δομή και λογική με το IC με τη διαφορά ότι ένας κόμβος προκειμένου να γίνει ενεργός πρέπει το άθροισμα των πιθανοτήτων επιρροής προς αυτόν από τους υπολοίπους γείτονες να υπερβαίνει ένα κατώφλι θ . Καταληκτικά το κύριο συστατικών των δυο μοντέλων είναι ότι κάθε κόμβος μόλις γίνεται ενεργός δεν απενεργοποιείται για αυτό αυτοί οι αλγόριθμοι ακολουθούν μονότονη ανάπτυξη.

Σε έρευνα του Kempe et. al [33] έχει αποδειχθεί ότι τα παραπάνω συστήματα είναι μονότονα, δηλαδή εάν ένας χρήστης εισαχθεί ως κόμβος επιρροής στο σύνολο S τότε το σύνολο S θα αυξηθεί, δηλαδή:

$$\sigma_M(S \cup \{V\}) \geq \sigma_M$$

Ξέροντας λοιπόν ότι πρόκειται για μονότονες συναρτήσεις μπορούμε να χρησιμοποιήσουμε την έρευνα του Nemhauser et al. [34] που αποδεικνύει ότι οποιαδήποτε μονότονη συνάρτηση μπορεί να λυθεί χρησιμοποιώντας άπληστους αλγορίθμους (greedy algorithms). Η μέθοδος αυτή έχει δυο προϋποθέσεις:

- Έναν κατευθυνόμενο γράφο $G(V, E)$.
- Τις πιθανότητες επιρροής (influence probabilities) για κάθε κόμβο που ανήκει στον γράφο.

Η πρώτη προϋπόθεση είναι δεδομένη καθώς έχουμε πρόσβαση σε όλο τον γράφο και τους γείτονες του κάθε κόμβου αλλά η δεύτερη, δηλαδή οι υπολογισμένες πιθανότητες επιρροής δεν είναι τετριμμένη διαδικασία. Απαιτεί την εξόρυξη πληροφοριών από την Καταγραφή Δράσεων (Action Log) [35] που αποτελείται από την τριπλέτα Χρήστης, Δράση, Χρόνος ή $Action(u, a, t)$ που σημαίνει ότι ένας χρήστης $u \in V$ έκανε μια ενέργεια a σε έναν χρόνο t .

Τα πιο πολλά συστήματα όπως το LT και IC θεωρούν ότι μια ενέργεια μπορεί να προκαλέσει μόνο θετικό αντίκτυπο. Ειδικότερα όσο αυξάνονται οι φίλοι ενός χρήστη οι οποίοι για παράδειγμα έχουν αγοράσει ένα συγκεκριμένο προϊόν τόσο περισσότερο αυξάνεται και η πιθανότητα να το αγοράσει και εκείνος.

Στην πραγματικότητα όμως δεν συμβαίνει αυτό καθώς μια ενέργεια ενός γείτονα του χρήστη ενδέχεται να μειώσει την θέληση του χρήστη να ενεργήσει με τον ίδιο τρόπο επειδή δεν τον εμπιστεύεται. Επομένως υπάρχει και το φαινόμενο της αρνητικής επιρροής και στη συνέχεια παρουσιάζουμε μια τέτοια προσέγγιση, το TGT (Trust-General Threshold) [32] μοντέλο, που δεν ακολουθεί μονότονη. Αυτό είναι καθοριστικό καθώς όπως είπαμε οι μονότονες μέθοδοι χρησιμοποιούν απλούς αλγορίθμους οι οποίοι είναι πολύ ακριβοί στην εκτέλεση τους και το πιο ακριβό βήμα τους είναι η αναζήτηση των κόμβων με τη μεγαλύτερη επιρροή.

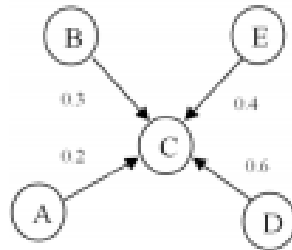
4.1.2 Μοντέλο Διάχυσης με Θετικό και Αρνητικό Αντίκτυπο στον Χρήστη

Το TGT μοντέλο γενικεύει τα προαναφερθέντα LT και IC. Αρχικά υπολογίζεται η πιθανότητα $P_u(S)$ να υιοθετήσει ο χρήστης την ενέργεια που έχουν ήδη κάνει οι ενεργοί γείτονες του (active members), δηλαδή οι γείτονες που έχουν εκτελέσει μια συγκεκριμένη ενέργεια [35], σύμφωνα με τον τύπο:

$$p_u(S) = 1 - \prod_{v \in S} (1 - p_{v,u})$$

όπου $p_{v,u}$ η επιρροή που ασκεί ο v στον u και S το σύνολο των ενεργών γειτόνων.

Ο παραπάνω τύπος λαμβάνει υπόψη μόνο για τους γείτονες που εμπιστεύεται ο χρήστης και όχι εκείνους που δεν εμπιστεύεται.



Σχήμα 1: Απεικόνιση γράφου με τις πιθανότητες επιρροής ως βάρη στις ακμές

Θεωρούμε δύο σενάρια για τον κόμβο C του σχήματος 1, όπου επίσης ισχύει ότι ο κόμβος C δεν εμπιστεύεται τον κόμβο A. Στο πρώτο σενάριο θεωρούμε ότι ο κόμβος E ενεργοποιείται και ότι η πιθανότητα επιρροής $p_{E,C}=0.3$. Τότε σύμφωνα με την παραπάνω τύπο $P(\{E\}) = 1 - (1 - 0.3) = 0.3$. Στο δεύτερο σενάριο, θεωρούμε ότι οι E, A ενεργοποιούνται. Αφού ο E είναι ο μόνος κόμβος που εμπιστεύεται ο C, η πιθανότητα, η πιθανότητα να ενεργοποιηθεί ο κόμβος C είναι και εδώ $P(\{E\}) = 1 - (1 - 0.3) = 0.3$. Στο δεύτερο σενάριο θα έπρεπε όμως να ληφθεί υπόψη και η αρνητική επιρροή από τον A, αφού ο C δεν εμπιστεύεται τον A,. Έτσι προκειμένου να υλοποιήσουμε ένα πιο ακριβές μοντέλο πρέπει να ενσωματώσουμε και την έννοια της αρνητικής επιρροής (negative influence probability), η οποία θα συμβολίζεται με $p_{v,u}$. Αν η θετική πιθανότητα επιρροής ισούται με μηδέν τότε ο χρήστης u δεν εμπιστεύεται τον χρήστη v ενώ αν η αρνητική είναι μηδέν το αντίστροφο. Αν S^+ είναι όλοι οι εμπιστευόμενοι κόμβοι του κόμβου u και S^- οι μη εμπιστευόμενοι αντίστοιχα με $S = S^+ \cup S^-$ τότε οι πιθανότητες υπολογίζονται ως εξής:

$$p_u(S^+) = 1 - \prod_{v \in S^+} (1 - p_{v,u}^+)$$

$$p_u(S^-) = 1 - \prod_{v \in S^-} (1 - p_{v,u}^-)$$

Δεδομένων των πιθανοτήτων επιρροής η κοινή ή συναθροισμένη πιθανότητα επιρροής (joint influence probability) είναι θετική εάν η τιμή της είναι μεγαλύτερη από ένα κατώφλι θ ή

αρνητική εάν είναι μικρότερη από το τετράγωνο του κατωφλιού θ όπως φαίνεται στις παρακάτω συνθήκες:

$$p_u(S^+) > \theta^1 \text{ AND} \\ p_u(S^-) < \theta^2$$

Στη πρώτη περίπτωση σημαίνει ότι ο χρήστης u είναι πολύ πιθανόν να δράσει με τον ίδιο τρόπο μέσα από τη συνολική που του ασκούν οι γείτονες του ενώ στη δεύτερη το αντίθετο, δηλαδή να μην δράσει με τον ίδιο τρόπο.

Έχοντας προσθέσει λοιπόν την δυνατότητα της αρνητικής επιρροής με τους παραπάνω τύπους μένει να ορίσουμε τη μέθοδο εύρεσης του αριθμού των κόμβων που θα επηρεαστούν σύμφωνα με αυτό το μοντέλο διάχυσης. Για να τους εντοπίσουμε πρέπει να υπολογίσουμε το $\sigma TGT(S)$ ακολουθώντας τα ακόλουθα βήματα [36][37] :

1. Βρες την τρέχουσα ομάδα των ενεργών χρηστών, την οποία συμβολίζουμε με H και αρχικά είναι το σύνολο S (όλοι οι ενεργοί κόμβοι).
2. Για κάθε γείτονα στο S υπολόγισε τις θετικές και αρνητικές πιθανότητες επιρροής.
3. Κάθε κόμβος u , εκτός του S , που γίνεται ενεργός προστίθεται στο H .
4. Για κάθε γείτονα του u που γίνεται ενεργός στο προηγούμενο βήμα, η θετική και αρνητική πιθανότητα επιρροής υπολογίζονται όπως στο βήμα 2 και 3.
5. Αν ενεργοποιηθούν και άλλοι χρήστες επαναλαμβάνεται για αυτούς το 3ο βήμα. Αν δεν ενεργοποιηθούν άλλοι τότε η διαδικασία της διάχυσης (process of diffusion) σταματάει εδώ. Η μετάδοση της επιρροής είναι $\sigma(S)$ είναι ο συνολικός αριθμός των ενεργοποιημένων χρηστών του συνόλου H .

