ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ

ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

«ΠΛΗΡΟΦΟΡΙΚΗ ΚΑΙ ΔΙΟΙΚΗΣΗ»

ΤΜΗΜΑΤΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΟΙΚΟΝΟΜΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

Αλγόριθμοι Συστάσεων στο Twitter

Διπλωματική Εργασία της

Γεωργίας Παπαγιάννη (ΑΕΜ: 505)

Εξεταστική Επιτροπή

Επιβλέπων: Συμεωνίδης Παναγιώτης

Μέλη: Βακάλη Αθηνά

Μπουτσούκη Χριστίνα

ΘΕΣΣΑΛΟΝΙΚΗ
Φεβρουάριος 2016

Πρόλογος

Στα πλαίσια του Διατμηματικού Προγράμματος Μεταπτυχιακών Σπουδών «Πληροφορική και Διοίκηση» του Αριστοτελείου Πανεπιστημίου Θεσσαλονίκης, πραγματοποιήθηκε η εκπόνηση αυτής της Διπλωματικής Εργασίας. Αντικειμενικός σκοπός της παρούσας Διπλωματικής Εργασίας είναι να εκθέσει, να περιγράψει και να αναλύσει τις διάφορες τεχνικές και αλγορίθμους που υπάρχουν στη βιβλιογραφία για την πραγματοποίηση συστάσεων στο κοινωνικό δίκτυο Twitter, έπειτα από το χτίσιμο ενός απαραίτητου θεωρητικού υποβάθρου που σχετίζεται με τα συστήματα συστάσεων, τις διάφορες κατηγορίες τους, καθώς και τις διαφορετικές τεχνικές που χρησιμοποιούνται για την υλοποίηση συστάσεων με βάση το κάθε είδος συστήματος συστάσεων.

Ωστόσο, δε θα μπορούσα να παραγνωρίσω ότι η επιτυχής εκπόνηση της παρούσας Διπλωματικής Εργασίας δεν οφείλεται καθαρά και μόνο σε δική μου προσπάθεια. Για αυτό το λόγο, θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου κ. Συμεωνίδη Παναγιώτη για όλη την καθοδήγηση και τις πολύτιμες συμβουλές του, το διδακτικό προσωπικό του Μεταπτυχιακού για τις χρήσιμες γνώσεις που μου χάρισε, καθώς επίσης και την οικογένειά μου για την αμέριστη συμπαράστασή τους σε όλη την προσπάθειά μου.

Παπαγιάννη Γεωργία

Φεβρουάριος 2016

Περιεχόμενα

Περιεχόμενα Εικόνων

1 Εισαγωγή

Η ανάπτυξη των σύγχρονων τεχνολογιών με τον εκσυγχρονισμό των ασύρματων δικτύων, τη δημιουργία νέων και πιο πολύπλοκων εφαρμογών, την παγκοσμιοποίηση των αγορών αλλά και τη γενικότερη ανάπτυξη της έρευνας πάνω σε θέματα πληροφοριακών συστημάτων έχει ως αποτέλεσμα την ολοένα και πιο ενεργή συμμετοχή των χρηστών στο διαδίκτυο καθώς και στο διαδικτυακό κοινωνικό γίγνεσθαι.

Η πρόσβαση κάθε χρήστη, οποιασδήποτε ηλικίας, μορφωτικού επιπέδου και κοινωνικής τάξης, στο διαδίκτυο έχει γίνει πιο εύκολη από ποτέ. Παράλληλα η προσβασιμότητα από δεκάδες διαφορετικά μέσα, οποιαδήποτε ώρα και από οποιοδήποτε σημείο του πλανήτη είχε ως αναπόφευκτο αποτέλεσμα τη ραγδαία ανάπτυξη του Web 2.0. Η έμφυτη ανάγκη του ανθρώπου να επικοινωνήσει, να αλληλεπιδράσει και να συμμετάσχει στην κοινωνική ζωή μαζί με άλλα άτομα είναι αδιαμφισβήτητη. Στη σύγχρονη εποχή λοιπόν, έχοντας ως δεδομένη αυτή την έμφυτη ανάγκη του ανθρώπου σε συνδυασμό με την ταχύτατη ανάπτυξη του διαδικτύου, η έλευση των κοινωνικών δικτύων και κοινωνικών μέσων στο διαδίκτυο δε θα πρέπει να μας προκαλεί εντύπωση.

Σήμερα, εκατομμύρια χρήστες καθημερινά αλληλεπιδρούν με άλλους σε κάθε άκρη της γης οποιαδήποτε ώρα της ημέρας μέσα από την τεράστια γκάμα μέσων που διαθέτουν. Τα κοινωνικά δίκτυα με τη μορφή που τα αναγνωρίζουμε σήμερα άρχισαν να κάνουν τα πρώτα τους βήματα την πρώτη δεκαετία της νέας χιλιετίας, ενώ πολλά από αυτά γεννήθηκαν από μια απλή ιδέα περιοριζόμενη στο campus ενός πανεπιστημίου ή ακόμα σε πιο στενά όρια όπως μια ομάδα μέσα σε ένα πανεπιστημιακό ίδρυμα. Έκτοτε, τα κοινωνικά δίκτυα αναπτύχθηκαν και έλαβαν μορφή πέρα από τη μορφή με την οποία άρχισαν να λειτουργούν για πρώτη φορά.

Τα κοινωνικά δίκτυα σήμερα συναντώνται σε όλες τις πτυχές της καθημερινής κοινωνικής ζωής του ατόμου. Επόμενο είναι λοιπόν, οι χρήστες αυτών των κοινωνικών δικτύων να είναι εκατομμύρια. Μέσα σε αυτόν το χαοτικό χάρτη χρηστών και κοινωνικών δικτύων δημιουργείται έντονα το πρόβλημα της υπερπληροφόρησης κάτι

που κάνει τον κάθε χρήστη να βρίσκεται μπροστά σε κάτι που δεν μπορεί να αντιμετωπίσει μεμονωμένα και φυσικά χωρίς τη χρήση των νέων τεχνολογιών.

Το πρόβλημα του βομβαρδισμού πληροφοριών ο σύγχρονος άνθρωπος το συναντά συνέχεια στην ενασχόλησή του με το διαδίκτυο, είτε το χρησιμοποιεί για να έρθει σε επαφή με άλλους χρήστες και να αλληλεπιδράσει, είτε για να πραγματοποιήσει τις on line αγορές του, είτε για να ακούσει μουσική, είτε να αναζητήσει ποια ταινία θα παρακολουθήσει, ακόμα και με τι θα μπορούσε να συνοδεύσει την ταινία που θα δει. Με άλλα λόγια, δημιουργείται το πρόβλημα για τον χρήστη του τι να διαλέξει ανάμεσα σε μια πληθώρα επιλογών που έχει στη διάθεσή του. Βέβαια, το πρόβλημα δεν έγκειται μόνο στο τι θα διαλέξει ο χρήστης, αλλά και στο πως αυτό που θα διαλέξει θα τον ικανοποιεί περισσότερο από κάτι άλλο. Συγκεντρωτικά λοιπόν, ο χρήστης αντιμετωπίζει δυο βασικά και πρωταρχικά προβλήματα:

1. Τι θα επιλέξει μέσα από μια τεράστια γκάμα δυνατών εναλλακτικών επιλογών
2. Πως αυτό που θα επιλέξει θα συμβάλει στη μεγιστοποίηση της ικανοποίησής του και θα ταιριάζει καλύτερα στις προτιμήσεις του

Την απάντηση στα παραπάνω δυο βασικά προβλήματα που αντιμετωπίζει ο χρήστης σήμερα έρχονται να δώσουν τα συστήματα συστάσεων (Recommendation Systems). Τα συστήματα αυτά έρχονται ύστερα από πολλές μελέτες και επίπονη εργασία πάνω στον τομέα των πληροφοριακών συστημάτων έτσι ώστε να προσφέρουν στο χρήστη ένα τρόπο διαχείρισης της πληθώρας των δεδομένων που έχουν στη διάθεσή τους για την δημιουργία επιλογών.

Τα συστήματα συστάσεων προσπαθούν μέσα από διάφορους αλγορίθμους και τεχνολογίες που έχουν αναπτυχθεί να προτείνουν σε έναν χρήστη, ανάλογα με το που βρίσκεται, εκείνο το προϊόν να αγοράσει, εκείνο το βιβλίο να διαβάσει, εκείνη την ταινία να παρακολουθήσει, εκείνους τους χρήστες να κάνει φίλους κτλ. Ωστόσο, το σύστημα συστάσεων δεν κάνει τυχαία την επιλογή της πρότασης. Πίσω από το τελικό προϊόν της σύστασης βρίσκεται μια πολύπλοκη και ύστερα από πολλές μελέτες διαδικασία που θα καταλήξει στη σύσταση ενός συγκεκριμένου αντικειμένου προς ένα χρήστη στόχο που θα τον ικανοποιεί, δεδομένου ότι τα συστήματα συστάσεων λαμβάνουν υπόψη και τις προτιμήσεις του χρήστη.

Τα συστήματα συστάσεων τυγχάνουν εφαρμογής σε πολλές και διαφορετικές περιπτώσεις που μπορούμε να συναντήσουμε στο διαδίκτυο. Όπως ήδη αναφέραμε τα

συστήματα συστάσεων μπορούν να χρησιμοποιηθούν σε κοινωνικά δίκτυα, σε διαδικτυακούς ιστότοπους αγορών, σε διάφορες εφαρμογές για μουσική, ταινίες, βιβλία και πολλά άλλα. Στην παρούσα διπλωματική εργασία θα επιχειρήσουμε να περιγράψουμε πως τα συστήματα συστάσεων χρησιμοποιούνται στο διαδικτυακό τόπο του κοινωνικού δικτύου Twitter προκειμένου να δημιουργήσουν συστάσεις για τους χρήστες του.

Με την εξάπλωση των κοινωνικών δικτύων, επακόλουθο ήταν να αναπτυχθούν και συστήματα που θα είχαν ως στόχο τη διευκόλυνση των χρηστών στην περιήγησή τους μέσα σε ένα κοινωνικό δίκτυο, αφού τα συστήματα και οι αλγόριθμοι συστάσεων προτείνουν στους χρήστες ενός κοινωνικού δικτύου, κατά κύριο λόγο με ποιους άλλους χρήστες να συνδεθούν, δηλαδή να γίνουν φίλοι ή να ακολουθήσουν και να αλληλεπιδράσουν με αυτούς.

Στην παρούσα διπλωματική εργασία αρχικά ξεκινάμε με ένα εισαγωγικό κεφάλαιο το οποίο προσπαθεί να περιγράψει τα κοινωνικά δίκτυα. Πιο συγκεκριμένα στο κεφάλαιο 2 της παρούσας διπλωματικής δίνουμε τον ορισμό του κοινωνικού δικτύου και των κοινωνικών μέσων προχωρώντας στον ευδιάκριτο διαχωρισμό μεταξύ αυτών των δύο εννοιών. Έπειτα, αναφερόμαστε και περιγράφουμε τα βασικότερα και πιο δημοφιλή κοινωνικά δίκτυα που συναντούμε σήμερα, όπως το Facebook, το Twitter, το LinkedIn κα, ενώ παράλληλα περιγράφουμε και ορισμένες από τις λειτουργίες και δυνατότητες που μπορούμε να αξιοποιήσουμε μέσω ενός κοινωνικού δικτύου.

Στη συνέχεια, στο κεφάλαιο 3 της παρούσας διπλωματικής εργασίας κάνουμε μια εκτενή αναφορά, μέσα από την αναλυτική περιγραφή των διάφορων ειδών συστημάτων συστάσεων που συναντούμε στη βιβλιογραφία, κάνοντας παράλληλα μια σύντομη αξιολόγηση του κάθε συστήματος μέσω της παράθεσης των πλεονεκτημάτων και μειονεκτημάτων που συναντάμε σε κάθε κατηγορία συστημάτων συστάσεων.

Έπειτα, στο επόμενο κεφάλαιο 4 επιχειρείται μια διεξοδική παρουσίαση των μεθόδων και των διάφορων αλγορίθμων που συναντάμε στο Twitter για την παραγωγή συστάσεων. Στην ανάλυσή μας έχουμε υιοθετήσει ένα διαχωρισμό των συστάσεων που συναντάμε στο Twitter ανάλογα με τη λειτουργία του, δηλαδή συστάσεις για followers, για followees, για tweets κτλ.

Στο κεφάλαιο 5 επιχειρείται να περιγραφεί μια ανάλυση που σχετίζεται με την σύγκριση διαφορετικών αλγορίθμων και κατ' επέκταση στην αξιολόγησή τους αναφορικά με την εξαγωγή συστάσεων. Συγκεκριμένα συγκρίνουμε topology – based,

content – based και collaborative filtering αλγορίθμους που χρησιμοποιούνται για την εξαγωγή συστάσεων.

Στο κεφάλαιο 6 της διπλωματικής εργασίας παραθέτουμε αναλυτικά τα συμπεράσματα στα οποία καταλήγουμε έπειτα από τη γενικότερη θεώρηση των παραπάνω κεφαλαίων και της ανάλυσης που έχει γίνει καθώς επίσης, δεδομένης και της σημαντικότητας του συγκεκριμένου πεδίου έρευνας, προτείνουμε περαιτέρω προτάσεις για ερεύνα και ανάλυση.

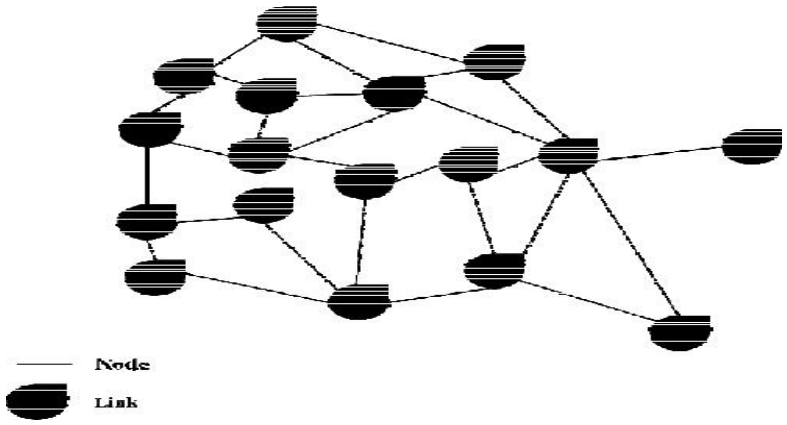
Τέλος, παρατίθεται η βιβλιογραφία πάνω στην οποία στηρίχθηκε η παρούσα διπλωματική εργασία.

2 Κοινωνικά Δίκτυα

Στην ενότητα αυτή θα παρουσιάσουμε τον ορισμό των κοινωνικών μέσων, καθώς και του κοινωνικού δικτύου και θα περιγράψουμε εν συντομία τα δημοφιλέστερα κοινωνικά δίκτυα που υπάρχουν σήμερα. Έπειτα θα διεισδύσουμε σε θέματα σχετικά με τα κοινωνικά δίκτυα και θα περιγράψουμε ορισμένες από τις λειτουργίες που μπορεί κάποιος χρήστης να συναντήσει σε ένα κοινωνικό δίκτυο, καθώς επίσης θα περιγράψουμε πως τα συστήματα συστάσεων (Recommendation Systems) συνδέονται με τα κοινωνικά δίκτυα.

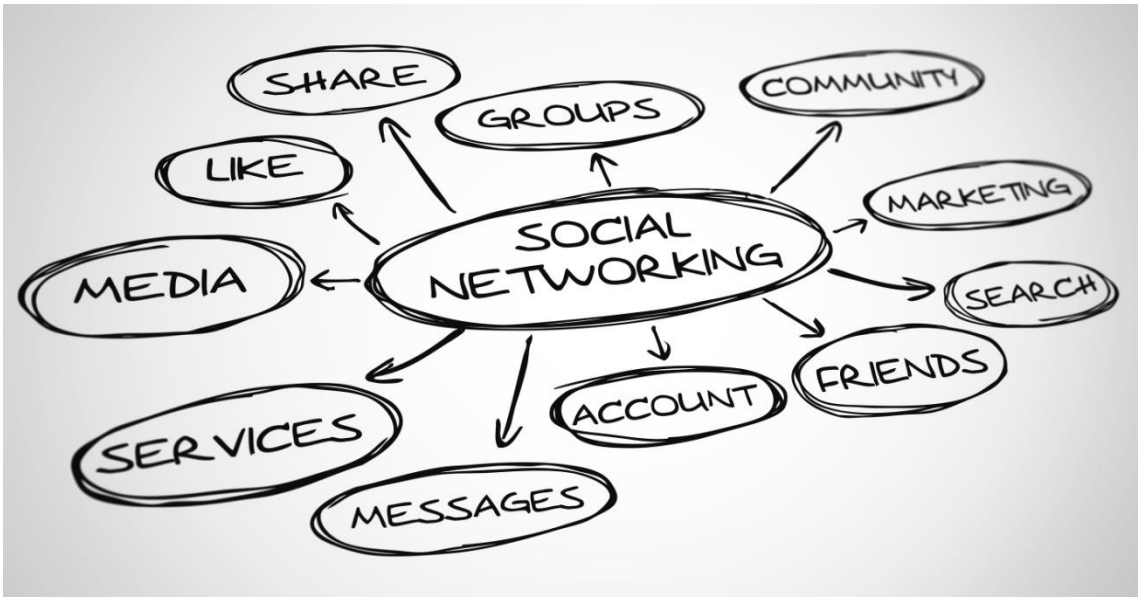
2.1 Γενικά για τα κοινωνικά δίκτυα

Ένα κοινωνικό δίκτυο (social network) θα μπορούσε να ορισθεί σαν ένας ιστός που περιλαμβάνει συνδέσεις κόμβων, οι οποίοι κόμβοι τείνουν να αντιπροσωπεύουν άτομα, καταστάσεις και οργανισμούς. Οι συνδέσεις των κόμβων ουσιαστικά αναπαριστούν τις σχέσεις οι οποίες συνδέουν τις κοινωνικές μονάδες του δικτύου. Οι σχέσεις που διέπουν τα συστατικά στοιχεία του κοινωνικού δικτύου μπορεί να είναι με τη μορφή φιλίας, κοινών ενδιαφερόντων, παροχή πληροφοριών, εμπορικών συναλλαγών και γενικότερα οτιδήποτε μπορεί να συνδέει δύο ή περισσότερους ανθρώπους. Θα μπορούσαμε λοιπόν να ορίσουμε τα κοινωνικά δίκτυα σαν *τα πολυδιάστατα συστήματα επικοινωνίας και διαμόρφωσης της ανθρώπινης πρακτικής και της κοινωνικής ταυτότητας* [1,2].



Εικόνα 1: Αναπαράσταση κοινωνικού δικτύου με κόμβους και συνδέσεις [77]

Πολύ συχνά τείνουν να ταυτίζονται και να εξομοιώνονται δύο κατά τα άλλα διαφορετικοί όροι που σχετίζονται με την κοινωνική δικτύωση. Οι όροι αυτοί είναι τα κοινωνικά μέσα (social media) και κοινωνικά δίκτυα (social networks). Ο όρος κοινωνικά μέσα αναφέρεται στα μέσα (εργαλεία) διάδοσης της πληροφορίας, των δεδομένων και της επικοινωνίας στο κοινό, ενώ ο όρος κοινωνικό δίκτυο αναφέρεται στη δημιουργία και την αξιοποίηση των κοινοτήτων για τη διασύνδεση ανθρώπων με κοινά ενδιαφέροντα. Επομένως, ο όρος κοινωνικά μέσα αναφέρεται στα εργαλεία και μέσα ενημέρωσης κοινωνικής δικτύωσης, ενώ ο όρος κοινωνικό δίκτυο στη διαδικασία της κοινωνικής δικτύωσης [78,79].



Εικόνα 2: Χαρακτηριστικά ενός κοινωνικού δικτύου [78]

Σχετικά με τα κοινωνικά δίκτυα, θα μπορούσαμε να διακρίνουμε ορισμένα χαρακτηριστικά όπως ότι υποστηρίζουν ποικιλία μορφών περιεχομένου, όπως κείμενο, βίντεο, φωτογραφίες, ήχο κτλ., ενώ πολλά από αυτά κάνουν χρήση περισσοτέρων του ενός από αυτές τις επιλογές ως προς το περιεχόμενο. Επιπλέον, τα κοινωνικά δίκτυα επιτρέπουν αλληλεπιδράσεις μεταξύ των χρηστών μέσω διαμοιρασμού email και ποικίλων μηνυμάτων, ενώ χαρακτηρίζονται από διαφορετικά επίπεδα εμπλοκής του χρήστη οι οποίοι μπορούν να δημιουργήσουν, να σχολιάσουν ή να παρακολουθήσουν διάφορα αντικείμενα σε ένα δίκτυο.

Ένα από τα σημαντικότερα ίσως χαρακτηριστικά ενός κοινωνικού δικτύου, ιδιαίτερα στη σύγχρονη εποχή, είναι ότι συμβάλουν στην απλοποίηση, στη βελτίωση, στην ταχύτητα καθώς και στο εύρος διάδοσης των πληροφοριών, προσφέροντας ταυτόχρονα

ενός-προς-ένα, ενός-προς-πολλούς και πολλών-προς-πολλούς επικοινωνία, επιτρέποντας την επικοινωνία αυτή να πραγματοποιείται είτε σε πραγματικό χρόνο ή ασύγχρονα. Τέλος, τα κοινωνικά δίκτυα είναι ανεξάρτητα της συσκευής με την έννοια ότι ο χρήστης μπορεί να χρησιμοποιήσει για τη συμμετοχή του σε ένα κοινωνικό δίκτυο, είτε έναν υπολογιστή είτε κινητές συσκευές όπως tablets και smartphones τα οποία είναι ευρύτατα διαδεδομένα στη σημερινή αγορά.

Τα κοινωνικά μέσα, όπως τα ορίσαμε και τα διαφοροποιήσαμε από τον όρο του κοινωνικού δικτύου παραπάνω, θα μπορούσαν να διακριθούν σε ορισμένες κατηγορίες ανάλογα με το σκοπό που επιδιώκουν ή τις λειτουργίες που προσφέρουν στους χρήστες που τα χρησιμοποιούν.

Εικόνα 3: Ορισμένα από τα δημοφιλή κοινωνικά μέσα [78]

Τα μέσα κοινωνικής δικτύωσης επομένως, θα μπορούσαν να χωριστούν στις εξής κατηγορίες [79]:

1. Social Networking

   Αυτή η κατηγορία περιλαμβάνει μέσα στα οποία ο χρήστης μπορεί να αλληλεπιδράσει με άλλους χρήστες που συμμετέχουν στο ίδιο δίκτυο μέσω της ανάρτησης μηνυμάτων, μέσω της επικοινωνίας (chatting) σε πραγματικό ή όχι χρόνο, μέσω της διάδοσης διάφορων πληροφοριών κτλ. Σε αυτήν την κατηγορία εντάσσονται κοινωνικά μέσα όπως το Facebook, το Twitter, το LinkedIn, το Hi5.

2. Conversation Apps

   Αυτή η κατηγορία κοινωνικών μέσων περιλαμβάνει εφαρμογές που επικεντρώνονται περισσότερο στη δημιουργία συζητήσεων και διάδοσης πληροφοριών, πολλές φορές συνδυάζοντας και εικόνα. Σε αυτήν την κατηγορία εντάσσονται εφαρμογές όπως το Skype, το Viber, το YahooMail, το Gmail.

3. Videos

   Σε αυτήν την κατηγορία των κοινωνικών μέσων εντάσσονται εφαρμογές που επιτρέπουν στους χρήστες είτε να βλέπουν και να ακούν συνδυασμούς εικόνων και ήχου, είτε να συμμετέχουν ενεργά στην εφαρμογή με το ανέβασμα εικόνων, μουσικής και βίντεο. Σε αυτήν την κατηγορία συναντάμε εφαρμογές όπως το Youtube, το MySpace, το Dailymotion και το Instagram.

2.2 Κοινωνικά Δίκτυα και Συστήματα Συστάσεων

Τα κοινωνικά συστήματα συστάσεων, δηλαδή τα συστήματα συστάσεων που εφαρμόζονται στην περίπτωση των κοινωνικών μέσων, έχουν ως βασικό στόχο να βοηθήσουν τον χρήστη να διαχειριστεί τη σωρεία πληροφοριών που δέχεται αδιάλειπτα. Με άλλα λόγια, τα συστήματα συστάσεων στα κοινωνικά δίκτυα παρουσιάζουν στους χρήστες όσο το δυνατόν πιο σχετικές με αυτούς πληροφορίες καθώς και πληροφορίες που πιθανώς θα ενδιαφέρουν τους χρήστες. Πέρα όμως από την παρουσίαση σχετικών ως προς τον χρήστη πληροφοριών, τα συστήματα συστάσεων στοχεύουν στην αύξηση της ενασχόλησης των χρηστών με τα κοινωνικά δίκτυα όπως επίσης και στην εμπλοκή νέων χρηστών στα κοινωνικά μέσα.

Τα συστήματα αυτά κάνουν συστάσεις για διάφορα αντικείμενα και θέματα όπως βιβλία, άρθρα στο διαδίκτυο, καταναλωτικά προϊόντα, φίλους σε ένα κοινωνικό δίκτυο, εικόνες και γενικώς οτιδήποτε θα μπορούσε να ενδιαφέρει έναν χρήστη και θα ήθελε να ενημερωθεί για αυτό. Κατά την περιήγηση ενός χρήστη στο διαδίκτυο, συνεχώς έρχεται αντιμέτωπος με συστήματα συστάσεων αφού κάθε διαφήμιση που προβάλει το google, κάθε προϊόν που συστήνεται στο amazon και το ebay, κάθε εικόνα, άρθρο, φίλο, δραστηριότητα, σύνδεσμο ή είδηση που προτείνεται στο facebook και στο twitter υπάρχει ένα καλά οργανωμένο σύστημα συστάσεων όπου πολύ προσεκτικά μελέτησε τις κινήσεις ενός χρήστη, τις συσχέτισε με άλλους χρήστες και προχώρησε στη σύσταση που θα ικανοποιούσε περισσότερο τον χρήστη. Γενικότερα, ο αντικειμενικός σκοπός ενός συστήματος συστάσεων είναι να διαχειρίζεται τον μεγάλο όγκο

πληροφοριών που διαχέονται στα κοινωνικά δίκτυα ώστε να βοηθήσει ένα χρήστη να επιλέξει την καταλληλότερη για αυτόν πληροφορία.

Η σχέση μεταξύ κοινωνικών δικτύων και συστημάτων συστάσεων είναι αμφίδρομη. Με την έννοια αμφίδρομη σχέση εννοούμε ότι τα κοινωνικά δίκτυα αποτελούν το πεδίο δράσης των συστημάτων συστάσεων αφού τους προσφέρουν όλα εκείνα τα μέσα που ένα σύστημα αξιοποιεί (tags, tweets, comments) προκειμένου να προβεί σε κάποια σύσταση. Από την άλλη πλευρά, τα συστήματα συστάσεων συμβάλουν στη διάδοση των κοινωνικών δικτύων αφού κάνουν τη χρήση τους περισσότερο ελκυστική με το να παρέχουν σε ένα χρήστη ακριβώς αυτό που θέλει.

Τα συστήματα συστάσεων που εφαρμόζονται στα κοινωνικά μέσα χρησιμοποιούν συνήθως τεχνολογίες που βασίζονται σε αξιολογήσεις των χρηστών για διάφορα αντικείμενα και βάση των αξιολογήσεων αυτών με τη χρήση ενός κατάλληλου αλγορίθμου καταλήγουν σε συστάσεις. Επιπλέον, τα συστήματα συστάσεων στα κοινωνικά μέσα ανακτούν πληροφορίες και αξιοποιούν δεδομένα από τον περίγυρο ενός χρήστη έτσι ώστε να καταφέρουν, μέσω της δραστηριότητας των ατόμων με τα οποία σχετίζεται και αλληλεπιδρά ένας χρήστης, να σκιαγραφήσουν και να δημιουργήσουν ένα προφίλ για ένα χρήστη-στόχο με τέτοιο τρόπο έτσι ώστε αυτό να αποτελέσει τη βάση πάνω στην οποία θα βασίζονται και οι μελλοντικές συστάσεις για τον χρήστη.

Συμπερασματικά, αυτό που κάνει ουσιαστικά ένα σύστημα συστάσεων σε ένα κοινωνικό δίκτυο, όπως θα δούμε και παρακάτω αναλυτικότερα, είναι να συγκρίνει τα στοιχεία από το προφίλ ενός χρήστη, που ενδεχομένως να έχει δημιουργήσει το ίδιο το σύστημα, με τα χαρακτηριστικά των αντικειμένων που κατόπιν έρευνας και ανάλυσης επιλέχτηκαν ότι θα ικανοποιούσαν καλύτερα τις ανάγκες του χρήστη αυτού. Σκοπός αυτής της σύγκρισης είναι να προταθούν εκείνα τα αντικείμενα σε ένα χρήστη που ταιριάζουν περισσότερο στις προτιμήσεις του. Εάν αυτές οι συστάσεις που προορίζονται προς τον χρήστη διαπιστωθεί από το σύστημα ότι τον ικανοποιούν, τότε αυτό θα αποτελέσει επιπλέον πληροφορία έτσι ώστε το σύστημα την επόμενη φορά να καταφέρει να προτείνει στον χρήστη ένα εξίσου ενδιαφέρον προς αυτόν αντικείμενο ή ακόμα και καλύτερο.

2.3 Τα δημοφιλέστερα κοινωνικά δίκτυα

Στην παρούσα ενότητα θα περιγράψουμε ορισμένα από τα πιο δημοφιλή κοινωνικά δίκτυα στα οποία τα συστήματα συστάσεων έχουν την μεγαλύτερη εφαρμογή. Θα περιγράψουμε το Facebook, το Twitter, το LinkedIn και το Youtube.

Facebook

Εμπνευστής της ιδέας αυτού του κοινωνικού δικτύου υπήρξε ο Μαρκ Ζάκερμπεργκ ως μέλος του Πανεπιστημίου Χάρβαρντ όπου μαζί με τους Εντουάρντο Σάβεριν, Άντριου Μακόλουμ, Ντάστιν Μάσκοβιτς και Κρίς Χούγες ίδρυσαν το Facebook τον Φεβρουάριο του 2004. Αρχικά δικαίωμα συμμετοχής είχαν μόνο οι φοιτητές του Χάρβαρντ ενώ αργότερα επεκτάθηκε η χρήση του και σε άλλα μεγάλα πανεπιστήμια της Αμερικής. Το 2006 η υπηρεσία του Facebook έγινε προσβάσιμη σε κάθε άνθρωπο του πλανήτη που η ηλικία του ξεπερνούσε τα 13 χρόνια. Το Facebook σήμερα έχει πάνω από 1 δισεκατομμύριο ενεργούς χρήστες αποτελώντας το νούμερο 1 διαδικτυακό τόπο παγκοσμίως και νούμερο 2 στις ΗΠΑ. Στην κατηγορία της κοινωνικής δικτύωσης το Facebook αναδεικνύεται σε ηγέτης [79, 80].

Εικόνα 4: Το λογότυπο του Facebook [80]

Το Facebook αποτελεί έναν τρόπο δικτύωσης με φίλους και γνωστούς ενώ επιτρέπει ασύγχρονη αλλά και σε πραγματικό χρόνο αλληλεπίδραση μεταξύ των χρηστών. Αυτό το κοινωνικό δίκτυο προσφέρει στους χρήστες του υπηρεσίες συνομιλίας, ανέβασμα φωτογραφιών, βίντεο, διάφορων κειμένων και οτιδήποτε άλλο είναι αποδεκτό από την πολιτική του Facebook και μπορεί να ενδιαφέρει τους χρήστες να κοινοποιήσουν στον περίγυρό τους.

Πέρα από τη λειτουργία του Facebook στα πλαίσια της κοινωνικής δικτύωσης, έχει μετατραπεί λόγω του υπερβολικά τεράστιου αριθμού χρηστών και σε ένα πολύτιμο και άκρως αποτελεσματικό εργαλείο προώθησης. Η προώθηση δεν περιορίζεται μόνο σε διαφημιζόμενα προϊόντα, αλλά εκτείνεται και σε προώθηση προσώπων, ιδεών,

καταστάσεων, διαδηλώσεων καθώς και οτιδήποτε άλλο μπορεί να ενδιαφέρει ή να ενεργοποιεί τους χρήστες.

Ωστόσο, το Facebook έχει παρατηρηθεί και μελετηθεί ότι επηρεάζει σε πάρα πολύ μεγάλο βαθμό τα επίπεδα της πραγματικής κοινωνικής ζωής, αφού πολλές φορές οι χρήστες επικεντρώνονται πιο πολύ στο τι γίνεται στη διαδικτυακή κοινωνία τους παρά στην πραγματική.

Twitter

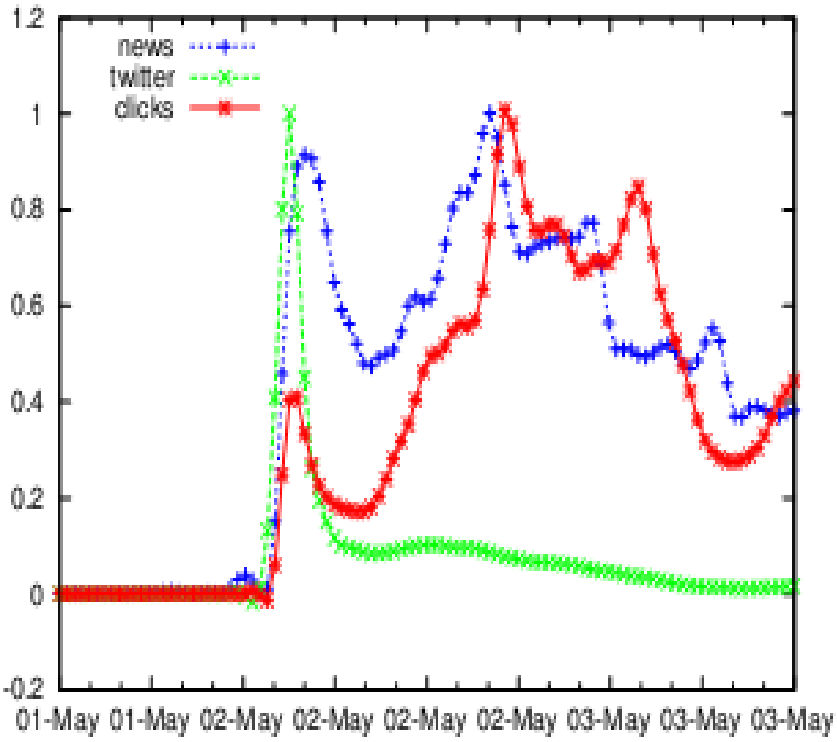
Το Twitter ήρθε μετά το Facebook για να κατακτήσει μία από τις πρώτες θέσεις στον χάρτη των κοινωνικών μέσων. Δημιουργήθηκε τον Μάρτιο του 2006 από τον Τζακ Ντόρσεϊ και είχε τη μορφή ενός microblog. Το Twitter σύμφωνα με στοιχεία του 2014 αριθμεί περί των 250 εκατομμυρίων χρηστών ανά τον κόσμο, ενώ κάθε μέρα δημοσιεύονται μέσω αυτού πάνω από 300 εκατομμύρια μηνύματα. Ενδεικτικό του μεγέθους αυτού του κοινωνικού δικτύου είναι ότι ανά ώρα δημοσιεύονται μηνύματα που αγγίζουν το εύρος των 400.000 με 1.400.000 μηνύματα.

Εικόνα 5: Το λογότυπο του Twitter [81]

Το Twitter παρέχει στους χρήστες του τη δυνατότητα ανάρτησης σύντομων μηνυμάτων που δεν μπορούν τα ξεπερνούν τους 140 χαρακτήρες. Τα μηνύματα αυτά ονομάζονται tweets. Επίσης ο κάθε χρήστης μπορεί να επιλέξει να συνδεθεί με άλλους χρήστες και έτσι να βλέπει τη δραστηριότητά τους και τα μηνύματα που αναρτώνται από αυτούς. Η διαδικασία της σύνδεσης ενός χρήστη με έναν άλλον ονομάζεται follow, ενώ αυτός που επιλέγει να ακολουθήσει κάποιον άλλο χρήστη ονομάζεται follower για τον χρήστη που ακολουθείται και followee για τον χρήστη που κάνει το follow. Για κάποιον χρήστη που θα δημοσιεύσει ένα μήνυμα, αυτό θα είναι ορατό στους ακολούθους του (followers) οι οποίοι μπορούν να αναδημοσιεύσουν το μήνυμα αυτό και να κάνουν το επονομαζόμενο retweet. Επιπλέον το Twitter παρέχει στους χρήστες του και άλλες λειτουργίες όπως την αναφορά ενός χρήστη σε κάποιο μήνυμα με την προσθήκη του συμβόλου @ ή την

χρήση του συμβόλου # για να κατηγοριοποιήσουν διάφορα μηνύματα και πράγματα ή καταστάσεις [79, 81].

Το Twitter έχει καταστεί σήμερα μέσο για τη διάδοση ειδήσεων, προώθηση προσώπων, προϊόντων κτλ. Ενδεικτική περίπτωση, που αποτυπώνει αντιπροσωπευτικά την ταχύτητα διάδοσης των ειδήσεων μέσω του Twitter αλλά και την άμεση ανταπόκριση του δικτύου αυτού στην ροή των εξελίξεων, παρουσιάζεται στην παρακάτω εικόνα.



Εικόνα 6: Ροή ειδήσεων σχετικά με τον Οσάμα Μπιν Λάντεν [24]

Στην παραπάνω εικόνα παρατηρούμε τις σχετικές ειδήσεις για τον Οσάμα Μπιν Λάντεν για την περίοδο 3 ημερών. Αρχικά βλέπουμε τα σχετικά με το θέμα tweets να προλαβαίνουν τις ειδήσεις, ενώ αρκετά υψηλά είναι και τα κλικς που έγιναν μέσω του Twitter σχετικά με το θέμα αυτό [24].