4.1.3 Υπολογισμός Πιθανότητας Επιρροής

Για να ολοκληρωθεί ο παραπάνω αλγόριθμος μας μένει να καθορίσουμε τον τύπο από τον οποίο θα προκύψει η πιθανότητα επιρροής, δηλαδή το $p_{v,u}$. Αναλυτικότερα θα εξορύξουμε τις πληροφορίες για τις ενέργειες του χρήστη σχετικά με την αντίδραση του στις ενέργειες των γειτόνων του μέσα από το ιστορικό ενεργειών του, action log, υπολογίζοντας τη θετική και αρνητική συχνότητα ενεργειών. Οι συχνότητες αυτές συμβολίζουν το ποσοστό των ίδιων ενεργειών που θα κάνει ένας χρήστης u αφού πρώτα τις έχει πράξει ένας γείτονας του v και χρησιμοποιούνται από τον τύπο υπολογισμού της πιθανότητας θετικής και αρνητικής επιρροής ως εξής:

$$p_{v,u}^+ = \begin{cases} 0 & \text{if } u \text{ distrust } v \\ \frac{A_{v,u}}{A_v} & \text{otherwise} \end{cases}$$

$$p_{v,u}^- = \begin{cases} 0 & \text{if } u \text{ trusts } v \\ \frac{A'_{v,u}}{A_v} & \text{otherwise} \end{cases}$$

όπου A_v : ο αριθμός των ενεργειών που έκανε ο χρήστης v , $A_{v,u}$: ο αριθμός των ενεργειών που έκανε ο χρήστης u επηρεασμένος από τον χρήστη v που έκανε προηγουμένως τις ίδιες ενέργειες, και $A'_{v,u}$ ο αριθμός των ενεργειών που δεν έκανε ο v από τις ενέργειες που έκανε ο u (τον οποίο δεν εμπιστεύεται). Για παράδειγμα εάν ο χρήστης v αγοράσει τα προϊόντα 1, 2, 3 και ο v τα 1, 2, τότε έχουμε ότι $A_{v,u} = 2$ ενώ $A_v = 3$ αφού 3 είναι το σύνολο των προϊόντων που έχει αγοράσει ο χρήστης v . Τελικά η θετική συχνότητα επιρροής είναι 0,66 ενώ η αρνητική το υπόλοιπο, δηλαδή 0,33.

4.1.4 Εύρεση Κόμβων με τη Μεγαλύτερη Επιρροή

Ο αλγόριθμος χρησιμοποιεί δύο κατώφλια θ_1 και θ_2 που έστω ότι έχουν τιμές 0.3 και 0.6 αντίστοιχα. και ο αριθμός S είναι 2. Αρχικά ο αλγόριθμος αναζητά τον κόμβο με την μεγαλύτερη επιρροή, οπότε εάν για παράδειγμα το σύνολο των κόμβων μας μαζί με την επιρροή που συνεισφέρει ο καθένας είναι το ακόλουθο:

Node v	u_1	u_2	u_3	u_4	u_5
$\sigma_{TGT}(\{v\})$	3	2	2	1	1

τότε στην αρχή του αλγορίθμου θα διαλέγαμε τον κόμβο u_1 και η διασπορά θα είχε τιμή

3. Στην συνέχεια ο αλγόριθμος επιλέγει τον κόμβο u_2 και το $\sigma_{TGT}(S+\{u_2\}) = 4 > 3$, οπότε αφού είναι η μεγαλύτερη τιμή δε θα συνεχίζει ο αλγόριθμος και οι κόμβοι του συνόλου $\{S\}$ δεν θα αλλάξουν αφού δεν υπάρχει άλλος συνδυασμός από τον οποίο θα προκύψει μεγαλύτερη διασπορά. Το προηγούμενο παράδειγμα μπορεί να υπολογιστεί με τον ακόλουθο αλγόριθμο:

Διαδικασία ΕξόρυξηΚόμβωνΕπιρροής (Γράφος g , ΑριθμόςΚόμβωνΕπιρροής)

1. $S = \text{NULL}$
2. Υπολόγισε διασποράς κάθε κόμβο $v \in V$, $\sigma\text{TGT}(\{v\})$ και διάλεξε εκείνο με τη μεγαλύτερη τιμή σ .
3. Διαμόρφωσε το S έτσι ώστε να έχει τη μεγαλύτερη διασπορά.
4. Αφαίρεσε κόμβο v έτσι ώστε $\sigma\text{TGT}(S - \{v\}) > \sigma\text{TGT}(S)$.
5. Πρόσθεσε τον κόμβο u , $u \in V - S$ ώστε $|S| < \text{ΑριθμόςΚόμβωνΕπιρροής}$ και $\sigma\text{TGT}(S + \{u\}) > \sigma\text{TGT}(S)$.
6. Εναλλαγή v , $v \in S$ με τον u έτσι ώστε $\sigma\text{TGT}(S + \{v\} - \{u\}) > \sigma\text{TGT}(S)$.
7. Συνέχισε το βήμα 4 μέχρι η τιμή της διασποράς να είναι μεγαλύτερη από την προηγούμενη.
8. Επέστρεψε το σύνολο S .

Τέλος Διαδικασίας

4.2. Μοντέλο Υπολογισμού της Επιρροής Σχεδιασμένο για Blogs

Στην υποενότητα αυτή θα αναλύσουμε ένα μοντέλο υπολογισμού της επιρροής σχεδιασμένο για κοινωνικά δίκτυα και blogs.

4.2.1 Μοντέλο Εύρεσης Κόμβων Επιρροής σε blogs

Η ανάρτηση άρθρων (blogging) είναι ένας ιδιαίτερα δημοφιλής τρόπος έκφρασης της άποψης του κάθε χρήστη μέσω του διαδικτύου καθώς έχουν την δυνατότητα να μοιράζονται τα νέα της καθημερινότητας, να παράσχουν συμβουλές, να σχηματίζουν ομάδες στις οποίες μπορούν να αναπτύσσουν επικοινωνιακούς διαλόγους δημιουργώντας έτσι έναν αποτελεσματικό τρόπο επικοινωνίας. Έρευνες [32], [38] έχουν δείξει πως στην πραγματικότητα το 83% των ανθρώπων συμβουλευονται τον κύκλο τους πριν δοκιμάσουν ένα εστιατόριο, το 71% πριν αγοράσει μια φαρμακευτική αγωγή και το 61% πριν δει μια ταινία. Επομένως πριν προβούμε σε κάποια κίνηση δεχόμαστε την επιρροή του κύκλου μας. Η ίδια διαδικασία συμβαίνει και στις ηλεκτρονικές κοινότητες, που παρέχουν μια αναπαράσταση της πραγματικότητας, οι οποίες είναι ένας χώρος στον οποίο μπορούμε να επεξεργαστούμε τις δραστηριότητες των χρηστών με μεγαλύτερη ευχέρεια βγάζοντας κρίσιμα συμπεράσματα. Τα συμπεράσματα αυτά έχουν μεγάλη αξία και

χρησιμοποιούνται για την διάδοση προϊόντων, την εξακρίβωση των τάσεων των καταναλωτών και της αγοράς, την ακριβή πρόβλεψη των πωλήσεων καθώς και την δημιουργία πηγών πληροφοριών. Προκειμένου να γίνει αυτό, είναι αναγκαίο και καθοριστικής σημασίας να εντοπίσουμε ανάμεσα σε αυτούς τους “bloggers” εκείνους που ασκούν μεγάλη επιρροή στους υπολοίπους οι οποίοι στρέφουν της προσοχή στις αναρτήσεις τους και επηρεάζονται με οδηγώντας τους στην λήψη κάποιας ενέργειας.

4.2.2 Bloggers με την Μεγαλύτερη Επιρροή

Το μοντέλο που θα αναλύσουμε και διατυπώθηκε από τους Nitin Agarwal, Huan Liu και Lei Tang [39] επισημαίνει πως ένας κόμβος επιρροής δεν είναι αναγκαίο να είναι ενεργός (active user), δηλαδή να καταχωρεί αναρτήσεις με μεγάλη συχνότητα, αλλά μπορεί να είναι και ανενεργός (inactive user) χωρίζοντας έτσι τους χρήστες στις ακόλουθες κατηγορίες:

- Ενεργός κόμβος επιρροής (Active influent user)
- Ανενεργός κόμβος επιρροής (Inactive influent user)
- Ενεργός κόμβος μη επιρροής (Active non influent user)
- Ανενεργός κόμβος μη επιρροής (Inactive non influent user)

Για παράδειγμα ένας χρήστης μπορεί να έχει μεγάλη συχνότητα καταχώρησης αναρτήσεων αλλά το περιεχόμενο τους να μην είναι καινοτόμο είτε να περιέχει μικρής έκτασης κείμενο το οποίο να μην προσφέρει κάποια σημαντική πληροφορία στους χρήστες, ενώ αντίθετα ένας χρήστης που δεν είναι ενεργός ενδέχεται να δημιουργεί αναρτήσεις, το περιεχόμενο των οποίων να έχει προέλθει από τις γνώσεις, τις πληροφορίες του ή ακόμα και προσωπικές εμπειρίες του χρήστη προσελκύοντας την προσοχή των μελών της κοινότητας. Επίσης ο τρόπος αναζήτηση αυτών των κόμβων θα εφαρμοστεί σε κοινότητες στις οποίες ο κάθε χρήστης μπορεί να ξεκινήσει ένα θέμα όπου μπορούν και άλλοι να απαντήσουν, να διαφωνήσουν ή να συμφωνήσουν. Ενδεικτικά παραδείγματα τέτοιων ηλεκτρονικών κοινοτήτων είναι το επίσημο blog της Google (googleblog.blogspot.com) και το ανεπίσημο blog της Apple (tuaw.com).

Σε αυτή την κατηγορία των blogs κάθε ανάρτηση (post) περιέχει κάποια μεταδεδομένα όπως τον δημιουργό του, την ημερομηνία δημιουργίας του και την εμβέλεια του (score). Επιπλέον περιέχει τις ακόλουθες στατιστικές πληροφορίες:

1. Τις αναρτήσεις άλλων χρηστών που έχει συμπεριλάβει στο post του ο χρήστης (outlinks).
2. Τον αριθμό των χρηστών που έχουν συμπεριλάβει το δικό του στα δικές τους αναρτήσεις (inlinks).

3. Το μέγεθος της ανάρτησης (post length), ως προς τη χωρητικότητα ή/και τον αριθμό των λέξεων του.
4. Τον μέσο αριθμό των σχολίων της ανάρτησης.
5. Τον ρυθμό με τον οποίο καταχωρούνται τα σχόλια στην ανάρτηση.

Έστω ότι με τον συνδυασμό των παραπάνω πληροφοριών ορίζουμε μια ανάρτηση επιρροής. Ένας απλός τρόπος για να διαπιστώσουμε εάν ένας blogger έχει δύναμη επιρροής είναι να ελέγξουμε αν διαθέτει τουλάχιστον μια ανάρτηση επιρροής. Δεδομένων λοιπόν των βαθμολογιών επιρροής $I(pi)$ για κάθε ανάρτηση pi του χρήστη bk , μπορούμε να τις ταξινομήσουμε και να ορίσουμε ως μέγιστη βαθμολογία επιρροής ανάμεσα στις N , $1 \leq i \leq N$, που έχει δημιουργήσει την τιμή $iIndex(bk)$, μια δομή που περιέχει για κάθε χρήστη τη μέγιστη βαθμολογία επιρροής από τις αναρτήσεις του και η αναζήτηση της γίνεται με κλειδί την ταυτότητα (id) του χρήστη. Επομένως αν ταξινομήσουμε τη δομή $iIndex$ και θέλουμε τους K κόμβους με τη μεγαλύτερη επιρροή τότε συλλέγουμε τους δημιουργούς των K πρώτων τιμών της δομής. Επίσης για να προστεθεί κάποιος χρήστης σε αυτό το σύνολο πρέπει να πληροί την συνθήκη $I(pi) \geq iIndex(bk)$. Έχοντας λοιπόν ορίσει το πρόβλημα του εντοπισμού των κόμβων με επιρροή ακολουθεί μια ανάλυση του τρόπου μελέτης των χαρακτηριστικών εκείνων που θα μας βοηθήσουν να ορίσουμε το $iIndex$ και I .

4.2.3 Ιδιότητες Εντοπισμού των Κόμβων Επιρροής

Ένας κόμβος έχει επιρροή εάν είναι αναγνωρίσιμος από τους υπόλοιπους χρήστες, αν έχει καινοτόμες ιδέες και μπορεί να δημιουργήσει δραστηριότητες που συγκεντρώνουν την προσοχή [32], [39]. Παρακάτω αναλύουμε πως αυτές οι ιδιότητες μπορούν να υπολογιστούν μέσω στατιστικών τεχνικών:

- *Αναγνώριση* - Μια ανάρτηση p με μεγάλη επιρροή σημαίνει ότι θα αναγνωρίζεται και από πολλούς. Αυτό συμβαίνει εάν η ανάρτηση p αναφέρεται ως πηγή σε πολλές άλλες, δηλαδή όταν ο αριθμός των inlinks (i) είναι μεγάλος. Επίσης όσο περισσότερο επηρεάζουν τα posts που αναφέρουν την p τόσο περισσότερο αυξάνεται η επιρροή του αναφερόμενου post.
- *Δημιουργία Δραστηριοτήτων* - Η ικανότητα της κάθε ανάρτησης να δημιουργεί κίνηση και την δραστηριότητα των υπολοίπων χρηστών μπορεί να μετρηθεί έμμεσα από τον αριθμό των σχολίων που λαμβάνει. Αυτό σημαίνει πως ένα post με λίγα η καθόλου σχόλια δεν έχει δύναμη επιρροής. Επομένως ένας μεγάλος αριθμός από σχόλια (γ) σηματοδοτεί πως η συγκεκριμένη ανάρτηση επηρεάζει πολλούς κόμβους, αφού ενέργησαν και γράψανε την απάντηση/σχόλιο τους. Αξίζει να αναφερθεί πως αν και υπάρχει το γνωστό φαινόμενο του spam, υπάρχουν αρκετές ερευνητικές δημοσιεύσεις οι οποίες προσφέρουν αξιόπιστες τεχνικές για την καταπολέμηση του προκειμένου να υπολογιστεί με ακρίβεια αυτή η παράμετρος [39].

- **Καινοτομία** - Οι καινούργιες ιδέες προκαλούν μεγαλύτερο αντίκτυπο και επιρροή στους χρήστες. Κατά συνέπεια εάν μια ανάρτηση περιέχει μεγάλο αριθμό από πηγές του διαδικτύου ή άλλες αναρτήσεις, δηλαδή outlinks (θ), γεγονός που δείχνει πως δεν είναι τόσο πιθανόν να είναι ένα καινοτόμο post. Συμπερασματικά ο αριθμός των πηγών που αναφέρουμε σε μια ανάρτηση συνδέεται αρνητικά με τον αριθμό των σχολίων σε αυτήν με αποτέλεσμα να μειώνεται ο αριθμός των χρηστών που προσελκύονται.
- **Μέγεθος Ανάρτησης** - Αν και είναι δύσκολο να διαπιστώσουμε εάν ο χρήστης δημιουργεί μακροσκελείς αναρτήσεις χωρίς λόγο, θεωρούμε ότι για να το κάνει υπάρχει κάποιος λόγος. Επομένως το μέγεθος της ανάρτησης (λ) είναι μια παράμετρος η οποία υποδεικνύει πως αν είναι μεγάλο τότε θα υπάρχουν και περισσότερα σχόλια.

4.2.4 Υπολογισμός των Κόμβων Επιρροής

Η επιρροή μιας ανάρτησης μπορεί αναπαρασταθεί με έναν κατευθυνόμενο γράφο όπου κάθε κόμβος του αποτελεί μια ανάρτηση (blog post) η οποία χαρακτηρίζεται από τις παραπάνω ιδιότητες i , θ , γ και λ . Δεδομένου του γράφου μπορούμε να ορίσουμε για ένα post p τη “Ροή Επιρροής” με τον τύπο:

$$InfluenceFlow(p) = w_{in} \sum_{m=1}^{|\iota|} I(p_m) - w_{out} \sum_{n=1}^{|\theta|} I(p_n)$$

όπου w_{out} και w_{in} είναι τα βάρη τα οποία μπορούν να χρησιμοποιηθούν για την στάθμιση του αποτελέσματος ανάλογα με τη σημασία που έχει για εμάς ο αριθμός των πηγών της ανάρτησης p αλλά και ο αριθμός που ενσωματώνεται ως πηγή η p αντίστοιχα. Επίσης το p_m είναι όλες οι αναρτήσεις στις οποίες αναφέρεται η ανάρτηση p , με $1 \leq m \leq |\iota|$ ενώ p_n είναι όλες οι αναρτήσεις τις οποίες αναφέρει το συγκεκριμένο post όπου $1 \leq n \leq |\theta|$. Η “Ροή Επιρροής” μετράει την διαφορά μεταξύ του i και λ που σημαίνει ότι όσο περισσότερα inlinks έχουμε τόσο αναγνωρίζεται το post και όσο αυξάνονται τα outlinks τόσο μειώνεται η αναγνωρισιμότητα του. Επιπρόσθετα υπάρχει, όπως εξηγήσαμε παραπάνω, η ιδιότητα της δημιουργίας συνεχής ροής (γ), δηλαδή η μεγάλη συχνότητα κοινοποίησης αναρτήσεων, η οποία συνυπολογίζεται στον τύπο μας με τον ακόλουθο τύπο:

$$w_{com}\gamma_p + InfluenceFlow(p)$$

όπου το w_{com} ορίζει το βάρος που μπορεί να εφαρμοστεί για να ρυθμίσει την συνεισφορά του αριθμού των σχολίων στον υπολογισμό της επιρροής. Τέλος εφαρμόζουμε στον παραπάνω τύπο την ιδιότητα (λ), δηλαδή το μέγεθος της ανάρτησης, η οποία επιβραβεύει

ή χαμηλώνει τη βαθμολογία της επιρροής του post. Για να την εφαρμόσουμε χρησιμοποιούμε τον ακόλουθο ολοκληρωμένο τύπο:

$$I(p) = w(\lambda) \times (w_{com}\gamma_p + InfluenceFlow(p))$$

Η συνάρτηση που υπολογίζει το βάρος για την παράμετρο του μεγέθους της ανάρτησης μπορεί να αντικατασταθεί από έναν αλγόριθμο ανάλυσης κειμένου προκειμένου να έχουμε ακριβέστερα αποτελέσματα. Έχοντας λοιπόν την παραπάνω εξίσωση έχουμε τη δυνατότητα να υπολογίσουμε τη βαθμολογία επιρροής για κάθε ανάρτηση. Επομένως για κάθε χρήστη θα υπολογίσουμε τις βαθμολογίες των post του, θα βρούμε εκείνη η οποία έχει τη μεγαλύτερη και θα την αντιστοιχίσουμε στη δομή *iIndex* με τον μοναδικό αριθμό της ταυτότητας του με τον τύπο:

$$iIndex(B) = \max(I(p_i))$$

όπου N όλες οι αναρτήσεις του χρήστη και $1 \leq i \leq N$. Μόλις ολοκληρωθεί η διαδικασία καταχώρησης της μεγαλύτερης βαθμολογίας για κάθε κόμβο στη δομή αυτή μπορούμε να την ταξινομήσουμε και οι k πρώτοι που θα προκύψουν θα είναι και εκείνοι με την μεγαλύτερη επιρροή ή εκείνοι που θα έχουν ικανοποιήσει ένα κατώφλι (threshold).