LinkedIn

Το LinkedIn αποτελεί ένα κοινωνικό μέσο που επικεντρώνεται στην επαγγελματική κοινωνική δικτύωση. Ιδρύθηκε τον Δεκέμβριο του 2002 από τον Ρέιντ Χόφμαν. Οι χρήστες του LinkedIn έχουν τη δυνατότητα να δημιουργήσουν το προσωπικό επαγγελματικό τους προφίλ, που ουσιαστικά πρόκειται για το βιογραφικό τους και να συνδεθούν με άλλους χρήστες, να αναζητήσουν εργασία, αλλά και να δημιουργήσουν πελατολόγιο. Μεταξύ των λειτουργιών που προσφέρονται στον χρήστη μέσω αυτού του

μέσου μπορούμε να διακρίνουμε τη δυνατότητα λεπτομερούς παρουσίασης της εργασιακής εμπειρίας και εκπαιδευτικού υπόβαθρου του χρήστη, τη δυνατότητα καταχώρησης προσωπικών πληροφοριών και ενδιαφερόντων, τη δικτύωση και την αλληλεπίδραση με τα συνδεδεμένα μέλη, τη δυνατότητα παροχής συστάσεων (endorsement) από συναδέλφους, συνεργάτες και γενικότερα μέλη του δικτύου με τα οποία είναι κανείς συνδεδεμένος καθώς και τη δημοσίευση αναρτήσεων, ειδήσεων και σχολίων [79, 82].

Youtube

Το Youtube αποτελεί μετά το Facebook τον δημοφιλέστερο ιστότοπο του διαδικτύου με την μεγαλύτερη επαναληψιμότητα καθημερινά. Το Youtube ιδρύθηκε τον Φεβρουάριο του 2005 από τους Τσαντ Χάρλεϊ, Στηβ Τσεν και τον Τζουντ Καρίμ δίνοντας τη δυνατότητα στους χρήστες του να αναπαράγουν, να αναζητήσουν και να αποθηκεύσουν ψηφιακά βίντεο που μπορεί να είναι είτε ταινίες, είτε αποσπάσματα, είτε μουσική είτε οτιδήποτε άλλο που μπορεί να βιντεοσκοπηθεί και να συνδυάσει εικόνα και ήχο.

Εικόνα 7: Το λογότυπο του Youtube [83]

Στο Youtube, για κάθε βίντεο που αναπαράγεται από κάποιον χρήστη υπάρχει ένας μετρητής των θεάσεων που έχει ένα συγκεκριμένο βίντεο. Το γεγονός ότι υπάρχουν βίντεο με εκατοντάδες εκατομμύρια views υποδηλώνει το μέγεθος αυτού του μέσου και την ευρύτατη χρήση από τους χρήστες του. Επίσης, κάθε χρήστης έχει τη δυνατότητα να σχολιάσει κάποιο βίντεο ή ακόμη και να εκφράσει την αρέσκεια ή δυσαρέσκειά του.

Επιπλέον, στο Youtube υπάρχει έντονο το στοιχείο της σύστασης νέων βίντεο καθώς ανάλογα με το τι παρακολουθεί ένας χρήστης του προτείνεται και τι άλλο να παρακολουθήσει αργότερα [79, 83].

2.4 Λειτουργίες και εργαλεία των κοινωνικών δικτύων

Τα κοινωνικά δίκτυα προκειμένου να ικανοποιήσουν με τον καλύτερο δυνατό τρόπο τους χρήστες τους και να είναι αποδοτικά και αποτελεσματικά, συνδυάζουν διάφορες λειτουργίες και εργαλεία. Πίσω από αυτές τις λειτουργίες που χρησιμοποιούν τα κοινωνικά δίκτυα βρίσκονται ποικίλοι αλγόριθμοι άλλοι απλοί και άλλοι πιο σύνθετοι. Πολλές φορές κάποια λειτουργία ενός κοινωνικού δικτύου, όπως για παράδειγμα μια αναζήτηση ή ένα Like, κρύβει πίσω της πολλές μελέτες, έρευνα αλλά και ώρες εργασίας για να υλοποιηθεί ο αλγόριθμος ή οι αλγόριθμοι που θα υλοποιούν την κάθε λειτουργία του κοινωνικού δικτύου. Στην παράγραφο αυτή επιχειρούμε να παραθέσουμε τις διάφορες λειτουργίες και αλγορίθμους που χρησιμοποιούν σήμερα τα διάφορα κοινωνικά δίκτυα [59, 77].

Στα περισσότερα κοινωνικά δίκτυα υπάρχουν φόρμες αναζήτησης που βασίζονται στον κοινωνικό γράφο (social graph) του ατόμου που προβαίνει στην αναζήτηση και βασίζεται στην ιδέα ότι αντικείμενα που ενδιαφέρουν τα άτομα του κοινωνικού μας περίγυρου θα είναι περισσότερο πιθανό να ενδιαφέρουν και εμάς σε σύγκριση με αντικείμενα που ενδιαφέρουν άτομα τα οποία εμείς δεν γνωρίζουμε. Κατά συνέπεια, τέτοια αντικείμενα θα εμφανιστούν πρώτα κατά τις αναζητήσεις του χρήστη.

Μια άλλη, ιδιαίτερα σημαντική και χρήσιμη λειτουργία, ειδικά σε πλατφόρμες ταινιών, μουσικής κτλ. είναι η αξιολόγηση διάφορων αντικειμένων που ουσιαστικά αντικατοπτρίζει την άποψη του χρήστη για ένα αντικείμενο, για ένα κείμενο, μια κατάσταση κτλ. Αυτό μπορεί να γίνει είτε με το πάτημα ενός κουμπιού που σηματοδοτεί αν μας αρέσει κάτι ή όχι είτε με την παροχή βαθμολογίας με βάση κάποια κλίμακα για κάτι που μας άρεσε και κατά πόσο μας άρεσε. Αυτό ακριβώς είναι που υλοποιεί το *rating* σε ένα κοινωνικό δίκτυο. Αν και η διαδικασία για το rating είναι πολύ απλή, τα δεδομένα, η επεξεργασία των δεδομένων και τα συμπεράσματα που μπορούμε να έχουμε μπορεί να είναι αρκετά περίπλοκα και με πολλές χρησιμότητες.

Πολλά κοινωνικά δίκτυα παρέχουν στους χρήστες τους πέρα από τα παραπάνω και τη δυνατότητα δημιουργίας ετικετών (tags). Κάθε χρήστης ενός κοινωνικού δικτύου μπορεί στα δεδομένα που ανεβάζει ή βλέπει να βάζει ένα tag το οποίο να προσδιορίζει την κατηγορία στην οποία ανήκουν. Κάθε φωτογραφία, μουσικό κομμάτι, σχόλιο, κείμενο, σύνδεσμος μπορεί να συνοδεύεται από μια ετικέτα που να προσδιορίζει το

περιεχόμενο του με αποτέλεσμα τυχόν αναζητήσεις και συσχετίσεις μεταξύ των δεδομένων αυτών να μπορούν να γίνουν με βάση αυτές τις ετικέτες.

Ίσως αυτό που χαρακτηρίζει τα κοινωνικά δίκτυα είναι ότι προσφέρουν τη συμμετοχή των χρηστών σε groups ποικίλων περιεχομένων έτσι ώστε να μοιράζονται κοινά ενδιαφέροντα και να συγκεντρώνουν και να οργανώνουν γνώση γύρω από τα θέματα που τους ενδιαφέρουν. Οι συζητήσεις, οι δημοσιεύσεις, οι ψηφοφορίες και γενικώς οι δραστηριότητες που μπορούν να έχουν οι χρήστες ενός group μπορούν να είναι δεδομένα με τα οποία θα παραχθεί γνώση.

Ένα άλλο κομμάτι στο οποίο τα κοινωνικά δίκτυα μπορούν να συμβάλουν καταλυτικά είναι ο τομέας του μάρκετινγκ και των πωλήσεων. Τα κοινωνικά δίκτυα μπορούν μέσω συγκεκριμένων αλγορίθμων να δημιουργήσουν ένα περιβάλλον μέσα στο οποίο θα αλληλεπιδρούν πελάτες και επιχειρήσεις με σκοπό οι πελάτες να γνωρίσουν τα προϊόντα των επιχειρήσεων και να δουν αν αυτά ταιριάζουν στις ανάγκες στους και οι επιχειρήσεις να μελετήσουν τις επιθυμίες και τις ανάγκες των καταναλωτών με σκοπό τη διαμόρφωση καταλληλότερων προϊόντων. Επίσης, το personalized marketing σε ένα κοινωνικό δίκτυο είναι ένας τρόπος για να έρθει ο χρήστης – καταναλωτής σε επαφή με προϊόντα που τον ενδιαφέρουν. Η παρουσία του χρήστη σε κάθε κοινωνικό δίκτυο δημιουργεί γνώση την οποία μπορεί να χρησιμοποιήσει μια τεχνολογία Personalized marketing προκειμένου να δώσει στον χρήστη αυτό που επιθυμεί περισσότερο.

3 Συστήματα Συστάσεων

Στο παρόν κεφάλαιο θα παρουσιάσουμε αναλυτικά τα διάφορα συστήματα συστάσεων που υπάρχουν με την παράλληλη αξιολόγηση του κάθε συστήματος.

3.1 Εισαγωγή

Σε ένα περιβάλλον, όπως αυτό της κοινωνικής ζωής, που χαρακτηρίζεται από σωρεία πληροφοριών κάθε είδους, κρίσιμος παράγοντας για να την ανάπτυξη και πρόοδό του, είναι η ύπαρξη ενός αποτελεσματικού συστήματος διάθεσης και διάχυσης των πληροφοριών και της γνώσης.

Στη σύγχρονη εποχή και στις εξελιγμένες κοινωνίες, το άτομο βρίσκεται συνεχώς αντιμέτωπο με προκλήσεις που απαιτούν λήψη αποφάσεων για περίπλοκα ζητήματα. Παράλληλα, υπάρχει μια τεράστια ποσότητα συσσωρευμένης γνώσης και ποικιλία πιθανών επιλογών και προτιμήσεων που κάνουν τον σύγχρονο άνθρωπο να στρέφεται στο στενό ή ευρύτερο κοινωνικό περίγυρο προκειμένου να λάβει τις απαραίτητες και εξειδικευμένες πληροφορίες, οι οποίες συνήθως έχουν την μορφή συστάσεων – προτάσεων. Κατά συνέπεια, ο κοινωνικός περίγυρος του ατόμου λειτουργεί σαν φίλτρο που επιτρέπει στις σχετικές πληροφορίες να φτάσουν στον ενδιαφερόμενο κάθε φορά, μπλοκάροντας ταυτόχρονα την πληθώρα των άχρηστων πληροφοριών.

Στα πλαίσια του διαδικτύου, το πρόβλημα αυτό των χρηστών (η επιλογή δηλαδή μέσα από εναλλακτικές λύσεις βάσει πληθώρας πληροφοριών) έρχονται να λύσουν τα Συστήματα Συστάσεων (Recommendation Systems) [18]. Τα συστήματα αυτά λύνουν τα προβλήματα των χρηστών παρέχοντάς τους αξιόπιστες συστάσεις για διάφορα προϊόντα, βιβλία, κείμενα, πληροφορίες, άλλους χρήστες κτλ. Τα Συστήματα Συστάσεων καταλήγουν σε συστάσεις μέσω διάφορων τεχνικών και είναι βασισμένα στο διαδίκτυο, ενώ οι συστάσεις που παρέχουν κατευθύνουν τον χρήστη μέσα σε ένα μεγάλο όγκο πληροφοριών.

Το πρώτο σύστημα συστάσεων τοποθετείται χρονικά στις αρχές της δεκαετίας του ’90 και ήταν το ονομαζόμενο “Tapestry”. Οι δημιουργοί του εισήγαγαν για πρώτη φορά τον όρο Συνεργατικό Φιλτράρισμα (Collaborative Filtering) για να περιγράψουν τη

δυνατότητα σύγκρισης των αξιολογήσεων ενός χρήστη με τις αξιολογήσεις των άλλων, με στόχο την παραγωγή συστάσεων οι οποίες να ικανοποιούν τις προτιμήσεις του. Το σύστημα αυτό λειτουργούσε στα πλαίσια μιας σχετικά μικρής ομάδας ανθρώπων με ύπαρξη διαπροσωπικών σχέσεων ανάμεσα στα μέλη της, όπως οι εργαζόμενοι σε ένα γραφείο ή μία επιχείρηση. Γενικά, τα συστήματα συστάσεων χρησιμοποιούν διάφορες τεχνικές για να συλλέξουν πληροφορίες σχετικά με τους χρήστες (προτιμήσεις, προτεραιότητες κ.α.) και τις εναλλακτικές επιλογές – προϊόντα, τα οποία βρίσκονται στη διάθεση των χρηστών (τεχνικά χαρακτηριστικά, βαθμολογία προερχόμενη από άλλους χρήστες, κ.α.). Τα στοιχεία αυτά αξιοποιούνται με σκοπό να συσχετιστούν οι χρήστες με αυτά τα αντικείμενα – προϊόντα τα οποία είναι πιο πιθανό να καλύπτουν τις απαιτήσεις τους. Η αξιοποίηση των στοιχείων αυτών και η εξαγωγή των συστάσεων, γίνονται μέσω της χρήσης κατάλληλων αλγορίθμων.

Κατηγορίες Συστημάτων Συστάσεων

Στη βιβλιογραφία χρησιμοποιούντα 6 διαφορετικές κατηγορίες συστημάτων συστάσεων τις οποίες αναφέρουμε παρακάτω [61]:

1. Συστήματα Content - Based: Τα συστήματα αυτά προβαίνουν σε συστάσεις αντικειμένων που είναι παρόμοια με εκείνα που είχε προτιμήσει ο χρήστης κατά το παρελθόν. Η ομοιότητα των αντικειμένων υπολογίζεται με βάση κοινά χαρακτηριστικά που έχουν τα συγκρινόμενα αντικείμενα. Για παράδειγμα, εάν κάποιος χρήστης αξιολόγησε στο παρελθόν θετικά μια ταινία που χαρακτηρίζεται ως κωμωδία, τότε ένα σύστημα συστάσεων βασισμένο στο περιεχόμενο θα προβεί σε σύσταση προς το χρήστη που θα προτείνει κάποια ταινία του ίδιου είδους, δηλαδή κωμωδία.
2. Συστήματα collaborative filtering: Τα συστήματα αυτής της κατηγορίας λειτουργούν αξιοποιώντας τις αξιολογήσεις και βαθμολογήσεις των χρηστών. Η βασική αρχή λειτουργίας αυτών των συστημάτων συνίσταται στο ότι οι χρήστες που δίνουν παρεμφερείς αξιολογήσεις για ορισμένα αντικείμενα, έχουν παρεμφερή ενδιαφέροντα και ανάγκες και κατά συνέπεια θα ενδιαφέρονται για τα ίδια αντικείμενα. Τα συστήματα αυτά δεν έχουν καμία γνώση για τα χαρακτηριστικά των αντικειμένων που προτείνουν καθώς ούτε και για τις ανάγκες των χρηστών τους [68].
3. Δημογραφικά Συστήματα Συστάσεων (Demographic Recommendation Systems): Τα συστήματα αυτά προτείνουν αντικείμενα με βάση το δημογραφικό

προφίλ κάθε χρήστη. Για παράδειγμα τέτοια συστήματα μπορεί να προβαίνουν σε συστάσεις ανάλογα με την ηλικία, τη γλώσσα, τη χώρα ενός χρήστη. Παρότι τέτοια συστήματα χρησιμοποιούντα ευρέως για σκοπούς marketing, δεν υπάρχει πολλή έρευνα για αυτά τα συστήματα στη βιβλιογραφία [47].

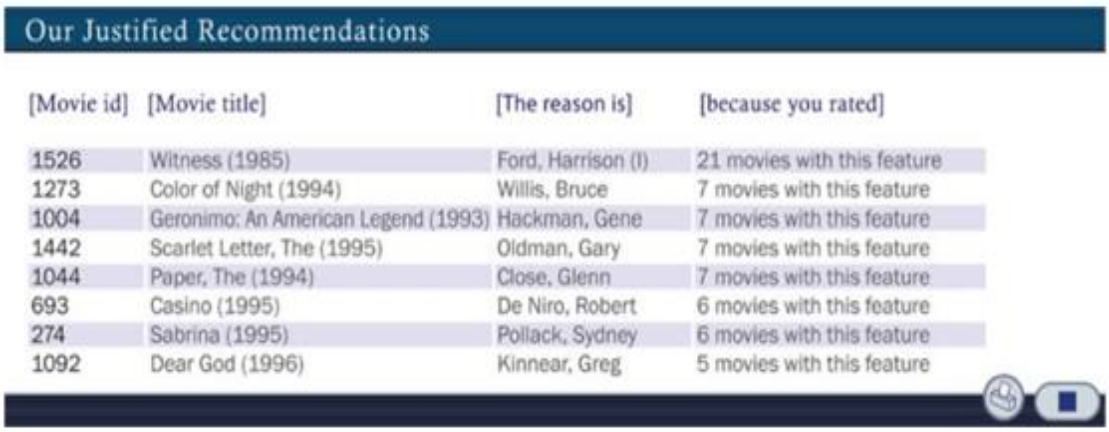
4. Συστήματα Knowledge – based: Τα συστήματα αυτά προβαίνουν σε συστάσεις με βάση τη γνώση για τα χαρακτηριστικά και τις ιδιότητες των αντικειμένων και πως αυτά θα μπορέσουν να ικανοποιήσουν τις ανάγκες και τις προτιμήσεις του χρήστη. Τα συστήματα αυτά διακρίνονται στα case – based systems και στα constraint – based systems [16,60]
5. Συστήματα Community – based: Τα συστήματα αυτής της κατηγορίας συστήνουν αντικείμενα σε ένα χρήστη βάσει των προτιμήσεων των φίλων αυτού - *δείξε μου το φίλο σου να σου πω ποιος είσαι* [4, 9]. Η πράξη έχει αποδείξει ότι τα άτομα τείνουν να επιλέγουν αντικείμενα που προτείνονται από τον κοινωνικό τους περίγυρο παρά από άτομα με τα οποία μπορεί να έχουν κοινά χαρακτηριστικά αλλά δε γνωρίζουν [4, 9]
6. Υβριδικά Συστήματα: Τα Υβριδικά Συστήματα είναι συστήματα που αξιοποιούν τις παραπάνω μεθόδους για να παράγουν μια νέα μέθοδο βασισμένη σε αυτές προκειμένου να άρει τις όποιες αδυναμίες των παραπάνω μεθόδων. Είναι δηλαδή συνδυασμός των παραπάνω συστημάτων [18, 19].

Παρακάτω θα παρουσιάσουμε μερικά αντιπροσωπευτικά παραδείγματα [71] των βασικών συστημάτων συστάσεων που κάνουν χρήση των Collaborative Filtering συστημάτων, των Content – based συστημάτων και των Υβριδικών συστημάτων.

Εικόνα 8: Σύσταση στο Amazon [71]

Στην παραπάνω εικόνα μπορούμε να παρατηρήσουμε πως πραγματοποιείται μια σύσταση στο amazon.com χρησιμοποιώντας Collaborative Filtering σύσταση. Η σύσταση προς τον χρήστη βασίζεται στις προτιμήσεις των άλλων χρηστών και έτσι το σύστημα προτείνει στον χρήστη αντικείμενα που έχουν προτιμηθεί από άλλους χρήστες με παρόμοια ενδιαφέροντα και προτιμήσεις με τον χρήστη στόχο.

Από την άλλη πλευρά, ένα αντιπροσωπευτικό παράδειγμα σύστασης με βάση την Content – based μέθοδο είναι συστάσεις ταινιών στο MovieExplain συνδυάζοντας το προφίλ του χρήστη με τα διάφορα χαρακτηριστικά των αντικειμένων. Η εικόνα παρακάτω δείχνει πως πραγματοποιείται μια σύσταση στο MovieExplain.



Our Justified Recommendations

[Movie id]	[Movie title]	[The reason is]	[because you rated]
1526	Witness (1985)	Ford, Harrison (I)	21 movies with this feature
1273	Color of Night (1994)	Willis, Bruce	7 movies with this feature
1004	Geronimo: An American Legend (1993)	Hackman, Gene	7 movies with this feature
1442	Scarlet Letter, The (1995)	Oldman, Gary	7 movies with this feature
1044	Paper, The (1994)	Close, Glenn	7 movies with this feature
693	Casino (1995)	De Niro, Robert	6 movies with this feature
274	Sabrina (1995)	Pollack, Sydney	6 movies with this feature
1092	Dear God (1996)	Kinnear, Greg	5 movies with this feature

Εικόνα 9: Συστάσεις στο MovieExplain [71]

Σε αυτήν την περίπτωση το σύστημα αιτιολογεί τη σύσταση που κάνει δείχνοντας πόσο ισχυρή είναι η αλληλεξάρτηση ενός χαρακτηριστικού του αντικειμένου (the reason why) με τα χαρακτηριστικά που συνθέτουν το προφίλ του χρήστη (because you rated).

Τέλος, ένα παράδειγμα που συνδυάζει τις αξιολογήσεις του χρήστη και τα χαρακτηριστικά των αντικειμένων που προτείνει είναι το HuffPost Social News όπου οι χρήστες του Facebook μπορούν να δουν τι έχουν διαβάσει και σχολιάσει οι φίλοι τους, δημιουργώντας έτσι μια πλατφόρμα συζητήσεων και ανταλλαγής απόψεων για διάφορα κείμενα ή κάθε άλλου είδους αντικείμενα που έχουν είτε διαβάσει, είτε σχολιάσει. Παρακάτω φαίνεται ένα απόσπασμα από την πλατφόρμα HuffPost Social News που δείχνει συστάσεις φίλων του χρήστη στο Facebook με το τι έχουν διαβάσει και σχολιάσει [71].

Εικόνα 10: Λειτουργία του HuffPost Social News [71]

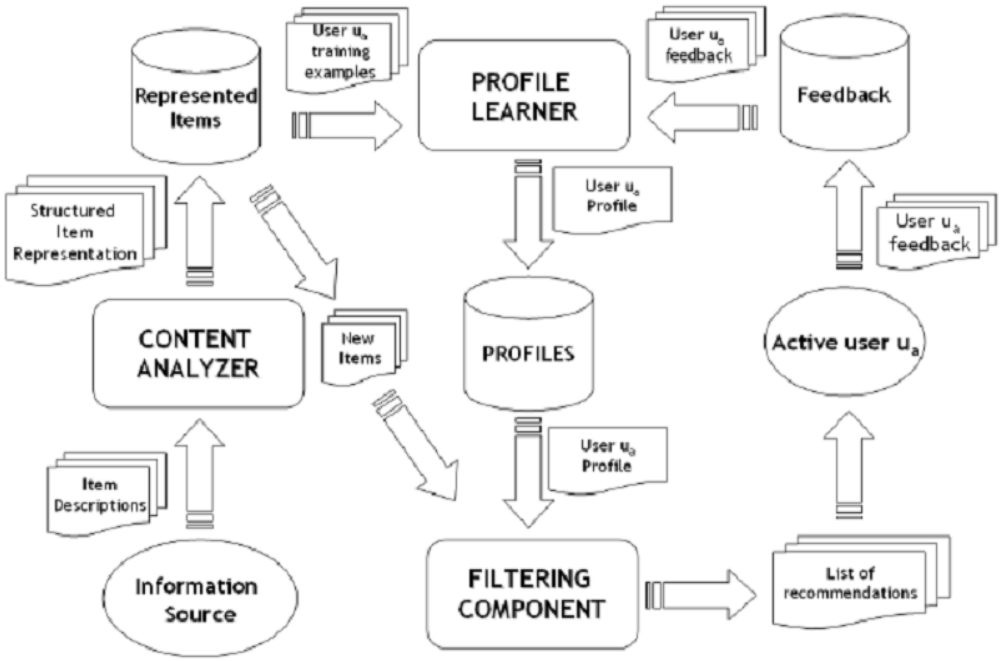
Στις ενότητες που ακολουθούν θα εξετάσουμε αναλυτικά τις κατηγορίες που αναφέραμε παραπάνω.

3.2 Content – Based Συστήματα

Τα συστήματα content – based εφαρμόζουν μια τεχνική βασισμένη στο περιεχόμενο με την έννοια ότι αναλύουν ένα σύνολο από κείμενα ή αντικείμενα τα οποία έχουν προηγουμένως αξιολογηθεί από κάποιον χρήστη, χτίζοντας έτσι, βάσει των αξιολογήσεων του χρήστη, ένα προφίλ των ενδιαφερόντων του. Το προφίλ αυτό ουσιαστικά πρόκειται για το χάρτη των ενδιαφερόντων του χρήστη. Η διαδικασία της σύστασης μέσα από αυτού του είδους τα συστήματα περιλαμβάνει το συσχετισμό των χαρακτηριστικών του χρήστη, έτσι όπως έχουν σκιαγραφηθεί μέσω του προφίλ του και των χαρακτηριστικών των διάφορων αντικειμένων που μπορεί να συσταθούν. Το αποτέλεσμα αυτού του συσχετισμού δείχνει ακριβώς το επίπεδο του ενδιαφέροντος του χρήστη με κάθε ένα αντικείμενο. Κατά συνέπεια, εάν το προφίλ του χρήστη είναι διαμορφωμένο με ακρίβεια, σχετικά με τα ενδιαφέροντά του, τότε η μέθοδος αυτή θα συστήσει στον χρήστη ακριβώς το αντικείμενο που επιθυμεί.

3.2.1 Λειτουργία της μεθόδου

Από τα παραπάνω, γίνεται εύκολα αντιληπτό ότι τέτοια συστήματα χρειάζονται κατάλληλες τεχνικές έτσι ώστε να μπορούν να αξιολογούν τα αντικείμενα για τη δημιουργία ενός συνεπούς προφίλ για τον χρήστη, όπως επίσης και κατάλληλες στρατηγικές προκειμένου να γίνει σωστά και αποδοτικά ο συσχετισμός των χαρακτηριστικών του χρήστη με αυτά των αντικειμένων. Η παρακάτω εικόνα παρουσιάζει την αρχιτεκτονική ενός συστήματος που βασίζεται στο περιεχόμενο.



Εικόνα 11: Αρχιτεκτονική συστημάτων βασισμένων στο περιεχόμενο [61]

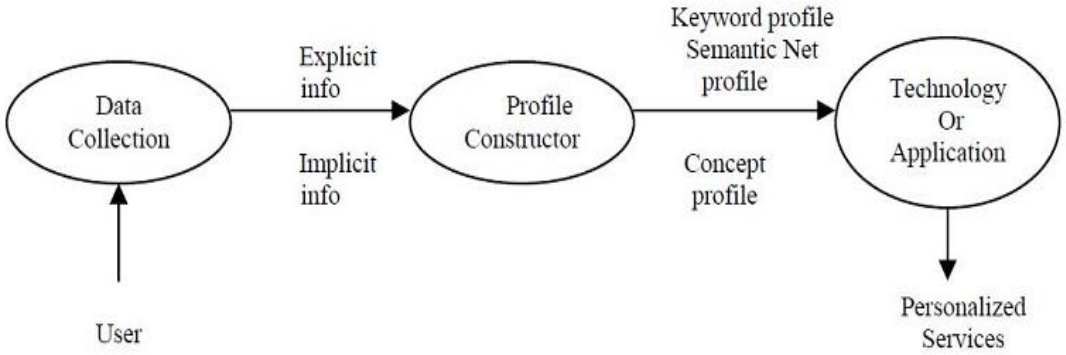
Η διαδικασία της σύστασης βασισμένη σε συστήματα content – based υλοποιείται σε τρία βήματα που περιγράφουμε παρακάτω:

Βήμα 1: Ανάλυση περιεχομένου.

Αυτό το πρώτο βήμα είναι πολύ σημαντικό στη διαδικασία της σύστασης γιατί τροφοδοτεί τα επόμενα δυο βήματα. Κάθε πληροφορία που υπάρχει διαθέσιμη για το σύστημα δεν είναι απαραίτητο ότι θα έχει συγκεκριμένη δομή ώστε να αξιολογηθεί (π.χ. ένα απλό κείμενο). Επομένως θα πρέπει να υπάρξει κάποιου είδους επεξεργασία της πληροφορίας ώστε να έρθει σε πιο δομημένη μορφή. Έτσι, σε αυτό το βήμα γίνεται η ανάλυση του περιεχομένου μιας γενικής πληροφορίας για ένα αντικείμενο ώστε να μπορέσει να αξιολογηθεί σε επόμενο βήμα [61].

Βήμα 2: Δημιουργία προφίλ.

Σε αυτό το βήμα το σύστημα προσπαθεί να συλλέξει όλα εκείνα τα χαρακτηριστικά που συνθέτουν τις προτιμήσεις ενός χρήστη ώστε να δημιουργήσει ένα κατάλληλο προφίλ για τον χρήστη αυτόν. Τη δημιουργία του προφίλ το σύστημα την επιτυγχάνει μέσω τεχνικών μηχανικής μάθησης οι οποίες αρχίζουν να μαθαίνουν στο σύστημα τι άρεσε και τι όχι σε έναν χρήστη κατά το παρελθόν [52, 61, 63]. Η εικόνα παρακάτω μας περιγράφει τον αλγόριθμο για τη δημιουργία του προφίλ ενός χρήστη.



Εικόνα 12: Αλγόριθμος για την κατασκευή προφίλ χρήστη [20]

Βήμα 3: Φιλτράρισμα των στοιχείων.

Σε αυτό το τρίτο και τελευταίο βήμα το σύστημα αξιοποιεί τις πληροφορίες για τις προτιμήσεις του χρήστη όπως προέρχονται από το προφίλ του, με το να τις ταυτοποιήσει και συσχετίσει με τα χαρακτηριστικά των αντικειμένων που προτείνονται για σύσταση. Το αποτέλεσμα αυτής της σύγκρισης είναι μια διακριτή ή συνεχής τιμή σχετικότητας του χρήστη με το αντικείμενο. Προφανώς, συστήνεται εκείνο το αντικείμενο για το οποίο η τιμή της σχετικότητας με τις προτιμήσεις του χρήστη είναι μεγαλύτερη [36, 61].

Στις παρακάτω ενότητες θα περιγράψουμε τους διάφορους αλγόριθμους που συναντάμε στη βιβλιογραφία προκειμένου να υλοποιηθεί κάθε βήμα της μεθόδου όπως περιγράψαμε παραπάνω.

3.2.2 Αλγόριθμοι Ανάλυσης Περιεχομένου

Όπως ήδη αναφέραμε, κάθε πληροφορία που καλείται το σύστημα να επεξεργαστεί και να αναλύσει προκειμένου είτε να χτίσει το προφίλ ενός χρήστη, είτε μετέπειτα να προβεί σε συστάσεις μπορεί να μην είναι πάντα σε επεξεργάσιμη μορφή. Το πλήθος των πληροφοριών που έχει στη διάθεσή του το σύστημα θα πρέπει να ανακτηθεί με τέτοιο τρόπο ή τουλάχιστον να επεξεργαστεί έτσι ώστε να είναι αξιοποιήσιμο. Για παράδειγμα, εάν η πηγή πληροφόρησης είναι ένα μήνυμα κειμένου πχ στο twitter τότε θα πρέπει να χρησιμοποιηθεί ένας αλγόριθμος έτσι ώστε από αυτό το απλό μήνυμα κειμένου να αντλήσει εκείνη την πληροφορία που θα δώσει στο σύστημα τη δυνατότητα να σκιαγραφήσει το ενδιαφέρον του χρήστη και να αποτελέσει τη βάση για κάποια σύσταση [61]. Παρακάτω θα περιγράψουμε το μοντέλο διανυσματικού χώρου (Vector Space Model VSM) το οποίο βασίζεται σε λέξεις κλειδιά προκειμένου να αντλήσει πληροφορίες από ένα απλό κείμενο.

Vector Space Model (VSM) βασισμένο σε λέξεις κλειδιά

Η τεχνική αυτή έχει ως στόχο να μοντελοποιήσει ένα έγγραφο σε ένα v-διάστατο διάνυσμα βαρών όρων. Με την έννοια βάρος (weight) νοείται ένας πραγματικός αριθμός που αντιπροσωπεύει το κατά πόσο συσχετίζεται ο εκάστοτε όρος με το έγγραφο. Έστω μια ομάδα εγγράφων η οποία αποτελείται από N έγγραφα:

$$D = \{d_1, d_2, \dots, d_N\} \quad (3.1)$$

Το βάρος w_{kj} του όρου t_k στο έγγραφο d_j εκφράζει τη συνάρτηση συχνότητας του όρου t_k στο έγγραφο d_j, του αριθμού των εγγράφων που περιλαμβάνουν τον όρο t_k και του συνολικού αριθμού εγγράφων του D.

Για τον υπολογισμό των βαρών χρησιμοποιείται η μέθοδος ανάκτησης πληροφορίας *Συχνότητα Όρου - Αντίστροφη Συχνότητα Εγγράφου (Term Frequency – Inverse Document Frequency TF-IDF)* η οποία βασίζεται σε τρεις υποθέσεις [64, 65]:

1) οι σπάνιοι όροι δεν είναι λιγότερο σχετικοί με το θέμα από τους όρους που χρησιμοποιούνται συχνότερα (ιδιότητα IDF),
2) οι συχνές αναφορές ενός όρου σε ένα έγγραφο δεν είναι περισσότερο σχετικές με το θέμα από μοναδικές αναφορές όρων (ιδιότητα TF) και

3) τα μεγάλης έκτασης έγγραφα δεν είναι προτιμότερα από τα μικρής έκτασης έγγραφα (ιδιότητα της ομαλοποίησης)

Έτσι, το βάρος ενός όρου υποδηλώνει ότι ένας όρος εμφανίζεται συχνότερα σε ένα συγκεκριμένο έγγραφο από ότι σε άλλα έγγραφα του D και άρα είναι πιο σχετικός με το θέμα που περιγράφεται στο έγγραφο. Επίσης, η ομαλοποίηση των διανυσμάτων βαρών εξασφαλίζει την απομόνωση των εγγράφων με μεγάλη έκταση.

Οι παραπάνω υποθέσεις μαθηματικά απεικονίζονται με τις παρακάτω συναρτήσεις.

Η συνάρτηση TF-IDF ορίζεται ως εξής:

$$TF - IDF(t_k, d_j) = TF(t_k, d_j) * \log\frac{N}{n_k} \qquad (3.2)$$

Όπου N είναι ο συνολικός αριθμός των εγγράφων του συνόλου D και n_k ο αριθμός των εγγράφων στα οποία εμφανίζεται τουλάχιστον μια φορά ο όρος t_k.

Επίσης ορίζεται ότι:

$$TF(t_k, d_j) = \frac{f_{k,j}}{\max_z f_{z,j}} \qquad (3.3)$$

Όπου το max υπολογίζεται πάνω στις συχνότητες $f_{z,j}$ όλων των όρων t_z που εμφανίζονται στο έγγραφο d_j.

Προκειμένου τα βάρη που υπολογίζονται από τη συνάρτηση TF-IDF να βρίσκονται στο διάστημα [0,1] και τα διανύσματα των εγγράφων να έχουν το ίδιο μήκος, τα βάρη κανονικοποιούνται με τη μέθοδο του συνημίτονου σύμφωνα με τον παρακάτω τύπο:

$$w_{kj} = \frac{TF(t_k,d_j) * \log\frac{N}{n_k}}{\sqrt{\sum_{s=1}^{|T|} TF-IDF(t_k,d_j)^2}} \qquad (3.4)$$

Η παραπάνω εξίσωση εκπληρώνει την Τρίτη υπόθεση όπως την αναφέραμε παραπάνω.

Προκειμένου να καθοριστεί η ομοιότητα μεταξύ δυο εγγράφων d_i, d_j, δηλαδή το κατά πόσο δυο έγγραφα συσχετίζονται, θα πρέπει να χρησιμοποιηθεί κάποιο κριτήριο ομοιότητας. Το συνηθέστερο κριτήριο ομοιότητας που χρησιμοποιείται για την

μέτρηση της ομοιότητας μεταξύ δυο εγγράφων είναι η ομοιότητα συνημίτονου σύμφωνα με τον παρακάτω τύπο:

$$sim(d_i, d_j) = \frac{\sum_k w_{ki} w_{kj}}{\sqrt{\sum_k {w_{ki}}^2} \sqrt{\sum_k {w_{kj}}^2}} \qquad (3.5)$$

Με βάση τα παραπάνω, στην περίπτωση συστάσεων αντικειμένων προς χρήστες η μέθοδος καταλήγει σε σύσταση μέσω της σύγκρισης της ομοιότητας βαρών συνημίτονου του διανύσματος των χαρακτηριστικών των προϊόντων και του διανύσματος του προφίλ του χρήστη [61].

3.2.3 Αλγόριθμοι δημιουργίας προφίλ

Τα συστήματα προτάσεων με βάση το περιεχόμενο όπως αναφέραμε παραπάνω, εφόσον ανακτήσουν τις απαραίτητες πληροφορίες για τον χρήστη, συνεχίζουν με τη δημιουργία ενός προφίλ για τον χρήστη που θα αποτελέσει τη βάση για να προβούν σε συστάσεις. Το σύστημα προσπαθεί με βάση το παρελθόν του χρήστη, δηλαδή με βάση τις επιλογές που έκανε ο χρήστης κατά το παρελθόν, τα προϊόντα που αγόρασε, τις αξιολογήσεις που έκανε, να χτίσει ένα προφίλ χαρακτηριστικών που θα βοηθήσουν το σύστημα να μάθει τον χρήστη και αντίστοιχα να του προτείνει εκείνα τα αντικείμενα που θα τον ικανοποιούν καλύτερα και θα τον ενδιαφέρουν.

Γενικά υπάρχουν δυο τρόποι μέσω των οποίων μπορεί να δημιουργηθεί το προφίλ του χρήστη. Ο πρώτος τρόπος είναι να γίνει χρήση διάφορων αλγορίθμων που θα αξιοποιούν διαφόρων ειδών πληροφορίες ώστε να μάθουν στο σύστημα τα ενδιαφέροντα του χρήστη που θα συνθέτουν το προφίλ του. Οι αλγόριθμοι που χρησιμοποιούνται για το σκοπό αυτό βασίζονται στο παρελθόν του χρήστη. Με βάση λοιπόν το παρελθόν του χρήστη μπορούν να χρησιμοποιήσουν πληροφορίες είτε από αγορές που έκανε ο χρήστης είτε πληροφορίες από αξιολογήσεις του χρήστη για διάφορα αντικείμενα. Ακόμη, μπορεί να χρησιμοποιηθούν και πληροφορίες από αποθηκευμένα αντικείμενα που μπορεί να έχει ο χρήστης σε διάφορες τοποθεσίες προκειμένου να αντληθούν στοιχεία για τα ενδιαφέροντά του. Αυτή η μέθοδος απαιτεί να μάθει το σύστημα τι αρέσει και τι όχι στον χρήστη κάτι το οποίο για να γίνει απαιτείται κόπος και επαναληψιμότητα όπως επίσης και διαθέσιμες πηγές πληροφόρησης όπως αναφέραμε παραπάνω.