5. ΠΕΡΙΓΡΑΦΗ ΥΛΟΠΟΙΗΣΗΣ

Η υλοποίηση μας έχει ως στόχο την ανάλυση δεδομένων που δέχεται απο το Twitter (<http://www.twitter.com>), ένα απο τα πιο δημοφιλή μέσα κοινωνικής δικτύωσης, με σκοπό τον υπολογισμό της επιρροής των χρηστών του καθώς και τον υπολογισμό της επιρροής μηνυμάτων του Twitter, που απο εδώ και στο εξής θα τα αποκαλούμε tweets, συσχετιζόμενα με μια συγκεκριμένη θεματική περιοχή. Για να μπορέσουμε να εντοπίσουμε χρήστες με επιρροή ή tweets θα πρέπει πρώτα να ορίσουμε τους παράγοντες εκείνους που συνδυάζοντας τους προκύπτει εάν το περιεχόμενο ενός tweet περιέχει συγκεκριμένες προϋποθέσεις ώστε να διαπιστώσουμε ότι έχει υψηλή δύναμη επιρροής. Στην ακόλουθη υποενότητα παρουσιάζουμε αυτούς ακριβώς τους παράγοντες.

5.1 Υπολογισμός Επιρροής ενός Μηνύματος στο Twitter (Influence of a Tweet)

Όπως αναλύθηκε και στην προηγούμενη ενότητα, ένα post με επιρροή χαρακτηρίζεται απο τέσσερις παράγοντες [39]. Για τις ανάγκες των κοινωνικών δικτύων, και ιδιαίτερα του Twitter, τους προσαρμόζουμε κατάλληλα καταλήγοντας στους παρακάτω:

Αναγνώριση (Recognition): Ένα tweet t , με επιρροή αναγνωρίζεται αδιαμφισβήτητα απο πολλούς χρήστες. Αυτή η χρήσιμη πληροφορία, την οποία αποκαλούμε και αριθμό των εσωτερικών συνδέσμων (number of inlinks) και συμβολίζουμε με i μπορεί να εξορυχθεί

βρίσκοντας τον συνολικό αριθμό των tweets τα οποία αναφέρονται στο tweet t . Οι αναφορές αυτές είναι γνωστές και ως retweets, δηλαδή αναδημοσίευση και αναπαραγωγή του tweet ενός άλλου χρήστη.

Προτίμηση (Preference): Αντί να χρησιμοποιήσουμε τον παράγοντα Αναπαραγωγής Δραστηριοτήτων που αναλύθηκε στη προηγούμενη ενότητα, χρησιμοποιούμε τον παράγοντα της προτίμησης (γ), που υπολογίζεται απο τον αριθμό εκείνο όπου ένα tweet t έχει μαρκαριστεί ή επιλεχθεί ως αγαπημένο (favorite) απο άλλους χρήστες. Με αυτή τη παράμετρο καταλαβαίνουμε πως ένας μεγάλος αριθμός γ υποδεικνύει ότι το tweet επηρεάζει τόσο πολύ τους χρήστες ώστε να το προσθέσουν στη συλλογή με τα αγαπημένα τους.

Καινοτομία (Novelty): Η καινοτομία είναι ένας καταλυτικός παράγοντας που πρέπει να εντοπίσουμε στα tweets. Οι καινοτόμες ιδέες υποδηλώνουν ότι ένα tweet είναι πολύ πιθανόν να έχει δύναμη επιρροής καθώς εκφράζει νέο και φρέσκο περιεχόμενο. Τον παράγοντα αυτόν τον συμβολίζουμε με θ . Πιο συγκεκριμένα στην υλοποίηση μας, εάν ένα tweet t αναφέρεται σε πολλά άλλα tweets ή σε άλλους συνδέσμους (links) τότε είναι λιγότερο πιθανό να περιέχει καινοτόμο περιεχόμενο καθώς ένα μεγάλο μέρος του έχει δανειστεί απο κάποια άλλη πηγή. Απο την άλλη μεριά ένα tweet το οποίο περιέχει πολυμέσα (multimedia), όπως για παράδειγμα εικόνες ή βίντεο είναι πιθανότερο να είναι καινοτόμο. Ακόμα και αν τα πολυμέσα έχουν δανειστεί απο κάποια άλλη πηγή γνωρίζουμε πως αυτά τα tweets προσελκύουν πέντε φορές μεγαλύτερη προσοχή σε σχέση με εκείνα που περιέχουν μόνο κείμενο. Ένας άλλος δείκτης καινοτομίας είναι όταν το t αναφέρεται σε άλλους χρήστες, γνωστός και ως user mentions, ή συγκεκριμένες θεματικές ενότητες, μια λειτουργία γνωστή στο Twitter ως hashtags (όπου κάθε χρήστης χρησιμοποιώντας το σύμβολο «#» πρίν απο τη λέξη που αντικατοπτρίζει το σχετικό θέμα π.χ. #Πανεπιστήμιο). Αυτός ο παράγοντας υποδηλώνει ότι το tweet στοχεύει μια ή περισσότερες κατηγορίες χρηστών ή/και θεματικών ενοτήτων προκειμένου να βρεθούν πιο εύκολα και αποδοτικά στη λειτουργία αναζήτησης tweets (searching for tweets functionality) απο τους άλλους χρήστες του κοινωνικού δικτύου.

Λόγιος (Eloquence): Ένα μακροσκελές tweet, που το συμβολίζουμε με l έχει συχνά μεγάλη επιρροή όπως παρουσιάσαμε και στη προηγούμενη ενότητα.

5.2 Υλοποίηση: Υπολογιστής Επιρροής για το Twitter (Twitter Influence Computer - TIC)

Παίρνοντας τους παραπάνω παράγοντες υπόψιν, παρουσιάζουμε στις ακόλουθες υποενότητες το σύστημα TIC (Twitter Influence Computer) αναλύοντας τον αλγόριθμο υπολογισμού της επιρροής ενός tweet, την σύγκριση του με το δημοφιλές σύστημα υπολογισμού της επιρροής στα κοινωνικά δίκτυα Klout (<http://www.klout.com>), τις διάφορες παραλλαγές και επιδράσεις του αλγορίθμου σταθμίζοντας διαφορετικά τα βάρη των παραμέτρων του καθώς και τον τρόπο πρόσβασης στα δεδομένα του Twitter.

5.2.1 Υπολογισμός Επιρροής ενός Μηνύματος στο Twitter (Influence of a Tweet)

Η επιρροή ενός tweet εκφράζεται μέσα από τις κοινωνικές δραστηριότητες και πράξεις άλλων χρηστών, μέσα από τον αποθήκευση ενός tweet ως αγαπημένο, από ένα retweet ή ακόμα και από την εξωστρέφεια που έχει ένα tweet (χρησιμοποίηση hashtags, user mentions, multimedia) που υποδηλώνει ότι θα έχει περισσότερες προβολές. Έτσι λοιπόν φτάνουμε στον τρόπο υπολογισμού της επιρροής ενός tweet. t Για να την υπολογίσουμε, δεδομένων των $(i, \theta, \lambda, \gamma)$ χρησιμοποιούμε με αντίστοιχο τρόπο όπως στην ενότητα 4 την εξίσωση:

$$I(t) = w_{\lambda}(\lambda) \times (w_{\gamma}(\gamma) + InfluenceFlow(t))$$

όπου w_{γ} και w_{λ} τα βάρη που χρησιμοποιούνται για να ρυθμίσουν την συνεισφορά του μεγέθους του tweet και του αριθμού των favorites αντίστοιχα για τον υπολογισμό της επιρροής. Από την άλλη μεριά, ο παράγοντας InfluenceFlow υπολογίζεται ως:

$$InfluenceFlow(t) = w_i(i) - w_{\theta}(\theta')$$

όπου w_i και w_{θ} τα βάρη για τον αριθμό των inlinks και των outlinks, που είναι ένας συνδυασμός των (i, θ, γ) , δηλαδή:

$$\theta' = (w_{\gamma}(\gamma) + w_i(i)) \times w_{\theta}(\theta)$$

Ο σκοπός του θ' είναι να παραχθεί ένας συγκρίσιμος αριθμός των outlinks σε σχέση με τα inlinks. Το βάρος του θ παίρνει τιμές $0 < w_{\theta} < 1$ ανάλογα τον βαθμό αυστηρότητας που θέλουμε να δώσουμε στο TIC. Εάν το w_{θ} είναι κοντά στο 1 τότε το σύστημα αποδίδει χαμηλά βάρη σε tweets τα οποία περιέχουν outlinks. Για παράδειγμα αν τα retweets και τα favorites είναι 3000 και 30 αντίστοιχα, όλα τα υπόλοιπα βάρη 1 εκτός από το w_{θ} που έχει τιμή 0.5 τότε η παράμετρος θ' έχει τιμή 1515 που μπορεί να συγκριθεί με την τιμή των retweets. Πρόκειται για μια κανονικοποίησης της τιμής των outlinks. Πριν αναλύσουμε τη μέθοδο για τον υπολογισμό της επιρροής ενός χρήστη, παρουσιάζουμε τον αλγόριθμο υπολογισμού ενός tweet χρησιμοποιώντας τη παραπάνω εξίσωση:

```

Result: Influence Score of a Tweet
w' ← weight for normalized theta;
w ← weight for i, gamma, l and theta;
i ← i * w;
i ← number of inlinks of tweet;
gamma ← number of favorites of tweet;
gamma ← gamma * w;
l ← length of tweet;
l ← l * w;
theta ← number of outlinks of tweet;
theta ← theta * w;
theta ← w' * (i + gamma) * theta;

InfluenceFlow ← i - theta;

if InfluenceFlow == 0 and gamma == 0 then
| InfluenceOfTweet ← l;
else
| InfluenceOfTweet ← l * (gamma + InfluenceFlow);
end

```

Εικόνα 2: Ψευδοκώδικας Υπολογισμού της Επιρροής ενός tweet

5.2.2 Υπολογισμός Επιρροής ενός Χρήστη στο Twitter (Influence of a User)

Βασιζόμενοι στην παραπάνω μέθοδο υπολογισμού της επιρροής ενός tweet, είμαστε στη θέση να υπολογίσουμε τη συνολική επιρροή ενός χρήστη u ακολουθώντας τα παρακάτω βήματα:

- 1.Ορίζουμε ένα κατώφλι k για κάθε βαθμολογία επιρροής που δίνουμε στα tweets τέτοιο ώστε αν $s \geq k$, όπου $k > 0$.
- 2.Αθροίζουμε όλα τα tweets του χρήστη, αυτά που ικανοποιούν το κατώφλι k και αυτά που δεν το ικανοποιούν και συμβολίζουμε τον αριθμό τους με a .
- 3.Αθροίζουμε μόνο τον αριθμό b των tweets του χρήστη που ικανοποιούν το κατώφλι k καθώς και το άθροισμα των βαθμολογιών τους c .
- 4.Η συνολική βαθμολογία επιρροής του χρήστη u δίνεται απο τον τύπο:

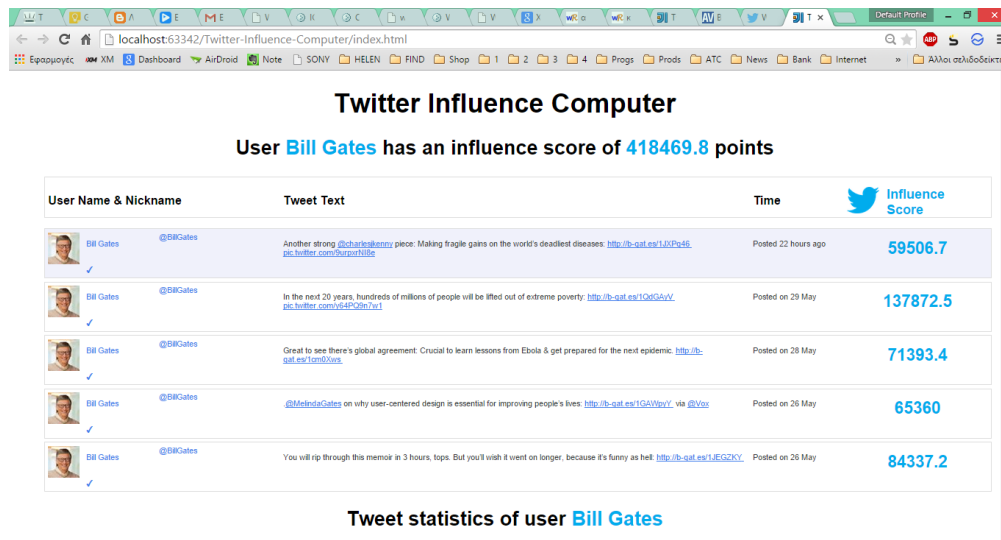
$$InfluenceScoreOfUser(u) = c \times \frac{b}{a}$$

Σε αντίθεση με το [40] όπου δεν υπάρχει διαχώριση μεταξύ των χρηστών που αναρτούν tweets ή posts με μεγάλη επιρροή σε αραιά χρονικά διαστήματα και αυτούς οι οποίοι συστηματικά αναρτούν και επηρεάζουν τους άλλους χρήστες ή ακόλουθους του κοινωνικού

δικτύου, το TIC τους επιβραβεύει με μεγαλύτερες βαθμολογίες. Χρησιμοποιώντας αυτόν τον κανένα εξασφαλίζουμε ότι στα αποτελέσματα του TIC θα εμφανίζονται χρήστες οι οποίοι θα τραβούν την προσοχή των χρηστών με μεγάλη συχνότητα και επομένως θα είναι πιο χρήσιμοι.

5.2.3 Τεχνολογίες και Εργαλεία που χρησιμοποιήθηκαν για το TIC

Υλοποιήσαμε το TIC σαν μια εφαρμογή διαδικτύου όπου προβάλλει το βαθμό επιρροής ενός χρήστη, tweet ή hashtags μαζί με την γραφική αναπαράσταση τους. Παρακάτω φαίνεται η υλοποίηση υπολογίζοντας την επιρροή του Bill Gates.



Εικόνα 3: Διεπαφή TIC - Υπολογισμός της Συνολικής Επιρροής ενός Χρήστη

Οι τεχνολογίες που χρησιμοποιήσαμε για την ανάπτυξη του TIC ήταν javascript σαν γλώσσα προγραμματισμού μαζί με την βιβλιοθήκη jQuery, HTML5 και CSS3. Τέλος χρησιμοποιήσαμε και τη βιβλιοθήκη ChartJs για την αναπαράσταση των αποτελεσμάτων με γραφήματα που δείχνουν την επιρροή ενός χρήστη με έναν πιο παραστατικό και κατανοητό τρόπο. Για να τρέξει κάποιος την εφαρμογή πρέπει να ακολουθήσει τα εξής βήματα:

1. Να συνδεθεί στον προσωπικό του λογαριασμό στο Twitter και να λάβει τον μοναδικό αριθμό id του χρήστη ή της θεματικής ενότητας απο τη κατηγορία Widgets που βρίσκεται στις ρυθμίσεις.
2. Για να πάρει τον κωδικό θα πρέπει να επιλέξει «Δημιουργία νέου widget».
3. Έπειτα θα διαλέξει χρονολόγιο χρήστη και θα πληκτρολογήσει το όνομα του χρήστη στο πεδίο «Όνομα Χρήστη»
4. Επιλέγουμε το κουμπί «Δημιούργησε widget».
5. Αποθηκεύουμε τον κωδικό αριθμό που βρίσκεται στο πεδίο data-widget-id στο κάτω μέρος της σελίδας.

6. Πληκτρολογούμε τον αριθμό αυτόν στο αρχείο `configurations.js` και συγκεκριμένα στο πεδίο «id» του αντικειμένου «`UserInfluenceConfiguration`».
7. Τέλος φορτώνουμε τη σελίδα `index.html`.

Για τον υπολογισμό επιρροής των tweets μια θεματικής ενότητας χρησιμοποιούμε τη κατηγορία «`Search`» και ακολουθώντας τα ίδια βήματα.

Για να λάβουμε τα δεδομένα απο το Twitter χρησιμοποιείται η ακόλουθη διεύθυνση πρόσβασης: `cdn.syndication.twimg.com/widgets/timelines/12345?&lang=english&callback=twitterFetcher.callback` που βρίσκεται στο αρχείο `twitterFetcher.js` όπου υλοποιείται ο πυρήνας της υλοποίησης, δηλαδή η απόκτηση των δεδομένων, σε γλώσσα `JavaScript`. Ο αριθμός 12345 είναι ο μοναδικός αριθμός, ταυτότητα, για μια θεματική ενότητα της οποίας θέλουμε να αποκτήσουμε πρόσβαση στα tweets της ή για έναν χρήστη του οποίου θέλουμε να υπολογίσουμε την επιρροή του. Η παράμετρος `lang` υποδηλώνει τη γλώσσα στην οποία θέλουμε να λάβουμε την απάντηση του αιτήματος μας για απόκτηση δεδομένων απο το Twitter και στο προκείμενο παράδειγμα έχει οριστεί σε Αγγλικά. Τέλος η τελευταία παράμετρος, `callback`, ορίζει τη συνάρτηση η οποία θα δεχτεί τα δεδομένα για την μετέπειτα επεξεργασία. Οι συναρτήσεις `callback`, αποτελούν κύριο γνώρισμα της `JavaScript`, και εκτελούνται μόλις συμβεί και ολοκληρωθεί κάποιο γεγονός. Αφού λάβει τα δεδομένα η συνάρτηση `callback` τα ανατρέχει και για κάθε tweet αποθηκεύει τις πληροφορίες του στους ακόλουθους πίνακες:

1. Τον αριθμό των retweets του tweet στον πίνακα `numberOfRetweets`.
1. Τον αριθμό των χαρακτήρων του tweet στον πίνακα `lengthOfTweets`.
2. Τον αριθμό των favorites του tweet στον πίνακα `numberOfFavorites`.
3. Εάν υπάρχει εικόνα αποθηκεύεται στον πίνακα `images`.
4. Τον χρήστη που δημοσίευσε το tweet στον πίνακα `authors`.
5. Εάν το tweet είναι retweet, δηλαδή ό χρήστης αναδημοσιεύει την ανάρτηση κάποιου άλλου χρήστη στον πίνακα `retweetedTweets`.
6. Την ώρα της δημοσίευσης του tweet.
7. Τους χρήστες που αναφέρονται στο tweet (`user mentions`) στον πίνακα `userMentionsOfTweet`.
8. Τα hashtags που αναφέρονται στο tweet στον πίνακα `hashtagsOfTweets`.