Η άλλη μέθοδος που υπάρχει και είναι ικανή να σκιαγραφήσει το προφίλ του χρήστη είναι πιο άμεση και σχετίζεται με την άμεση καταγραφή από τον χρήστη των προτιμήσεων του. Σε αυτήν την περίπτωση ο χρήστης μόνος του χτίζει το προφίλ του μέσω απαντήσεων σε διάφορες ερωτήσεις ή μέσω βαθμολογιών σε διάφορα αντικείμενα [61].

Παρακάτω θα περιγράψουμε ορισμένους δημοφιλείς αλγορίθμους που χρησιμοποιούνται για τη δημιουργία προφίλ ενός χρήστη και οι οποίοι ουσιαστικά μαθαίνουν μια συνάρτηση που μοντελοποιεί τις προτιμήσεις του χρήστη. Βασική προϋπόθεση για την υλοποίηση τέτοιων αλγορίθμων είναι η ύπαρξη προγενέστερων αξιολογήσεων από τον χρήστη για διάφορα αντικείμενα. Έπειτα, μέσω των αξιολογήσεων αυτών συμπεραίνουν το προφίλ του χρήστη στο οποίο στηρίζεται η σύσταση εκείνων των προϊόντων που θα ενδιαφέρουν περισσότερο τον χρήστη.

Μέθοδος Naïve Bayes

Η μέθοδος αυτή αποτελεί μια πιθανολογική προσέγγιση της επαγωγικής μάθησης ανήκοντας στη γενικότερη κατηγορία των Bayesian ταξινομητών. Μέσω αυτής της προσέγγισης υπολογίζεται ένα πιθανολογικό μοντέλο που βασίζεται σε παρελθοντικά παρατηρούμενα δεδομένα. Το μοντέλο αυτό υπολογίζει την εκ των υστέρων (a posteriori) πιθανότητα $P(c|d)$ ένα έγγραφο d να ανήκει στην κατηγορία c με βάση την εκ των προτέρων πιθανότητα $P(c)$ παρατήρησης κάποιου εγγράφου στην κατηγορία c, την πιθανότητα $P(d|c)$ παρατήρησης ενός εγγράφου d δεδομένης της κατηγορίας c και τέλος με βάση την πιθανότητα $P(d)$ παρατήρησης του εγγράφου d. Με βάση λοιπόν αυτές τις πιθανότητες υπολογίζεται η εκ των υστέρων πιθανότητα $P(c|d)$ με βάση το θεώρημα του Bayes ως εξής:

$$P(c|d) = \frac{P(c)P(d|c)}{P(d)} \qquad (3.6)$$

Η κατηγορία c στην οποία ανήκει το έγγραφο d προκύπτει από την τιμή c που μεγιστοποιεί την πιθανότητα $P(c|d)$. Δηλαδή:

$$c = \arg\max c_j \frac{P(c_j)P(d|c_j)}{P(d)} \qquad (3.7)$$

Για τον υπολογισμό της δεσμευμένης πιθανότητας με τον ταξινομητή Naïve Bayes, υποτίθεται ότι όλα τα χαρακτηριστικά του εγγράφου d, δηλαδή οι λέξεις ή τα σημεία (tokens) είναι όλα ανεξάρτητα μεταξύ τους. Έτσι, ένα σημαντικό πλεονέκτημα αυτών των ταξινομητών είναι ότι είναι αρκετά ισχυροί ώστε να απομονώνουν άσχετα χαρακτηριστικά ή τιμές που λείπουν.

Υπάρχουν δυο ευρέως διαδεδομένα μοντέλα των Naïve Bayes ταξινομητών που είναι το μοντέλο πολυμεταβλητών Bernoulli και το πολυωνυμικό μοντέλο [50]. Αυτές οι δύο μέθοδοι μοντελοποιούν ένα έγγραφο με τη μορφή ενός διανύσματος τιμών, όπου κάθε είσοδος στο διάνυσμα αντιπροσωπεύει αν μια λέξη εμφανίστηκε στο κείμενο ή όχι. Για την κατηγοριοποίηση κειμένων το μοντέλο πολυμεταβλητών Bernoulli θεωρεί ότι το κάθε έγγραφο εκπροσωπείται από ένα δυαδικό διάνυσμα στο πλαίσιο όλων των λέξεων ενός λεξιλογίου σε αντίθεση με το πολυωνυμικό μοντέλο που μετρά πόσες φορές εμφανίζεται μια λέξη σε ένα κείμενο. Στην πράξη έχει αποδειχτεί ότι το πολυωνυμικό μοντέλο έχει καλύτερες επιδόσεις σε σχέση με το πολυμεταβλητό μοντέλο Bernoulli, ειδικά σε περιπτώσεις όπου υπάρχουν μεγάλα λεξιλόγια.

Με βάση το πολυωνυμικό μοντέλο, η πιθανότητα $P(c_j|d_i)$ υπολογίζεται ως εξής:

$$P(c_j|d_i) = P(c_j) \prod_{w \in V_{d_j}} P(t_k|c_j)^{N_{(d_j,t_k)}} \qquad (3.8)$$

Όπου $N_{(d_j,t_k)}$ ο αριθμός των φορών που η λέξη t_k εμφανίζεται στο έγγραφο d_j.

Ωστόσο, παρά το γεγονός ότι το πολυωνυμικό μοντέλο υπερτερεί όπως περιγράψαμε και παραπάνω, πολλές φορές οι επιδόσεις του περιορίζονται στις παρακάτω περιπτώσεις:

1) Όταν τα έγγραφα έχουν διαφορετικό μέγεθος με αποτέλεσμα να γίνεται μια πρόχειρη εκτίμηση των παραμέτρων

2) Όταν υπάρχουν λίγες κατηγορίες και άλλα λίγα διαθέσιμα δεδομένα

Οι παραπάνω περιπτώσεις εμφανίζονται συχνά κατά τη δημιουργία προφίλ ενός χρήστη όπου δε μπορεί να γίνει κάποια υπόθεση για το μέγεθος των εγγράφων καθώς επίσης και στην περίπτωση που δεν υπάρχουν κατάλληλα αρνητικά παραδείγματα με την έννοια εκείνων των αντικειμένων που δε θα επέλεγε ο χρήστης έναντι εκείνων που θα προτιμούσε [28, 70, 40].

Αλγόριθμος του Rocchio

Η ανάκτηση πληροφοριών, όπως περιγράψαμε παραπάνω βάσει του μοντέλου Space Vector Model (VSM) βασισμένο σε λέξεις κλειδιά, είναι περισσότερο αποτελεσματική όταν υπάρχει η δυνατότητα να αξιολογεί αυτές τις πληροφορίες που ανακύπτουν. Η αξιολόγηση αυτή των πληροφοριών μπορεί να γίνει μέσω αρχικών ερωτήσεων που θέτει ο χρήστης με βάση ορισμένες λέξεις κλειδιά. Έτσι, προς την κατεύθυνση της αποτελεσματικότητας της ανάκτησης πληροφορίας, έχουν διαμορφωθεί μέθοδοι που επιτρέπουν στον χρήστη να τελειοποιεί τα αρχικά του ερωτήματα βάσει παλιότερων αποτελεσμάτων αναζήτησης. Αυτές οι μέθοδοι είναι γνωστές ως ανατροφοδότηση σχετικότητας (Relevance Feedback) και επιτρέπουν στον χρήστη να αξιολογεί τα αποτελέσματα του συστήματος σχετικά με το πόσο ανταποκρίνονται στις πληροφοριακές ανάγκες του χρήστη μέσω των βαθμολογήσεων που παρέχει ο χρήστης στα αποτελέσματα του συστήματος. Όπως θα δούμε και παρακάτω στο κομμάτι της ανατροφοδότησης, υπάρχουν άμεσοι και έμμεσοι τρόποι με τους οποίους μπορεί να πραγματοποιηθεί η ανατροφοδότηση σχετικότητας.

Τη διαδικασία που περιγράψαμε παραπάνω έρχεται να υλοποιήσει ο αλγόριθμος του Rocchio όντας μια προσαρμογή της ανατροφοδότησης σχετικότητας στην κατηγοριοποίηση κειμένου που εφαρμόζεται στο VSM. Ο αλγόριθμος του Rocchio μετατρέπει ένα αρχικό ερώτημα του χρήστη μέσω διαφορετικά σταθμισμένων πρωτοτύπων συναφών και μη εγγράφων. Έτσι, προκειμένου να διαμορφωθεί το προφίλ ενός χρήστη ή να εκπαιδευτεί ένας αλγόριθμος μάθησης που χρησιμοποιεί το προφίλ του χρήστη, αξιοποιείται η βαθμολόγηση, μέσω της ανατροφοδότησης του χρήστη, των εγγράφων που ανακτήθηκαν. Παρακάτω θα περιγράψουμε τα 6 βήματα που ακολουθεί ο αλγόριθμος του Rocchio.

1. Παρουσίαση των εγγράφων ως διανύσματα. Έγγραφα με παρόμοιο περιεχόμενο έχουν και παρόμοια διανύσματα.

2. Το κάθε διάνυσμα εγγράφου έχει ως στοιχεία του τις λέξεις του εγγράφου στο οποίο αντιστοιχεί.

3. Το βάρος του κάθε στοιχείου του διανύσματος υπολογίζεται με τη χρήση της μεθόδου υπολογισμού των βαρών Συχνότητα Όρου - Αντίστροφη Συχνότητα Εγγράφου (TF-IDF) όπως περιγράψαμε στην ενότητα παραπάνω.

4. Για τη μάθηση του αλγορίθμου γίνεται συνδυασμός των διανυσμάτων εγγράφων με ένα πρωτότυπο διάνυσμα της κάθε κατηγορίας.

5. Η ταξινόμηση ενός νέου εγγράφου D γίνεται με τον υπολογισμό της ομοιότητας για κάθε κατηγορία μεταξύ των πρωτότυπων διανυσμάτων και του διανύσματος που εκπροσωπεί το έγγραφο και στη συνέχεια το έγγραφο D συγκαταλέγεται στην κατηγορία της οποίας το διάνυσμα εγγράφου έχει την υψηλότερη τιμή ομοιότητας.

6. Δημιουργία δύο πρωτότυπων εγγράφων αθροίζοντας τα διανύσματα των σχετικών και των άσχετων εγγράφων.

Ο παρακάτω τύπος παρουσιάζει μαθηματικά [14] τον αλγόριθμο του Rocchio:

$$Q_{i+1} = aQ_i + \beta \sum_{rel} \frac{D_i}{|D_i|} - \gamma \sum_{nonrel} \frac{D_i}{|D_i|} \qquad (3.9)$$

Όπου Q ισοδυναμεί με το ερώτημα του χρήστη στην επανάληψη i, D το έγγραφο i και α, β, γ οι παράμετροι που ελέγχουν την επίδραση του αρχικού ερωτήματος και των δύο πρωτοτύπων στο τροποποιημένο ερώτημα που προκύπτει [52, 56].

3.2.4 Το στάδιο της ανατροφοδότησης (Feedback)

Με τα παραπάνω βήματα όπως είδαμε αυτή η μέθοδος καταλήγει σε ορισμένα αντικείμενα για σύσταση. Ωστόσο, για να διευρυνθεί ή να ανανεωθεί το προφίλ ενός χρήστη για να χρησιμοποιηθεί ως βάση για νέες συστάσεις, θα πρέπει να υπάρξει κάποιου είδους αξιολόγηση αυτών των συστάσεων που του έχουν ήδη γίνει. Αυτό το στάδιο κατά το οποίο το σύστημα συλλέγει τις αντιδράσεις του χρήστη για τα αντικείμενα που του έχουν προταθεί λέγεται *Feedback ή Ανατροφοδότηση*. Γενικά μπορούμε να διακρίνουμε δυο επίπεδα ανατροφοδότησης: της θετικής πληροφόρησης (που αναφέρεται σε αντικείμενα και χαρακτηριστικά που άρεσαν στον χρήστη) και της αρνητικής πληροφόρησης (που αναφέρεται σε αντικείμενα και χαρακτηριστικά που δεν άρεσαν στον χρήστη).

Για την καταγραφή τώρα της ανατροφοδότησης από τον χρήστη υπάρχουν δυο διαφορετικές τεχνικές ανάλογα με τη συμμετοχή του χρήστη. Όταν το σύστημα απαιτεί από τον χρήστη να αξιολογήσει άμεσα κάποιο αντικείμενο, τότε αυτή η τεχνική

ονομάζεται *άμεση ανατροφοδότηση (explicit feedback)* ενώ όταν δεν απαιτείται η ενεργή συμμετοχή του χρήστη στην αξιολόγηση, με την έννοια ότι η αξιολόγηση συνάγεται από την ανάλυση και τον έλεγχο της δραστηριότητας του χρήστη, τότε η τεχνική ανατροφοδότησης ονομάζεται *έμμεση ανατροφοδότηση (implicit feedback).*

Η άμεση ανατροφοδότηση υποδεικνύει πόσο σχετικό ή ενδιαφέρον είναι ένα αντικείμενο που έχει συστηθεί στον χρήστη και υπάρχουν τρεις κύριες προσεγγίσεις που συμβάλουν στην άμεση ανατροφοδότηση:

1. Μέτρηση της αρέσκειας (Like/dislike): Με την προσέγγιση αυτή ο χρήστης δηλώνει αν του αρέσει κάτι και άρα δείχνει πόσο σχετικό είναι με αυτόν, ή δεν του αρέσει. Η μέτρηση του like/dislike μπορεί να γίνει είτε με το πάτημα ενός απλού κουμπιού που δηλώνει like, όπως στο facebook είτε με την αξιολόγηση με βάση κάποια κλίμακα.
2. Βαθμολογήσεις (Ratings): Για την μέτρηση της ανατροφοδότησης μπορούν επίσης να υιοθετηθούν διαφόρων ειδών βαθμολογήσεις για ένα αντικείμενο και ανάλογα με τη βαθμολογία σε μια κλίμακα να δηλώνεται η αρέσκεια ενός αντικειμένου στον χρήστη. Παράδειγμα μια τέτοιας προσέγγισης είναι τα αστεράκια που ακολουθούν ένα αντικείμενο. Έτσι αν το αντικείμενο πάρει πέντε αστεράκια σημαίνει ότι ικανοποιεί πλήρως τον χρήστη ενώ αν πάρει ένα σημαίνει ότι απλά δεν θα ξαναχρησιμοποιηθεί.
3. Σχολιασμός (text comments): Μια άλλη προσέγγιση στην άμεση ανατροφοδότηση για ένα αντικείμενο είναι τα σχόλια που γράφονται από όλους τους χρήστες που έχουν χρησιμοποιήσει το συγκεκριμένο αντικείμενο. Ανάλογα με τα σχόλια που αφήνει κάθε χρήστης δηλώνει και άμεσα το κατά πόσο ήταν ή όχι ενδιαφέρον ένα συγκεκριμένο αντικείμενο. Παράδειγμα μιας τέτοιας προσέγγισης αποτελεί το Amazon ή το eBay όπου για κάθε προϊόν υπάρχουν σχόλια που βοηθούν άλλους χρήστες να αποφασίσουν αν θα αγοράσουν ή όχι κάποιο αντικείμενο. Ωστόσο, παρότι είναι πολύ χρήσιμη αυτή η μέθοδος στη διαμόρφωση μιας άποψης για ένα αντικείμενο, πολλές φορές τα σχόλια είναι τόσα πολλά και ίσως και αντικρουόμενα που μπορεί να μπερδέψουν τον χρήστη.

Από την άλλη πλευρά, η έμμεση ανατροφοδότηση βασίζεται στο να υπολογίζεται κάθε φορά μια βαθμολογία για ένα αντικείμενο ανάλογα με τη δραστηριότητα ενός χρήστη επί αυτού του αντικειμένου. Δηλαδή ανάλογα με τον αν αποθηκεύει, εκτυπώνει, σελιδοποιεί κτλ. ένα αντικείμενο κάποιος χρήστης υποδηλώνεται το κατά πόσο

προτιμάται από αυτόν. Το πλεονέκτημα της έμμεσης ανατροφοδότησης είναι ότι δεν απαιτεί την άμεση συμμετοχή του χρήστη, ωστόσο, είναι πολύ πιθανό ακριβώς επειδή δεν υπάρχει άμεση επαφή να δημιουργηθεί λανθασμένη εντύπωση για ένα προϊόν [13, 62, 55, 58].

3.2.5 Πλεονεκτήματα των content – based συστημάτων

1. Το βασικό πλεονέκτημα των content – based συστημάτων είναι ότι χρησιμοποιούν και αξιοποιούν απλές αξιολογήσεις από τον συγκεκριμένο χρήστη για τον οποίο προορίζεται η σύσταση, σε αντίθεση με τα collaborative – filtering συστήματα που όπως θα δούμε παρακάτω χρειάζονται αξιολογήσεις και από άλλους χρήστες ώστε να εντοπίσουν τον κοντινότερο γείτονα του χρήστη, δηλαδή εκείνους τους άλλους χρήστες που έχουν τις ίδιες προτιμήσεις με τον χρήστη αφού προβαίνουν σε παρόμοιες αξιολογήσεις για τα ίδια αντικείμενα. Κατά συνέπεια, μόνο τα αντικείμενα που προτιμούνται περισσότερο από τους γείτονές του προτείνονται στον χρήστη [61].
2. Ένα άλλο πλεονέκτημα αυτών των συστημάτων είναι η διαφάνεια. Με την έννοια διαφάνεια εννοούμε ότι μπορεί να αιτιολογηθεί μια σύσταση μέσω της κατηγοριοποίησης των χαρακτηριστικών ενός αντικειμένου που προτείνεται για σύσταση. Έτσι με βάση την κατηγοριοποίηση αυτή μπορεί να αιτιολογηθεί εάν είναι αξιόπιστη ή όχι μια σύσταση. Αυτό σε άλλα σύστημα όπως τα collaborative – filtering που βασίζονται σε ανώνυμες αξιολογήσεις δε συμβαίνει [61].
3. Τέλος, ένα επίσης πολύ σημαντικό πλεονέκτημα αυτών των συστημάτων είναι ότι έχουν τη δυνατότητα να συστήσουν και νέα αντικείμενα για τα οποία δεν υπάρχει καμία αξιολόγηση. Έτσι, τα συστήματα αυτά ξεπερνούν το πρόβλημα του first – rater που αντιμετωπίζουν τα collaborative – filtering συστήματα [61].

3.2.6 Μειονεκτήματα των content – based συστημάτων

1. Ένα από τα βασικά μειονεκτήματα ενός content – based συστήματος είναι ότι έχουν ένα φυσικό περιορισμό στον αριθμό αλλά και στον τύπο των χαρακτηριστικών που μπορούν να συσχετίσουν με τα αντικείμενα που συστήνουν. Για αυτό πολλές φορές είναι αναγκαίο να υπάρχει γνώση, όπως για

παράδειγμα στην περίπτωση σύστασης ταινιών που θα πρέπει να υπάρχει γνώση για τους ηθοποιούς, τους σκηνοθέτες κτλ. Κατά συνέπεια, τα συστήματα αυτά δε μπορούν να δώσουν σωστές συστάσεις εάν δεν έχουν αρκετή πληροφόρηση για το περιεχόμενο ενός αντικειμένου ώστε να διακρίνουν τα αντικείμενα σε αυτά που θα τάσσονταν με τα ενδιαφέροντα του χρήστη και σε αυτά που δεν θα τον ενδιέφεραν [61].

2. Ένα άλλο μειονέκτημα αυτών των συστημάτων είναι ότι, όπως περιγράψαμε και παραπάνω, θα συστήσουν σε κάποιο χρήστη ένα αντικείμενο του οποίου τα χαρακτηριστικά ταιριάζουν με τα χαρακτηριστικά που θα επιθυμούσε ο χρηστής βάσει του προφίλ που έχει σκιαγραφήσει το σύστημα, με αποτέλεσμα οι συστάσεις στις οποίες προβαίνει το σύστημα να έχουν περιορισμένο βαθμό καινοτομίας. Το πρόβλημα αυτό αναφέρεται στη βιβλιογραφία ως *serendipity* και υποδηλώνει την τάση των συστημάτων αυτών να προτείνουν μη καινοτομικά αντικείμενα. Για παράδειγμα, εάν ένας χρήστης έχει αξιολογήσει μόνο ταινίες από έναν συγκεκριμένο σκηνοθέτη, τότε το σύστημα θα προτείνει ταινίες του είδους αυτού του σκηνοθέτη [61].
3. Τέλος, τα συστήματα αυτά θα πρέπει να έχουν στη διάθεσή τους αρκετές αξιολογήσεις έτσι ώστε να αποτυπώσουν τις προτιμήσεις ενός χρήστη και να προβούν σε ακριβείς συστάσεις αντικειμένων. Συνεπώς, εάν υπάρχουν λίγες αξιολογήσεις σχετικά με έναν νέο χρήστη, τότε το σύστημα δε θα μπορεί να δώσει αξιόπιστη πρόταση [61].

3.3 Collaborative - filtering (CF) Συστήματα

Η μέθοδος Collaborative Filtering (CF) αποτέλεσε τη βάση πάνω στην οποία στηρίχτηκαν τα πρώτα συστήματα συστάσεων (Tapestry, Group Lens). Το βασικό χαρακτηριστικό αυτής της μεθόδου είναι ότι δε βασίζεται σε γνώση των χαρακτηριστικών των αντικειμένων προκειμένου να παράγει συστάσεις για κάποιον χρήστη, αλλά στην ομοιότητα του χρήστη αυτού με άλλους. Με την έννοια ομοιότητα εννοούμε την ομοιότητα που προκύπτει από τη σύγκριση των αξιολογήσεων του εν λόγω χρήστη με τις αξιολογήσεις άλλων χρηστών. Έτσι, κάθε φορά που βρίσκονται χρήστες με ίδιες αξιολογήσεις, τότε το σύστημα προβαίνει σε συστάσεις καθώς οι χρήστες των οποίων οι αξιολογήσεις ομοιάζουν εμφανίζουν μεγάλη πιθανότητα να επιλέξουν τα ίδια αντικείμενα. Ουσιαστικά, το σύστημα δεν ενδιαφέρεται και ούτε αναγνωρίζει τα χαρακτηριστικά και τις ιδιότητες των αντικειμένων, παρά μόνο αρκείται στις αξιολογήσεις που δέχονται από τους χρήστες.

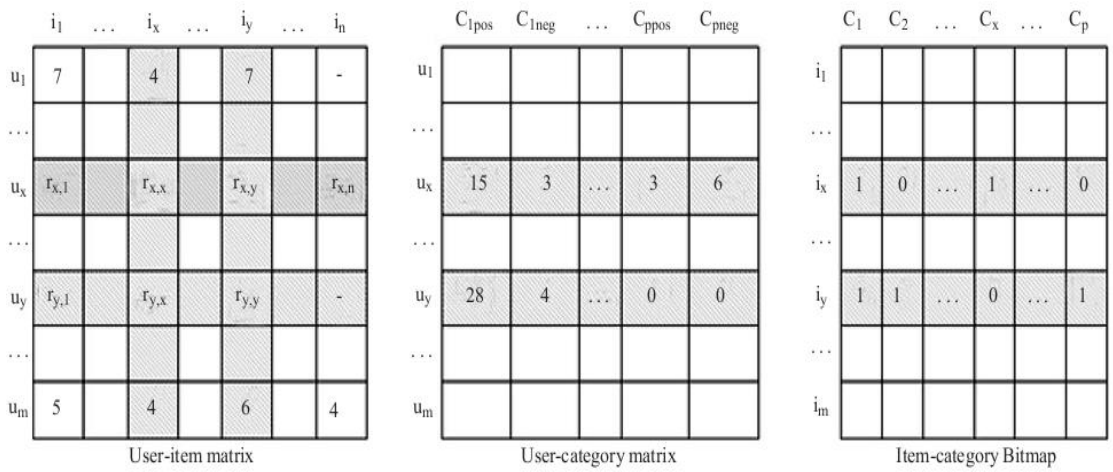
Η μέθοδος Collaborative Filtering βασίζεται κυρίως στη συλλογή ενός μεγάλου αριθμού αξιολογήσεων από τους χρήστες για διάφορα αντικείμενα, όπου οι αξιολογήσεις αυτές μπορεί να έχουν διάφορες μορφές. Η συνηθέστερη μορφή αξιολόγησης είναι αυτή της άμεσης αξιολόγησης με διακριτή κλίμακα βαθμολογίας (π.χ. από το 1 έως το 10). Με βάση αυτή τη βαθμολογία δημιουργείται ο πίνακας Collaborative Filtering Matrix (CFM) που περιέχει όλες τις αξιολογήσεις που έχουν δοθεί από όλους τους χρήστες για όλα τα αντικείμενα. Ο Collaborative Filtering Matrix είναι ένας m x n μεγέθους πίνακας, όπου m είναι ο αριθμός των αντικειμένων του συστήματος και n ο αριθμός των χρηστών. Ο πίνακας αυτός αποτελεί τη βάση λειτουργίας ολόκληρου του συστήματος.

Η διαδικασία της σύστασης με τα collaborative – filtering συστήματα υλοποιείται ακολουθώντας τα τρία βήματα που περιγράφουμε παρακάτω:

Βήμα 1: Όταν κάποιος χρήστης ζητήσει σύσταση, τότε το σύστημα συγκρίνει τις αξιολογήσεις που έχει δώσει ο χρήστης αυτός, με αυτές όλων των άλλων χρηστών που βρίσκονται στον ίδιο πίνακα αξιολογήσεων. Στην περίπτωση που η αξιολόγηση του συγκεκριμένου χρήστη δε βρίσκεται στον πίνακα αξιολογήσεων, τότε καταχωρείται εκείνη τη στιγμή. Το σύστημα επομένως προβαίνει σε σύγκριση των αξιολογήσεων και μέσω αυτής της σύγκρισης και με βάση κατάλληλο αλγόριθμο, το σύστημα βρίσκει έναν αριθμό n χρηστών που μοιάζουν περισσότερο στον συγκεκριμένο χρήστη. Οι

χρήστες αυτοί ονομάζονται *γείτονες* (neighbors) του χρήστη και το σύνολό τους *γειτονιά* του (neighborhood).

Βήμα 2: Στη συνέχεια, αφού καθοριστεί το ποιοι είναι οι κοντινότεροι *γείτονες* του χρήστη, το σύστημα ελέγχει αν υπάρχουν αντικείμενα που να μην έχουν αξιολογηθεί από τον συγκεκριμένο χρήστη ενώ ταυτόχρονα να έχουν λάβει υψηλές βαθμολογίες από τους *γείτονές* του. Το σύστημα στη συνέχεια προτείνει αυτά τα αντικείμενα στον χρήστη υπό μορφή συστάσεων είτε μέσω της μεθόδου των πιο συχνά προτιμούμενων αντικειμένων, είτε μέσω της μεθόδου των κανόνων συσχέτισης.



Εικόνα 13: Παράδειγμα μήτρας χρηστών, αντικειμένων και συσχετισμού και των δύο [84]

Κατηγορίες συστημάτων Collaborative Filtering

Οι Collaborative Filtering μέθοδοι μπορούν να διακριθούν σε δυο μεγάλες κατηγορίες [1, 10, 14, 25]:

1. Συστήματα συστάσεων Collaborative Filtering βασισμένα στη μνήμη (memory based ή neighborhood based συστήματα). Στα συστήματα αυτά οι βαθμολογήσεις χρηστών και αντικειμένων που είναι αποθηκευμένες στο σύστημα χρησιμοποιούνται άμεσα για την εξαγωγή βαθμολογιών σχετικά με νέα προϊόντα. Αυτό μπορεί να γίνει με δυο τρόπους:
    a. Μέσω της user based σύστασης
    b. Μέσω της item based σύστασης

    Τα συστήματα αυτά χαρακτηρίζονται για την απλότητά τους αφού είναι εύκολο να δημιουργηθούν και απλά στην εφαρμογή τους, αφού στην πιο απλή τους μορφή αρκεί μόνο μια παράμετρος να καθοριστεί για να μπορέσουν να

τρέξουν. Αυτή η παράμετρος είναι ο αριθμός των γειτόνων που θα χρησιμοποιηθούν για την εξαγωγή μιας σύστασης. Ένα άλλο πολύ σημαντικό κομμάτι που χαρακτηρίζει αυτά τα συστήματα είναι ότι μπορούν να αιτιολογήσουν στο χρήστη μια σύσταση. Με άλλα λόγια το σύστημα θα δώσει εξήγηση στον χρήστη για το πώς κατέληξε σε αυτήν τη σύσταση κάτι που είναι πολύ χρήσιμο για τον χρήστη αφού μπορεί άμεσα να συσχετίσει τη σύσταση αυτή με τις προτιμήσεις του. Επίσης τα συστήματα αυτά είναι σταθερά με την έννοια ότι δεν επηρεάζονται ή επηρεάζονται ελάχιστα από την προσθήκη νέων χρηστών, νέων αντικειμένων και νέων αξιολογήσεων στο σύστημά τους. Ακόμη, τα συστήματα αυτής της κατηγορίας έχουν ελάχιστες απαιτήσεις σε μνήμη ενώ είναι αρκετά αποδοτικά υπό την έννοια ότι δε χρειάζεται να εκπαιδεύσουν το σύστημα και άρα να δεσμεύσουν χρόνο και χρήμα.

2. Συστήματα συστάσεων Collaborative Filtering βασισμένα στο μοντέλο (model based συστήματα). Τα συστήματα αυτής της κατηγορίας αξιοποιούν τις βαθμολογήσεις των χρηστών ώστε να εκπαιδεύσουν ένα μοντέλο συστάσεων. Η βασική ιδέα πίσω από τα συστήματα αυτά είναι να μοντελοποιήσουν τις αλληλεπιδράσεις των χρηστών και αντικειμένων με τα χαρακτηριστικά των χρηστών και αντικειμένων που υπάρχουν στο σύστημα. Έπειτα, το μοντέλο αυτό χρησιμοποιεί τις πληροφορίες που είναι ήδη καταχωρημένες στο σύστημα για την εξαγωγή προτάσεων και συστάσεων. Στα βασικά πλεονεκτήματα των συστημάτων αυτής της κατηγορίας θα μπορούσαμε να εντάξουμε τη δυνατότητά τους να διαχειρίζονται αραιά δεδομένα χωρίς αυτό να μειώνει την αποδοτικότητα των συστημάτων αυτών και ως αποτέλεσμα τα model based συστήματα να μπορούν να χρησιμοποιηθούν σε εφαρμογές με μεγάλα datasets. Επιπλέον, όπως ήδη αναφέραμε, τα συστήματα αυτά μαθαίνουν το μοντέλο να κάνει συστάσεις και επομένως αυτή η συνεχής εκπαίδευσή του οδηγεί σε αύξηση της αποδοτικότητάς του που σε συνδυασμό με τη δυνατότητα να δίνουν διαισθητικές επεξηγήσεις για τα αποτελέσματα των συστάσεών τους τα κάνουν να τα εμπιστεύεται περισσότερο ο χρηστής. Ωστόσο τα συστήματα αυτά υστερούν έναντι των memory based συστημάτων καθώς η εκπαίδευση που απαιτούν τα κάνει να είναι πιο ακριβά και να απαιτούν κάθε τόσο τεχνικές μείωσης διαστάσεων, κάτι που μπορεί να θυσιάσει διαθέσιμες πληροφορίες που είναι ήδη αποθηκευμένες στο σύστημα.

3.3.1 User based recommendation

Οι συστάσεις με βάση τον χρήστη προβλέπουν τη βαθμολογία r_{ui} ενός χρήστη u για ένα νέο αντικείμενο i χρησιμοποιώντας τις βαθμολογήσεις που δόθηκαν για το ίδιο αντικείμενο από χρήστες παρόμοιους με τον χρήστη u δηλαδή από τους γείτονές του. Σε αυτήν την περίπτωση ορίζουμε ως w_{uv} την ομοιότητα προτιμήσεων μεταξύ του χρήστη u και του γείτονά του v υπό την συνθήκη ότι $u \neq v$ [26]. Η ομοιότητα μεταξύ δυο χρηστών βασίζεται στη βαθμολόγηση των ίδιων προϊόντων που αναπαριστώνται ως το σύνολο I_{uv} , δηλαδή I_u και I_v ενώ υπολογίζεται από τον τύπο του Pearson ως εξής:

$$PC(u,v) = \frac{\sum_{i \in I_{uv}} (r_{ui} - \overline{r_u})\,(r_{vi} - \overline{r_v})}{\sqrt{\sum_{i \in I_{uv}} (r_{ui} - \overline{r_u})^2}\,\sqrt{\sum_{i \in I_{uv}} (r_{vi} - \overline{r_v})^2}} \qquad (3.10)$$

Οι k κοντινότεροι γείτονες του χρήστη ορίζονται ως $N(u)$ και είναι οι k γείτονες v του χρήστη u με την υψηλότερη ομοιότητα w_{uv} με αυτόν. Ωστόσο, μόνο οι χρήστες που έχουν αξιολογήσει το συγκεκριμένο αντικείμενο i χρησιμοποιούνται για τον υπολογισμό της πρόβλεψης r_{ui} και για αυτό ορίζουμε ένα άλλο σύνολο γειτόνων που έχουν αξιολογήσει το αντικείμενο ως $N_i(u)$. Έτσι, η πρόβλεψη r_{ui} υπολογίζεται ως η μέση βαθμολογία που έχει δοθεί στο αντικείμενο από τους $N_i(u)$ γείτονες.

Δηλαδή:

$$\widehat{r_{ui}} = \frac{1}{|N_i(u)|} \sum_{v \in N_i(u)} r_{vi} \qquad (3.11)$$

Η παραπάνω εξίσωση δεν λαμβάνει ωστόσο υπόψη το γεγονός ότι οι γείτονες μπορεί να έχουν διαφορετικά επίπεδα ομοιότητας. Μια λύση σε αυτό το πρόβλημα είναι να σταθμιστεί η συνεισφορά κάθε γείτονα στην πρόβλεψη με βάση την ομοιότητά του ως προς τον χρήστη u. Έτσι η παραπάνω φόρμουλα παίρνει την παρακάτω μορφή:

$$\widehat{r_{ui}} = \sum_{v \in N_i(u)} w_{uv} r_{vi} \,/\, \sum_{v \in N_i(u)} |w_{uv}| \qquad (3.12)$$

Ωστόσο, και ο παραπάνω τύπος έχει ένα μειονέκτημα σε ότι αφορά τα διαφορετικά επίπεδα βαθμολογήσεων που θα δίνουν οι γείτονες για το αντικείμενο. Για παράδειγμα,

κάποιος χρήστης μπορεί να είναι πιο παρορμητικός και να αξιολογεί με υψηλές βαθμολογίες αντικείμενα που απλά του αρέσουν, ενώ κάποιοι είναι πιο συντηρητικοί και βαθμολογούν υψηλά μόνο όταν τους εκστασιάσει κάτι πάρα πολύ. Το πρόβλημα αυτό λύνεται με την κανονικοποίηση των βαθμολογιών των γειτόνων. Έτσι ο παραπάνω τύπος μετασχηματίζεται ως εξής:

$$\widehat{r_{ui}} = h^{-1}\left(\frac{\sum_{v \epsilon N_i(u)} w_{uv} h(r_{vi})}{\sum_{v \epsilon N_i(u)} |w_{uv}|}\right) \qquad (3.13)$$

[26]

3.3.2 Item based recommendation

Σε αντίθεση με την παραπάνω κατηγορία συστάσεων με βάση τον χρήστη και τους γείτονές του, στις συστάσεις με την προσέγγιση του αντικειμένου το σύστημα για να υπολογίσει την προβλεπόμενη βαθμολογία r_{ui} του χρήστη u για ένα αντικείμενο i λαμβάνει υπόψη τις βαθμολογήσεις που έχει δώσει ο χρήστης αυτός για άλλα j προϊόντα που είναι όμοια με το i. Η ομοιότητα των προϊόντων i και j και εδώ υπολογίζεται με τον τύπο του Pearson ως εξής:

$$PC(i,j) = \frac{\sum_{u \in U_{ij}} (r_{ui} - \overline{r_i})\,(r_{uj} - \overline{r_j})}{\sqrt{\sum_{u \in U_{ij}} (r_{ui} - \overline{r_i})^2} \sqrt{\sum_{u \in U_{ij}} (r_{uj} - \overline{r_j})^2}} \qquad (3.14)$$

Έτσι, αν το σύνολο των προϊόντων που έχουν βαθμολογηθεί από τον χρήστη u ορίζεται ως $N_u(i)$ τότε η πρόβλεψη της βαθμολογίας για το προϊόν i θα είναι ο σταθμισμένος με βάση την ομοιότητα μέσος όρος των βαθμολογιών του χρήστη u στα προϊόντα $N_u(i)$. Δηλαδή:

$$\widehat{r_{ui}} = \sum_{j \epsilon N_u(i)} w_{ij} r_{uj} \,/\, \sum_{j \epsilon N_u(i)} |w_{ij}| \qquad (3.15)$$

Όπως και στην προηγούμενη περίπτωση και εδώ ο παραπάνω τύπος πάσχει από το πρόβλημα των διαφορετικών βαθμολογήσεων. Ομοίως, το πρόβλημα αυτό ξεπερνιέται με την κανονικοποίηση των βαθμολογιών οπότε προκύπτει ο παρακάτω τύπος:

$$\widehat{r_{ui}} = h^{-1}\left(\frac{\sum_{j \epsilon N_u(i)} w_{ij} r_{uj}}{\sum_{j \epsilon N_u(i)} |w_{ij}|}\right) \qquad (3.16)$$

[25, 61]

Κριτήρια επιλογής user based συστημάτων και item based συστημάτων

Προκειμένου να επιλεγεί ποιο σύστημα (user based ή item based) θα εφαρμοστεί για την εξαγωγή συστάσεων είναι κρίσιμο να εξεταστούν οι παρακάτω 5 βασικοί παράγοντες [25, 61].

Ακρίβεια: Η ακρίβεια γενικά των συστημάτων που βασίζονται στη μνήμη (neighborhood recommendation systems) εξαρτάται από τον λόγο του αριθμού των χρηστών προς τα αντικείμενα που συμπεριλαμβάνονται στο σύστημα. Η ομοιότητα μεταξύ δυο χρηστών στην user based μέθοδο σύμφωνα με την οποία καθορίζονται και οι γείτονες ενός συγκεκριμένου χρήστη, υπολογίζεται συγκρίνοντας τις βαθμολογήσεις αυτών των χρηστών για το ίδιο αντικείμενο. Αντίθετα, με βάση την άλλη προσέγγιση (item based προσέγγιση) υπολογίζεται η ομοιότητα δυο αντικειμένων με βάση τη σύγκριση των αξιολογήσεων που έχουν δοθεί από τον ίδιο χρήστη για τα δυο αυτά αντικείμενα. Κατά συνέπεια, είναι προτιμότερο ένα σύστημα να βασίζεται σε λίγους αλλά πολύ όμοιους γείτονες, παρά σε πολλούς και λιγότερο όμοιους. Σε περιπτώσεις όπου ο αριθμός των χρηστών ξεπερνά κατά πολύ τον αριθμό των αντικειμένων τα συστήματα item based τείνουν να είναι πιο αποτελεσματικά. Από την άλλη πλευρά, σε περιπτώσεις όπου ο αριθμός των αντικειμένων υπερβαίνει κατά πολύ τον αριθμό των χρηστών, είναι προτιμότερο ένα σύστημα user based [32].

Αποτελεσματικότητα: Η έννοια της αποτελεσματικότητας έγκειται στη χρήση λιγότερης μνήμης και υπολογιστικής ισχύος. Η ποσότητα της μνήμης και της ισχύος που χρειάζεται ένα σύστημα εξαρτάται ομοίως από τον λόγο χρήστες προς αντικείμενα. Κατά συνέπεια, σε περιπτώσεις όπου ο αριθμός των αντικειμένων ξεπερνά τον αριθμό των χρηστών, η χρήση ενός item based συστήματος για τον υπολογισμό των βαρών ομοιότητας απαιτεί λιγότερη θυσία σε μνήμη και ισχύ από ότι ένα σύστημα βασισμένο στον χρήστη με αποτέλεσμα τα πρώτα να είναι πιο αποτελεσματικά.

Σταθερότητα: Η επιλογή ανάμεσα σε ένα σύστημα με βάση τον χρήστη και ένα σύστημα με βάση το αντικείμενο εξαρτάται, πέρα από τα παραπάνω, και από την συχνότητα αλλαγής των χρηστών και των αντικειμένων στο σύστημα, καθώς επίσης

από το μέγεθος των αλλαγών αυτών. Έτσι, σε περιπτώσεις όπου στο σύστημα ο αριθμός των αντικειμένων παραμένει σταθερός ενώ οι χρήστες εναλλάσσονται με μεγάλη συχνότητα, είναι προτιμότερα συστήματα που βασίζονται στο αντικείμενο καθώς ο υπολογισμός της ομοιότητας μεταξύ των αντικειμένων, εφόσον δεν αλλάζουν συχνά, θα είναι πιο σταθερός. Αντίθετα, σε περιπτώσεις όπου τα αντικείμενα εναλλάσσονται με μεγάλη συχνότητα ενώ οι χρήστες παραμένουν σταθεροί, συστήματα βασισμένα στους χρήστες τείνουν να είναι περισσότερο σταθερά.

Αιτιολόγηση: Ένα πλεονέκτημα των συστημάτων που βασίζονται στο αντικείμενο είναι ότι μπορούν να αιτιολογήσουν την εξαγωγή μιας σύστασης. Έτσι, η λίστα των προτεινόμενων γειτονικών προϊόντων όπως επίσης και η ομοιότητά τους μπορούν να χρησιμοποιηθούν ως αιτιολόγηση προς τον χρήστη για την σύσταση που έκανε το σύστημα. Με αυτόν τον τρόπο, ο χρήστης μπορεί άμεσα να συμμετέχει ενεργά στη διαδικασία της σύστασης είτε μέσω της τροποποίησης της λίστας των γειτονικών προϊόντων, είτε μέσω της αλλαγής των βαρών ομοιότητάς τους. Αντίθετα, στα συστήματα με βάση τον χρήστη αυτή η δυνατότητα δεν είναι εφικτή καθώς ο χρήστης δε γνωρίζει τους γείτονές του και έτσι δεν μπορεί να τροποποιήσει τη λίστα τους.

Καινοτομία: Η καινοτομία με την έννοια της σύστασης νέων και διαφορετικών προϊόντων προς τον χρήστη είναι χαρακτηριστικό των user based συστημάτων και όχι των item based. Αυτό συμβαίνει γιατί τα συστήματα που βασίζονται στα αντικείμενα εκτιμούν τη βαθμολογία για ένα συγκεκριμένο αντικείμενο με βάση τις βαθμολογίες που έχει δώσει ένας συγκεκριμένος χρήστης για άλλα παρόμοια προϊόντα με το αντικείμενο για το οποίο εκτιμάται η βαθμολογία. Συνεπώς, το σύστημα τείνει να προτείνει στον χρήστη αντικείμενα που εντάσσονται στις προτιμήσεις του χρήστη και είναι παρόμοια με αυτά που έχει επιλέξει στο παρελθόν. Έτσι, παρά το γεγονός ότι αυτό οδηγεί σε μια ασφαλή σύσταση προς τον χρήστη, με την έννοια ότι θα προταθεί κάτι στον χρήστη που σίγουρα θα τον ενδιαφέρει, δε βοηθά τον χρήστη να ανακαλύψει κάποιο αντικείμενο το οποίο ναι μεν τον ενδιαφέρει αλλά είναι περισσότερο διαφορετικό από εκείνα που έχει ήδη επιλέξει κατά το παρελθόν.

3.3.3 Επιλογή των γειτόνων (Neighbors selection)

Ένας πολύ κρίσιμος παράγοντας στην επιτυχή σύσταση του συστήματος είναι η διαδικασία επιλογής των γειτόνων. Από τη σωστή επιλογή των γειτόνων ενός χρήστη

ουσιαστικά κρίνεται η ποιότητα της σύστασης που θα του προταθεί. Στη διαδικασία επιλογής των γειτόνων διακρίνουμε δυο βήματα [32]:

1. Αρχικό φιλτράρισμα των υποψηφίων γειτόνων κατόπιν του οποίου χρησιμοποιούνται οι πλέον κατάλληλοι με βάση την μεγαλύτερη ομοιότητα με τον χρήστη στόχο.
2. Χρησιμοποίηση των γειτόνων που έχουν επιλεχθεί από το βήμα 1 στη διαδικασία προβλέψεων και εξαγωγής συστάσεων.

Φιλτράρισμα των γειτόνων

Σε περιπτώσεις όπου το σύστημα έχει να διαχειριστεί έναν πραγματικά τεράστιο όγκο αντικειμένων και χρηστών η διαδικασία να ξεχωρίσει τους εντελώς ανόμοιους χρήστες, δηλαδή χρήστες των οποίων η ομοιότητα είναι μηδενική, δεν είναι εφικτή εξαιτίας των υπερβολικών απαιτήσεων σε μνήμη του συστήματος. Επιπλέον, έστω και αν αυτό ήταν δυνατό, τότε και πάλι δε θα ήταν εποικοδομητικό καθώς δε θα συνείσφερε ουσιαστικά στη διαδικασία της εξαγωγής σύστασης από το σύστημα αφού για την πραγματοποίηση των προβλέψεων χρησιμοποιούνται οι πλέον όμοιοι χρήστες ή αντικείμενα. Ωστόσο, το φιλτράρισμα των πιο όμοιων χρηστών ή αντικειμένων είναι σκόπιμο έτσι ώστε το σύστημα να μπορεί εύκολα μέσα από ένα τεράστιο πλήθος πληροφοριών να εξάγει τα βάρη ομοιότητας περιορίζοντας έτσι τον αριθμό των γειτόνων που θα χρησιμοποιηθούν στη σύσταση και αυξάνοντας με αυτόν τον τρόπο την αποτελεσματικότητα του συστήματος.

Υπάρχουν διάφοροι τρόποι μέσω των οποίων μπορεί να γίνει το φιλτράρισμα των γειτόνων και τους τρόπους αυτούς τους παρουσιάζουμε παρακάτω:

Τοπ - Ν φιλτράρισμα

Αυτή η μέθοδος φιλτράρει τους γείτονες ενός χρήστη παίρνοντας μια λίστα από τους Ν πιο κοντινούς γείτονες του συγκεκριμένου χρήστη. Η επιλογή των κοντινότερων γειτόνων φυσικά γίνεται με βάση την ομοιότητα, όπως περιγράψαμε παραπάνω ότι υπολογίζεται.

Ωστόσο, σε αυτή τη μέθοδο θα πρέπει να δοθεί ιδιαίτερη προσοχή στην παράμετρο Ν, δηλαδή στον αριθμό των κοντινότερων γειτόνων που θα χρησιμοποιηθούν. Εάν ο αριθμός Ν είναι πολύ μεγάλος, τότε αυτό σημαίνει ότι το σύστημα θα χρειαστεί αρκετή μνήμη για να αποθηκεύσει τα βάρη ομοιότητας των χρηστών και επιπλέον, εξαιτίας του μεγάλου αριθμού των γειτόνων η ταχύτητα εξαγωγής της σύστασης από το σύστημα θα είναι μικρή. Αντίθετα, αν ο αριθμός Ν των κοντινότερων γειτόνων είναι πολύ μικρός, τότε υπάρχει η πιθανότητα η κάλυψη της σύστασης να είναι περιορισμένη κάτι που σημαίνει ότι ορισμένα αντικείμενα μπορεί να μην προταθούν ποτέ.

Φιλτράρισμα με βάση μια τιμή κατώφλι

Με βάση αυτή τη μέθοδο, αντί να επιλέγεται ένας συγκεκριμένος αριθμός γειτόνων με την υψηλότερη ομοιότητα με τον χρήστη, επιλέγονται γείτονες κατόπιν της σύγκρισης της ομοιότητάς τους με μια τιμή κατώφλι. Με άλλα λόγια, με βάση αυτή τη μέθοδο θα επιλεγούν εκείνοι οι γείτονες ενός χρήστη των οποίων η ομοιότητα είναι πάνω από μια κριτική τιμή που τίθεται από το σύστημα. Το βασικό μειονέκτημα αυτού του τρόπου επιλογής των γειτόνων είναι η δυσκολία του να καθοριστεί μια τέτοια τιμή. Ωστόσο, ο τρόπος αυτός γενικά είναι πιο ευέλικτος και πιο αποτελεσματικός όσον αφορά την επιλογή των γειτόνων έναντι της προηγούμενης μεθόδου.

Αρνητικό φιλτράρισμα

Το αρνητικό φιλτράρισμα των γειτόνων γενικά θεωρείται ότι μειονεκτεί έναντι του θετικού φιλτραρίσματος των γειτόνων. Αυτό συμβαίνει καθώς, ενώ με το θετικό φιλτράρισμα παίρνουμε ομάδες χρηστών που έχουν τις ίδιες προτιμήσεις και άρα είναι ξεκάθαρο ότι ανήκουν στην ίδια ομάδα, στην περίπτωση του αρνητικού φιλτραρίσματος των γειτόνων ουσιαστικά υποδεικνύει ότι οι χρήστες έχουν διαφορετικές προτιμήσεις και άρα ανήκουν σε διαφορετικές κατηγορίες, ωστόσο, αυτό από μόνο του δεν προσθέτει επιπλέον πληροφόρηση καθώς δε διευκρινίζεται πόσο διαφορετικές είναι αυτές οι ομάδες στις οποίες ανήκουν οι διαφορετικοί χρήστες. Εμπειρικές μελέτες έχουν δείξει ότι το αρνητικό φιλτράρισμα των γειτόνων δε συμβάλει περισσότερο στη δημιουργία προβλέψεων και εξαγωγή συστάσεων, ωστόσο, η απόρριψη ενός αρνητικού φιλτραρίσματος ως αναποτελεσματική μέθοδο επιλογής γειτόνων εξαρτάται κάθε φορά από τα δεδομένα που έχει στη διάθεσή του το σύστημα.

Στο σημείο αυτό θα πρέπει να επισημάνουμε, ύστερα και από την παραπάνω παράθεση των τρόπων επιλογής των γειτόνων, ότι οι παραπάνω μέθοδοι δεν είναι αμοιβαία αποκλειόμενες με την έννοια ότι μπορεί να χρησιμοποιηθεί συνδυασμός των παραπάνω μεθόδων ώστε να γίνει το αρχικό φιλτράρισμα των γειτόνων. Για παράδειγμα, μπορεί να χρησιμοποιηθεί το αρνητικό φιλτράρισμα και για εκείνους τους χρήστες που θα προκύψουν να γίνει χρήση της μεθόδου φιλτραρίσματος με βάση μια τιμή κατώφλι [32,37].

Χρήση γειτόνων στις προβλέψεις

Εφόσον πραγματοποιηθεί το αρχικό φιλτράρισμα των γειτόνων, το επόμενο στάδιο είναι να χρησιμοποιηθούν οι *k* κοντινότεροι γείτονες για την εξαγωγή προβλέψεων βαθμολογίας για τα αντικείμενα. Επισημαίνουμε ότι οι *k* κοντινότεροι γείτονες είναι εκείνοι που έχουν την μεγαλύτερη ομοιότητα με τον χρήστη για τον οποίο προορίζεται η σύσταση. Το πρόβλημα που υπάρχει εδώ για τον υπολογισμό των προβλέψεων είναι το μέγεθος *k* των κοντινότερων γειτόνων, δηλαδή με άλλα λόγια ο αριθμός των γειτόνων που θα χρησιμοποιηθούν για τον υπολογισμό της πρόβλεψης. Ένας μικρός αριθμός κοντινότερων γειτόνων, για παράδειγμα $k < 20$, έχει ως αποτέλεσμα να δημιουργεί προβλήματα αξιοπιστίας για την πρόβλεψη και έτσι το σύστημα συστάσεων να πάσχει από αξιοπιστία και ακρίβεια στη σύσταση. Καθώς ο αριθμός των κοντινότερων γειτόνων αυξάνει, η ακρίβεια και η αξιοπιστία του συστήματος και των συστάσεων βελτιώνεται ολοένα. Ωστόσο, η ακρίβεια του συστήματος αρχίζει να φθίνει καθώς χρησιμοποιούνται όλο και περισσότεροι κοντινότεροι γείτονες, για παράδειγμα $k > 50$. Έχει παρατηρηθεί στη βιβλιογραφία ότι ένας αριθμός κοντινότερων γειτόνων μεταξύ 20 και 50 είναι ικανός να προσδώσει στο σύστημα τα χαρακτηριστικά της ακρίβειας και της αξιοπιστίας στις προβλέψεις και στις συστάσεις. Ωστόσο, ένας πιο επιστημονικός τρόπος είναι χρήσιμο να χρησιμοποιηθεί για τον καθορισμό του βέλτιστου αριθμού των κοντινότερων γειτόνων. Συνηθέστερα, η μέθοδος που χρησιμοποιείται για τον καθορισμό του βέλτιστου *k* είναι η μέθοδος του *cross validation*, σύμφωνα με την οποία το σύστημα τρέχει τη διαδικασία εξαγωγής πρόβλεψης για διάφορες τιμές *k* και κάθε φορά αξιολογεί την αποτελεσματικότητα της πρόβλεψης. Φυσικά, επιλέγεται εκείνος ο αριθμός των κοντινότερων γειτόνων που συμβάλει στην μέγιστη αποτελεσματικότητα του συστήματος [25, 61].

Με βάση την παραπάνω περιγραφή των συστημάτων Collaborative Filtering στο σημείο αυτό θα ήταν χρήσιμο να αξιολογήσουμε συνολικά αυτή τη μέθοδο. Παρακάτω λοιπόν παραθέτουμε συνολικά τα πλεονεκτήματα και μειονεκτήματα των Collaborative Filtering συστημάτων.

3.3.4 Πλεονεκτήματα των collaborative – filtering συστημάτων

1. Το βασικό πλεονέκτημα της μεθόδου αυτής είναι ότι δεν υπάρχει ανάγκη διατήρησης βάσης γνώσης για τα χαρακτηριστικά των αντικειμένων. Τα συστήματα αυτά βασίζονται μόνο στη συσχέτιση χρηστών μεταξύ τους, μέσω των αξιολογήσεων που έχουν δώσει για τα διάφορα αντικείμενα. Έτσι, δεν απαιτείται μεγάλη και πολύπλοκη βάση δεδομένων που να στηρίζεται στη γνώση. Στην πράξη, τα συστήματα αυτά δε χρειάζεται να έχουν καμία γνώση για τα αντικείμενα που προτείνουν και αξιολογούν κάνοντας αυτά τα συστήματα εύκολα στην ανάπτυξή τους και γρήγορα στη χρήση τους.
2. Ένα άλλο πλεονέκτημα είναι ότι η ποιότητα των συστάσεων του συστήματος δεν είναι στατική αλλά βελτιώνεται με τον χρόνο. Ο πίνακας των αντικειμένων που περιλαμβάνει τις αξιολογήσεις για αυτά από τους χρήστες και στον οποίο βασίζονται τα συστήματα collaborative filtering προκειμένου να βρουν τους *γείτονες* κάθε τρέχοντος χρήστη, εμπλουτίζεται διαρκώς με καινούριες αξιολογήσεις. Έτσι η ποιότητα των συστάσεων που μπορούν να δώσουν αυτά τα συστήματα βελτιώνεται συνεχώς με τη χρήση τους.

3.3.5 Μειονεκτήματα των collaborative – filtering συστημάτων

1. Το πρόβλημα *ramp–up* [11,12]. Το πρόβλημα αυτό προκύπτει από το γεγονός ότι η ποιότητα των συστάσεων βασίζεται σε ένα μεγάλο σύνολο δεδομένων. Δηλαδή, προκειμένου το σύστημα να δώσει αξιόπιστες συστάσεις, πρέπει να υπάρχει ήδη ένας αρκετά μεγάλος αριθμός αξιολογήσεων από χρήστες. Ωστόσο, ο πίνακας στον οποίο βασίζεται η μέθοδος, λαμβάνει αξιολογήσεις προοδευτικά με αποτέλεσμα το σύστημα να μη μπορεί να δώσει αξιόπιστες συστάσεις όταν δεν υπάρχουν αρκετές αξιολογήσεις στον πίνακα από τους χρήστες, ιδιαίτερα στα πρώτα στάδια λειτουργίας του συστήματος.
2. Το πρόβλημα *sparsity* [11, 12]]. Το πρόβλημα αυτό έγκειται στην ύπαρξη πολλών αντικειμένων και λίγων χρηστών. Όταν κάθε χρήστης του συστήματος δεν αξιολογεί μεγάλο αριθμό αντικειμένων, τότε το σύστημα θα αντιμετωπίσει

το πρόβλημα *sparsity*. Ο πίνακας αντικειμένων-χρηστών, ο οποίος περιέχει τις αξιολογήσεις θα παραμένει αραιός δυσκολεύοντας με αυτόν τον τρόπο την εύρεση των κατάλληλων γειτόνων των χρηστών για να βασιστεί η σύσταση.

3. Το πρόβλημα *first rater* [15]. Τα συστήματα αυτής της κατηγορίας προβαίνουν σε συστάσεις αντικειμένων προς τους χρήστες ανάλογα με τις αξιολογήσεις που έχουν δεχθεί αυτά από άλλους χρήστες. Κατά συνέπεια, ένα νέο αντικείμενο δε μπορεί να συσταθεί αμέσως, παρά μόνο όταν δεχθεί ικανό αριθμό αξιολογήσεων από τους χρήστες του συστήματος.
4. Τα συστήματα Collaborative Filtering εμφανίζουν δυσκολία στις συστάσεις προς κάποιο χρήστη με ιδιαίτερες απαιτήσεις (popularity bias) [15]. Για όλες τις μεθόδους αυτών των συστημάτων ισχύει ότι όσο πιο συνηθισμένες προτιμήσεις έχει ένας χρήστης, τόσο πιο εύκολη θα είναι η εύρεση γειτόνων και κατ' επέκταση, η παροχή συστάσεων προς έναν τέτοιο χρήστη θα είναι εύκολη. Ωστόσο, για έναν χρήστη με ιδιαίτερες και ασυνήθιστες προτιμήσεις και ενδιαφέροντα, θα είναι πολύ δύσκολο να βρεθεί μία γειτονιά με αρκετά μέλη ώστε να λειτουργήσουν σωστά οι αλγόριθμοι ομοιότητας όπως του Pearson.
5. Τα Collaborative Filtering συστήματα συνήθως έχουν ανάγκη από μεγάλη υπολογιστική ισχύ [59] με την έννοια ότι οι αλγόριθμοι που εκτελούνται για να εντοπίσουν τους κοντινότερους γείτονες ενός χρήστη, έχουν ανάγκη σημαντικής υπολογιστικής ισχύος. Η ανάγκη αυτή είναι μεγαλύτερη όσο περισσότερους χρήστες και αντικείμενα καλείται να εξετάσει και να διαχειριστεί το σύστημα.
6. Τα συστήματα αυτά εμφανίζονται να μην είναι πολύ ευαίσθητα σε τυχόν αλλαγές των προτιμήσεων των χρηστών [11]. Εφόσον ο πίνακας αντικειμένων και χρηστών περιέχει αξιολογήσεις των αντικειμένων από τους χρήστες, οι αξιολογήσεις αυτές αποτελούν ουσιαστικά ένα ιστορικό των προτιμήσεων του χρήστη. Έτσι αυτές οι αξιολογήσεις συνθέτουν ένα ιστορικό των προτιμήσεων του χρήστη με αποτέλεσμα κάποια αλλαγή να μην έχει σημαντική επίδραση σε αυτό το ιστορικό αξιολογήσεων του χρήστη. Επομένως, η ευαισθησία του συστήματος στις αλλαγές προτιμήσεων των χρηστών είναι αρκετά μικρή, και τόσο μικρότερη, όσο πιο εκτεταμένο χρονικά είναι το αρχείο των αξιολογήσεων που διατηρείται.
7. Τέλος, ένα άλλο πρόβλημα που αντιμετωπίζουν τα συστήματα αυτά είναι το πρόβλημα της σύστασης συχνά προτιμώμενων αντικειμένων (πρόβλημα της

μπανάνας) *(banana problem)* [61]. Το πρόβλημα αυτό συνδέεται στενά με τον τρόπο λειτουργίας των συστημάτων collaborative filtering δεδομένου ότι τα συστήματα αυτής της κατηγορίας έχουν την τάση να συστήνουν περισσότερο αντικείμενα που προτιμούνται συχνά από πολλούς χρήστες.

3.4 Δημογραφικά συστήματα συστάσεων

Τα δημογραφικά συστήματα συστάσεων βασίζονται στα δημογραφικά χαρακτηριστικά των χρηστών έτσι ώστε να προβούν σε συστάσεις. Με τον όρο δημογραφικά χαρακτηριστικά νοούνται το φύλο, η ηλικία, το μορφωτικό επίπεδο, η καταγωγή, η κοινωνική τάξη κτλ. Ένα τέτοιο σύστημα λοιπόν θα διαμορφώσει ένα προφίλ για κάθε χρήστη ανάλογα με αυτά τα δημογραφικά χαρακτηριστικά λαμβάνοντας υπόψη του τις προτιμήσεις του χρήστη ανά χαρακτηριστικό. Έτσι θα διαμορφωθεί ένας δημογραφικός χάρτης μέσα από τον οποίο θα παράγονται συστάσεις σε ένα χρήστη ανάλογα με την ηλικία του, το μορφωτικό του επίπεδο και τα άλλα δημογραφικά του χαρακτηριστικά λαμβάνοντας υπόψη τις συνολικές προτιμήσεις αυτής της εκάστοτε δημογραφικής ομάδας [61].

3.4.1 Πλεονεκτήματα των Δημογραφικών συστημάτων

1. Το βασικό πλεονέκτημα της δημογραφικής προσέγγισης σύστασης είναι ότι τα συστήματα αυτά δε βασίζονται σε πρότερες αξιολογήσεις του χρήστη και έτσι οι νέοι χρήστες μπορούν να πάρουν συστάσεις πριν αξιολογήσουν οποιοδήποτε αντικείμενο.
2. Ένα άλλο πλεονέκτημα αυτών των συστημάτων είναι ότι δεν απαιτείται γνώση σχετικά με τα αντικείμενα και τα χαρακτηριστικά τους για να προβούν σε μια σύσταση.

3.4.2 Μειονεκτήματα των Δημογραφικών συστημάτων

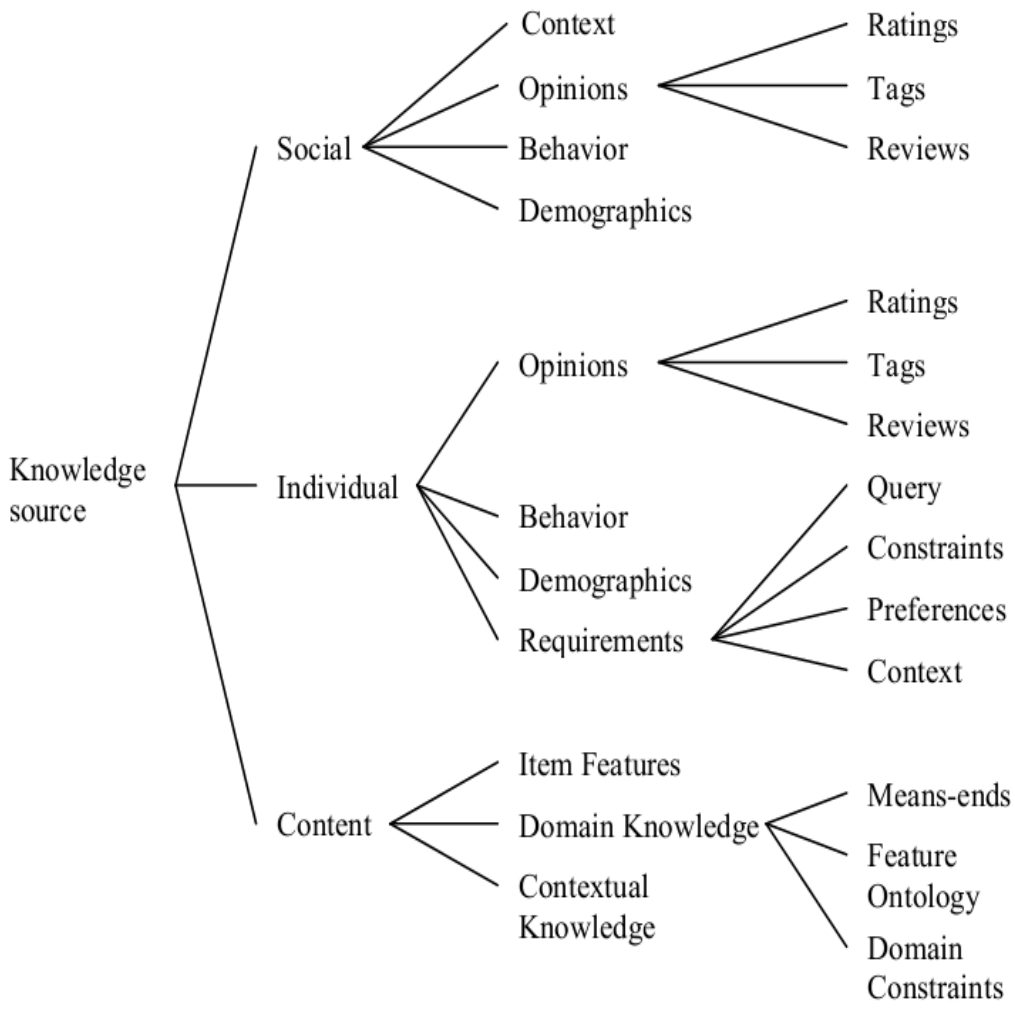
1. Το βασικότερο μειονέκτημα των συστημάτων αυτών είναι ότι δεδομένου του τρόπου σύστασης με τη συλλογή στοιχείων για τα δημογραφικά χαρακτηριστικά των χρηστών προκύπτουν ζητήματα προστασίας της ιδιωτικότητας των χρηστών.

2. Επίσης, ένα άλλο μειονέκτημα είναι ότι η δημογραφική κατηγοριοποίηση είναι πάρα πολύ αργή για εξατομικευμένες συστάσεις με αποτέλεσμα να μειώνεται έτσι η αποδοτικότητα αυτών των συστημάτων.
3. Τα συστήματα αυτά αδυνατούν να δώσουν αξιόπιστες συστάσεις για εξειδικευμένους χρήστες και αντικείμενα.
4. Τέλος, τα συστήματα αυτά αντιμετωπίζουν δυσκολίες στη δημιουργία συστάσεων κάθε φορά που αλλάζουν τα δημογραφικά χαρακτηριστικά του χρήστη.

3.5 Knowledge - based Συστήματα

Τα Knowledge – based συστήματα συστάσεων αξιοποιούν τη γνώση σχετικά με τα αντικείμενα ώστε να προβούν σε συστάσεις. Με άλλα λόγια, τα συστήματα αυτά προσπαθούν να δημιουργήσουν γνώση για τα χαρακτηριστικά των αντικειμένων που είναι υποψήφια για σύσταση ή γνώση για τους χρήστες για τους οποίους προορίζεται η σύσταση και βάσει αυτής της αποκτημένης γνώσης για τα χαρακτηριστικά τους να προβούν σε συστάσεις. Η γνώση πάνω στην οποία βασίζονται τα συστήματα αυτής της κατηγορίας μπορεί να διακριθεί σε τρία επίπεδα [18, 19]:

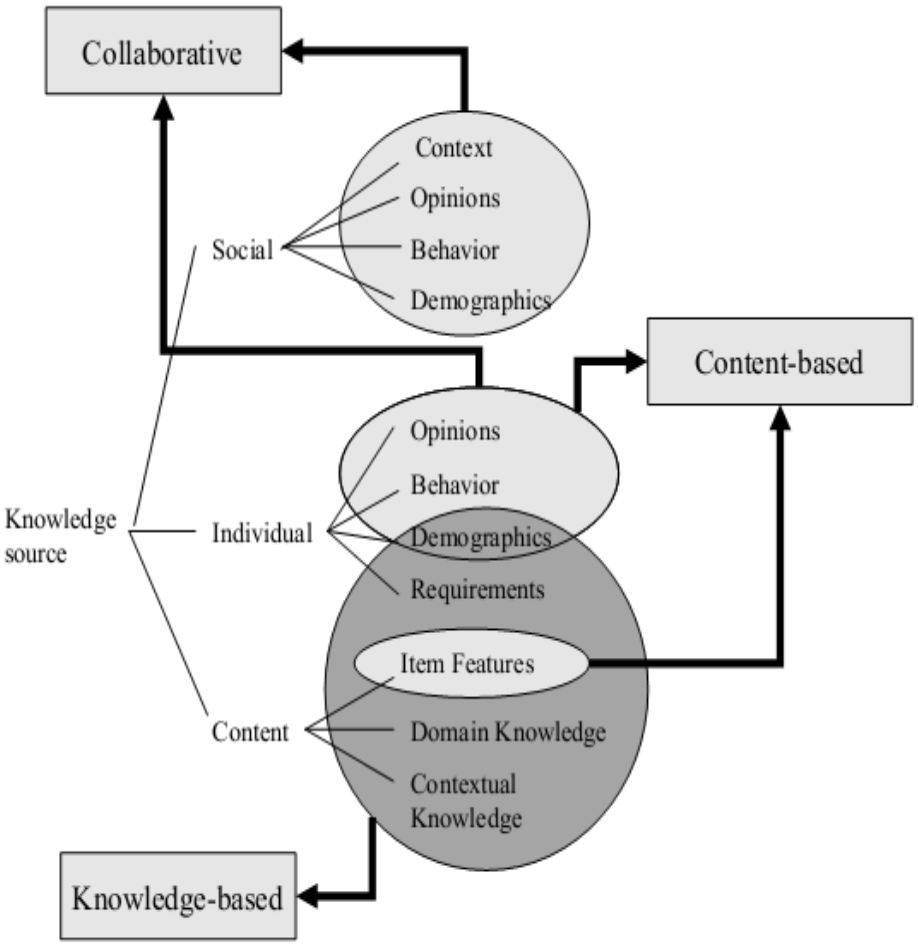
- Επίπεδο 1: Γνώση για τα αντικείμενα: Αυτό το επίπεδο γνώσης αναφέρεται στη γνώση που πρέπει να έχει το σύστημα σχετικά με τα αντικείμενα, δηλαδή με τα ιδιαίτερα χαρακτηριστικά και ιδιότητες των αντικειμένων καθώς επίσης και γνώση για το πως τα αντικείμενα σχετίζονται μεταξύ τους.
- Επίπεδο 2: Γνώση για τους χρήστες: Το σύστημα πέρα από τη γνώση για τα αντικείμενα θα πρέπει να έχει και γνώση σχετικά με τους χρήστες. Θα πρέπει δηλαδή να γνωρίζει τα χαρακτηριστικά και τις επιθυμίες του κάθε χρήστη και αυτό πραγματοποιείται με τη δημιουργία ενός προφίλ για κάθε χρήστη που αντανακλά ακριβώς τη γνώση που έχει το σύστημα για τον συγκεκριμένο χρήστη.
- Επίπεδο 3: Γνώση για τον τρόπο κάλυψης των αναγκών: Αυτό το επίπεδο γνώσης προκύπτει από τα δυο παραπάνω επίπεδα αφού σχετίζεται με τη γνώση του τρόπου με τον οποίο μπορούν τα διάφορα αντικείμενα να ικανοποιήσουν συγκεκριμένους χρήστες. Με άλλα λόγια τα συστήματα έχουν γνώση για το ποια αντικείμενα ταιριάζουν στο προφίλ κάθε χρήστη.



Εικόνα 14: Κατηγοριοποίηση της γνώσης [61]

Η παραπάνω εικόνα αποτυπώνει την κατηγοριοποίηση της γνώσης όπως αυτή χρησιμοποιείται από τα διάφορα συστήματα για την εξαγωγή συστάσεων [61].

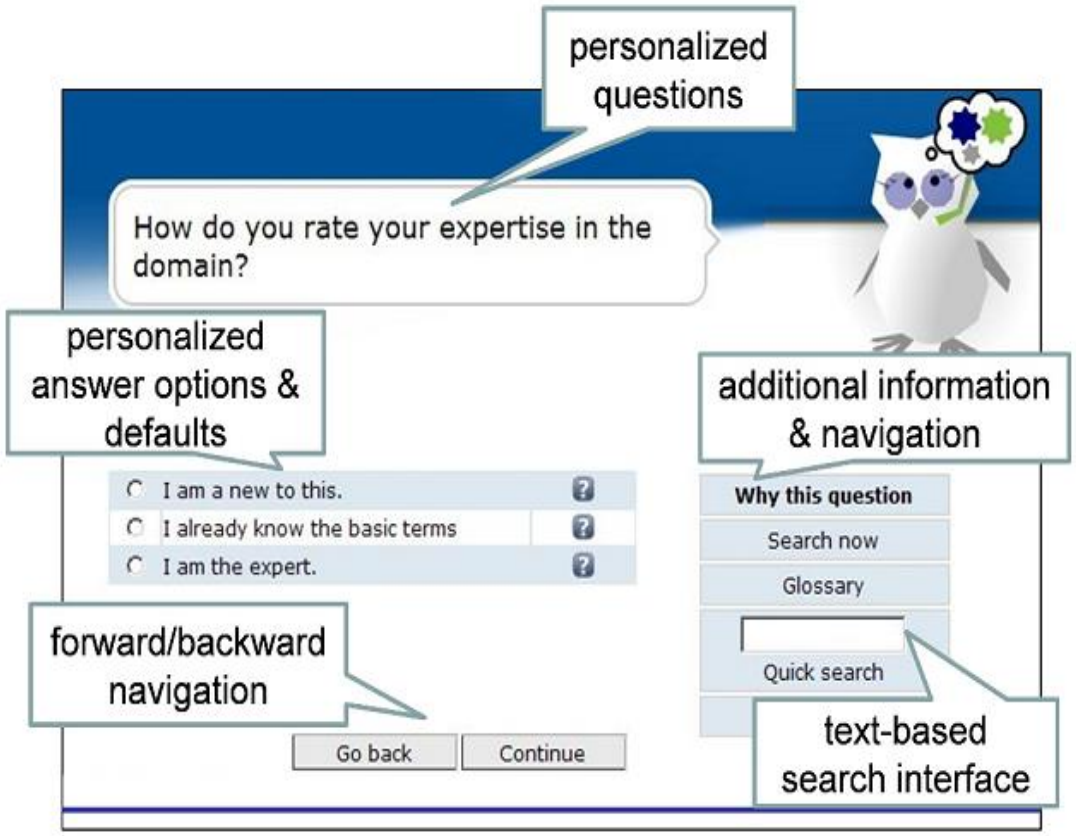
Σε συνέχεια της παραπάνω εικόνας, και ανάλογα με το είδος της γνώσης που χρησιμοποιεί ένα σύστημα εντάσσεται και στην ανάλογη κατηγορία. Παρακάτω μπορούμε να παρατηρήσουμε αυτόν τον διαχωρισμό των συστημάτων συστάσεων σε κατηγορίες ανάλογα με τη γνώση που χρησιμοποιούν για να προβούν σε προβλέψεις και εξαγωγή συστάσεων χρηστών ή αντικειμένων.



Εικόνα 15: Τύποι συστημάτων συστάσεων ανάλογα με την κατηγορία της γνώσης που χρησιμοποιούν [61]

Το πως θα αποκτηθεί η γνώση που περιγράψαμε παραπάνω αλλά και ο τρόπος με τον οποίο θα χρησιμοποιηθεί είναι χαρακτηριστικά των διαφόρων ειδών συστημάτων αυτής της κατηγορίας. Η γνώση σχετικά με τα αντικείμενα μπορεί να αντλείται με έμμεσο τρόπο ή άμεσο τρόπο.

- Ο έμμεσος τρόπος άντλησης γνώσης απαιτεί από το σύστημα τη χρήση τεχνικών εξόρυξης γνώσης (data mining) που εφαρμόζονται σε μεγάλους όγκους δεδομένων όπου το σύστημα διακρίνει ποια αντικείμενα θα συμπεριλάβει στη βάση δεδομένων και με ποια χαρακτηριστικά.
- Ο άμεσος τρόπος άντλησης γνώσης απαιτεί τα αντικείμενα και τα χαρακτηριστικά του χρήστη να εισάγονται στο σύστημα απευθείας.



Εικόνα 16: Παράδειγμα άμεσου τρόπου δημιουργίας γνώσης στο σύστημα από τον χρήστη [61]

Με τους δύο παραπάνω τρόπους το σύστημα καταφέρνει να δημιουργήσει γνώση με την έννοια ότι δημιουργεί μια βάση δεδομένων αναφορικά με τα αντικείμενα και τα χαρακτηριστικά τους έτσι ώστε να μπορέσει να τα συσχετίσει μεταξύ τους αλλά και με τους χρήστες ώστε να προβεί σε συστάσεις.