Εάν το tweet ανήκει στον πίνακα `retweetedTweets`, σημαίνει ότι δεν ανήκει στον χρήστη τον οποίο εξετάζουμε και συνεπώς παραλείπεται και ο αλγόριθμος συνεχίζει με το επόμενο. Αν δεν ανήκει όπως εξηγήσαμε και στις προηγούμενες συναρτήσεις υπολογίζουμε την επιρροή του χρήστη χρησιμοποιώντας τη συνάρτηση `computeInfluenceOfUser`, που βρίσκεται στο αρχείο `influentialComputations.js`, η οποία δέχεται ως παραμέτρους τους πίνακες 1, 2, 3, 4, 8 και 9. Η `computeInfluenceOfUser` αυτή για κάθε tweet χρησιμοποιεί τη συνάρτηση `computeInfluenceOfTweet` για να υπολογίζει τη επιρροή του κάθε tweet δεχόμενη τις παραμέτρους `I`, `gamma`, `I` και `theta`. Αφού υπολογιστεί η συνολική βαθμολογία του χρήστη την εκτυπώνουμε μέσω της `html` χρησιμοποιώντας συναρτήσεις της `javascript` οι οποίες χειρίζονται το μοντέλο της σελίδας (`DOM – Document Object Manipulation`), το οποίο εμπεριέχει όλα τα δεδομένα και στοιχεία της σελίδας.

5.2.4 Πρόσβαση στα Δεδομένα του Twitter

Η πρόσβαση στα δεδομένα του Twitter επιτυγχάνεται μέσω του Widget plugin [44] που έχει αναπτύξει για την ενσωμάτωση των προφίλ χρηστών σε ιστοσελίδες τρίτων. Οι κατηγορίες των Widgets είναι οι ακόλουθες:

- Χρονολόγια χρηστών (user timelines) τα οποία προβάλλουν δημόσια tweets απο οποιοδήποτε χρήστη.
- Η κατηγορία των αγαπημένων (favorites category) δείχνει tweets απο έναν συγκεκριμένο χρήστη τα οποία έχουν μαρκαριστεί ως αγαπημένα.
- Λίστες (Lists), οι οποίες δείχνουν δημόσια tweets απο δημόσιες λίστες στις οποίες ο χρήστης έχει εγγραφεί.
- Αναζήτηση (Search)

Στην υλοποίηση μας χρησιμοποιήσαμε την πρώτη και την τρίτη κατηγορία των Twitter Widgets. Η πρώτη χρησιμοποιήθηκε για να καθοριστεί η επιρροή ενός συγκεκριμένου χρήστη ενώ η τρίτη για την επιρροή των tweets τα οποία ανήκουν σε μια συγκεκριμένα θεματική ενότητα. Ο λόγος για τον οποίο χρησιμοποιήσαμε αυτόν τον τρόπο για να έχουμε πρόσβαση στα δεδομένα του Twitter είναι γιατί μας δίνει επιπλέον την δυνατότητα με μια απλή και εύκολη διαδικασία να ενσωματώνεται το TIC σε κάθε ιστοσελίδα. Επιπρόσθετα προσπερνά την πολύπλοκη διαδικασία πιστοποίησης του προγραμματιστή, αποκτώντας ένα ενδεικτικό πρόσβασης (access token) απο το Twitter, προκειμένου να αποκτήσει πρόσβαση στα δεδομένα του.

6. ΕΛΕΓΧΟΣ

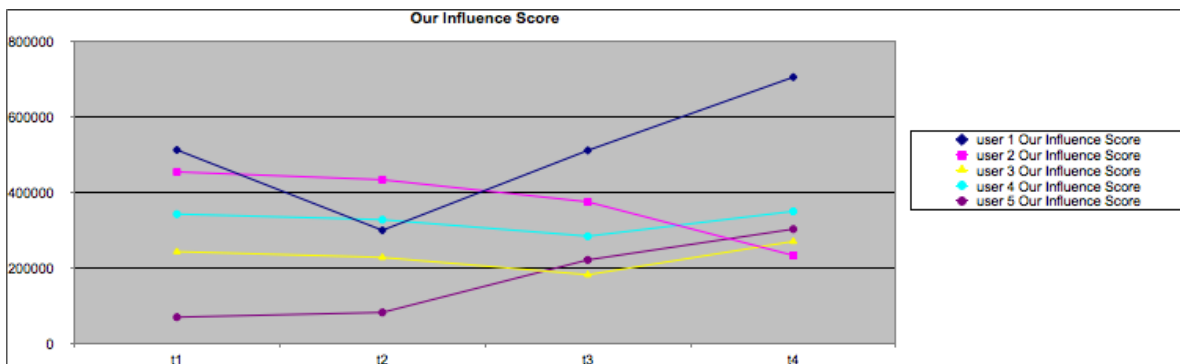
Σε αυτή την ενότητα παρουσιάζουμε την αξιολόγηση και τα συμπεράσματα των πειραμάτων μας .

6.1 Σύγκριση με το Klout

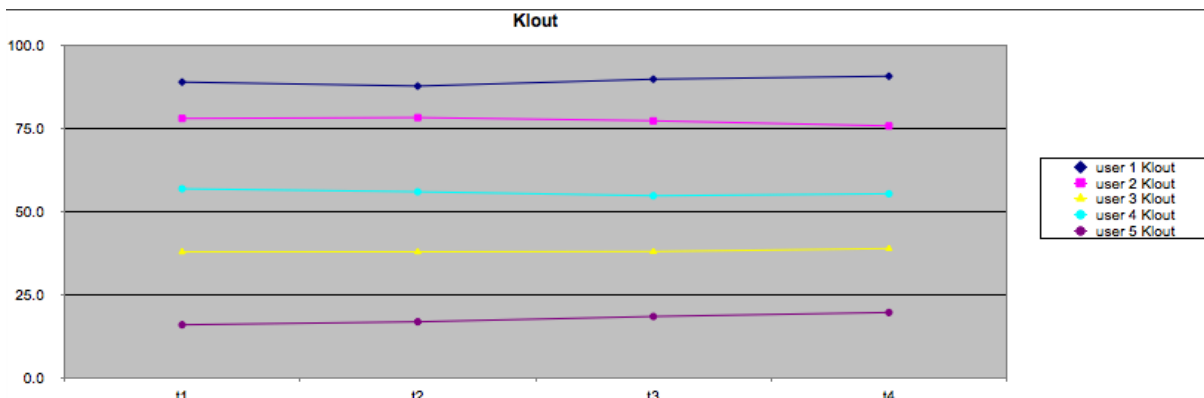
Για την αξιολόγηση του TIC κάναμε κάποια πειράματα σύγκρισης με το σύστημα υπολογισμού της επιρροής χρηστών κοινωνικών δικτύων, Klout [44]. Για την σύγκριση με το Klout εξετάσαμε τη εξέλιξη της επιρροής του TIC μέσα απο ένα δείγμα χρηστών και το συγκρίναμε με την εξέλιξη του Klout για τους ίδιους χρήστες. Χρησιμοποιήσαμε το Klout για τη σύγκριση μιας και θεωρείται το ενδεδειγμένο εργαλείο υπολογισμού επιρροής για τα κοινωνικά δίκτυα. Πιο συγκεκριμένα το Klout

χρησιμοποιεί εργαλεία υπολογισμού της επιρροής απο όλα τα μεγάλα κοινωνικά δίκτυα όπως το Twitter, Facebook και LinkedIn για να κατατάξει τους χρήστες τους σύμφωνα με την κοινωνική επιρροή τους μέσω της «Klout Βαθμολόγηση» (Klout Score), που

αναπαρίσταται με μια αριθμητική ένδειξη απο το 1 μέχρι το 100. Το Klout, ενδεικτικά, μετράει την επιρροή απο το Twitter χρησιμοποιώντας τον αριθμό των ακόλουθων (followers) που έχει κάθε χρήστης, τον αριθμό των retweets, των ανενεργών λογαριασμών που ακολουθούν τον χρήστη καθώς και πόση δύναμη επιρροής έχουν οι άνθρωποι που αναπαράγουν τα tweets του χρήστη. Για να συγκρίνουμε το Klout με το TIC απομονώνουμε το Klout Score που έχει προκύψει αποκλειστικά για το Twitter του χρήστη. Στο πρώτο μας πείραμα διαλέγουμε πέντε χρήστες απο το Twitter ($U_1, \dots, U_5$), οι δυο απο τους οποίους είναι δημοφιλή πρόσωπα ενώ οι άλλοι τρεις είναι τυχαίοι χρήστες. Επιπλέον ρυθμίζουμε όλα τα βάρη με τη τιμή 1 ώστε να συνεισφέρουν όλα το ίδιο και μετράμε την επιρροή τους σε τέσσερις χρονικές στιγμές ($t_1, \dots, t_4$) με διαφορά μιας εβδομάδας. Το ίδιο πείραμα πραγματοποιούμε και για το Klout. Το αποτέλεσμα αναπαρίσταται στα γραφήματα που ακολουθούν.