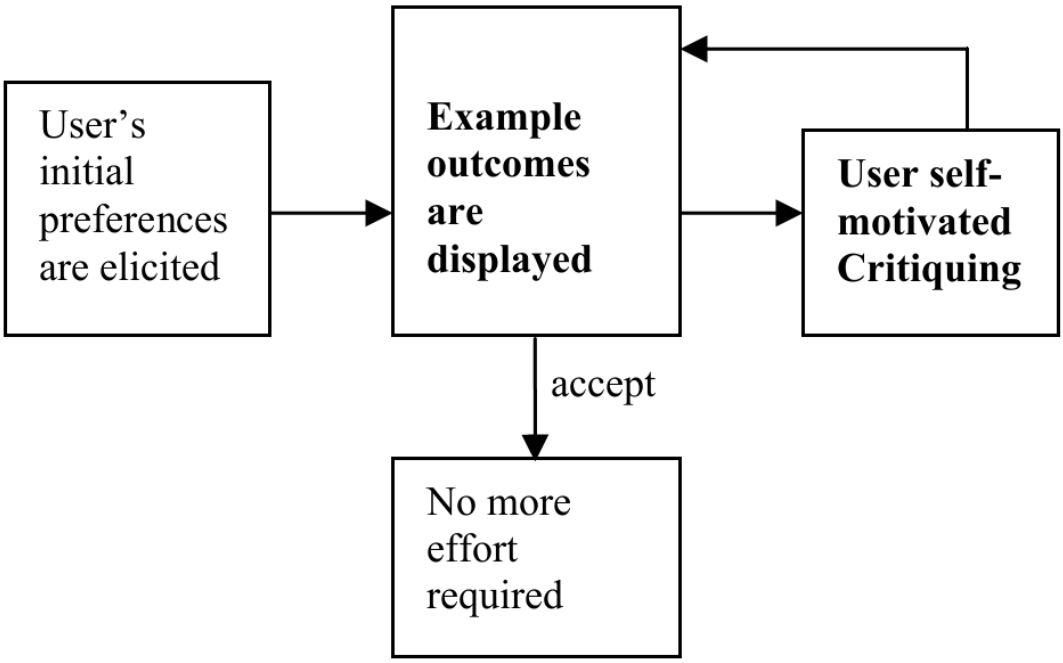
Παρακάτω θα περιγράψουμε δυο είδη των συστημάτων που βασίζονται στη γνώση καθένα από τα οποία δημιουργεί κάποιο είδος γνώσης όπως περιγράψαμε παραπάνω για να προβεί σε συστάσεις.

3.5.1 Critique-based recommendation

Τα Critique-based recommendation systems είναι συστήματα που αξιοποιούν το δεύτερο είδος γνώσης μέσω της κατασκευής ενός προφίλ για τον χρήστη [22]. Τα συστήματα critique – based προσπαθούν να εντοπίσουν εκείνα τα αντικείμενα που έχουν συγκεκριμένα χαρακτηριστικά και τα οποία χαρακτηριστικά είναι ικανά να ικανοποιήσουν τις προτιμήσεις του χρήστη, όπως αυτές καθορίζονται από τον ίδιο. Τα συστήματα αυτά υλοποιούνται ακολουθώντας τρία βήματα που περιγράφονται παρακάτω:

1. Το πρώτο βήμα της λειτουργίας ενός Critique-based recommendation system είναι να καταγράψει τις προτιμήσεις του χρήστη. Η καταγραφή αυτή γίνεται μέσα από ερωτήσεις προς τον χρήστη σχετικά με το τι του αρέσει και πόσο και γενικότερα ποια θα ήταν τα χαρακτηριστικά που θα ήθελε να έχει ένα αντικείμενο ώστε να τον ικανοποιήσει. Κατά συνέπεια, μέσα από αυτές τις ερωτήσεις ποσοτικοποιούνται οι προτιμήσεις του χρήστη, παίρνουν δηλαδή μια μορφή επεξεργάσιμη από το σύστημα. Συνήθως, ο χρήστης καλείται να απαντήσει με βάση μια κλίμακα που περιέχει σύντομες απαντήσεις που όμως σκιαγραφούν τις προτιμήσεις του. Έτσι, αυτή η καταγραφή των προτιμήσεων του χρήστη ουσιαστικά συνθέτει και το προφίλ του που θα χρησιμοποιηθεί για το συσχετισμό με αντικείμενα για την εξαγωγή συστάσεων.
2. Στο δεύτερο βήμα της λειτουργίας ενός συστήματος Critique-based και αφού έχει δημιουργηθεί το προφίλ του χρήστη, το σύστημα εξετάζει τα ιδιαίτερα χαρακτηριστικά των αντικειμένων που έχει στη βάση του και δεδομένου της γνώσης που έχει για το πως αυτά τα ιδιαίτερα χαρακτηριστικά των αντικειμένων καλύπτουν τις ανάγκες των χρηστών, εντοπίζει εκείνα τα αντικείμενα που ταιριάζουν καλύτερα στον συγκεκριμένο χρήστη και του τα κοινοποιεί με τη μορφή συστάσεων. Σε αυτό το σημείο, εάν ο χρήστης δηλώσει ικανοποιημένος από την συγκεκριμένη σύσταση μέσω της ανατροφοδότησης καθώς και αν δεν υπάρχει κανένα ενδιαφέρον από την πλευρά του για το προϊόν της σύστασης που του έγινε, τότε η διαδικασία ολοκληρώνεται.
3. Στο τρίτο και τελευταίο βήμα της διαδικασίας ο χρήστης έχοντας ήδη εξετάσει τα αντικείμενα που του συστάθηκαν στο δεύτερο βήμα, μπορεί να μεταβάλλει αναλόγως τις προτιμήσεις του με αποτέλεσμα το σύστημα να δέχεται αυτές τις καινούριες προτιμήσεις και να επαναλαμβάνει το δεύτερο βήμα.

Παρακάτω παρουσιάζεται αντιπροσωπευτικά ένα σχήμα που αναπαριστά τη λειτουργία ενός Critique-based recommendation system βάσει των τριών βημάτων που περιγράψαμε αναλυτικά παραπάνω.



Εικόνα 17: Λειτουργία ενός Critique-based recommendation system [22]

3.5.2 Goal-based recommendation

Τα συστήματα αυτής της κατηγορίας χρησιμοποιούν την ομοιότητα έτσι ώστε να προσδιορίσουν αντικείμενα που εξυπηρετούν τον ίδιο σκοπό με αυτά που είχε επιλέξει ο χρήστης στο παρελθόν. Τα Goal–based systems αναφέρονται επίσης και με τον όρο *Συστήματα Findme* [19], και αυτό γιατί πολλές φορές κάνουν χρήση διάφορων παραδειγμάτων για να υποβοηθήσουν την επικοινωνία τους με τον χρήστη. Η λειτουργία των συστημάτων αυτών περιλαμβάνει τρία βήματα:

1. Στο πρώτο βήμα της διαδικασίας ενός Goal–based system το σύστημα επιδιώκει να καταγράψει και να καθορίσει τις προτιμήσεις ενός χρήστη με βάση τα αντικείμενα που έχει επιλέξει ο χρήστης στο παρελθόν. Τα αντικείμενα που έχουν επιλεγεί από τον χρήστη παλαιότερα αποτελούν μια ισχυρή ένδειξη του τι ενδιαφέρει τον χρήστη αυτόν και άρα υποδηλώνει τις προτιμήσεις του αλλά και το σκοπό του. Για αυτό το λόγο λοιπόν το σύστημα προσπαθεί να βρει παρόμοια αντικείμενα.
2. Το δεύτερο βήμα περιλαμβάνει την εύρεση αντικειμένων που είναι παρόμοια με αυτά που είχε προτιμήσει ο χρήστης στο παρελθόν. Με τον όρο παρόμοια, εννοούμε αντικείμενα που εξυπηρετούν παρόμοιους

σκοπούς. Στο βήμα αυτό χρησιμοποιείται εκτεταμένα η μέθοδος των κριτηρίων ομοιότητας έτσι ώστε να προσδιοριστούν τα αντικείμενα που θα χρησιμοποιηθούν στη σύσταση.

3. Το τρίτο βήμα της διαδικασίας ενός Goal–based system, είναι, όπως συμβαίνει και στα Critique-based recommendation systems, η ανατροφοδότηση του χρήστη, ο οποίος μπορεί να μην ικανοποιείται από τη σύσταση που θα λάβει κατόπιν του συσχετισμού των προτιμήσεών του με τα αντικείμενα. Η μέθοδος *tweak* που στρέφει το σύστημα σε μία διαφορετική κατεύθυνση από την αρχική είναι η μέθοδος που χρησιμοποιείται για την υλοποίηση της ανατροφοδότησης σε αυτό το βήμα. Η στροφή του συστήματος σε διαφορετική κατεύθυνση μπορεί να γίνει με άμεση βαθμολόγηση των αντικειμένων που προτείνονται στον χρήστη. Ένας άλλος τρόπος για την ανατροφοδότηση του χρήστη, είναι το σύστημα να παρακολουθήσει τη συμπεριφορά του χρήστη μετά από τη σύσταση και να εξάγει έμμεσα συμπεράσματα για το αν τον ενδιέφερε ή όχι. Συνεπώς, πέρα από την αρχική σύσταση που θα δεχτεί ο χρήστης από το σύστημα, μπορεί ανάλογα με την ανατροφοδότηση που θα παράσχει στο σύστημα να δεχτεί και μια δεύτερη σύσταση που να εξυπηρετεί ενδεχομένως καλύτερα τις ανάγκες του.

3.5.3 Πλεονεκτήματα των Knowledge – based Systems

1. Τα συστήματα Knowledge – based δεν αντιμετωπίζουν όπως τα Collaborative filtering συστήματα το πρόβλημα *Ramp–up* ή *cold start* για το λόγο ότι το σύστημα δε βασίζεται σε μία σειρά αξιολογήσεων των αντικειμένων από τους χρήστες, οι οποίες αξιολογήσεις μπορεί να είναι χρονοβόρες. Ένα σύστημα βασισμένο στη γνώση, ακριβώς επειδή δε στηρίζεται σε συγκέντρωση βαθμολογιών των αντικειμένων, μπορεί να δώσει αξιόπιστες συστάσεις από την αρχή της λειτουργίας του απορροφώντας άμεσα όλες τις αλλαγές των προτιμήσεων των χρηστών.
2. Τα συστήματα αυτής της κατηγορίας προκειμένου να ανακτήσουν πληροφορίες σχετικά με τις προτιμήσεις ενός χρήστη, έχουν την δυνατότητα να τις ανακτούν άμεσα από τον χρήστη με διάφορους τρόπους, όπως για παράδειγμα με τη μορφή ερωτήσεων. Το γεγονός ότι το σύστημα μπορεί άμεσα να δέχεται τις προτιμήσεις των χρηστών συμβάλει στην προσαρμοστικότητα τους σε

αναπάντεχες αλλαγές των προτιμήσεων του χρήστη, κάτι που αυξάνει την αποτελεσματικότητα των συστημάτων αυτών.

3. Ένα άλλο πολύ σημαντικό πλεονέκτημα των knowledge – based συστημάτων συστάσεων είναι ότι δεν έχουν υψηλές απαιτήσεις σε ότι αφορά την υπολογιστική ισχύ προκειμένου να υπολογίσουν σε πραγματικό χρόνο τις αλληλεπιδράσεις μεταξύ χρηστών και αντικειμένων. Αυτό είναι ένα πολύ σημαντικό πλεονέκτημα έναντι των Collaborative Filtering συστημάτων συστάσεων καθώς αυτά χρειάζονται μεγάλη υπολογιστική ισχύ ώστε να εντοπίσουν τους κοντινότερους γείτονες προκειμένου να προβούν σε συστάσεις.
4. Τα συστήματα αυτά δεν επηρεάζονται από στατιστικές ανωμαλίες στο δείγμα χρηστών με την έννοια ότι λόγω της γνώσης που έχουν αποκτήσει μπορούν να απομονώσουν εύκολα κάποια ακραία τιμή.
5. Τα συστήματα knowledge –based μπορούν να κάνουν συστάσεις και σε χρήστες με ιδιαίτερες απαιτήσεις. Ένας χρήστης που έχει προτιμήσεις ή απαιτήσεις που διαφέρουν πολύ από αυτές των περισσοτέρων άλλων χρηστών θα αντιμετωπίσει πρόβλημα αν χρησιμοποιήσει ένα σύστημα Collaborative filtering καθώς τα συστήματα αυτά προσπαθούν να εντοπίσουν τους γείτονες του χρήστη για να εξάγουν συστάσεις. Ωστόσο, ένα σύστημα που βασίζεται στη γνώση, θα μπορέσει να προσφέρει αξιόπιστες συστάσεις στον χρήστη ακόμη και αν δεν υπάρχει άλλος όμοιός του αφού ο ίδιος ο χρήστης πληροφορεί το σύστημα με τι του αρέσει και τι όχι.
6. Ένα άλλο σημαντικό πλεονέκτημα έναντι των Collaborative filtering συστημάτων είναι ότι εφόσον τα συστήματα με βάση τη γνώση δε βασίζονται σε αξιολογήσεις των αντικειμένων, μπορούν να συστήσουν και αντικείμενα για τα οποία δεν υπάρχουν αρκετές αξιολογήσεις. Το γεγονός αυτό δίνει τη δυνατότητα στα συστήματα με βάση τη γνώση να μπορούν να συστήσουν σε ένα χρήστη νέα και μη αξιολογημένα αντικείμενα.
7. Τέλος, ένα σύστημα βασισμένο στη γνώση μπορεί να αιτιολογήσει τις συστάσεις του μέσω των χαρακτηριστικών των αντικειμένων που προτάθηκαν σε έναν συγκεκριμένο χρήστη. Στα συστήματα αυτά, ο χρήστης μπορεί να ζητήσει από το σύστημα να του παρουσιάσει τη διαδικασία που ακολούθησε ώστε να καταλήξει σε κάποια συγκεκριμένη σύσταση και αυτή ακριβώς η

διαδικασία είναι που λειτουργεί σαν αιτιολόγηση του συστήματος προς τον χρήστη για την σύσταση που έκανε.

3.5.4 Μειονεκτήματα των Knowledge based Systems

1. Όπως αναφέραμε παραπάνω, για να υλοποιηθεί ένα σύστημα βασισμένο στη γνώση απαιτείται η κατασκευή μιας κατάλληλης βάσης γνώσης για το σύστημα η οποία είναι μια δύσκολη και επίπονη διαδικασία. Αυτή η συγκεκριμένη βάση δεδομένων πάνω στην οποία βασίζεται ένα σύστημα knowledge – based θα πρέπει να περιλαμβάνει πληροφορίες για τις επιθυμίες των χρηστών, για το πώς αυτές οι επιθυμίες καλύπτονται από τα διάφορα αντικείμενα, ποια είναι τα ιδιαίτερα χαρακτηριστικά των αντικειμένων που μπορούν να ικανοποιήσουν τις ανάγκες ενός χρήστη και φυσικά πληροφορίες για τα ίδια τα αντικείμενα. Γίνεται λοιπόν εύκολα αντιληπτό ότι μια τέτοια βάση απαιτεί μια αρκετά δύσκολη διαδικασία για να δημιουργηθεί. Ωστόσο, υπάρχουν και περιπτώσεις όπου το πρόβλημα αυτό δεν είναι τόσο έντονο. Τα συστήματα του *Find Me* που περιγράψαμε και παραπάνω, δεν απαιτούν ιδιαίτερα μεγάλες προσπάθειες για την κατασκευή βάσης γνώσης αφού χρησιμοποιούν στοιχεία που ήδη έχουν καταγραφεί.
2. Σε συνέχεια του προηγούμενου μειονεκτήματος των συστημάτων συστάσεων βασισμένων στη γνώση, ένα ακόμη μειονέκτημα είναι ότι τα συστήματα αυτά προβαίνουν σε συστάσεις που μπορεί να χαρακτηρίζονται από στατικότητα. Αυτό συμβαίνει γιατί, όπως είπαμε, τα συστήματα αυτά απαιτούν τη δημιουργία μιας μεγάλης βάσης δεδομένων για χρήστες και αντικείμενα, γεγονός που καθιστά έντονο το πρόβλημα της ανανέωσης αυτής της μεγάλης βάσης δεδομένων. Έτσι, το γεγονός της δυσκολίας ενημέρωσης και επικαιροποίησης της βάσης αυτής μπορεί να οδηγεί σε εξαγωγή συστάσεων βασισμένων πάνω στα ίδια δεδομένα και πληροφορίες και άρα να είναι στατικές.
3. Τέλος, άλλο ένα μειονέκτημα των συστημάτων αυτών είναι ότι θα πρέπει να ποσοτικοποιηθούν τα χαρακτηριστικά των αντικειμένων. Όπως ήδη έχουμε δει και περιγράψει, τα συστήματα αυτά για να εξάγουν συστάσεις βασίζονται στη γνώση που έχουν για τα χαρακτηριστικά των

αντικειμένων και πως αυτά μπορούν να ικανοποιήσουν τις ιδιαίτερες προτιμήσεις των χρηστών. Προκύπτει επομένως το πρόβλημα της ποσοτικοποίησης αυτών των χαρακτηριστικών των αντικειμένων έτσι ώστε να μπορέσει το σύστημα να τα αξιοποιήσει και να εξάγει κάποια σύσταση.

3.6 Συστήματα Community – based

Τα συστήματα αυτά λειτουργούν στη βάση του ότι κάποιος χρήστης θα επιλέξει εκείνο το αντικείμενο που επιλέγουν και οι φίλοι του ή άτομα από τον κοινωνικό του περίγυρο. Έτσι οι χρήστες τείνουν να στηρίζονται σε συστάσεις περισσότερο από άτομα που εμπιστεύονται. Κατά συνέπεια, το γεγονός ότι τα κοινωνικά δίκτυα είναι ανοιχτά, με την έννοια ότι μπορεί οποιοσδήποτε να τα χρησιμοποιήσει, έχει οδηγήσει τα συστήματα συστάσεων να δίνουν ιδιαίτερη σημασία στον παράγοντα της εμπιστοσύνης, και για την εξαγωγή συστάσεων προς ένα χρήστη να λαμβάνουν υπόψη και την εμπιστοσύνη του χρήστη ως προς διάφορους παράγοντες ή άλλους χρήστες.

Οι συστάσεις που πραγματοποιούνται μέσω των συστημάτων αυτών βασίζονται σε πληροφορίες που προέρχονται από ένα δίκτυο εμπιστοσύνης του χρήστη. Το δίκτυο εμπιστοσύνης του χρήστη είναι ένα κοινωνικό δίκτυο στο οποίο τα μέλη του συνδέονται με σχέσεις εμπιστοσύνης, δηλαδή εκφράζει το βαθμό στον οποίο τα μέλη εμπιστεύονται το ένα το άλλο. Έτσι, τα συστήματα συστάσεων με βάση την εμπιστοσύνη στηρίζονται στη γνώση που προέρχεται από το δίκτυο εμπιστοσύνης ενός χρήστη έτσι ώστε να προβούν σε πιο εξατομικευμένες συστάσεις, με αποτέλεσμα ένας χρήστης να λαμβάνει συστάσεις αντικειμένων τα οποία έχουν λάβει υψηλή βαθμολογία από τα μέλη που ανήκουν στον ιστό εμπιστοσύνης του χρήστη (*Web of Trust , WOT)* ή ακόμη και από χρήστες που εμπιστεύονται τα μέλη του ιστού εμπιστοσύνης του χρήστη στόχου [61].

Σχετικά με τα συστήματα συστάσεων με βάση την εμπιστοσύνη, υπάρχουν δυο βασικές κατηγορίες αυτών των μοντέλων:

1. Το πιθανοτικό μοντέλο *(probabilistic)*
2. Η σταδιακή προσέγγιση *(gradual approach)*

Στην πρώτη κατηγορία, το πιθανοτικό μοντέλο υπολογίζει την πιθανότητα για κάποιον χρήστη να είναι άξιος εμπιστοσύνης, με την εμπιστοσύνη να λαμβάνει δυο τιμές, την εμπιστοσύνη σε έναν χρήστη ή όχι.

Στην άλλη περίπτωση, η σταδιακή προσέγγιση εκτιμά την τιμή της εμπιστοσύνης για έναν χρήστη, δεδομένου ότι ένας χρήστης μπορεί να εμπιστεύεται έναν άλλον σε κάποιο βαθμό και όχι όπως παραπάνω απόλυτα ή καθόλου. Η σταδιακή προσέγγιση στην εκτίμηση της εμπιστοσύνης για έναν χρήστη είναι πιο ρεαλιστική, καθώς είναι αυτό που συμβαίνει και στην πραγματική ζωή.

Όπως ήδη αναφέραμε, ο κάθε χρήστης έχει έναν ιστό εμπιστοσύνης βάσει του οποίου το σύστημα λαμβάνει πληροφορίες για να εξάγει συστάσεις. Ωστόσο, ο ιστός εμπιστοσύνης ενός χρήστη είναι πολύ περιορισμένος σχετικά με το σύνολο της διαδικτυακής κοινότητας και κατά συνέπεια το σύστημα θα ήταν περισσότερο αποτελεσματικό αν μπορούσε να αξιοποιήσει τη γνώση που προέρχεται από ένα μεγαλύτερο υποσύνολο του πληθυσμού των χρηστών για να προβεί σε συστάσεις.

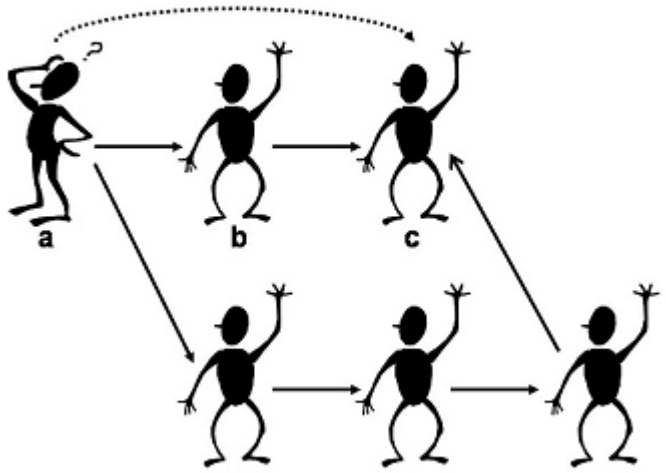
Στο σημείο αυτό θα πρέπει να αναφέρουμε δυο σημαντικές έννοιες που σχετίζονται με τα συστήματα με βάση την εμπιστοσύνη. Οι έννοιες αυτές είναι η διάδοση και η άθροιση. Η έννοια της μετάδοσης βασίζεται στην υπόθεση ότι η εμπιστοσύνη είναι μεταδοτική. Αυτό μπορεί εύκολα να ερμηνευτεί με το παρακάτω παράδειγμα: Έστω ότι έχουμε τρεις χρήστες Α, Β και Γ των οποίων θα αναλύσουμε την εμπιστοσύνη για ένα αντικείμενο με το ίδιο περιεχόμενο. Αν ο Α εμπιστεύεται τον Β και ο Β εμπιστεύεται τον Γ για το ίδιο αντικείμενο, τότε με μεγάλη πιθανότητα και ο Α θα εμπιστεύεται τον Γ. Το παραπάνω παράδειγμα παρουσιάστηκε στην πιο απλή του μορφή για να γίνει κατανοητή η έννοια της μετάδοσης. Παρακάτω θα παρουσιάσουμε το ίδιο παράδειγμα αλλά με μια πιο σύνθετη ανάλυση αφού θα εισάγουμε και την έννοια της δυσπιστίας, δηλαδή την έλλειψη εμπιστοσύνης ενός χρήστη ως προς έναν άλλον. Στο συγκεκριμένο παράδειγμα λοιπόν διακρίνουμε τρεις διαφορετικές περιπτώσεις:

1. Αν ο Α δεν εμπιστεύεται τον Β και ο Β με τη σειρά του δεν εμπιστεύεται τον Γ, τότε ο Α μπορεί να εμπιστεύεται τον Γ με την έννοια ότι εμπιστεύεται κάποιον χρήστη που δεν τον εμπιστεύονται εκείνοι που δεν εμπιστεύεται ο χρήστης και άρα μπορεί να τον εμπιστευτεί (πολλαπλασιαστική διάδοση μη εμπιστοσύνης)

2. Αν ο Α δεν εμπιστεύεται τον Β και ο Β με τη σειρά του δεν εμπιστεύεται τον Γ, τότε και ο Α δεν εμπιστεύεται τον Γ το οποίο είναι αντίστροφο στο

παράδειγμα που παραθέσαμε παραπάνω στην περίπτωση της εμπιστοσύνης (προσθετική διάδοση μη εμπιστοσύνης)

3. Αν ο Α δεν εμπιστεύεται τον Β και ο Β με τη σειρά του δεν εμπιστεύεται τον Γ τότε δεν μπορεί να εξαχθεί κάποιο συμπέρασμα για τη σχέση μεταξύ Α και Γ καθώς ο Α δε θα πρέπει να λάβει υπόψη του πληροφορίες από τον Β, ένα άτομο που δεν εμπιστεύεται.



Εικόνα 18: Παράδειγμα διάδοσης [61]

Από την άλλη πλευρά, η έννοια της άθροισης βασίζεται στην υπόθεση ότι σε μεγάλα δίκτυα υπάρχουν πολλά μονοπάτια για τον υπολογισμό του επιπέδου εμπιστοσύνης ενός χρήστη. Με την μέθοδο της άθροισης, προκειμένου να υπολογιστεί ο βαθμός εμπιστοσύνης ενός χρήστη, αθροίζονται όλοι οι βαθμοί εμπιστοσύνης που προκύπτουν για έναν χρήστη από όλες τις αλληλεπιδράσεις του με άλλους χρήστες.

Αυτός ο συνδυασμός των διάφορων μονοπατιών εμπιστοσύνης σε μια μοναδική και συνολική τιμή εμπιστοσύνης για έναν χρήστη γίνεται μέσω της μέσης σύνθεσης και της σταθμισμένης μέσης σύνθεσης. Μέσω της μέσης σύνθεσης υπολογίζεται ο μέσος όρος των τιμών εμπιστοσύνης που προκύπτει από κάθε μονοπάτι εμπιστοσύνης, σύμφωνα με τον παρακάτω τύπο:

$$T_{S \to T} = \frac{\sum_{i=1}^{p} T_{S \xrightarrow{p_i} T}}{p} \qquad (3.17)$$

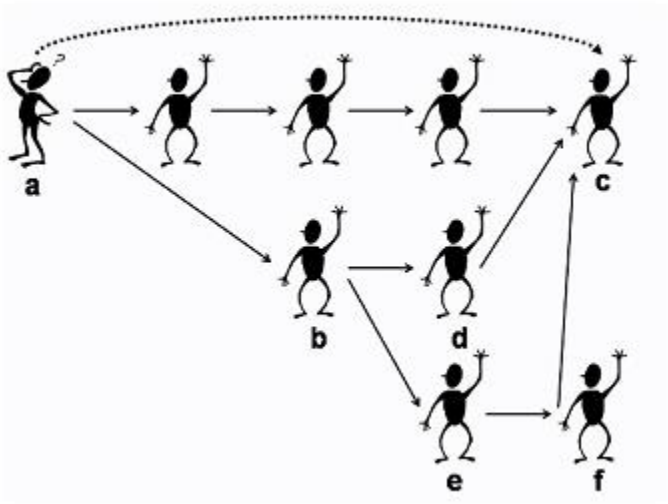
Όπου p ο αριθμός των μονοπατιών εμπιστοσύνης.

Με βάση τον άλλο τρόπο υπολογισμού, τη σταθμισμένη μέση σύνθεση, υπολογίζεται ο σταθμισμένος μέσος όρος των βαθμών εμπιστοσύνης από τα διάφορα μονοπάτια εμπιστοσύνης χρησιμοποιώντας ως συντελεστές στάθμισης τη διαδιδόμενη εμπιστοσύνη κάθε έμμεσης συσχέτισης μεταξύ δυο χρηστών, S και T.

Η τιμή της μέσης σταθμισμένης εμπιστοσύνης δίνεται από τον παρακάτω τύπο:

$$T_{S\to T} = \sum_i^p \frac{C_{S\xrightarrow[p_i]{}T}}{\sum_{i=1}^p C_{S\xrightarrow[p_i]{}T}} \; T_{S\xrightarrow[p_i]{}T} \qquad (3.18)$$

Όπου $C_{S\xrightarrow[p_i]{}T}$ η διαδιδόμενη εμπιστοσύνη ενός μονοπατιού που συνδέει τους χρήστες S και T με ενδιάμεσους κόμβους.



Εικόνα 19: Παράδειγμα άθροισης [61]

3.7 Υβριδικά Συστήματα Συστάσεων

Τα συστήματα της κατηγορίας αυτής ουσιαστικά δεν είναι κάποια νέα συστήματα που έρχονται να αναλύσουν πληροφορίες για χρήστες, αντικείμενα και χαρακτηριστικά χρηστών και αντικειμένων ώστε να προβούν σε συστάσεις, αλλά αποτελούν συνδυασμό ή συνδυασμούς όλων των παραπάνω συστημάτων που αναλύσαμε.

Τα υβριδικά συστήματα συστάσεων με άλλα λόγια αποτελούν ένα μείγμα των παραδοσιακών συστημάτων συστάσεων που έρχονται να καλύψουν τις αδυναμίες των παραδοσιακών συστημάτων αλλά και να εκμεταλλευτούν τα δυνατά τους σημεία. Ανάλογα με την περίπτωση για την οποία προορίζονται τα συστήματα συστάσεων, γίνεται και ο κατάλληλος συνδυασμός ώστε να είναι το πλέον αποτελεσματικό υβριδικό σύστημα για την εξαγωγή συστάσεων. Σχεδόν όλοι οι σύγχρονοι τρόποι σύστασης και στις περισσότερες εφαρμογές, στηρίζονται σε υβριδικά συστήματα συστάσεων καθώς προσπαθούν να άρουν τα προβλήματα που αντιμετωπίζουν τα Collaborative Filtering συστήματα συστάσεων ή τα Content – based συστήματα συστάσεων.

Υπάρχουν διάφορες κατηγορίες υβριδικών συστημάτων συστάσεων [18,19]:

- σταθμισμένα συστήματα (weighted)
- συστήματα εναλλαγής (switching)
- μικτά συστήματα (mixed)
- συστήματα συνδυασμού χαρακτηριστικών (feature combination)
- συστήματα καταρράχτη (cascade)
- συστήματα επαυξημένων χαρακτηριστικών (feature augmentation)
- συστήματα μεταφοράς μοντέλου (meta level)

3.7.1 Σταθμισμένα συστήματα

Στα σταθμισμένα υβριδικά συστήματα το αποτέλεσμα της σύστασης βασίζεται στην στάθμιση των αποτελεσμάτων των επιμέρους συστημάτων στα οποία βασίζεται το υβριδικό σύστημα. Με άλλα λόγια, το δημιουργούμενο σύστημα λαμβάνει τις βαθμολογίες που έχουν δοθεί για ένα προτεινόμενο αντικείμενο από τα επιμέρους συστήματα και με βάση συγκεκριμένους συντελεστές στάθμισης των βαθμολογιών αυτών, το υβριδικό σύστημα καταλήγει σε κάποια συγκεκριμένη σύσταση. Ωστόσο, ένα πολύ σημαντικό μειονέκτημα αυτής της κατηγορίας είναι ότι βαθμολογίες για ένα αντικείμενο με βάση κάποιο συγκεκριμένο σύστημα μπορεί να υστερούν έναντι των βαθμολογιών άλλων συστημάτων. Ενδεικτικό παράδειγμα είναι τα αποτελέσματα για μια σύσταση μέσω του Collaborative filtering συστήματος όταν δεν υπάρχουν πολλές αξιολογήσεις για ένα αντικείμενο [19].

3.7.2 Συστήματα εναλλαγής

Μια άλλη κατηγορία υβριδικών συστημάτων είναι αυτή που εναλλάσσει δυο επιμέρους συστήματα συστάσεων ανάλογα με το ποιο μπορεί να αποδώσει καλύτερα σε κάθε περίπτωση με βάση κάποια συγκεκριμένα κριτήρια. Η λειτουργία ενός τέτοιου συστήματος έγκειται στην χρησιμοποίηση των διαφορετικών συστημάτων για την εξαγωγή συστάσεων με την τελική αξιολόγηση των αποτελεσμάτων κάθε συστήματος να δίνει την τελική σύσταση του υβριδικού συστήματος. Το μειονέκτημα ενός συστήματος εναλλαγής συστημάτων σύστασης είναι ότι αυξάνει πολύ το βαθμό πολυπλοκότητας στη διαδικασία της σύστασης αφού θα πρέπει να καθοριστούν τα κριτήρια εναλλαγής μεταξύ των συστημάτων. Ωστόσο, ένα τέτοιο υβριδικό σύστημα μπορεί να ξεπεράσει τα προβλήματα που εμφανίζουν τα επιμέρους συστήματα στα οποία βασίζεται αφού ανάλογα με την περίπτωση χρησιμοποιείται και το καταλληλότερο σύστημα συστάσεων [19].

3.7.3 Μικτά συστήματα

Τα υβριδικά συστήματα της κατηγορίας αυτής συνήθως χρησιμοποιούνται σε περιπτώσεις όπου απαιτείται η ταυτόχρονη παρουσίαση πολλών αποτελεσμάτων συστάσεων. Ένα μικτό υβριδικό σύστημα συστάσεων παρουσιάζει στον χρήστη τα αποτελέσματα που προέρχονται από τη διαδικασία της σύστασης που ακολουθούν τα επιμέρους συστήματα συστάσεων τα οποία λειτουργούν ανεξάρτητα μεταξύ τους. Το πλεονέκτημα ενός τέτοιου συστήματος είναι ότι παρέχει στον χρήστη συστάσεις μέσα από τις οποίες μπορεί να επιλέξει ποια είναι προτιμότερη σε αυτόν και ταιριάζει καλύτερα στις ανάγκες του. Από την άλλη πλευρά όμως, τα συστήματα αυτά πάσχουν στην περίπτωση που απαιτούνται λίγες συστάσεις από έναν χρήστη καθώς και όταν οι διαφορετικές συστάσεις που προέρχονται από τα επιμέρους συστήματα συστάσεων έρχονται σε σύγκρουση [19].

3.7.4 Συστήματα συνδυασμού χαρακτηριστικών

Αυτή η κατηγορία υβριδικών συστημάτων συστάσεων κάνει συνδυασμό των μεθόδων Collaborative Filtering και Content-based. Η λειτουργία ενός τέτοιου συστήματος έγκειται στην παρακάτω διαδικασία: το σύστημα προσθέτει στα αντικείμενα προς αξιολόγηση για την εξαγωγή συστάσεων και πληροφορίες που προέρχονται από τη διαδικασία που ακολουθεί η Collaborative Filtering τεχνική και έπειτα με βάση αυτές τις επιπλέον πληροφορίες για τα αντικείμενα εφαρμόζεται η μέθοδος συστάσεων με

βάση το περιεχόμενο για την εξαγωγή συστάσεων. Η πρόσθετη πληροφόρηση που ενσωματώνεται στα αντικείμενα ουσιαστικά πρόκειται για τις αξιολογήσεις που αντλεί το σύστημα Collaborative Filtering για τα αντικείμενα από τους χρήστες. Το πλεονέκτημα αυτού του υβριδικού συστήματος είναι ότι ενσωματώνει στα αντικείμενα επιπλέον πληροφορίες και επομένως υπερνικά το πρόβλημα που αντιμετωπίζουν τα Collaborative Filtering συστήματα στην περίπτωση της έλλειψης αξιολογήσεων για ένα αντικείμενο [19].

3.7.5 Συστήματα καταρράχτη

Τα υβριδικά συστήματα συστάσεων τύπου καταρράχτη κάνουν συνδυασμό δύο επιμέρους συστημάτων συστάσεων έτσι ώστε να κάνουν μια ανακατανομή και επαναταξινόμηση των συστάσεων που έχουν προκύψει από τη διαδικασία ενός εκ των δύο συστημάτων. Τα υβριδικά συστήματα αυτά λειτουργούν σε δύο στάδια. Το πρώτο στάδιο περιλαμβάνει την εξαγωγή συστάσεων από το ένα σύστημα αλλά με ένα πιο μαζικό και πρόχειρο τρόπο. Στο δεύτερο στάδιο της λειτουργίας ενός τέτοιου υβριδικού συστήματος, επεμβαίνει το άλλο επιμέρους σύστημα συστάσεων για να κάνει την τελική ταξινόμηση των συστάσεων και να καταλήξει στις τελικές συστάσεις. Σε αντίθεση με όλα τα παραπάνω διαφορετικά υβριδικά συστήματα συστάσεων, σε ένα υβριδικό σύστημα τύπου καταρράχτη υπάρχει αλληλεπίδραση μεταξύ των συστημάτων σε αντίθεση με την ανεξάρτητη λειτουργία των συστημάτων των παραπάνω κατηγοριών. Δυο σημαντικά σημεία που πρέπει να επισημάνουμε σε ότι αφορά τα υβριδικά συστήματα αυτής της κατηγορίας είναι ότι δεδομένου του επιπλέον φιλτραρίσματος από το δεύτερο σύστημα των ήδη δημιουργημένων συστάσεων από το πρώτο σύστημα, αυτό το υβριδικό σύστημα εμφανίζεται περισσότερο αποτελεσματικό έναντι του σταθμισμένου υβριδικού συστήματος που εφαρμόζει όλες του τις τεχνικές στο σύνολο των αντικειμένων και όχι σε μια επιμέρους τελική ομάδα συστάσεων. Επίσης, το επιμέρους σύστημα του υβριδίου έρχεται στο δεύτερο στάδιο της διαδικασίας να βελτιώσει τις ήδη δημιουργημένες συστάσεις από το πρώτο επιμέρους σύστημα στο πρώτο στάδιο της διαδικασίας, χωρίς όμως να μπορεί να ανατρέψει τις συγκεκριμένες συστάσεις [19].

3.7.6 Συστήματα επαυξημένων χαρακτηριστικών

Στα υβριδικά συστήματα συστάσεων αυτής της κατηγορίας γίνεται χρήση ενός συστήματος για να παράγει κάποια αποτελέσματα που βασίζονται σε πληροφορίες από

διάφορες πηγές έτσι ώστε να αποτελέσουν εισροές για το δεύτερο επιμέρους σύστημα. Η κατηγορία αυτών των υβριδικών συστημάτων διαφέρει από τα υβριδικά συστήματα συνδυασμού χαρακτηριστικών τα οποία βασίζονται σε ακατέργαστες μορφές πληροφορίας και έπειτα συνδυάζονται από το σύστημα για την εξαγωγή συστάσεων. Επιπλέον τα συστήματα αυτά διαφοροποιούνται και από τα συστήματα καταρράκτη καθώς το δεύτερο επιμέρους σύστημα ενός υβριδικού συστήματος επαυξημένων χαρακτηριστικών αξιοποιεί την εκροή στην οποία καταλήγει το πρώτο επιμέρους σύστημα σε αντίθεση με τα συστήματα καταρράκτη στα οποία τα αποτελέσματα συστάσεων των δυο επιμέρους συστημάτων συνδυάζονται με μορφή προτεραιότητας [19].

3.7.7 Συστήματα μεταφοράς μοντέλου

Τα υβριδικά συστήματα συστάσεων μεταφοράς μοντέλου χρησιμοποιούν το ένα επιμέρους σύστημα ως μοντέλο για τη λειτουργία και εξαγωγή συστάσεων μέσω του δεύτερου επιμέρους μοντέλου. Πιο συγκεκριμένα, το πρώτο σύστημα δημιουργεί ένα μοντέλο που λειτουργεί ως είσοδος στη διαδικασία του δεύτερου επιμέρους συστήματος το οποίο προχωρά στη δημιουργία συστάσεων βασισμένο στο μοντέλο αυτό. Αυτά τα υβριδικά συστήματα διαφέρουν από την αμέσως παραπάνω κατηγορία υβριδικών συστημάτων στη βάση ότι τα συστήματα μεταφοράς μοντέλου δίνουν ως εισροή στο δεύτερο σύστημα ένα ολόκληρο μοντέλο και όχι απλά επεξεργασμένα χαρακτηριστικά αντικειμένων όπως συμβαίνει στην περίπτωση των υβριδικών συστημάτων επαυξημένων χαρακτηριστικών. Τα βασικό πλεονέκτημα ενός τέτοιου συστήματος είναι ότι προσφέρει στο δεύτερο σύστημα ένα μοντέλο των προτιμήσεων των χρηστών, γεγονός που αυξάνει την αποτελεσματικότητα σε σύγκριση με την αξιοποίηση ακατέργαστων κομματιών πληροφορίας [19].

4 Recommendation Systems και Twitter

4.1 Εισαγωγή

Το Twitter με περισσότερους από εκατό εκατομμύρια χρήστες οι οποίοι αναπαράγουν καθημερινώς πάνω από τριακόσια εκατομμύρια μηνύματα (tweets) αποτελεί σήμερα ένα από τα πιο δημοφιλή κοινωνικά δίκτυα και παράλληλα βάση για πολλές έρευνες και μελέτες σχετικά με τη λειτουργία του σε ερευνητικά κέντρα, εταιρίες και πανεπιστήμια. Οι χρήστες του Twitter καθημερινά βομβαρδίζονται από σωρεία πληροφοριών καθώς και από έναν τεράστιο αριθμό άλλων χρηστών με τους οποίους μπορούν να αλληλεπιδράσουν. Προκειμένου να ξεπεραστεί αυτό το πρόβλημα της υπερπληροφόρησης έχουν μελετηθεί διάφορα συστήματα συστάσεων έτσι ώστε να βοηθήσουν τους χρήστες να προβούν σε σωστές επιλογές. Σε αυτό το κομμάτι της εργασίας θα περιγράψουμε αναλυτικά τα συστήματα συστάσεων που χρησιμοποιούνται στο Twitter για τη διεκπεραίωση συγκεκριμένων λειτουργιών που είναι διαθέσιμες σε αυτό το κοινωνικό δίκτυο. Προτού όμως περιγράψουμε τα συστήματα συστάσεων για κάθε λειτουργία, θα πρέπει να ορίσουμε το ποιες είναι αυτές οι λειτουργίες του Twitter.

Οι λειτουργίες του Twitter [8, 39, 43] που θα απαριθμήσουμε είναι 4 και περιγράφονται παρακάτω:

- Follow: Η λειτουργία *Follow* αναφέρεται στη σύνδεση μεταξύ δύο χρηστών που έχει ως αποτέλεσμα την κοινοποίηση των tweets αυτού που δέχεται τη σύνδεση σε αυτόν που πραγματοποιεί τη σύνδεση. Ο χρήστης που πραγματοποιεί τη σύνδεση ονομάζεται *follower* και αυτός με τον οποίο συνδέεται *followee.*
- *Tweet:* Η λειτουργία του *Tweet* αναφέρεται στην ανάρτηση ενός μηνύματος έως 140 χαρακτήρες. Το περιεχόμενο αυτών των μηνυμάτων ποικίλει ανάλογα με τον χρήστη και μπορεί να είναι οτιδήποτε, κείμενο ή ακόμα και ειδήσεις. Πολλές φορές τέτοια μηνύματα μπορεί να περιέχουν και συνδέσμους URLs

που παραπέμπουν σε άλλες σελίδες του διαδικτύου ή και *hashtags*[1] για να συσχετίσουν παρόμοια πράγματα.

- *Retweet:* Η λειτουργία του *Retweet* αναφέρεται στην προώθηση ενός *Tweet* από ένα χρήστη στους ακολούθους του. Αυτή η λειτουργία είναι η επικρατέστερη στο Twitter για τη διάδοση πληροφοριών.
- *Mention:* Η λειτουργία του *Mention* αναφέρεται στην αναφορά ενός ή περισσότερων χρηστών σε ένα μήνυμα tweet με την πρόσθεση μπροστά από το όνομα του χρήστη ή των χρηστών που θέλουμε να αναφέρουμε του συμβόλου @. Οι χρήστες που αναφέρονται λαμβάνουν αντίστοιχα και το tweet και αυτός είναι ένας τρόπος για να αρχίσει μια συζήτηση ή γενικά να αρχίσουν να αλληλεπιδρούν οι χρήστες μεταξύ τους.

Στην ανάλυση που ακολουθεί παρακάτω θα περιγράψουμε τα συστήματα συστάσεων που χρησιμοποιούνται στην περίπτωση των *followee, follower, tweet, retweet, hashtag* και *ειδήσεις.*

4.2 Μέθοδοι Συστάσεων για followee

Στο Twitter οι χρήστες συνήθως ενδιαφέρονται όχι μόνο να συνδεθούν με τους φίλους τους, αλλά παράλληλα και να διευρύνουν τον κύκλο τους με το να συνδεθούν και με άτομα που μπορεί ναι μεν να μη γνωρίζουν προσωπικά αλλά έχουν κοινά ενδιαφέροντα και θα μπορούσαν να έχουν και άλλα κοινά χαρακτηριστικά. Αυτοί οι (άγνωστοι) χρήστες μπορεί να αποτελέσουν πηγή πληροφοριών για τον χρήστη. Ανάλογα τώρα με τις ανάγκες του χρήστη, διάφοροι αλγόριθμοι συστάσεων μπορούν να χρησιμοποιηθούν. Για παράδειγμα, αλγόριθμοι μπορεί να αξιοποιούν πληροφορίες από τους κοινούς φίλους, ή από το γενικό προφίλ του χρήστη ή ακόμα και από τις αξιολογήσεις που έχει κάνει ο χρήστης, προκειμένου να προβεί σε κάποια σύσταση. Στο Twitter υπάρχει η λειτουργία *Who to follow* και η οποία προβαίνει σε συστάσεις *followees* ανάλογα με την ομοιότητα των ήδη υπαρχόντων *followees* του χρήστη και κατ' επέκταση *followees* αυτών των *followees.* Παρακάτω θα περιγράψουμε τις μεθόδους που χρησιμοποιούνται για την σύσταση followee στο Twitter.

[1] *Hashtag* είναι μια λέξη κλειδί η οποία αναπαρίσταται με μια δίεση (#) για να συνοψίσει ένα θέμα και να το συσχετίσει με κάποιο γεγονός, άτομο, κατάσταση κτλ.

4.2.1 Topology – Based Methods

Σχετικά με αυτήν την κατηγορία μεθόδων σύστασης έχουν προταθεί προσεγγίσεις που βασίζονται στις Collaborative Filtering και Content – Based μεθόδους συστάσεων [5, 6, 7]. Η χρήση της collaborative filtering μεθόδου έγκειται στο ότι παρόμοιοι χρήστες συστήνονται βάσει του κοινωνικού τους γράφου. Η υπόθεση πάνω στην οποία βασίζεται η μέθοδος είναι ότι ο χρήστης παρουσιάζει πολλές ομοιότητες με τους followers των followees του. Έτσι οι υποψήφιοι για σύσταση followees κατατάσσονται με βάση τον αριθμό των κοινών followees που έχουν με τον συγκεκριμένο χρήστη και ο πρώτος σε κατάταξη με βάση τη μεθοδολογία που περιγράψαμε, είναι και αυτός που συστήνεται.

Από την άλλη πλευρά με βάση την προσέγγιση της content – based μεθόδου, η σύσταση ενός χρήστη γίνεται με βάση το περιεχόμενο των μηνυμάτων που αναρτά. Δηλαδή, σε ένα χρήστη θα προταθεί ένας followee του οποίου το περιεχόμενο των tweets είναι παρόμοιο με το περιεχόμενο των tweets των ήδη followees του χρήστη [51].

Collaborative – Filtering Σύσταση

Ο αλγόριθμος που έχει αναπτυχθεί και υλοποιεί αυτού του είδους τη σύσταση βασίζεται στην πρόταση για δυνητικούς followees από τη γειτονιά του χρήστη. Παρακάτω περιγράφουμε πως λειτουργεί αυτός ο αλγόριθμος σε 4 βήματα προκειμένου να δώσει σύσταση ενός ή περισσοτέρων followees σε έναν χρήστη που ορίζουμε u_T.

Βήμα 1: Για τον χρήστη u_T ορίζουμε ένα σύνολο έστω S που περιλαμβάνει όλους τους χρήστες που ακολουθεί:

$$S(u_T) = \bigcup_{\forall\, f \in followees(u_T)} f \qquad (4.1)$$

Βήμα 2: Για κάθε στοιχείο του συνόλου S παίρνουμε το σύνολο, έστω L των followers του:

$$L(u_T) = \bigcup_{\forall\, s \in S} followers\ (s) \qquad (4.2)$$

Βήμα 3: Στο τρίτο βήμα παίρνουμε για κάθε στοιχείο του συνόλου L ένα σύνολο T που αποτελείται από τους followees των στοιχείων του L:

$$T(u_T) = \bigcup_{\forall\, l \in L} followees\ (l) \qquad (4.3)$$

Βήμα 4: Στο τέταρτο βήμα απαλείφουμε από το σύνολο T εκείνους τους χρήστες που ο χρήστης u_T ήδη ακολουθεί και το αποτέλεσμα είναι ένα σύνολο R που περιλαμβάνει υποψήφιους followees για τον χρήστη u_T:

$$R = T - S \qquad (4.5)$$

Τα παραπάνω βήματα στοχεύουν να δημιουργήσουν ένα σύνολο R που θα περιλαμβάνει χρήστες που έχουν όμοια ενδιαφέροντα με τον χρήστη u_T. Το επόμενο στάδιο λοιπόν του αλγορίθμου, είναι να ταξινομήσει κάθε χρήστη του συνόλου R έτσι ώστε να συστήσει τον πλέον κατάλληλο followee στον χρήστη u_T.

Για να ταξινομηθούν οι επικρατέστεροι χρήστες του συνόλου R έστω $u_c \in R$ χρησιμοποιείται η παρακάτω εξίσωση:

$$score(u_c) = \frac{Occurences(u_c, R)}{|R|} \; x \; \frac{|followers(u_c)|}{|followees(u_c)|} \; x \; \frac{|mentions(u_c)|}{M} \qquad (4.6)$$

Ο πρώτος όρος της παραπάνω εξίσωσης υποδηλώνει τον αριθμό των εμφανίσεων ενός χρήστη u_c στην τελική λίστα με τους πιθανούς followees. Ο δεύτερος όρος δίνει τον λόγο των χρηστών που ακολουθούν τον χρήστη u_c προς τους χρήστες που ο χρήστης u_c ακολουθεί. Τέλος, ο τρίτος παράγοντας δηλώνει τον αριθμό που αναφέρεται ένας χρήστης u_c σε μηνύματα προς μια σταθερά M που δηλώνει τον μέγιστο αριθμό αναφορών για ένα χρήστη και ορίζεται στο 100.

Κατά συνέπεια, λαμβάνοντας υπόψη την παραπάνω εξίσωση και δεδομένου του σκορ που λαμβάνει κάθε χρήστης u_c προτείνεται στον χρήστη u_T ως followee εκείνος με το μεγαλύτερο σκορ [6].

Content – based σύσταση

Οι βασικές υποθέσεις πάνω στις οποίες βασίζεται ο αλγόριθμος που υλοποιεί μια content – based σύσταση είναι ότι οι χρήστες επιλέγουν να ακολουθήσουν εκείνους τους άλλους χρήστες που είναι πιθανότερο να κοινοποιήσουν τα μηνύματά τους, ενώ παράλληλα τα ενδιαφέροντα του χρήστη σκιαγραφούνται από το περιεχόμενο που έχουν τα μηνύματα που αναρτούν τα άτομα που ακολουθεί. Μαθηματικά τα παραπάνω περιγράφονται με τους ακόλουθους τύπους.

Ορίζουμε ως

$$tweets(u_f) = \{t_1, t_2 \dots t_k\} \qquad (4.7)$$

το σύνολο των μηνυμάτων του χρήστη u_f που είναι ο κάθε followee του χρήστη u_T και για τους οποίους ορίζουμε $followees(u_T) = \{f_1, f_2 \dots f_l\}$ και επίσης ορίζουμε ως $\text{profile}_{\text{base}}(u_f)$ το διάνυσμα των $tweets(u_f) = \{t_1, t_2 \dots t_k\}$ και $followees(u_T) = \{f_1, f_2 \dots f_l\}$. Τέλος, ορίζουμε ότι το προφίλ του χρήστη για τον οποίο προορίζεται η σύσταση ως την ένωση των διανυσμάτων των ακολούθων του:

$$profile(u_T) = \bigcup_{\vee\ u_f \in followees(u_T)} \text{profile}_{\text{base}}(u_f) \qquad (4.8)$$

Ο συγκεκριμένος αλγόριθμος, προκειμένου να προβεί σε σύσταση ενός χρήστη δε λαμβάνει υπόψη του χρήστες από τη γειτονιά του χρήστη όπως ο παραπάνω αλγόριθμος που περιγράψαμε, αλλά προσπαθεί να εντοπίσει δυνητικούς ακολούθους με βάση τα μηνύματα των χρηστών που εμφανίζονται δημόσια στο χρονολόγιο του Twitter (Twitter's public timeline).

Εφόσον έχουμε ορίσει τα βασικά στοιχεία που χρησιμοποιεί ο συγκεκριμένος αλγόριθμος, παρακάτω περιγράφουμε τη διαδικασία που ακολουθεί για να προβεί σε σύσταση.

Βήμα 1: Δημιουργούμε το σύνολο U με U=$\{u_1, uf_2 \dots u_m\}$ που αποτελείται από τους πιο πρόσφατους χρήστες που έχουν δημοσιεύσει κάποιο μήνυμα στο χρονολόγιο.

Βήμα 2: Έπειτα για κάθε χρήστη $u_c \in U$ δημιουργούμε το διάνυσμα $\text{profile}_{\text{base}}(u_c)$

Βήμα 3: Στο τρίτο βήμα υπολογίζουμε την τιμή

$$sim(u_c, u_T) = max_{\vee\, i:\, f_i \in followees(u_T)} sim_{cos}[\text{profile}_{\text{base}}(f_i), \text{profile}_{\text{base}}(u_c)] \qquad (4.9)$$

Όπου sim_{cos} η ομοιότητα συνημίτονου των δύο διανυσμάτων.

Από τον παραπάνω υπολογισμό, αν $sim(u_c, u_T) > \gamma$, όπου γ κάποια σταθερά, τότε ο χρήστης u_c λογίζεται ως δυνητικός ακόλουθος του χρήστη u_T.

Από όλους τους δυνητικούς ακολούθους όπως προκύπτουν από το βήμα 3 θα συσταθεί εκείνος που έχει τη μεγαλύτερη τιμή ομοιότητας με τον χρήστη u_T [6].

4.2.2 Σταθμισμένες Content – Based Methods

Στην κατηγορία αυτή προσδιορίζονται ορισμένα χαρακτηριστικά στα οποία μπορεί να βασιστεί ένα σύστημα για να προβεί σε κάποια σύσταση [30]. Τέτοια χαρακτηριστικά είναι η δημοσιότητα, το πόσο δραστήριος είναι κάποιος χρήστης, η τοποθεσία, οι κοινοί φίλοι και το περιεχόμενο των μηνυμάτων. Από τα παραπάνω χαρακτηριστικά προκύπτει ότι σημαντική επίδραση για κάποια σύσταση έχουν η δημοσιότητα ενός followee και το πόσο ενεργός είναι [30]. Έτσι αν ένας χρήστης είναι δημοφιλής και δραστήριος, μόνο δημοφιλείς και δραστήριοι followees θα του προτείνονται. Με τον όρο δημοσιότητα νοείται ο λόγος followers προς followees και ένας χρήστης χαρακτηρίζεται ως δημοφιλής συγκρίνοντας την τιμή αυτού του λόγου με μια κριτική τιμή. Από την άλλη πλευρά, η δραστηριότητα μετράται με τον αριθμό των tweets που έχει αναρτήσει κάποιος χρήστης από την εγγραφή του στο Twitter και ανάλογα με τον αριθμό και τη σύγκριση του με κάποια κριτική τιμή, ο χρήστης χαρακτηρίζεται ως ενεργός ή μη. Έτσι το σύστημα εξετάζοντας τις τιμές ενός χρήστη για τη δημοσιότητα και τη δραστηριότητα προβαίνει σε συστάσεις followees που έχουν αντίστοιχες τιμές πάνω από κάποια κριτική τιμή που ορίζεται έμμεσα από τον συγκεκριμένο χρήστη.

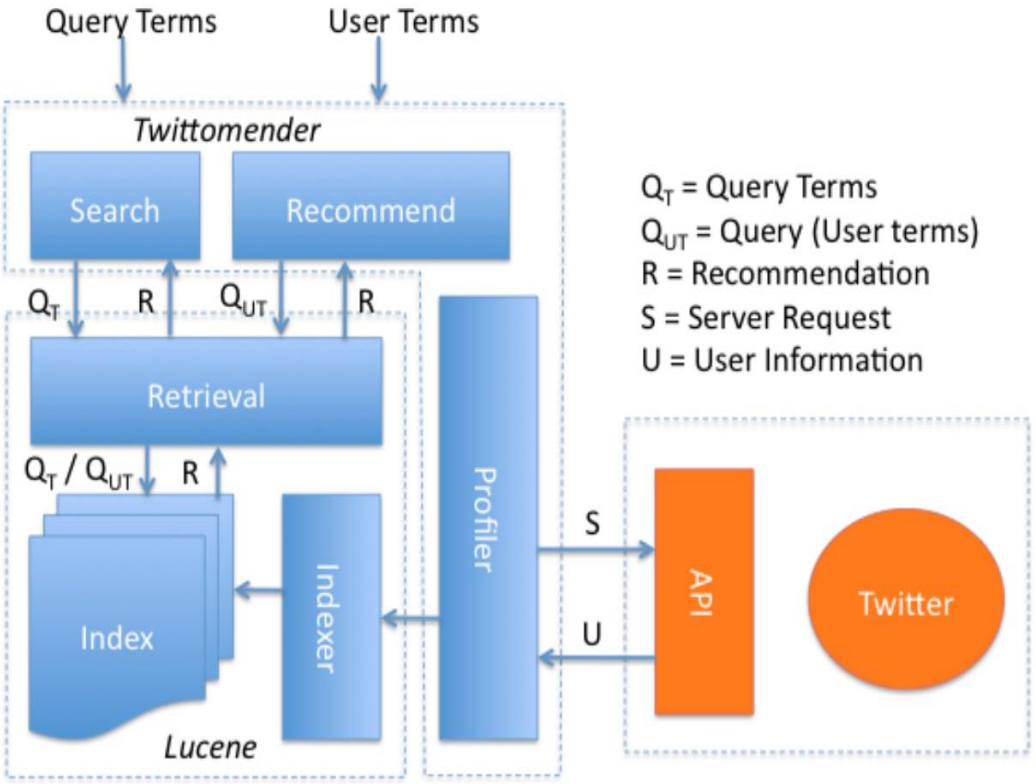
4.2.3 Σύστημα Twittomender

Το σύστημα αυτό προτείνει followees κάνοντας χρήση μεθόδων content – based και collaborative filtering [33, 34, 35]. Από τη σκοπιά της content – based μεθόδου, ο χρήστης προσδιορίζεται από 4 βασικά χαρακτηριστικά: τα tweets που αναρτά, τα tweets των followees του, τα tweets των followers του και τον συνδυασμό των παραπάνω. Έτσι προτείνονται στον χρήστη followees των οποίων τα tweets είναι όμοια με του χρήστη, είτε followees των οποίων οι followees αναρτούν παρόμοια tweets με τα tweets των followees του χρήστη κοκ.

Από τη σκοπιά της collaborative – filtering μεθόδου, ο χρήστης προσδιορίζεται από τα IDs των followees του, τα IDs των followers του ή συνδυασμό των δυο παραπάνω. Έτσι τα IDs αποτελούν το σημείο αναφοράς και κάθε χρήστης προσδιορίζεται από ένα συνδυασμό IDs των followees του. Κάνοντας, στη συνέχεια, χρήση του σχήματος στάθμισης TF-IDF προσδιορίζονται χρήστες με παρόμοιους συνδυασμούς IDs των followees τους.

Αρχιτεκτονική του συστήματος

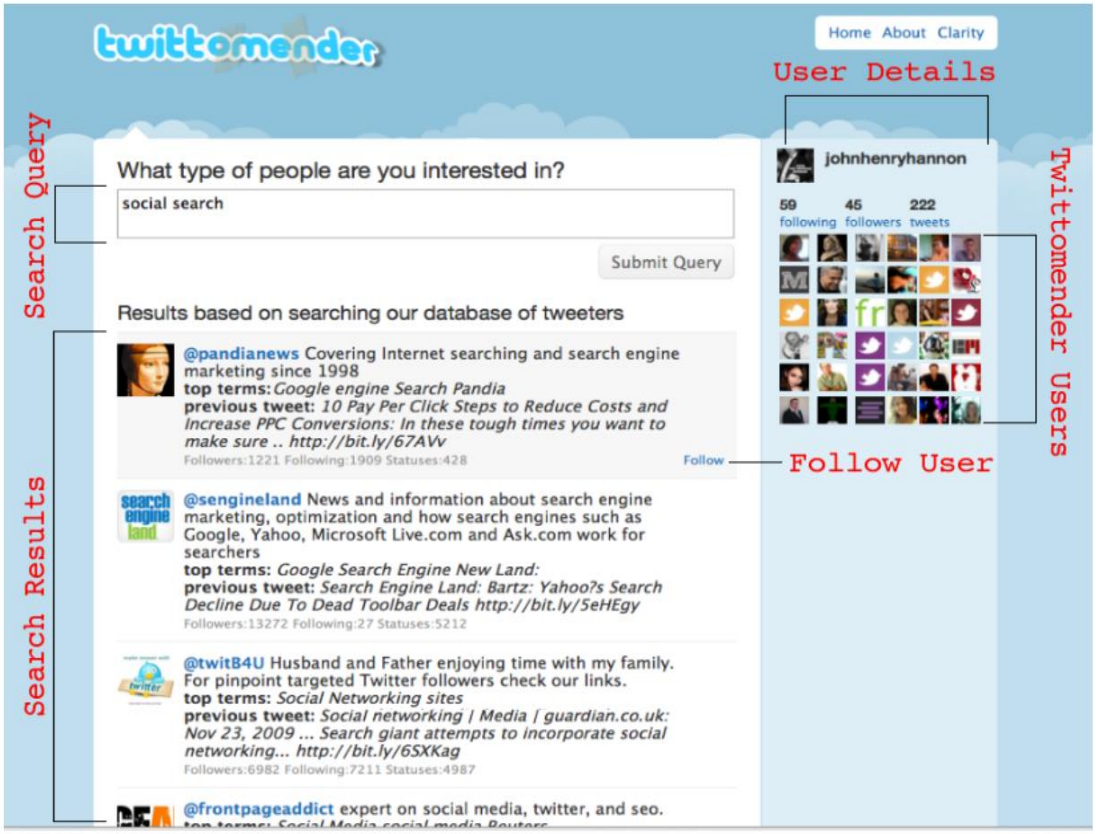
Η αρχιτεκτονική αυτού του συστήματος μπορεί να παρουσιαστεί εύκολα με το παρακάτω σχεδιάγραμμα.



Εικόνα 20: Αρχιτεκτονική του συστήματος [34]

Το σύστημα κάνει χρήση του Twitter API έτσι ώστε να συλλέξει πληροφορίες για τους χρήστες, τα μηνύματα που αναρτούν, τους followers και τους followees. Κάθε χρήστης προσδιορίζεται από διάφορα χαρακτηριστικά που σχετίζονται με τη δραστηριότητά του στο Twitter. Το σύστημα παρέχει στον χρήστη την πρόσβαση σε δυο λειτουργίες που σχετίζονται με την αναζήτηση ενός χρήστη (User search) καθώς και με τη σύσταση ενός χρήστη (User Recommendation). Παρακάτω περιγράφουμε αυτές τις δύο λειτουργίες.

User Search: Μέσω αυτής της λειτουργίας ο χρήστης μέσω λέξεων κλειδιά λαμβάνει μια λίστα με χρήστες σχετικούς ως προς την αναζήτησή του. Η εικόνα παρακάτω παρουσιάζει το αποτέλεσμα για την αναζήτηση με τον όρο *social search.* Το αποτέλεσμα της αναζήτησης είναι χρήστες διατεταγμένοι ανάλογα με τη σχετικότητα της δραστηριότητάς τους μέσω μηνυμάτων, αναφορών κτλ. στο Twitter με την αναζήτηση του χρήστη.



Εικόνα 21: Παράδειγμα αποτελέσματος αναζήτησης με λέξη κλειδί [34]

Από τα αποτελέσματα που θα δώσει η αναζήτηση, ο χρήστης μπορεί να επιλέξει έναν ή περισσότερους χρήστες για να συνδεθεί μαζί τους.

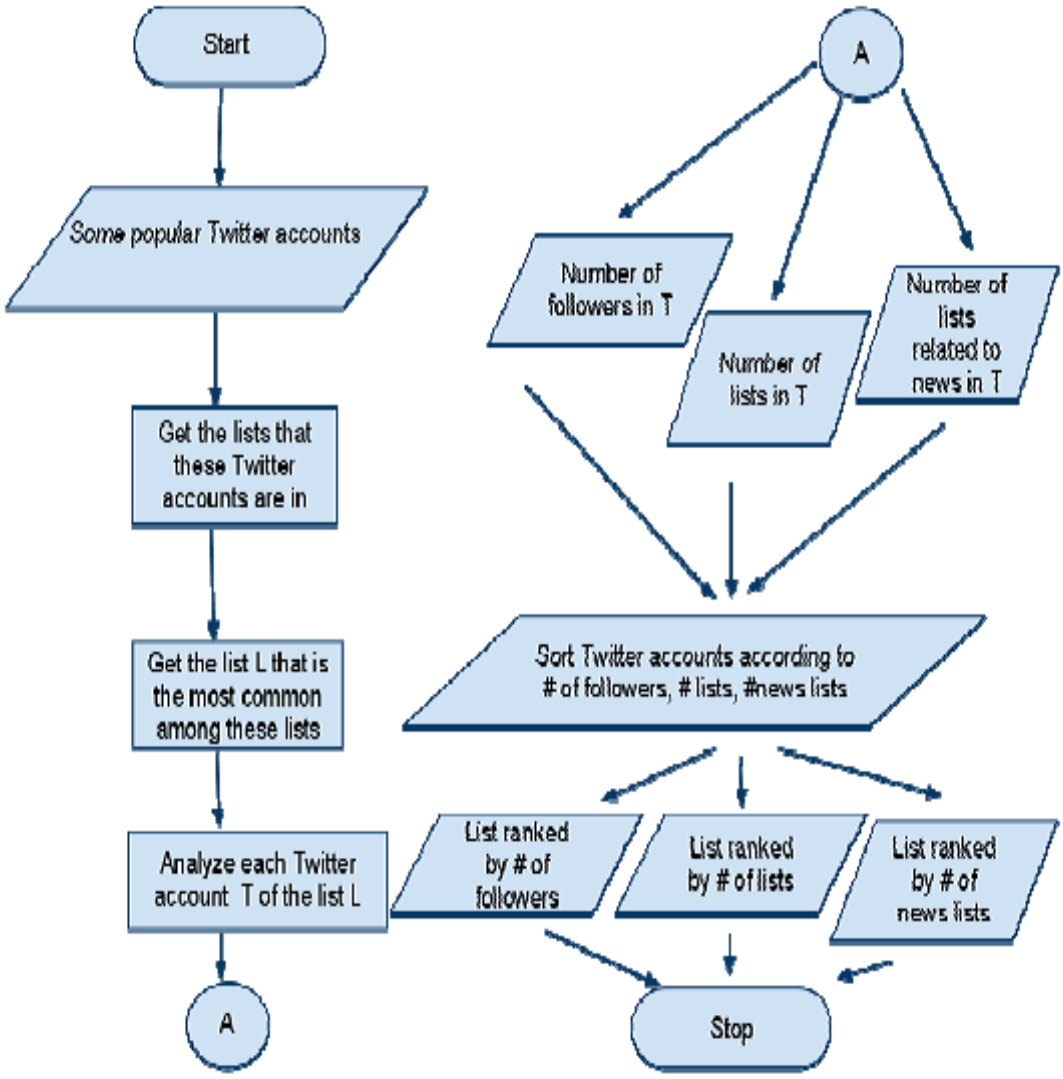
User Recommendation: Μέσω αυτής της λειτουργίας ο χρήστης μπορεί να χρησιμοποιήσει το προφίλ του ώστε να βρει άτομα που θα ταιριάζουν με αυτόν και να συνδεθεί. Αυτό γίνεται καταχωρώντας στο πεδίο αναζήτησης το προφίλ του χρήστη και το σύστημα προσπαθεί να εντοπίσει όλους εκείνους τους άλλους χρήστες που εμφανίζουν κοινά χαρακτηριστικά με το προφίλ του χρήστη που προβαίνει στην αναζήτηση. Έτσι, μέσα από αυτή τη μεθοδολογία το σύστημα καταλήγει σε μια λίστα που περιλαμβάνει σχετικούς χρήστες που θα ενδιέφεραν τον χρήστη που κάνει την αναζήτηση να συνδεθεί. Βασική προϋπόθεση για να παρέχει το συγκεκριμένο σύστημα αξιόπιστες προτάσεις σε έναν χρήστη είναι να προβεί σε συγχρονισμό του προφίλ του με την εφαρμογή, έτσι ώστε να αντλούνται όλες εκείνες οι πληροφορίες που απαιτούνται για μια αξιόπιστη σύσταση. Στην παρακάτω εικόνα μπορούμε να παρατηρήσουμε πως υλοποιείται αυτή η λειτουργία παρουσιάζοντας όλα όσα αναφέραμε παραπάνω στην περιγραφή.

Εικόνα 22: Αποτελέσματα αναζήτησης με βάση το προφίλ [34]

4.2.4 Συστάσεις βάσει followers και lists

Η τελευταία μέθοδος που θα περιγράψουμε σχετικά με τη σύσταση followees σε έναν χρήστη βασίζεται στον αριθμό των followers που έχει ο συγκεκριμένος χρήστης καθώς και στον αριθμό των γκρουπ που συμμετέχει. Η συγκεκριμένη μέθοδος προτείνει στον χρήστη followees με αντίστοιχο αριθμό followers ή συμμετοχές σε γκρουπ ενώ έχει διερευνηθεί ότι σημαντικότερο ρόλο παίζει ο αριθμός των followers [42].

Το παρακάτω σχεδιάγραμμα περιγράφει τον τρόπο με τον οποίο λειτουργεί ο αλγόριθμος που βασίζεται στις λίστες και στους followers για να προβεί σε συστάσεις.

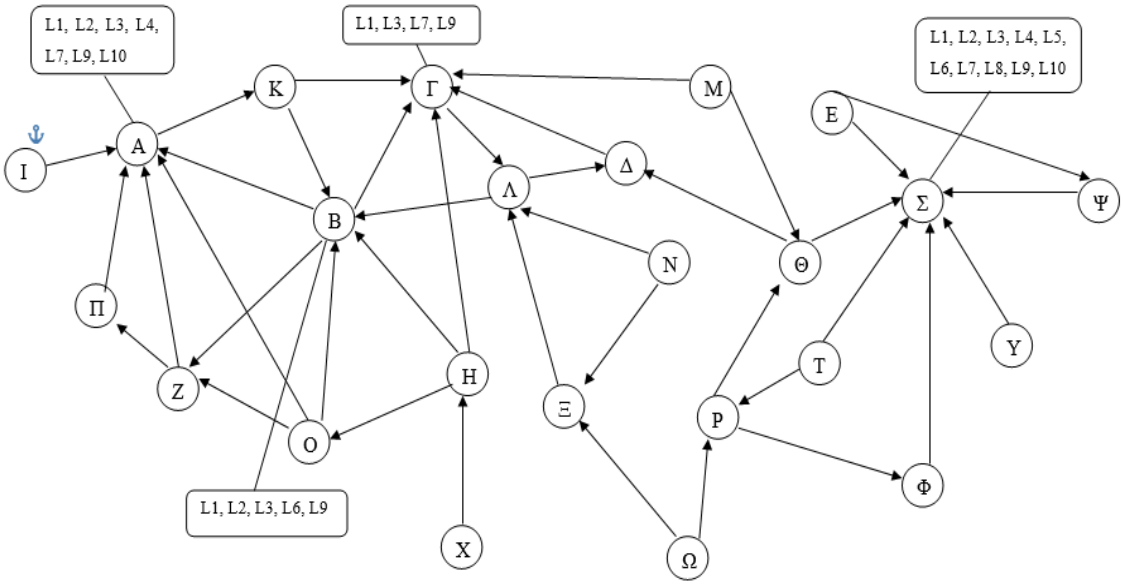


Εικόνα 23: Λειτουργία αλγορίθμου [42]

4.2.5 Ένα toy example για τους παραπάνω αλγόριθμους

Σε αυτήν την ενότητα θα παρουσιάσουμε ένα toy example που θα εφαρμοστεί πάνω στους αλγόριθμους που περιγράψαμε για τη σύσταση followees σε έναν χρήστη και πιο συγκεκριμένα στον αλγόριθμο που κάνει χρήση του collaborative filtering όπως παρουσιάστηκε στην ενότητα 4.2.1 καθώς επίσης και για τον αλγόριθμο που βασίζεται στους followers και στις λίστες έτσι όπως παρουσιάστηκε στην ενότητα 4.2.4.

Παρακάτω παρουσιάζεται η εικόνα του δικτύου με τα κυκλικά σχήματα να αναπαριστώνται οι χρήστες, με τα παραλληλόγραμμα οι λίστες στις οποίες συμμετέχει ένας χρήστης, ενώ η σχέση A→ B χρησιμοποιείται για να υποδείξει ότι ο A ακολουθεί τον B και κατά συνέπεια ο A είναι follower για τον B και ο B followee για τον A.



Εικόνα 24 : Toy Example

Εφαρμογή του Toy Example για τον αλγόριθμο collaborative filtering

Έστω ότι θέλουμε να συστήσουμε έναν followee για τον χρήστη Ο. Ακολουθώντας τη διαδικασία που περιγράφει ο αλγόριθμος collaborative – filtering στην ενότητα 4.2.1 έχουμε:

Βήμα 1: Το σύνολο S των followees του χρήστη Ο θα είναι:

$$S(O) = A, B, Z$$

Βήμα 2: Το σύνολο L των followers για κάθε στοιχείο του $S(O)$ θα είναι:

$$L(A) = I, \Pi, Z, O, B$$

$$L(B) = H, K, \Lambda, O$$

$$L(Z) = B, O$$

Συνεπώς το σύνολο L= $L(A)\ U\ L(B)\ U\ L(Z) = B, Z, H, I, K, \Lambda, O, \Pi$

Βήμα 3: Το σύνολο T των followees κάθε στοιχείου του L θα είναι:

$$T(B) = A, \Gamma, Z$$

$$T(I) = A$$

$$T(\Pi) = A$$

$$T(Z) = A, \Pi$$

$$T(O) = A, B, Z$$

$$T(K) = B, \Gamma$$

$$T(\Lambda) = B, \Delta$$

$$T(H) = B, \Gamma, O$$

Συνεπώς το σύνολο Τ είναι:

Τ= $A, \Gamma, Z, A, A, A, \Pi, A, B, Z, B, \Gamma, B, \Delta, B, \Gamma, O$

Βήμα 4: Εφαρμόζοντας την εξίσωση 4.5 έχουμε ότι:

$$R = T - S = \Gamma, \Delta, \Pi$$

Σύμφωνα με τα παραπάνω, καταλήγουμε στο σύνολο R που περιλαμβάνει τους δυνητικούς followees για τον χρήστη Ο. Το τελευταίο στάδιο προκειμένου να συσταθεί ο καταλληλότερος χρήστης στον Ο είναι να υπολογίσουμε το score κάθε δυνητικού followee σύμφωνα με την εξίσωση 4.6. Προκειμένου να εφαρμόσουμε την εξίσωση αυτή θα κάνουμε ορισμένες υποθέσεις σχετικά με τον αριθμό των αναφορών κάθε χρήστη σε μηνύματα, ενώ η σταθερά Μ ισούται με 100. Έτσι έχουμε:

Για τον χρήστη Γ

$Occurences(\Gamma, R) = 1$

$|followers(\Gamma)|$= 5

$|followees(\Gamma)|$= 1

$|mentions(\Gamma)|$= 200

Άρα εφαρμόζοντας την εξίσωση 4.6 έχουμε:

$$score(\Gamma) = \frac{1}{3}\ x\ \frac{5}{1}\ x\ \frac{200}{100}\ = 3{,}33$$

Για τον χρήστη Δ

$Occurences(\Delta, R) = 1$

$|followers(\Delta)|$= 2

$|followees(\Delta)|$= 1

$|mentions(\Delta)|$= 80

Άρα εφαρμόζοντας την εξίσωση 4.6 έχουμε:

$$score(\Delta) = \frac{1}{3}\ x\ \frac{2}{1}\ x\ \frac{80}{100}\ = 0{,}53$$

Για τον χρήστη Π

$Occurences(\Pi, R) = 1$

$|followers(\Pi)|= 1$

$|followees(\Pi)|= 1$

$|mentions(\Pi)|= 40$

Άρα εφαρμόζοντας την εξίσωση 4.6 έχουμε:

$$score(\Pi) = \frac{1}{3}\ x\ \frac{1}{1}\ x\ \frac{40}{100}\ = 0{,}13$$

Κατά συνέπεια, σύμφωνα με τα παραπάνω από την σύγκριση των scores προκύπτει ότι ο χρήστης που θα συσταθεί ως followee στον χρήστη Ο είναι ο χρήστης Γ που συγκέντρωσε το μεγαλύτερο score σύμφωνα με την εξίσωση 4.6

Εφαρμογή του Toy Example για τον αλγόριθμο με βάση τους follower lists

Στο δίκτυο όπως παρουσιάζεται παραπάνω παρατηρούμε ότι οι δημοφιλέστεροι χρήστες, δηλαδή αυτοί που ακολουθούνται από τους πιο πολλούς χρήστες είναι οι Α, Β, Γ, Σ. Έπειτα παίρνουμε το σύνολο L που αποτελείται από τις λίστες στις οποίες συμμετέχουν αυτοί οι δημοφιλής χρήστες και έτσι έχουμε ότι L=L1,L2,L3,L4,L5,L6,L7,L8,L9,L10. Η λίστα που εμφανίζεται να είναι κοινή σε όλους τους χρήστες είναι η L1. Επομένως υπάρχουν τρία κριτήρια με βάση τα οποία θα κατηγοριοποιηθούν οι χρήστες και ανάλογα το κριτήριο θα προταθεί ο κατάλληλος χρήστης ως followee. Για παράδειγμα, με βάση τον αριθμό των ακολούθων εύκολα διαπιστώνουμε ότι προτείνεται ο χρήστης Σ γιατί έχει τους περισσότερους, ενώ με βάση τον αριθμό των λιστών στις οποίες συμμετέχει βλέπουμε ότι προτείνεται ξανά ο Σ αφού συμμετέχει στις περισσότερες λίστες.