Σχήμα 3: Βαθμολογίες Επιρροής του TIC για το Δείγμα Χρηστών



Σχήμα 4: Βαθμολογίες Επιρροής του Klout για το Δείγμα Χρηστών

Με μια πρώτη ματιά, παρατηρούμε στα παραπάνω γραφήματα ότι οι τιμές των δυο συστημάτων για το ίδιο δείγμα χρηστών είναι αρκετά διαφορετικές καθώς το Klout Score παίρνει τιμές απο ανάμεσα στο 1 και στο 100 ενώ το TIC δεν έχει κάποιο συγκεκριμένο όριο τιμών. Επιπλέον οι βαθμολογίες του Klout δεν έχουν μεγάλες διακυμάνσεις σε βραχυπρόθεσμα διαστήματα (π.χ. από την μια μέρα στην επόμενη) ενώ το TIC ενδέχεται να έχουν καθώς υπολογίζει την επιρροή παίρνοντας υπόψιν τα τελευταία είκοσι tweets του χρήστη. Παρά τις διαφορές αυτές των δυο συστημάτων μπορούμε να παρατηρήσουμε όμως ότι τα αποτελέσματα του TIC, τα οποία είναι βραχυπρόθεσμα, δείχνουν την τάση της

επιρροής που θα έχει ο κάθε χρήστης, όπως αναπαρίσταται στο γράφημα του Klout, μακροπρόθεσμα. Το γεγονός αυτό δείχνει πως οι χρήστες οι οποίοι συστηματικά επηρεάζουν το κοινό τους έχουν τα ίδια επίπεδα επιρροής τόσο βραχυπρόθεσμα όσο και μακροπρόθεσμα.

6.2 Επιδράσεις και Διαφορετικές Χρήσεις των Βαρών (Weights)

Όπως αναλύσαμε στη προηγούμενη ενότητα το TIC στηρίζεται σε πέντε μεταβλητές. Οι μεταβλητές αυτές συνεισφέρουν στον υπολογισμό της επιρροής από πέντε βάρη αντίστοιχα. Όλα τα βάρη παίρνουν πραγματικές τιμές στο πεδίο $[0,1]$. Σε αυτή την υποενότητα πειραματιζόμαστε με αυτές τις παραμέτρους και δείχνουμε τι αλλαγές θα επιφέρουν στο σύστημα υπολογισμού της επιρροής. Αρχικά βλέπουμε πως w_l απλώς μεγαλώνει ή μικραίνει το συνολικό score του tweet, δηλαδή δεν αναμένεται ότι θα αλλάξει την κατάταξη των χρηστών με τη μεγαλύτερη επιρροή. Η παρατήρηση αυτή επιβεβαιώνεται από πειράματα που κάναμε στα οποία διατηρούσαμε σταθερά τα άλλα τρία βάρη και αλλάζαμε μόνο τη τιμή του w_l . Για τα υπόλοιπα βάρη διατηρούμε σταθερά δυο ή τρία από αυτά και παρατηρούμε τις αλλαγές στα εναπομείναντα. Κρατώντας σταθερό το w_l και το w_i και αλλάζοντας τα w_θ και w_θ' από το 0.0 έως το 1.0 αυξάνοντας κάθε φορά κατά 0.1 παρατηρούμε ότι τα αποτελέσματα του μοντέλου σταθεροποιούνται για $w_\theta \leq 0.5$ $w_\theta' \leq 0.7$, δηλαδή δεν υπάρχουν μεγάλες διακυμάνσεις και διαφορές στις κατατάξεις των χρηστών με επιρροή. Αλλάζοντας το w_i , παρατηρούμε πως το μοντέλο δείχνει μια συνέπεια στα αποτελέσματα του για $w_i \geq 0.8$ για το w_l σταθεροποιείται όταν $w_l \geq 0.8$. Συνοψίζοντας κατά τη διάρκεια των πειραμάτων διαπιστώσαμε πως τα αποτελέσματα του TIC είναι πιο ακριβή όταν $w_i \geq 0.9$, $w_\theta \leq 0.4$ $w_\theta' \leq 0.7$ και $w_l \geq 0.9$. Τέλος βλέπουμε πως αλλάζοντας τα πέντε αυτά βάρη προκύπτουν διαφορετικά αποτελέσματα. Για παράδειγμα θέτοντας το w_i και w_θ σε 0, παίρνουμε βαθμολογίες βασιζόμενες σε tweets τα οποία είναι μακροσκελή και έχουν έναν μεγάλο αριθμό από favorites. Ένα άλλο σύστημα παραμέτρων θα έδινε χαμηλότερη βαρύτητα στα w_θ και w_θ' τα οποία αυξάνουν την προτεραιότητα των tweets τα οποία περιλαμβάνουν πολλά outlinks. Δηλαδή οι παράμετροι αυτοί μπορούν να σταθμιστούν με πολλούς τρόπους ανάλογα τον λόγο για τον οποίο χρησιμοποιείται.

7. ΣΥΜΠΕΡΑΣΜΑΤΑ

Τα συστήματα φήμης αποτελούν καθοριστικά μέρη των κοινωνικών δικτύων και γενικότερα των διαδικτυακών εφαρμογών οι οποίες προσφέρουν προϊόντα προς πώληση ή υπηρεσίες καθώς δημιουργούν σχέσεις εμπιστοσύνης. Στην εργασία αυτή αναλύσαμε τις διάφορες κατηγορίες των συστημάτων φήμης και συστάσεων αναλύοντας τα βασικά χαρακτηριστικά τους και τις ποικίλες προσεγγίσεις και μεθοδολογίες που μπορεί να ακολουθήσει κάποιος για να το δημιουργήσει αποτελεσματικά. Συγκεκριμένα είδαμε τα συστήματα φήμης που χρησιμοποιούν collaborative filtering, τα οποία χρησιμοποιούν έμμεσα στοιχεία και πληροφορίες προκειμένου να υπολογιστεί η φήμη του αντικειμένου, χρήστη ή υπηρεσίας με τη χρήση τεχνικών σημασιολογικού κοινωνικού συνεργατικού φιλτραρίσματος, συστημάτων CF βασισμένα στη μνήμη, στο μοντέλο ή ακόμα και συνδυασμών των δυο. Ακόμα παραθέσαμε τις βασικές αρχές ενός υβριδικού συστήματος

CF που χρησιμοποιεί στη πρώτη φάση υπολογισμού της φήμης έναν αλγόριθμο CBS, ο οποίος όπως αναλύσαμε αξιοποιεί τα δεδομένα, τις πληροφορίες και το ιστορικό του χρήστη αποκλειστικά και είναι πιο αποδοτικός από τις υλοποιήσεις βασιζόμενες μόνο στο CF. Έπειτα εστιάσαμε στα κοινωνικά δίκτυα όπως το Twitter επεξηγώντας τις βασικές αρχές και παράγοντες για τον σχεδιασμό ενός συστήματος φήμης για το Twitter. Επιπρόσθετα παρουσιάσαμε την δική μας υλοποίηση υπολογισμού της δύναμης της επιρροής των χρηστών ή των tweets συγκεκριμένων θεματικών ενοτήτων βασιζόμενοι στα στοιχεία που αποτελούν ένα tweet. Τέλος παραθέτουμε τα πειράματα που διεξείγαμε για την υλοποίηση μας συγκρίνοντας το με το πιο αναγνωρισμένο εργαλείο υπολογισμού της επιρροής χρηστών σε κοινωνικά δίκτυα, Klout, όπου να εξετάσαμε την αποδοτικότητα και χρησιμότητα του καταλήγοντας πως αν και πρόκειται για δυο διαφορετικά συστήματα υπολογισμού επιρροής παρουσιάζουν ορισμένα σημεία ομοιότητας.

ΠΙΝΑΚΑΣ ΟΡΟΛΟΓΙΑΣ

Ξενόγλωσσος όρος	Ελληνικός Όρος
Reputation systems	Συστήματα φήμης
Commitment	Δέσμευση
Recommendations	Συστάσεις
Belief	Πεποίθηση
Loyalty	Πίστη
Content poisoning	Μόλυνση Περιεχομένου
Collaborative filtering	Συνεργατικό φιλτράρισμα
Content based recommendation systems	Συστήματα συστάσεων βασιζόμενα στο περιεχόμενο
E-commerce	Ηλεκτρονικό Εμπόριο
Peer to peer networks	Δίκτυα ομότιμων χρηστών
Server	Εξυπηρετητής
Servant	Υπηρέτης
Expertise	Εξειδίκευση
Access Control List(s)	Λίστες Ελέγχου Πρόσβασης
Client	Πελάτης
Personality Diagnosis	Ανάλυση Προσωπικότητας
Overspecialization	Υπερεξειδίκευση
New user problem	Πρόβλημα νέου χρήστη
Clusters	Συστάδες
Hierarchical	Ιεραρχικές
Density based	Βασιζόμενες στη πυκνότητα
Partition	Διαχωρισμός
Decision tree	Δέντρο απόφασης
Training Phase	Φάση εκμάθησης
Rating	Βαθμολογία
Constraint	Περιορισμός
Ranks	Κατατάξεις
Item	Αντικείμενο
Value	Αξία
Similarity	Ομοιότητα
User	Χρήστης
Influence	Επιρροή
Computer	Υπολογιστής
Preference	Προτίμηση
Recognition	Αναγνώριση
Eloquence	Λόγιος
Novelty	Καινοτομία
Multimedia	Πολυμέσα
Widgets	Πρόσθετες Εφαρμογές
Timeline	Χρονολόγιο
Document Object Model	Αντικείμενο Μοντέλο Εγγράφου

ΣΥΝΤΜΗΣΕΙΣ – ΑΡΚΤΙΚΟΛΕΞΑ – ΑΚΡΩΝΥΜΙΑ

P2P	Peer to Peer
CF	Collaborative Filtering
CBS	Content Based (Reputation) Systems
ΣΦ	Συνεργατικό φιλτράρισμα
ACL	Access Control List(s)
PD	Personality Diagnosis
TIC	Twitter Influence Computer
DOM	Document Object Manipulation

ΑΝΑΦΟΡΕΣ

- [1] Marsh, S. Formalising Trust as a Computational Concept. Ph.D. thesis, University of Stirling, Department of Mathematics and Computer Science, 1944.
- [2] Castelfranchi, C., & Falcone, R. Principles of trust for mas: Cognitive anatomy, social importance, and quantification. 3rd International Conference on Multi Agent Systems, 1998.
- [3] Castelfranchi, C., & Falcone, R. Social trust: A cognitive approach. Trust and Deception in Virtual Societies. Cristiano Castelfranchi AND Yao-Hua Tan, Eds., Kluwer Academic Publishers, 2002.
- [4] D. Harrison McKnight and Norman L. Chervany. Trust and Distrust Definitions: One Bite at a Time.
- [5] Deutsch, M. Cooperation and trust. some theoretical notes. Jones, M.R. ED. Nebraska Symposium on Motivation. Nebraska University Press, 1962.
- [6] Sztompka, P. Trust: A Sociological Theory. Cambridge University Press, 1999.
- [7] S. Tadelis. Firm Reputation with Hidden Information. Economic Theory, 2003.
- [8] Kruk, S.R., Decker, S., Zieborak, L.: JeromeDL - Reconnecting Digital Libraries and the Semantic Web. In: DEXA, 2005.
- [9] Maltz, D., Ehrlich, K.: Pointing the way: active collaborative filtering. In: Proceedings of the Conference on Computer-Human Interaction, 1995.
- [10] Sebastian Ryszard Kruk, Stefan Decker: Semantic Social Collaborative Filtering with FOAFRealm Digital Enterprise Research Institute, Galway, Ireland
- [11] Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl, "Grouplens: an open architecture for collaborative filtering of netnews," in Proceedings of the ACM Conference on Computer Supported Cooperative Work, pp. 175–186, New York, NY, USA, 1994.
- [12] B. M. Sarwar, G. Karypis, J. A. Konstan, and J. Riedl, "Item-based collaborative filtering recommendation algorithms," in Proceedings of the 10th International Conference on World Wide Web (WWW '01), pp. 285–295, May 2001.
- [13] Xiaoyuan Su and Taghi M. Khoshgoftaars: A Survey of Collaborative Filtering Techniques, Department of Computer Science and Engineering, Florida Atlantic University, USA.
- [14] G. Salton and M. McGill, Introduction to Modern Information Retrieval, McGraw-Hill, New York, NY, USA, 1983.
- [15] J. L. Herlocker, J. A. Konstan, A. Borchers, and J. Riedl, "An algorithmic framework for performing collaborative filtering," in Proceedings of the Conference on Research and Development in Information Retrieval (SIGIR '99), pp. 230–237, 1999.
- [16] R. Greinemr, X. Su, B. Shen, and W. Zhou, "Structural extension to logistic regression: discriminative parameter learning of belief net classifiers," Machine Learning, vol. 59, no. 3, pp. 297–322, 2005.
- [17] B. Shen, X. Su, R. Greiner, P. Musilek, and C. Cheng, "Discriminative parameter learning of general Bayesian network classifiers," in Proceedings of the 15th IEEE International Conference on Tools with Artificial Intelligence, pp. 296–305, Sacramento, Calif, USA, November 2003.
- [18] J. Han and M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, San Francisco, Calif, USA, 2001.
- [19] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in Proceedings of the 5th Symposium on Math, Statistics, and Probability, pp. 281–297, Berkeley, Calif, USA, 1967.
- [20] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in Proceedings of the International Conference on Knowledge Discovery and Data Mining (KDD '96), 1996.
- [21] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander, "OPTICS: ordering points to identify the clustering structure," in Proceedings of ACM SIGMOD Conference, pp. 49–60, June 1999.
- [22] B. M. Sarwar, G. Karypis, J. A. Konstan, and J. Riedl, "Recommender systems for large-scale E-commerce: scalable neighborhood formation using clustering," in Proceedings of the 5th International Conference on Computer and Information Technology (ICCIT '02), December 2002.
- [23] M. O'Connor and J. Herlocker, "Clustering items for collaborative filtering," in Proceedings of the ACM SIGIR Workshop on Recommender Systems (SIGIR '99), 1999.
- [24] S. Vucetic and Z. Obradovic, "Collaborative filtering using a regression-based approach," Knowledge and Information Systems, vol. 7, no. 1, pp. 1–22, 2005.
- [25] Zhu, R. Greiner, and G. Haubl, "Learning a model of a web user's interests," in Proceedings of the 9th

- International Conference on User Modeling (UM '03), vol. 2702, pp. 65–75, Johnstown, Pa, USA, June 2003.
- [26] K. Pearson, “On lines and planes of closest fit to systems of points in space,” *Philosophical Magazine*, vol. 2, pp. 559–572, 1901.
- [27] M. K. Condiff, D. D. Lewis, D. Madigan, and C. Posse, “Bayesian mixed-effects models for recommender systems,” in *Proceedings of ACM SIGIR Workshop of Recommender Systems: Algorithm and Evaluation*, 1999.
- [28] P. Melville, R. J. Mooney, and R. Nagarajan, “Content- boosted collaborative filtering for improved recommenda- tions,” in *Proceedings of the 18th National Conference on Artificial Intelligence (AAAI '02)*, pp. 187–192, Edmonton, Canada, 2002.
- [29] A. Ansari, S. Essegaier, and R. Kohli, “Internet recommendation systems,” *Journal of Marketing Research*, vol. 37, no. 3, pp. 363–375, 2000.
- [30] K. Yu, A. Schwaighofer, V. Tresp, X. Xu, and H.-P. Kriegel, “Probabilistic memory-based collaborative filtering,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 1, pp. 56–69, 2004.
- [31] D. M. Pennock, E. Horvitz, S. Lawrence, and C. L. Giles, “Collaborative filtering by personality diagnosis: a hybrid memory- and model-based approach,” in *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence (UAI '00)*, pp. 473–480, 2000.
- [32] Sabbir Ahmed and C.I. Ezeife: Discovering Influential Nodes from Trust Network. University of Windsor, School of Computer Science, Windsor, Ontario.
- [33] Kempe, D., Kleinberg, J., and Tardos, E. 2003. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. KDD '03*. ACM, New York, NY, USA, 137–146.
- [34] Nemhauser, G. L., Wolsey, L. A. and Fisher, M. L. An analysis of approximations for maximizing submodular set functions, *Mathematical Programming* 14, 1978, 265–294.
- [35] Goyal, A., Bonchi, F., and Lakshmanan, L. V. 2010. Learning influence probabilities in social networks. In *Proceedings of the third ACM international conference on Web search and data mining. WSDM '10*. ACM, New York, NY, USA, 241–250.
- [36] Kimura, M. and Saito, K. 2006. Tractable models for information diffusion in social networks. In *Knowledge Discovery in Databases: PKDD 2006. Lecture Notes in Computer Science*, vol. 4213. Springer Berlin / Heidelberg, 259–271. 10.1007/11871637.
- [37] Ed Keller and Jon Berry. One American in ten tells the other nine how to vote, where to eat and, what to buy. They are The Influentials. The Free Press, 2003.
- [38] P. Kolari, T. Finin, and A. Joshi. SVMs for the blogosphere: Blog identification and splog detection. In *AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs*, 2006.
- [39] Nitin Agarwal and Philip S. Yu: Identifying the Influential Bloggers in a Community. *WSDM '08 Proceedings of the International Conference on Web Search and Data Mining*, 2008, pages 207-218
- [40] YouTube, www.youtube.com, πρόσβαση στις 10/5/2015
- [41] Facebook, www.facebook.com, πρόσβαση στις 10/5/2015
- [42] Twitter, www.twitter.com, πρόσβαση στις 10/5/2015
- [43] Twitter Widgets, www.dev.twitter.com/web/embedded-timelines, πρόσβαση στις 10/5/2015
- [44] Klout , www.klout.com, πρόσβαση στις 10/5/2015