4.3 Μέθοδοι συστάσεων για follower

Παρότι η πλειοψηφία των χρηστών στο twitter ενδιαφέρεται να ανακαλύψει άτομα να ακολουθήσει (followee recommendation) υπάρχουν και χρήστες που ενδιαφέρονται να προσελκύσουν ακολούθους, δηλαδή followers, που θα μπορούν να αναδημοσιεύσουν τα μηνύματά τους ή τις ειδήσεις που κάνουν tweet. Τέτοιοι χρήστες είναι συνήθως πολιτικοί αλλά και επιχειρηματίες που επιδιώκουν να προωθήσουν είτε ιδέες είτε προϊόντα. Παρακάτω θα περιγράψουμε τη μέθοδο Tadvise [53] που χρησιμοποιείται για τη σύσταση followers.

4.3.1 Μέθοδος σύστασης Tadvise

Το Tadvise είναι μια εφαρμογή που βοηθά τους χρήστες του twitter ώστε να βρουν εκείνους τους followers που θα μπορούσαν να διαδώσουν τα μηνύματα του χρήστη με τον πλέον αποδοτικό τρόπο σε μεγαλύτερα κοινά όμοιων ενδιαφερόντων με τον χρήστη. Το Tadvise μέσω του αλγορίθμου που χρησιμοποιεί για την υλοποίησή του, προσπαθεί να προσθέσει στα μηνύματα ενός χρήστη εκείνους του ακολούθους των οποίων οι ακόλουθοι είναι πιο πιθανό ότι θα συμμεριστούν τα ενδιαφέροντα του αρχικού χρήστη από ότι άλλοι. Το Tadvise είναι ένα πολύ χρήσιμο εργαλείο για εκείνους τους χρήστες του twitter που έχουν ως στόχο να διαδώσουν μια είδηση, να ανακοινώσουν ένα γεγονός και γενικότερα να μοιραστούν πληροφορίες μέσα σε ένα σύνολο από άτομα με παρόμοια ενδιαφέροντα [53].

Περιγραφή της μεθόδου

Το Tadvise δημιουργεί προφίλ χρηστών στο twitter έτσι ώστε να συστήσει tweets ή retweets τα οποία θα μπορούσαν να είναι σχετικά στον κύκλο των ακολούθων του χρήστη. Για να γίνει αυτό, ο χρήστης έστω u θα πρέπει να εγγραφεί στο Tadvise με το να ακολουθήσει τον λογαριασμό του Tadvise στο twitter (@Tadvise). Αφού γίνει αυτό, το Tadvise ανακτά το κοινωνικό δίκτυο που αποτελείται από τους followers του χρήστη u και δημιουργεί προφίλ των ακολούθων του χρήστη. Μετά τη δημιουργία των προφίλ, μια ενέργεια που πραγματοποιείται offline, το Tadvise ενημερώνει τον χρήστη με ένα άμεσο μήνυμα ότι είναι έτοιμο να παρέχει *συμβουλή*. Έτσι, ο χρήστης μπορεί να επισκεφθεί τη σελίδα του Tadvise και μέσω αυτού να αναρτήσει διάφορα μηνύματα.

Το Tadvise χρησιμοποιεί στο γραφικό του περιβάλλον τρία διαφορετικά χρώματα ως ενδείξεις. Η πράσινη ένδειξη υποδηλώνει ότι η πλειοψηφία των ακολούθων του χρήστη u προστέθηκαν με hashtags στο μήνυμα του χρήστη. Η κόκκινη ένδειξη υποδηλώνει ότι κανένας από τους ακόλουθους του χρήστη δεν προστέθηκε στο μήνυμα, ενώ η κίτρινη ένδειξη υποδηλώνει ότι μερικοί από τους ακολούθους του χρήστη προστεθήκαν.

Το Tadvise λειτουργεί σε τρία στάδια. Το πρώτο στάδιο περιλαμβάνει την ανάκτηση πληροφοριών (crawler), το δεύτερο στάδιο περιλαμβάνει τη δημιουργία προφίλ (user profile builder) και το τρίτο στάδιο περιλαμβάνει τον μηχανισμό παροχής σύστασης. (advice engine). Παρακάτω θα περιγράψουμε αναλυτικά κάθε στάδιο, προτού όμως, θα ορίσουμε το σύστημα πάνω στο οποίο θα περιγράψουμε τα παραπάνω στάδια.

Ορίζουμε ένα σύστημα S που αποτελείται από n χρήστες έτσι ώστε:

$$U = \{u_1, u_2 \dots u_n\} \qquad (4.10)$$

Σε αυτό το σύνολο υπάρχουν R αμφίδρομες σχέσεις μεταξύ των χρηστών έτσι ώστε αν ένας χρήστης u_i συνδεθεί με έναν χρήστη u_j με μια σχέση $r_{i,j} \epsilon R$ καλούμε τον u_i follower του u_j και τον u_j followee του u_i. Τη σχέση αυτή την ορίζουμε μαθηματικά ως $u_i \rightarrow u_j$. Επίσης, υποθέτουμε ότι το σύστημα S είναι ανοιχτό έτσι ώστε κάθε χρήστης να μπορεί να συνδεθεί με κάποιον άλλο χρήστη. Το σύνολο των followers / followees ενός χρήστη u_i ορίζεται ως U_i^{fr} και U_j^{fo} αντίστοιχα. Κάθε χρήστης u_i μπορεί να αναθέσει κανένα ή περισσότερα tags $\{t_1, t_2 \dots t_m\}$ σε καθένα από τους followees του. Τέλος, ορίζεται μια συνάρτηση που λαμβάνει ως εισροή έναν χρήστη u_j και δίνει ως εκροή ένα ζεύγος (u_i, t_k) υποδεικνύοντας ότι ο χρήστης u_i έχει προσθέσει τον χρήστη u_j με το t_k tag.

Ανάκτηση Πληροφοριών

Αυτό το στάδιο του Tadvise υλοποιείται σε 2 βήματα. Το πρώτο βήμα περιλαμβάνει την ανάκτηση, μέσω του API του twitter, του δικτύου των followers ενός χρήστη και το δεύτερο βήμα, ανακτά τις λίστες στις οποίες συμμετέχουν οι followers έτσι όπως έχουν ανακτηθεί από το πρώτο βήμα.

Δημιουργία προφίλ των χρηστών

Προκειμένου να εξεταστεί η σχετικότητα ενός μηνύματος με έναν μεμονωμένο χρήστη u_j, το Tadvise δημιουργεί σταθμισμένα προφίλ των χρηστών με βάση τα ενδιαφέροντά τους, τα δίκτυά τους κτλ. Για τη δημιουργία ενός σταθμισμένου με βάση τα χαρακτηριστικά του χρήστη προφίλ, θα πρέπει προηγουμένως να υπολογιστεί ο βαθμός των tags ενός χρήστη. Αυτό γίνεται με την παρακάτω εξίσωση:

$$rank(u_i) = \log(\# U_j^{fo}) \qquad (4.11)$$

Από την παραπάνω σχέση μπορούμε να γενικεύσουμε και να γράψουμε ότι

$$rank(u_i) = \sum_{u_j \in U_i^{fo}} rank(u_j) \qquad (4.12)$$

Η συνεισφορά μιας λίστας στο προφίλ ενός χρήστη υπολογίζεται από το άθροισμα των βαθμών εκείνων που έχουν αναθέσει τη συγκεκριμένη λίστα στον χρήστη u_j , δηλαδή

$$weight(t_k, u_j) = \sum_{(u_i, t_k) \in lists(u_j)} rank(u_j) \qquad (4.13)$$

Μηχανισμός παραγωγής συστάσεων

Σε αυτό το στάδιο, ο μηχανισμός παραγωγής συστάσεων λαμβάνει ως εισροές τα προφίλ των χρηστών και τα tweets και παράγει δυο ειδών συστάσεις σε πραγματικό χρόνο:

1. Το σύνολο των ακολούθων που σχετίζονται και προστίθενται σε ένα συγκεκριμένο tweet
2. Το σύνολο εκείνων των χρηστών που είναι πιθανότερο να διαδώσουν τα μηνύματα του αρχικού χρήστη

```
input  : Directed graph (g)
         Integer k // number of recommended hubs
output: candidates ⊂ g
candidates ← ∅;
covered ← ∅;
while size(candidates)!=k do
    calculate hubs in g and sort them based on hubs scores;
    node ← get the node with the highest score of hubs, so that
    followers(node) ∩ covered is minimum;
    candidates ← candidates ∪ node ;
    covered ← covered ∪ followers(node) ;
    g ← g − followers(node) − node ;
    if g == root(g) then break;
end
return candidates;
```

Εικόνα 25 : Αλγόριθμος παραγωγής συστάσεων [53]

Ο παραπάνω αλγόριθμος υλοποιείται για να δώσει ως σύσταση εκείνους τους πιθανούς ως followers για έναν χρήστη που είναι πιο σχετικοί με τα μηνύματα που αναρτά αυτός ο συγκεκριμένος χρήστης. Ο συγκεκριμένος αλγόριθμος λαμβάνει ως είσοδο έναν γράφο g ο οποίος αποτελείται από όλους τους followers ενός χρήστη u. Επίσης, το g περιλαμβάνει και όλους εκείνους τους χρήστες που είναι οι followees των followers του χρήστη για τον οποίο προορίζεται η σύσταση. Ο λόγος για τον οποίο συμπεριλαμβάνονται οι followees των followers του χρήστη είναι γιατί κάθε φορά που ένας follower του χρήστη δημοσιεύει ένα μήνυμα τότε αυτό μπορεί να κοινοποιηθεί και να χρησιμοποιηθεί ως σύνδεσμος από έναν followee του αφού τα μηνύματα των followers είναι ορατά στους followees τους.

Μετά την εισαγωγή της βάσης των χρηστών βάσει των οποίων θα γίνει η σύσταση, ο αλγόριθμος βρίσκει τους k χρήστες που είναι πιο σχετικοί με το περιεχόμενο όσο το δυνατόν περισσότερων μηνυμάτων έτσι ώστε να μην υπάρχει ή να υπάρχει όσο το δυνατόν μικρότερη αλληλοκάλυψη μεταξύ των ακολούθων ενός χρήστη. Ο αλγόριθμος έχει μια σταθερή τιμή για το k που είναι ίση με 3. Στον αλγόριθμο το *hub score* καταδεικνύει τον αριθμό εκείνων των χρηστών που είναι πιο σχετικοί και έχουν μεγαλύτερη πιθανότητα να δεχτούν ένα μήνυμα μέσω ενός σημείου αναφοράς.

Τέλος, επειδή το κάθε μήνυμα στο twitter έχει ένα συγκεκριμένο αριθμό χαρακτήρων, ο αλγόριθμος λαμβάνει υπόψη το μέγεθος του ονόματος του χρήστη που θα υποδείξει ως σύσταση στον χρήστη στόχο. Αυτό σημαίνει ότι αν εντοπιστούν δυο πανομοιότυποι χρήστες ως προς το ενδιαφέρον σχετικά με ένα μήνυμα και υπό την προϋπόθεση ότι και οι δυο αυτοί χρήστες θα αποτελούσαν πιθανές συστάσεις από το σύστημα προς τον χρήστη για τον οποίο προορίζεται η σύσταση, τότε ο αλγόριθμος θα επέλεγε και θα πρότεινε στον χρήστη στόχο εκείνον τον χρήστη ως πιθανό follower του οποίου το όνομα θα είχε το μικρότερο μήκος.

4.4 Μέθοδοι συστάσεων για Hashtags

Παραπάνω αναλύσαμε συστήματα και μεθόδους που κάνουν συστάσεις σε ένα χρήστη σχετικά με τον ποιον να ακολουθήσουν ή από ποιους να ακολουθηθούν. Ωστόσο, στο Twitter είναι ευρύτατα διαδεδομένη και ιδιαίτερα συχνή η χρήση hashtags. Τον ορισμό της έννοιας αυτής τον δώσαμε παραπάνω και ουσιαστικά ένα hashtag χρησιμοποιείται από τον χρήστη είτε για να κατηγοριοποιήσει τα μηνύματα που αναρτά, είτε για να μεταδώσει μαζικά ειδήσεις. Δεδομένου του τεράστιου αριθμού των χρηστών του

Twitter και κατ' επέκταση του τεράστιου αριθμού των μηνυμάτων που αναρτώνται με χιλιάδες διαφορετικά hashtags και παράλληλα, δεδομένου του ότι το hashtag δεν είναι σαφώς ορισμένο όπως μια λίστα ή ένα γκρουπ στο Twitter, η προσπάθεια του χρήστη να βρει το κατάλληλο hashtag για κάποιο μήνυμά του είναι ιδιαίτερα δύσκολη. Προς αυτήν την κατεύθυνση, την υποβοήθηση στην εύρεση του κατάλληλου hashtag, προσανατολίζονται οι μέθοδοι σύστασης hashtags [35, 38, 45,].

Παρακάτω θα περιγράψουμε δύο τρόπος με τους οποίους υλοποιούνται τέτοια συστήματα συστάσεων.

4.4.1 Σύσταση hashtags με το σχήμα TF-IDF

Ο τρόπος αυτός σύστασης υποθέτει ότι ο βασικός σκοπός χρήσης ενός hashtag είναι η κατηγοριοποίηση των μηνυμάτων ώστε να είναι πιο εύκολη η αναζήτησή τους [76]. Έτσι, προτείνονται hashtags κάθε φορά με βάση το περιεχόμενο των μηνυμάτων που συνήθως αναρτά ένας χρήστης χωρίς να δίνεται σημασία στην προτίμηση ενός συγκεκριμένου hashtag από τον χρήστη. Η μέθοδος αυτή ακολουθεί τη διαδικασία που περιγράφουμε αμέσως παρακάτω. Αρχικά, όταν ένας χρήστης γράφει ένα μήνυμα το σύστημα ανακτά ένα σύνολο μηνυμάτων, εκ των οποίων τα όμοια μηνύματα ανακτώνται και κατατάσσονται με το σχήμα διαβάθμισης ανάκτησης πληροφορίας TF-IDF. Έπειτα τα hashtags των μηνυμάτων κατατάσσονται ανάλογα:

i. Με το πόσες φορές εμφανίζονται στο σύνολο των μηνυμάτων
ii. Με το πόσες φορές εμφανίζονται στο σύνολο των όμοιων μηνυμάτων
iii. Με την ομοιότητα των μηνυμάτων

Έχει παρατηρηθεί ότι από τα παραπάνω πιο αποτελεσματικό είναι το iii και επομένως συστήνεται το hashtag των περισσότερο όμοιων μηνυμάτων.

4.4.2 Σύσταση hashtags με το μοντέλο του Bayes

Αυτός ο τρόπος σύστασης ομοιάζει με τον τρόπο που αναφέραμε παραπάνω, με την έννοια ότι βασίζεται και αυτός στο περιεχόμενο των μηνυμάτων του χρήστη, αλλά για να προβεί σε σύσταση κάνει χρήση του μοντέλου του Bayes προσδιορίζοντας την πιθανότητα χρήσης ενός hashtag. Για τον υπολογισμό της πιθανότητας γίνεται χρήση του παρακάτω μοντέλου [49]:

$$p(C_i|x_{1,} \ldots\ldots, x_{n,}) = \frac{p(C_i)p(x_1|C_i) \ldots\ldots p(C_i)p(x_n|C_i)}{p(x_1 \ldots\ldots x_n)} \quad (4.14)$$

Όπου,

C_i το i-οστό hashtag

$x_{1,} \ldots\ldots, x_{n,}$ οι λέξεις του μηνύματος

Ο παράγοντας $p(C_i|x_{1,} \ldots\ldots, x_{n,})$ μας δίνει την πιθανότητα να χρησιμοποιηθεί το C_i hashtag δεδομένων των λέξεων που χρησιμοποιεί ο χρήστης. Φυσικά, το hashtag με τη μεγαλύτερη τιμή πιθανότητας θα είναι και αυτό που θα συσταθεί στον χρήστη.

4.5 Μέθοδοι συστάσεων για Tweets

Αν παρατηρήσουμε την αρχική σελίδα ενός χρήστη στο Twitter θα διαπιστώσουμε ότι δέχεται σωρεία διαφορετικών μηνυμάτων διάφορων θεματολογιών. Ειδικά όταν κάποιος χρήστης ακολουθεί μεγάλο αριθμό followees, είναι πολύ πιθανό να μην εντοπίζει εύκολα tweet που του είναι πιο ενδιαφέροντα αφού θα χάνονται μέσα στο πλήθος των μηνυμάτων που εμφανίζονται. Έτσι, συστήματα συστάσεων μπορούν να χρησιμοποιηθούν για να φιλτράρουν τα μηνύματα που δέχεται ο χρήστης και να τα κατατάξουν με σειρά εμφάνισης ανάλογα με τις προτιμήσεις του χρήστη. Στη βιβλιογραφία προτείνονται δύο μέθοδοι για αυτό το σκοπό που σχετίζονται με την ταξινόμηση των tweets και των χρηστών [75].

4.5.1 Σύσταση Tweets μέσω ταξινόμησης

Σύμφωνα με την ταξινόμηση των tweets, για κάθε χρήστη τα μηνύματα που του εμφανίζονται κατατάσσονται ανάλογα με την πιθανότητα που έχει κάθε μήνυμα να αναδημοσιευτεί από τον συγκεκριμένο χρήστη. Από την άλλη πλευρά, με βάση την ταξινόμηση των χρηστών, για κάθε tweet οι χρήστες κατατάσσονται ανάλογα με την πιθανότητα που έχουν να το αναδημοσιεύσουν. Κατά συνέπεια, στον χρήστη συστήνεται εκείνο το tweet που είναι πιθανότερο να κάνει retweet, δηλαδή να το αναδημοσιεύσει.

Όσον αφορά τη λειτουργία του retweet όπως την περιγράψαμε στην εισαγωγική ενότητα αυτού του κεφαλαίου, παρότι στη βιβλιογραφία και στην πράξη δεν έχουν αναπτυχθεί ξεχωριστά συστήματα μέσω των οποίων να προτείνεται η σύστασή τους, με βάση την παραπάνω ανάλυση, μπορούμε να ισχυριστούμε ότι μέθοδος σύστασης

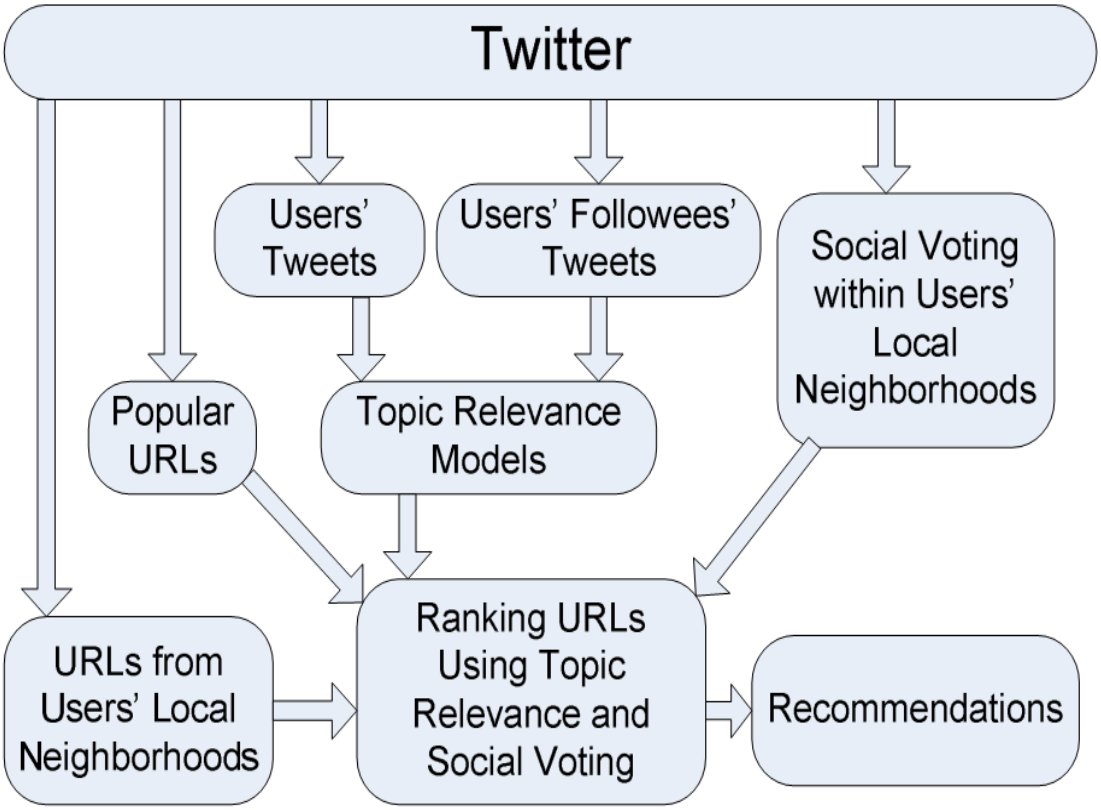
retweet μπορεί να αποτελέσει η μέθοδος που περιγράψαμε στην ενότητα 4.5.1 αφού ουσιαστικά η σύσταση ενός tweet σε έναν χρήστη αποτελεί ταυτόχρονα και σύσταση retweet. Αυτό συμβαίνει γιατί ένα συγκεκριμένο μήνυμα που θα προταθεί στον χρήστη προτείνεται ακριβώς γιατί έχει μεγαλύτερη πιθανότητα να γίνει retweet, δηλαδή πρόκειται για έμμεση μέθοδο σύστασης tweet. Ωστόσο, στην περίπτωση που το μήνυμα που προτείνεται σε κάποιο χρήστη είναι ήδη αναδημοσιευμένο από κάποιον άλλο χρήστη, τότε στην ουσία πρόκειται για σύσταση ενός retweet.

4.6 Μέθοδοι συστάσεων συνδέσμων από Tweets

Όπως έχουμε ήδη περιγράψει, ένα tweet ή retweet γράφεται από τον χρήστη ή διαδίδεται αντίστοιχα από αυτόν. Τα μηνύματα λοιπόν που συνθέτει ή μεταδίδει ο χρήστης αντανακλούν τις προτιμήσεις και τα ενδιαφέροντά του. Κατά συνέπεια διάφορα συστήματα μπορούν να στηριχτούν σε αυτό ώστε να προβούν σε συστάσεις. Παρακάτω θα περιγράψουμε δυο μεθόδους που κάνουν χρήση των tweets και retweets ώστε να προτείνουν συνδέσμους (URLs) σχετικούς με τα ενδιαφέροντα ενός χρήστη.

4.6.1 URL Recommendation system

Το σύστημα αυτό [21] προβαίνει σε σύσταση συνδέσμων που είναι σχετικοί με τα ενδιαφέροντα του χρήστη βασισμένο σε ένα σύνολο υποψήφιων συνδέσμων από διάφορες πηγές και έπειτα κατάταξη αυτών των συνδέσμων με βάση διαφορετικά κριτήρια. Η παρακάτω εικόνα παρουσιάζει σχηματικά το μοντέλο του συστήματος για σύσταση συνδέσμων.



Εικόνα 26: Λειτουργία συστήματος συστάσεων συνδέσμων [21]

Επιλογή υποψήφιων συνδέσμων

Για την επιλογή των συνδέσμων γίνεται χρήση δυο διαφορετικών μεθόδων. Η μια μέθοδος είναι η μέθοδος FoF, followee-of-followees. Με βάση αυτή τη μέθοδο επιλέγονται σύνδεσμοι που δημοσιεύονται στη γειτονιά του χρήστη, δηλαδή σύνδεσμοι που δημοσιεύονται από τους followees ενός χρήστη και από τους followees των followees του χρήστη. Η βασική ιδέα πίσω από την επιλογή αυτών των συνδέσμων είναι ότι ο χρήστης συνήθως βρίσκει ενδιαφέροντες συνδέσμους που δημοσιεύονται από άτομα που συνδέονται με τον κύκλο του και άρα έχουν κοινά ενδιαφέροντα παρά από άτομα που του είναι άγνωστα και ενδεχομένως να έχουν ή όχι κοινά ενδιαφέροντα. Η δεύτερη μέθοδος που χρησιμοποιείται για την επιλογή των συνδέσμων σχετίζεται με την δημοφιλία των συνδέσμων (popularity). Αυτή η μέθοδος στηρίζεται στην ιδέα ότι διάσημοι σύνδεσμοι που δημοσιεύονται συχνά και από μεγάλο αριθμό χρηστών, σίγουρα θα ενδιαφέρουν πολύ περισσότερο έναν χρήστη από ότι σύνδεσμοι που δεν αναπαράγονται από κανέναν ή λίγους χρήστες.

Κατάταξη συνδέσμων

Αφού έχει καθοριστεί το σετ των συνδέσμων που θα χρησιμοποιηθούν, το επόμενο βήμα του συστήματος αυτού είναι να κατατάξει τους συνδέσμους με βάση διάφορες μεθόδους και διάφορα κριτήρια. Υπάρχουν δυο μέθοδοι που συμβάλλουν στην κατάταξη των συνδέσμων.

Κατάταξη με βάση τη σχετική θεματολογία

Η πρώτη μέθοδος είναι η κατάταξη με βάση τη σχετικότητα θέματος και κάνει χρήση δύο κριτηρίων για την κατάταξη των συνδέσμων:

- Το πρώτο κριτήριο είναι η ομοιότητα των tweets που εμπεριέχουν έναν υποψήφιο σύνδεσμο, όπως έχει προσδιοριστεί από τις παραπάνω μεθόδους, με τα tweets του χρήστη για τον οποίο προορίζεται η σύσταση.
- Το δεύτερο κριτήριο είναι η ομοιότητα των μηνυμάτων που εμπεριέχουν υποψήφιους συνδέσμους με τα μηνύματα των followees του συγκεκριμένου χρήστη.

Κατάταξη με βάση την κοινωνικότητα

Με βάση αυτό το κριτήριο οι υποψήφιοι σύνδεσμοι κατατάσσονται με βάση τη συχνότητα με την οποία αναρτώνται στο δίκτυο καθώς και με βάση τη δύναμη αυτού που τον αναρτά. Με την έννοια δύναμη ενός χρήστη εδώ εννοείται ο λόγος των followers ενός χρήστη σε σχέση με τη συχνότητα που αυτός ο χρηστής αναρτά συνδέσμους. Έτσι, όσο πιο πολύ αναπαράγεται ένας σύνδεσμος από κάποιον αξιόπιστο χρηστή, τόσο πιο πολύ θα ενδιαφέρει κάποιον.

Έτσι με βάση την παραπάνω μεθοδολογία προτείνεται εκείνος ο σύνδεσμος από το σετ των υποψήφιων συνδέσμων που έχει συγκεντρώσει τη μεγαλύτερη βαθμολογία, δηλαδή έχει καταταχτεί πρώτος. Εμπειρικά έχει αποδειχτεί [22] ότι η πιο αξιόπιστη πηγή υποψηφίων συνδέσμων είναι οι σύνδεσμοι που δημοσιεύονται στον κύκλο ενός χρήστης (FoF) ενώ σχετικά με τα κριτήρια της κατάταξης πιο αξιόπιστα αποτελέσματα τείνουν να δίνουν το κριτήριο της ομοιότητας των μηνυμάτων που περιέχουν υποψήφιους συνδέσμους με τα μηνύματα ενός συγκριμένου χρήστη καθώς επίσης και το κριτήριο της κοινωνικότητας όπως το περιγράψαμε παραπάνω.

Η παρακάτω εικόνα ομαδοποιεί και παραθέτει σχηματικά τον αλγόριθμο που χρησιμοποιεί το σύστημα που περιγράψαμε με την επιλογή των συνδέσμων και την κατάταξή τους προκειμένου να εξάγει συστάσεις συνδέσμων για έναν χρήστη.

Design Dimension	Possible Design Choices		
CandidateSet: Selecting Candidate Set	**FoF** (followee-of-followees)	**Popular**	
Ranking-Topic: Ranking Using Topic Relevance	**Self-Topic** score	**Followee-Topic** score	**None**
Ranking-Social: Ranking Using Social Voting	**Vote** score	**None**	

Εικόνα 27: Περιγραφή του αλγορίθμου που υλοποιεί τη σύσταση συνδέσμων [21]

4.6.2 News recommendation system

Το σύστημα που περιγράφεται στην ενότητα αυτή αξιοποιεί τα tweets ενός χρήστη για να χτίσει ένα προφίλ για τον χρήστη ώστε να του προτείνει ειδήσεις από το Yahoo που θα τον ενδιαφέρουν [24]. Το σύστημα αυτό χρησιμοποιεί έναν αλγόριθμο για να κατατάξει τις ειδήσεις που θα συσταθούν στον χρήστη. Αυτός ο αλγόριθμος περιγράφεται από την παρακάτω σχέση:

$$R_T\ (u,n) = \ a \sum_T (u,n) + \ \beta\ \Gamma_T(u,n) + \ \gamma\ \Pi_T(n) \qquad (4.15)$$

Όπου

$R_T\ (u,n)$: η κατάταξη των ειδήσεων n για τον χρήστη u

$\sum_T(u,n)$: η σχετικότητα περιεχομένου ανάμεσα στον χρήστη u και την είδηση n στον χρόνο T

$\Gamma_T(u,n)$: η κοινωνική σχετικότητα του χρήστη u με την είδηση n στον χρόνο T

$\Pi_T(n)$: η δημοφιλία της είδησης n στον χρόνο T

α, β, γ: οι συντελεστές που δείχνουν τη στάθμιση κάθε παράγοντα.

Όπως μπορούμε να παρατηρήσουμε από τα παραπάνω, η σύσταση μια είδησης σε έναν χρήστη μέσω του συστήματος που αξιοποιεί πληροφορίες από tweets προκύπτει από την κατάταξη αυτής της είδησης που προτείνεται σε σχέση με τις υπόλοιπες που

αξιολογούνται. Η κατάταξη λοιπόν των ειδήσεων βασίζεται σε τρία κριτήρια. Το πρώτο κριτήριο είναι ο βαθμός συσχέτισης του χρήστη με την είδηση. Ο βαθμός συσχέτισης υπολογίζεται με βάση του κατά πόσο το περιεχόμενο των μηνυμάτων του χρήστη σχετίζεται με κάποια είδηση. Εάν συμβαίνει αυτό, τότε η είδηση θα ενδιαφέρει τον χρήστη. Το άλλο κριτήριο είναι ο βαθμός συσχέτισης των μηνυμάτων που δημοσιεύουν οι followees ενός χρήστη με μια συγκεκριμένη είδηση. Το κριτήριο αυτό λαμβάνεται υπόψη καθώς ο χρήστης ενδιαφέρεται για τα αντικείμενα που αναρτούν οι followees του να εντάσσονται στα ενδιαφέροντά του. Κατά συνέπεια, αν μια είδηση είναι σχετική με τα μηνύματα που αναρτώνται στον κύκλο του χρήστη τότε η είδηση αυτή θα ενδιαφέρει περισσότερο τον χρήστη αυτό. Το τρίτο κριτήριο είναι η δημοσιότητα μιας είδησης. Το κριτήριο αυτό είναι ανεξάρτητο από τον κύκλο του χρήστη αφού σημαντικές ειδήσεις και ειδήσεις που έχουν πολλούς θιασώτες είναι περισσότερο πιθανό να ενδιαφέρουν και έναν συγκεκριμένο χρήστη, ακριβώς γιατί η συγκεκριμένη είδηση προτιμάται από πολλούς άλλους χρηστές.

5 Αξιολόγηση Αλγορίθμων Συστάσεων στο Twitter

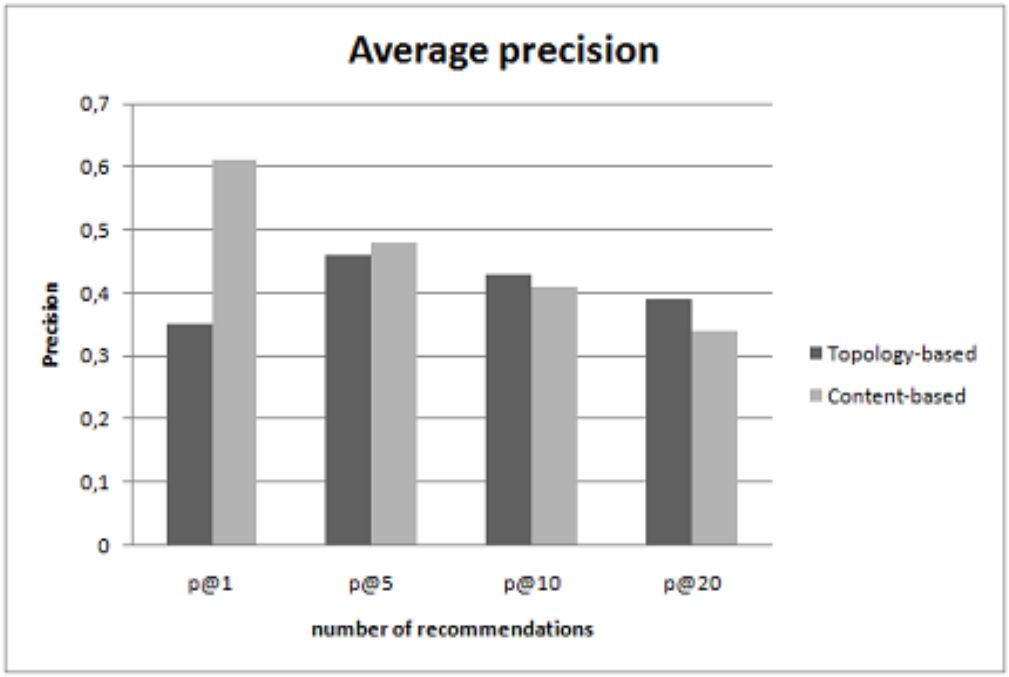
Στην παρούσα ενότητα θα προχωρήσουμε στην αξιολόγηση ορισμένων από τους αλγορίθμους που περιγράψαμε παραπάνω σχετικά με την παραγωγή συστάσεων στο Twitter. Συγκεκριμένα θα αξιολογήσουμε τους αλγόριθμους Topology – based και Content – based που όπως περιγράψαμε παραπάνω χρησιμοποιούνται για την εξαγωγή συστάσεων followees. Επίσης θα προχωρήσουμε και στην αξιολόγηση του αλγορίθμου που περιγράφηκε στην ενότητα 4.2.3 μέσα από τη σύγκριση των 9 τεχνικών βάσει content – based και collaborative filtering προσεγγίσεων.

5.1 Σύγκριση Topology – based και Content – based αλγορίθμων

Για τη σύγκριση αυτών των δύο αλγορίθμων επιχειρήθηκε μια πειραματική προσέγγιση ώστε να αξιολογηθούν οι συστάσεις με βάση καθένα αλγόριθμο. Επιλέχτηκε ένα δείγμα από 26 χρήστες, 20 άνδρες και 6 γυναίκες που ήταν τελειόφοιτοι στο πανεπιστήμιο και είχαν επιλέξει ως μάθημα επιλογής κάποιο μάθημα σχετικό με τα συστήματα συστάσεων. Οι χρήστες αυτοί φυσικά ήταν εξοικειωμένοι με το Twitter και τους ζητήθηκε να δημιουργήσουν ο καθένας έναν λογαριασμό στο Twitter και να ακολουθήσει τουλάχιστον 20 άλλους χρήστες. Η επιλογή των followees έγινε με βάση κάποια κριτήρια που έθετε ο κάθε χρήστης. Δηλαδή, οι χρήστες επέλεξαν followees σχετικούς με ένα θέμα ή σχετικούς με περισσότερα από ένα θέματα που ορίστηκαν από αυτούς. Έπειτα, τους ζητήθηκε να αναζητήσουν συστάσεις μέσω των δυο αυτών αλγορίθμων. Οι αλγόριθμοι παρείχαν 20 συστάσεις χρηστών για καθένα από τους οποίους έδειχναν στον χρήστη το όνομα, την περιγραφή, την εικόνα προφίλ και τον σύνδεσμο για την αρχική σελίδα της κάθε σύστασης. Έτσι, ο καθένας από τους 26 χρήστες θα έπρεπε να αξιολογήσει κατά πόσο οι συστάσεις που δόθηκαν με καθένα αλγόριθμο ήταν σχετικές και σύμφωνα με τα κριτήρια που είχαν επιλέξει αρχικά οι χρήστες [5,6,7].

Στο παρακάτω διάγραμμα μπορούμε να παρατηρήσουμε τα αποτελέσματα των δυο αλγορίθμων που δείχνουν την ακρίβεια, δηλαδή το πόσο σχετικές με τους αρχικούς χρήστες ήταν οι συστάσεις που δόθηκαν.

Από το διάγραμμα παρατηρούμε τα αποτελέσματα της ακρίβειας για διάφορους αριθμούς συστάσεων. Συγκεκριμένα αναλύονται τα επίπεδα για 1, 5, 10 και 20 συστάσεις.



Εικόνα 28: Αποτελέσματα συγκρινόμενων αλγορίθμων [6]

Από την παραπάνω εικόνα παρατηρούμε ότι οι δυο αλγόριθμοι δίνουν περίπου την ίδια ακρίβεια συστάσεων με εξαίρεση την περίπτωση της μιας σύστασης όπου ο content – based αλγόριθμος δίνει πιο ακριβή αποτελέσματα. Παράλληλα, παρατηρούμε ότι καθώς ο αριθμός των συστάσεων αυξάνεται, η ακρίβεια του content – based αλγορίθμου μειώνεται ενώ πιο ακριβής στις συστάσεις φαίνεται να είναι ο topology – based αλγόριθμος.

5.2 Σύγκριση content –based τεχνικών και collaborative filtering τεχνικών

Για την αξιολόγηση του Twittomender χρησιμοποιήθηκε μια offline σύγκριση αλγορίθμων βάσει ενός δείγματος. Αυτή η offline αξιολόγηση περιλάμβανε τη συλλογή δεδομένων 20.000 χρηστών του Twitter καθώς και συλλογή δεδομένων 100 tweets ανά

χρήστη. Στη συνέχεια το δείγμα χωρίστηκε σε 1.000 χρήστες που συμπεριφέρονταν ως test users και στους υπόλοιπους 19.000 που αποτελούσαν ένα training set. Μετά τον καθορισμό των δεδομένων και του δείγματος πραγματοποιήθηκε σύγκριση 9 διαφορετικών αλγορίθμων καθένας από τους οποίους έκανε χρήση content –based και collaborative filtering τεχνικών για να προβεί σε συστάσεις [33,34,35]. Αναλυτικότερα χρησιμοποιήθηκαν 4 αλγόριθμοι που βασίζονται σε content – based τεχνικές και οι οποίοι παρουσιάζονται παρακάτω:

S1: συστάσεις που βασίζονται στα tweets κάθε χρήστη

S2: συστάσεις που βασίζονται στα tweets των followees κάθε χρήστη

S3: συστάσεις που βασίζονται στα tweets των followers κάθε χρήστη

S4: συστάσεις που βασίζονται σε συνδυασμό των S1, S2, S3

Επίσης χρησιμοποιήθηκαν 3 αλγόριθμοι που βασίζονται σε collaborative filtering τεχνικές και οι οποίοι είναι:

S5: συστάσεις με βάση τα IDs των followees των χρηστών

S6: συστάσεις με βάση τα IDs των followers των χρηστών

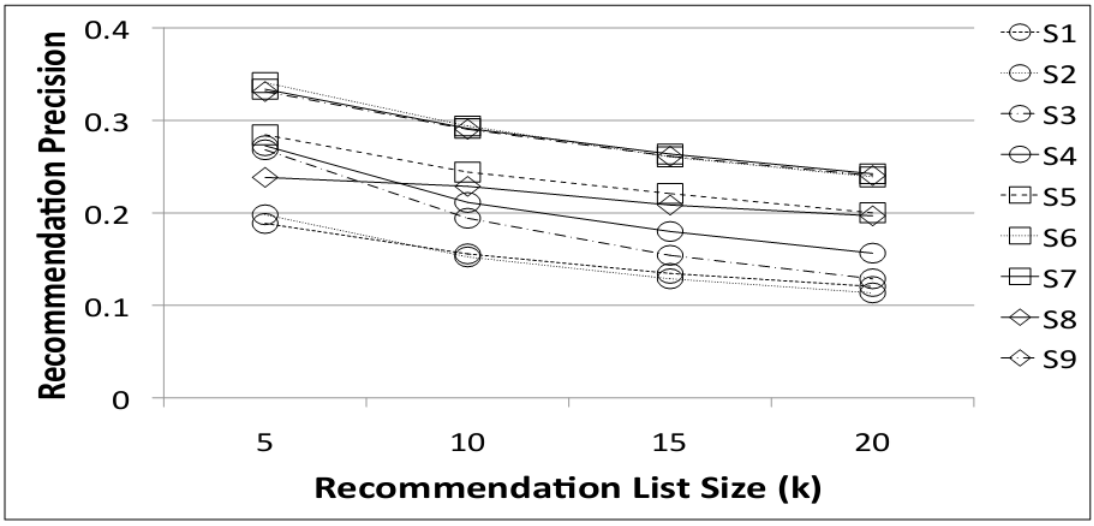
S7: συστάσεις που βασίζονται σε συνδυασμό των S5, S6

Επιπλέον χρησιμοποιήθηκαν και 2 υβριδικοί αλγόριθμοι που ουσιαστικά έπαιρναν τις συστάσεις από τους παραπάνω και τις κατέτασσαν ώστε να δώσουν τις δικές τους συστάσεις. Οι αλγόριθμοι αυτοί παρουσιάζονται παρακάτω:

S8: συστάσεις με βάση το σκορ των συστάσεων που δόθηκαν από τους S1 και S6

S9: συστάσεις με βάση τη θέση που έχει ένας χρήστης σε μια λίστα συστάσεων

Στο παρακάτω διάγραμμα παρουσιάζονται τα αποτελέσματα των παραπάνω τεχνικών συγκριτικά με την ακρίβεια των συστάσεων για διαφορετικούς αριθμούς συστάσεων.

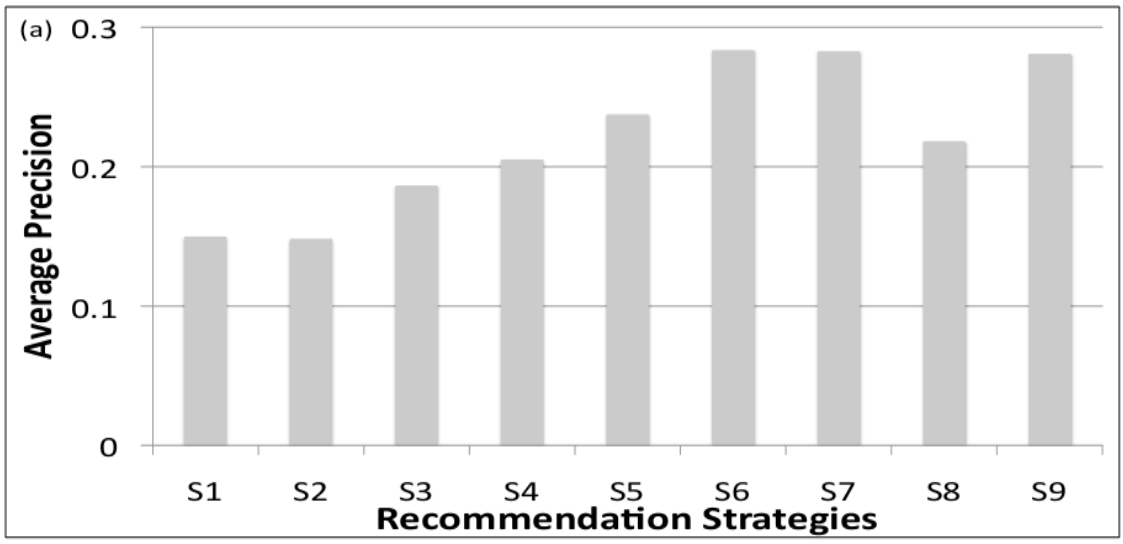


Εικόνα 29: Αποτελέσματα ακρίβειας από τη σύγκριση των 9 αλγορίθμων [34]

Από το παραπάνω παρατηρούμε ότι όλοι οι αλγόριθμοι παρουσιάζουν ακρίβεια μεταξύ 11% και 35%, κάτι που δείχνει αρκετά καλή επίδοση και συνεπώς οι συστάσεις προς τους χρήστες με βάση όλους τους αλγορίθμους είναι σχετικές με τους χρήστες. Παρατηρούμε ότι όσο το μέγεθος των συστάσεων αυξάνεται (k) η ακρίβεια φαίνεται να μειώνεται σε όλες τις περιπτώσεις.

Παράλληλα, μπορούμε να παρατηρήσουμε ότι οι αλγόριθμοι που βασίζονται σε collaborative filtering τεχνικές είναι πιο συνεπείς με τις προτιμήσεις των χρηστών σε όλα τα επίπεδα αριθμών συστάσεων. Επίσης, παρατηρούμε ότι συστάσεις με βάση τα tweets των followers των χρηστών είναι πιο ακριβείς σε σχέση με τις συστάσεις που παρέχονται σε ένα χρήστη με βάση τα tweets των followees του.

Επιπροσθέτως, στο παρακάτω γράφημα μπορούμε να παρατηρήσουμε τη μέση τιμή που λαμβάνει η ακρίβεια των συστάσεων με βάση την κάθε τεχνική που παρουσιάσαμε παραπάνω.



Εικόνα 30: Μέση τιμή ακρίβειας ανά αλγόριθμο [34]

Όπως παρατηρούμε, τη μεγαλύτερη μέση ακρίβεια φαίνεται να έχουν οι αλγόριθμοι S6, S7 και S9, ενώ τη μικρότερη κυρίως αλγόριθμοι που βασίζονται σε content – based τεχνικές όπως οι S1 και S2.

6 Συμπεράσματα

Στην παρούσα ενότητα θα παρουσιάσουμε τα συμπεράσματα στα οποία καταλήγουμε έπειτα από την παράθεση του θεωρητικού υποβάθρου σχετικά με τα συστήματα συστάσεων και την περιγραφή και καταγραφή των αλγορίθμων που χρησιμοποιούνται προκειμένου να εξαχθούν συστάσεις στο Twitter.

Αρχικά έγινε μια σύντομη αναφορά σε βασικές έννοιες όπως αυτές του κοινωνικού δικτύου και των κοινωνικών μέσων, προχωρώντας έπειτα στην ανάλυση της σχέσης ανάμεσα στα συστήματα συστάσεων και τα κοινωνικά δίκτυα. Όπως περιγράψαμε η σχέση μεταξύ συστημάτων συστάσεων και κοινωνικών δικτύων είναι αμφίδρομη αφού τα μεν κοινωνικά δίκτυα αποτελούν το πεδίο δράσης για τα συστήματα αυτά και τα δε συστήματα συστάσεων συμβάλλουν στο να κάνουν τα κοινωνικά δίκτυα πιο ελκυστικά και αποτελεσματικά ως προς τον χρήστη αφού του δίνουν εκείνες τις πληροφορίες από ένα τεράστιο σύνολο, που θα του είναι πιο χρήσιμες.

Έπειτα προχωρήσαμε στην περιγραφή των βασικών και πιο δημοφιλών κοινωνικών μέσων περιγράφοντας τις βασικές λειτουργίες αυτών των μέσων αλλά και γενικότερα τις λειτουργίες που μπορεί να συναντήσει κάποιος χρήστης μέσω ενός κοινωνικού δικτύου.

Έπειτα, επιχειρήθηκε μια πιο τεχνική προσέγγιση των συστημάτων συστάσεων μέσω της περιγραφής των διάφορων κατηγοριών τους. Συνολικά περιγράψαμε 6 κατηγορίες συστημάτων:

1. Content – based συστήματα
2. Collaborative Filtering συστήματα
3. Knowledge – based συστήματα
4. Δημογραφικά συστήματα
5. Community – based συστήματα
6. Υβριδικά συστήματα

Καθένα από τα παραπάνω συστήματα λειτουργεί με ξεχωριστό τρόπο για να καταλήξει στη σύσταση αντικειμένων προς έναν χρήστη. Οι δυο βασικές κατηγορίες συστημάτων ουσιαστικά είναι τα Content – based συστήματα και τα Collaborative Filtering συστήματα. Τα συστήματα με βάση το περιεχόμενο (Content – based) επιχειρούν να δημιουργήσουν ένα προφίλ για τον χρήστη βάσει διάφορων αλγορίθμων, έτσι ώστε να δημιουργήσουν γνώση για τις προτιμήσεις του και να μπορέσουν να τις συσχετίσουν με τα χαρακτηριστικά των αντικειμένων βάσει ομοιότητας ώστε να καταλήξουν σε εκείνες τις συστάσεις που ικανοποιούν καλύτερα τον χρήστη.

Από την άλλη πλευρά, τα Collaborative Filtering συστήματα βασίζονται σε αξιολογήσεις των αντικειμένων που έχουν γίνει από τον χρήστη είτε από άλλους χρήστες παρόμοιους με τον χρήστη στόχο. Τα συστήματα αυτά προσπαθούν να εξετάσουν την ομοιότητα του χρήστη με άλλους χρήστες και βάσει του τι επιλέγουν αυτοί οι άλλοι χρήστες, να προτείνουν και στον χρήστη για τον οποίο προορίζεται η σύσταση. Ωστόσο, λόγω της λειτουργίας των συστημάτων αυτών ανακύπτουν διάφορα προβλήματα όπως αυτό της έλλειψης αξιολογήσεων για κάποιο αντικείμενο.

Έπειτα, τα συστήματα που βασίζονται στη γνώση εμφανίζονται να είναι πιο ευέλικτα αφού προσπαθούν να δημιουργήσουν μια βάση γνώσης για τον χρήστη και τα αντικείμενα ώστε να γνωρίζουν κάθε φορά τι θα ήταν εκείνο που θα ικανοποιούσε καλύτερα έναν χρήστη. Τα συστήματα αυτά πλεονεκτούν έναντι των Collaborative Filtering συστημάτων καθώς το γεγονός ότι δεν βασίζονται σε αξιολογήσεις αντικειμένων άρει το πρόβλημα της δυσκολίας στη σύσταση όταν υπάρχουν λίγες αξιολογήσεις. Τα Knowledge – based συστήματα χρησιμοποιούν άμεσα τους χρήστες ώστε να δημιουργήσουν το προφίλ τους και έτσι να δημιουργήσουν γνώση για τις ιδιαίτερες προτιμήσεις τους, ενώ ταυτόχρονα μπορούν να προβούν σε συστάσεις και για τους πλέον ιδιαίτερους χρήστες.

Τα παραπάνω συστήματα παρουσιάζουν ωστόσο πολλά μειονεκτήματα στη χρήση τους κάτι που τα κάνει να είναι αναποτελεσματικά σε ορισμένες περιπτώσεις όπου χρήζουν ιδιαίτερης μεταχείρισης. Σε αυτήν την περίπτωση έρχονται τα υβριδικά συστήματα συστάσεων για να συνδυάσουν παραδοσιακά συστήματα συστάσεων και να υπερνικήσουν έτσι τις όποιες αδυναμίες που αυτά μπορεί να εμφανίζουν. Όπως περιγράψαμε, υπάρχουν διάφορα είδη υβριδικών συστημάτων που βασίζονται κυρίως σε content –based, collaborative filtering και knowledge – based συστήματα.

Έτσι, μετά την παρουσίαση του βασικού θεωρητικού υποβάθρου που σχετίζεται με διάφορα συστήματα συστάσεων, προχωρήσαμε στην καταγραφή των μεθόδων που χρησιμοποιούνται προκειμένου να γίνουν συστάσεις στο Twitter. Γενικά μπορούμε να παρατηρήσουμε ότι όλες οι εφαρμοσμένες μέθοδοι για συστάσεις στο Twitter βασίζονται σε μία ή περισσότερες από τις κλασσικές μεθόδους συστάσεων (content – based, collaborative filtering και knowledge – based). Βέβαια, λόγω του τεράστιου όγκου δεδομένων που έχουν να διαχειριστούν τα συστήματα στα πλαίσια του Twitter, συνήθως γίνεται χρήστη ενός υβριδικού μοντέλου σύστασης που ουσιαστικά βασίζεται στις παραπάνω μεθόδους.

Λόγω της συνθετότητας ενός κοινωνικού μέσου με εκατομμύρια χρήστες και πάρα πολλές λειτουργίες, προχωρήσαμε την ανάλυσή μας σχετικά με αλγορίθμους συστάσεων στο Twitter χωρίζοντας τις βασικές λειτουργίες αυτού του κοινωνικού μέσου και παρουσιάζοντας ένα σύνολο τρόπων και μεθόδων που χρησιμοποιούνται για την παραγωγή σύστασης για κάθε λειτουργία.

Ειδικότερα, παρουσιάσαμε αλγορίθμους για συστάσεις followees δηλαδή αλγορίθμους που χρησιμοποιούνται προκειμένου να συστήσουν σε έναν χρήστη ποιον ή ποιους άλλους χρήστες να ακολουθήσει. Τέτοιοι αλγόριθμοι κάνουν χρήση content – based και collaborative filtering μεθόδων προκειμένου να προβούν σε συστάσεις χρηστών με βάση το περιεχόμενο των μηνυμάτων που αναρτά ο χρήστης στόχος και σε σύσταση προς αυτούς χρηστών σχετικών με τα ενδιαφέροντα του χρήστη στόχου. Σχετικά με τη σύσταση followee παρουσιάσαμε μια πολύ χρήσιμη εφαρμογή που έχει αναπτυχθεί στο Twitter, το Twittomender, μέσω της οποίας ο χρήστης μπορεί να καταχωρήσει σε μια φόρμα κάποιες λέξεις κλειδιά και έτσι το σύστημα να του παράγει συστάσεις με βάση το προφίλ του.

Επιπρόσθετα, παρουσιάστηκαν άλλες μέθοδοι που χρησιμοποιούνται για την παραγωγή συστάσεων σχετικά με λίστες, hashtags, tweets, retweets οι οποίες βασίζονται στην ταξινόμηση με βάση την ομοιότητα του περιεχομένου με το προφίλ του χρήστη. Και σε αυτές τις περιπτώσεις υπάρχει συνδυασμός των μεθόδων content – based και collaborative filtering συστάσεων, αφού αρχικά εντοπίζονται μέσω αυτών των διαδικασιών τα προτεινόμενα αντικείμενα και έπειτα με έναν κατάλληλο αλγόριθμο κατάταξης με βάση την ομοιότητα ταξινομούν τα αντικείμενα έτσι ώστε να συσταθεί στον χρήστη εκείνο ή εκείνα που θα ταιριάζουν περισσότερο στο προφίλ του.

Ακόμη, παρουσιάστηκε μια αρκετά χρήσιμη μέθοδος που προορίζεται για τη σύσταση followers. Πολλές φορές στο Twitter κάποιος χρήστης μπορεί να ενδιαφέρεται να αναζητήσει χρήστες που να τον ακολουθήσουν και έτσι να διαδώσουν τα μηνύματα του χρήστη στόχου. Στο Twitter αυτό μπορεί να επιτευχθεί με το Tadvise, μια μέθοδο που βάσει του περιεχομένου των μηνυμάτων που αναρτούν οι followers του χρήστη στόχου καθώς και οι followees των followers του χρήστη στόχου, το σύστημα με βάση την ομοιότητα προτείνει εκείνους τους χρήστες στον χρήστη στόχο που θα είναι πιο πιθανό να μεταδώσουν τα μηνύματα του χρήστη στόχου.

Επιπρόσθετα, λόγω του ότι το Twitter αποτελεί πηγή ειδήσεων και πληροφοριών, καταγράψαμε και μεθόδους που ακολουθούνται προκειμένου να προβούν σε συστάσεις συνδέσμων και ειδήσεων στο Twitter. Αυτές οι μέθοδοι, ανακτούν πληροφορίες από τα μηνύματα που αναρτά ένας χρήστης και του προτείνουν συνδέσμους και ειδήσεις που το περιεχόμενό τους είναι παρόμοιο με το περιεχόμενο των μηνυμάτων που αναρτά ο χρήστης. Έτσι, είναι πιθανότερο μια είδηση ή σύνδεσμος να ενδιαφέρει τον χρήστη αφού θα ταιριάζει στην ιδιοσυγκρασία του καθώς και στις προτιμήσεις του.

Τέλος, τις περισσότερες φορές, αν όχι πάντα, είναι χρήσιμο να επιχειρείται μια αξιολόγηση του κάθε αλγορίθμου έτσι ώστε να διαπιστωθεί κατά πόσο αυτός ο αλγόριθμος μπορεί να προβεί σε συστάσεις που είναι χρήσιμες για έναν χρήστη. Επίσης, η σύγκριση των διάφορων αλγορίθμων για την εξαγωγή συστάσεων είναι σκόπιμη προκειμένου να διαπιστωθεί ποιος αλγόριθμος είναι περισσότερο αποδοτικός για την παροχή συστάσεων σε έναν χρήστη. Στην παρούσα εργασία περιγράψαμε τη σύγκριση μεταξύ των topology – based αλγορίθμων και των content – based αλγορίθμων για την παροχή συστάσεων ακολούθων σε έναν χρήστη. Η ανάλυση έδειξε ότι και οι δυο μέθοδοι αποδίδουν αποτελεσματικά με τον content – based αλγόριθμο να αποδίδει περισσότερο όταν ο αριθμός των συστάσεων είναι μικρός.

Μια ακόμη σύγκριση στην οποία προχωρήσαμε ήταν σχετικά με το Twittomender για την αξιολόγηση του οποίου έγινε μια offline αξιολόγηση 9 διαφορετικών αλγορίθμων βάσει προσεγγίσεων που βασίζονται σε collaborative filtering, content – based και υβριδικές τεχνικές. Από την ανάλυση προέκυψε ότι οι αλγόριθμοι με βάση το collaborative filtering αποδίδουν καλύτερα αφού η ακρίβεια των συστάσεων είναι μεγαλύτερη όποιος και αν είναι ο αριθμός των συστάσεων.

Το πεδίο εφαρμογής των διαφόρων συστημάτων συστάσεων στα κοινωνικά δίκτυα χωρεί ακόμη μεγάλη ανάπτυξη και έρευνα. Αυτό συμβαίνει γιατί τα κοινωνικά δίκτυα

δεν είναι στατικά αλλά ενέχουν δυναμικά χαρακτηριστικά που τα κάνουν συνεχώς να αναπτύσσονται και επομένως και τα συστήματα συστάσεων θα πρέπει να αναπτύσσονται ώστε να μπορούν να εξυπηρετήσουν καλύτερα τους σκοπούς των κοινωνικών δικτύων. Σχετικά με τους αλγορίθμους συστάσεων στο Twitter επιπλέον προτάσεις για έρευνα θα ήταν η περεταίρω αξιολόγηση αυτών των αλγορίθμων. Είναι σημαντικό να γνωρίζουμε κατά πόσο αυτοί οι αλγόριθμοι που περιγράψαμε όντως εμφανίζονται αποτελεσματικοί στην παραγωγή συστάσεων.

Επίσης, θα ήταν χρήσιμο να εξεταστούν και οι τρόποι που μπορούν να άρουν τα προβλήματα που ενδεχομένως να παρουσιάζονται κατά την δημιουργία συστάσεων στο Twitter. Βέβαια, λόγω του ότι σχεδόν οι περισσότερες από τις μεθόδους που παρουσιάσαμε κάνουν χρήση των content – based και collaborative filtering μεθόδων θα πρέπει να προσδιοριστούν οι τρόποι με τους οποίους μπορούν να αρθούν οι αδυναμίες αρχικά αυτών των συστημάτων και έπειτα στοχευμένα και ανάλογα με την κάθε περίπτωση να επεκταθεί αυτή η ανάλυση και στα ειδικότερα συστήματα και αλγορίθμους που χρησιμοποιούνται και είναι προσαρμοσμένοι για τις ανάγκες του Twitter.

Βιβλιογραφία

[1] Adomavicius, G., Tuzhilin, A.: *Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions*. IEEE Transactions on Knowledge and Data Engineering17(6), pp. 734–749, 2005

[2] Aggarwal, C.: *An introduction to social network data analytics*. Springer US, 2011

[3] Aiello, L. M., Petkos, G., Martin, C. et al.: *Sensing trending topics in Twitter*. 2013

[4] Arazy, O., Kumar, N., Shapira, B.: *Improving social recommender systems*. IT Professional 11(4), pp. 38–44, 2009

[5] Armentano, M.G., Godoy, D.L., Amandi, A.A.: *Recommending Information Sources to Information Seekers in Twitter*. In: International Workshop on Social Web Mining

[6] Armentano, M.G., Godoy, D.L., Amandi, A.A.: *Towards a Followee Recommender System for Information Seeking Users in Twitter*. In: The 2nd International Workshop on Semantic Adaptive Social Web

[7] Armentano, M.G., Godoy, D.L., Amandi, A.A.: *A Topology-Based Approach for Followees Recommendation in Twitter*. In: 9th Workshop on Intelligent Techniques for Web Personalization and Recommender Systems. Barcelona, Spain, July 2011

[8] Bakshy, E., Hofman, J. M., Mason, W. A., Watts, D. J.: *Identifying influencers on twitter*. In Fourth ACM International Conference on Web Search and Data Mining (WSDM), 2011

[9] Ben-Shimon, D., Tsikinovsky, A., Rokach, L., Meisels, A., Shani, G., Naamani, L.: *Recommender system from personal social networks*. In: AWIC, Advances in Soft Computing, vol. 43, pp. 47–55. Springer, 2007

[10] Bell, R., Koren, Y., Volinsky, C.: *Modeling relationships at multiple scales to improve accuracy of large recommender systems*. In: KDD ’07: Proc. of the 13th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, pp. 95–104. ACM, New York, NY, USA, 2007

[11] Bell, R.M., Koren, Y.: *Scalable collaborative filtering with jointly derived neighborhood interpolation weights*. In: ICDM '07: Proc. of the 2007 Seventh IEEE Int. Conf. on Data Mining, pp. 43–52. IEEE Computer Society, Washington, DC, USA, 2007

[12] Billsus, D., Pazzani, M.J.: *Learning collaborative information filters*. In: ICML '98: Proc. Of the 15th Int. Conf. on Machine Learning, pp. 46–54. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1998

[13] Billsus, D., Pazzani, M.J.: *A Hybrid User Model for News Story Classification*. In: Proceedings of the Seventh International Conference on User Modeling. Banff, Canada, 1999

[14] Billsus, D., Pazzani, M.J.: *User modeling for adaptive news access*. User Modeling and User-Adapted Interaction 10(2-3), pp. 147–180, 2000

[15] Breese, J.S., Heckerman, D., Kadie, C.: *Empirical analysis of predictive algorithms for collaborative filtering*. In: Proc. of the 14th Annual Conf. on Uncertainty in Artificial Intelligence, pp. 43–52. Morgan Kaufmann, 1998

[16] Bridge, D., Goker, M., McGinty, L., Smyth, B.: *Case-based recommender systems*. The Knowledge Engineering review20(3), pp. 315–320, 2006

[17] Boiy, E. and Moens, M.: A *Machine Learning Approach to Sentiment Analysis in Multilingual Web Texts*. Information Retrieval, 12(5) pp. 526-558, 2009.

[18] Burke, R.: *Hybrid recommender systems: Survey and experiments*. User Modeling and User-Adapted Interaction, 12(4): pp. 331–370, 2002.

[19] Burke, R.: *Hybrid web recommender systems*. In: The Adaptive Web, pp. 377–408. Springer Berlin / Heidelberg, 2007

[20] Chandramouli, A., Micarelli, A.: *User Profiles for Personalized Information Access*, 2007

[21] Chen, J., Nairn, R., Nelson, L., Bernstein, M., Chi, E.: *Short and Tweet: Experiments on Recommending Content from Information Streams.* In: The 28th International Conference on Human Factors in Computing Systems, 2010

[22] Chen, L. and Pu, P. *Hybrid critiquing-based recommender systems*, January 28–31, Honolulu, Hawaii, USA, 2007

[23] Choudhary, A., Hendrix, W., Lee, K., Palsetia, D., Liao, W.K.: *Social media evolution of the Egyptian revolution*. Communications of ACM 55(5), pp. 74-80, May 2012

[24] De Francisci Morales, G., Gionis, A., Lucchese, C.: *From Chatter to Headlines: Harnessing the Real-Time Web for Personalized News Recommendation*. In: The 5th ACM International Conference on Web Search and Data Mining, 2012

[25] Deshpande, M., Karypis, G.: *Item-based top-N recommendation algorithms*. ACM Transaction on Information Systems 22(1), pp.143–177, 2004

[26] Destoisers C., Karypis, G.: *A Comprehensive Survey of Neighborhood-based Recommendation Methods,* Survey of Recommendation Systems Handbook, pp. 104-140

[27] Ding, X., B. Liu, Yu, P.: *A Holistic Lexicon-based Approach to Opinion Mining*. In: Proceedings of the Conference on Web Search and Web Data Mining (WSDM-2008), 2008.

[28] Domingos, P., Pazzani, M.J.: O*n the Optimality of the Simple Bayesian Classifier under Zero-One Loss*. Machine Learning29(2-3), pp. 103–130, 1997

[29] Fouss, F., Renders, J.M., Pirotte, A., Saerens, M.: *Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation*. IEEE Transactions on Knowledge and Data Engineering 19(3), pp.355–369, 2007

[30] Garcia, R., Amatriain, X.: *Weighted Content Based Methods for Recommending Connections in Online Social Networks*. In: The 2nd ACM Workshop on Recommendation Systems and the Social Web. Barcelona, Spain, June 2010

[31] Golder, S.A., Marwick, A., Yardi, S., Boyd, D.: *A structural approach to con-tact recommendations in online social networks.* In: Workshop on Search in Social Media, In conjunction with ACM SIGIR Conference on Information Retrieval

[32] Good, N., Schafer, J.B., Konstan, J.A., Borchers, A., Sarwar, B., Herlocker, J., Riedl, J.: *Combining collaborative filtering with personal agents for better recommendations*. In: AAAI '99/IAAI '99: Proc. of the 16th National Conf. on Artificial Intelligence, pp. 439–446. American Association for Artificial Intelligence, Menlo Park, CA, USA, 1999

[33] Hannon, J., Bennett, M., Smyth, B.: *Recommending Twitter Users to Follow Using Content and Collaborative Filtering Approaches*. In: The 4th ACM Conference on Recommender Systems, 2010

[34] Hannon, J., McCarthy, K., Smyth, B.: *Finding Useful Users on Twitter: Twittomender the Followee Recommender*. In: The 33rd European Conference on Advances in Information Retrieval, 2011

[35] Hannon, J., McCarthy, K., Smyth, B.: *The Pursuit of Happiness: Searching for Worthy Followees on Twitter*. In: The 22nd Irish Conference on Artificial Intelligence and Cognitive Science, August 2011

[36] Herlocker, L., Konstan, J.A., Terveen, L.G., Riedl, J.T.: *Evaluating Collaborative Filtering Recommender Systems*. ACM Transactions on Information Systems22(1), pp.5–53, 2004

[37] Herlocker, J.L., Konstan, J.A., Borchers, A., Riedl, J.: *An algorithmic framework for perform-ing collaborative filtering.* In: SIGIR '99: Proc. of the 22nd Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, pp. 230–237. ACM, New York, NY, USA, 1999

[38] Huang, J., Thornton, K.M., Efthimiadis, E.N.: *Conversational Tagging in Twitter*. In: The 21st ACM Conference on Hypertext and Hypermedia. pp. 173-178, 2010

[39] Java, A., Song, X., Finin, T., Tseng, B.: *Why We Twitter: Understanding Microblogging Usage and Communities.* In: The 9th Web KDD and 1st SNA - KDD 2007 Workshop on Web Mining and Social Network Analysis. pp. 56-65, 2007

[40] Kim, S.B., Han, K.S., Rim, H.C., Myaeng, S.H.: *Some Effective Techniques for Naıve Bayes Text Classification*. IEEE Trans. Knowl. Data Eng.18(11), pp. 1457–1466, 2006

[41] Koren, Y.: *Factorization meets the neighborhood: a multifaceted collaborative filtering model*. In: KDD'08: Proceeding of the 14th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, pp. 426–434. ACM, New York, NY, USA, 2008

[42] Krutkam, W., Saikeaw, K., Chaosakul, A.: *Twitter Accounts Recommendation Based on Followers and Lists*. In: 3rd Joint International Information and Communication Technology, 2010

[43] Kwak, H., Lee, C., HosungPark, Moon, S.: *What is Twitter, a social network or a news media?* In: The 19th International Conference on World Wide Web, 2010

[44] Lang, K.: *News Weeder: Learning to filter netnews*. In: Proc. of the 12th Int. Conf. on Machine Learning, pp. 331–339. Morgan Kaufmann publishers Inc.: San Mateo, CA, USA, 1995

[45] Li, T., Yu Wu, Y.Z.: *Twitter hash tag prediction algorithm*. In: World Congress in Computer Science, Computer Engineering, and Applied Computing, 2011

[46] Liu, F., Lee, H. J.: *Use of social network information to enhance collaborative filtering performance*. Expert Systems with Applications, 37, pp. 4772-4778, 2010.

[47] Mahmood, T., Ricci, F.: *Towards learning user-adaptive state models in a conversational recommender system.* In: A. Hinneburg (ed.) LWA 2007: Lernen - Wissen - Adaption, Halle, September 2007, Workshop Proceedings, pp. 373–378. Martin-Luther-University Halle-Wittenberg, 2007

[48] Mathioudakis, M., Koudas, N.: *Twittermonitor: trend detection over the twitter stream*. In Proceedings of the 2010 ACM SIGMOD International Conference on Management of data pp. 1155-1158, ACM, 2010.

[49] Mazzia, A., Juett, J.: *Suggesting hashtags on twitter*. EECS 545 (Machine Learning) Couse Project Report

[50] McCallum, A., Nigam, K.: *A Comparison of Event Models for Naıve Bayes Text Classification*. In: Proceedings of the AAAI/ICML-98 Workshop on Learning for Text Categorization, pp. 41–48. AAAI Press, 1998

[51] McPherson, M., Smith-Lovin, L., Cook, J.M.: *Birds of a Feather: Homophily in Social Networks*

[52] Mitchell, T.: *Machine Learning*. McGraw-Hill, New York, 1997

[53] Nasirifard, P., Hayes, C.: *Tadvise: A Twitter Assistant Based on Twitter Lists*. In: The 3rd International Conference on Social Informatics. pp. 153-160, 2011

[54] Paranjpe, D.: *Learning Document Aboutness from Implicit User Feedback and Document Structure.* In: ACM Conference on Information and Knowledge Management, 2009

[55] Pazzani, M.J., Muramatsu, J., Billsus, D.: *Syskill and Webert: Identifying Interesting Web Sites*. In: Proceedings of the Thirteenth National Conference on Artificial Intelligence and the Eighth Innovative Applications of Artificial Intelligence Conference, pp. 54–61. AAAI Press / MIT Press, Menlo Park, 1996

[56] Pazzani M., Billsus, D. : *Content-based Recommendation Systems*, 1997

[57] Pham, M. C., Cao, Y., Klamma, R., Jarke, M.: *A clustering approach for collaborative filtering recommendation using social network analysis*. J. UCS, 17 (4), pp. 583-604, 2011

[58] Picard, R.W.: *Affective Computing*. MIT Press, 2000

[59] Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., Riedl, J.: *GroupLens: An open architecture for collaborative filtering of netnews*. In: CSCW '94: Proc. of the 1994 ACM Conf. on Computer Supported Cooperative Work, pp. 175–186. ACM, New York, NY, USA, 1994

[60] Ricci, F., Cavada, D., Mirzadeh, N., Venturini, A.: *Case-based travel recommendations*. In: D.R. Fesenmaier, K. Woeber, H. Werthner (eds.) Destination Recommendation Systems: Behavioral Foundations and Applications, pp. 67–93. CABI, 2006

[61] Ricci, F., Rokach, L., Shapira, B.: *Recommender Systems Handbook*. Springer, 2011

[62] Rich, E.: *User Modeling via Stereotypes*. Cognitive Science, pp. 329–354, 1979

[63] Rocchio, J.*: Relevance Feedback Information Retrieval*. In: G. Salton (ed.) The SMART retrieval system - experiments in automated document processing, pp. 313–323. Prentice-Hall, Englewood Cliffs, NJ, 1971

[64] Salton, G.: *Automatic Text Processing*. Addison-Wesley, 1989

[65] Salton, G., McGill, M.: *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, 1983

[66] Salton, G., Buckley, C.: *Term-Weighting Approaches in Automatic Text Retrieval*. Information Processing and Management 24(5), pp. 513 – 523, 1988

[67] Schafer, J.B., Konstan, J.A., Riedl, J.: *E-Commerce Recommendation Applications*. Data Mining and Knowledge Discovery 5(1-2), pp. 115-153, 2001

[68] Schafer, J.B., Frankowski, D., Herlocker, J., Sen, S.: *Collaborative filtering recommender systems*. In: The Adaptive Web, pp. 291–324. Springer Berlin / Heidelberg, 2007

[69] Schwab, I., Kobsa, A., Koychev, I.: *Learning User Interests through Positive Examples using Content Analysis and Collaborative Filtering*, 2001

[70] Shardanand, U., Maes, P.: *Social Information Filtering: Algorithms for Automating "Word of Mouth"*. In: Proceedings of ACM CHI'95 Conference on Human Factors in Computing Systems, vol. 1, pp. 210–217, 1995

[71] Symeonidis, P.: *Content-based Dimensionality Reduction for Recommender Systems*. In: C. Preisach, H. Burkhardt, L. Schmidt-Thieme, R. Decker (eds.) Data Analysis, Machine Learning and Applications, Studies in Classification, Data Analysis, and Knowledge Orga-nization, pp. 619–626. Springer Berlin Heidelberg, 2008

[72] Symeonidis, P., Tiakas, E., Manolopoulos, Y.: *Product recommendation and rating prediction based on multi-modal social networks*. In Proceedings of the fifth ACM conference on Recommender systems, pp. 61-68. ACM, 2011

[73] Symeonidis, P. et al., *Recommender Systems for Location-based Social Networks*. Springer Briefs in Electrical and Computer Engineering, pp. 16-18, 2014

[74] Takacs, G., Pilaszy, I., Nemeth, B., Tikk, D.: *Major components of the gravity recommendation system*. SIGKDD Exploration Newsletter 9(2), pp. 80–83, 2007

[75] Uysal, I., Croft, B.W.: *User Oriented Tweet Ranking: a Filtering Approach to Microblogs*. In: The 20th ACM International Conference on Information and Knowledge Management, 2011

[76] Zangerle, E., Gassler, W.: *Recommending #-Tags in Twitter*. In: Workshop on Semantic Adaptive Social Web 2011, in connection with the 19th International Conference on User Modeling, Adaptation and Personalization, 2011

[77] Zuber, M.: *A Survey of Data Mining Techniques for Social Network Analysis*. International Journal of Research in Computer Engineering and Electronics. VOl : 1 ISSUE :1, 2012

[78] www.google.com

[79] www.wikipedia.com

[80] www.facebook.com

[81] www.twitter.com

[82] www.linkedin.com

[83] www.youtube.com

[84] Papagelis, M., Plexousakis, D.: *Qualitative analysis of user-based and item-based prediction algorithms for recommendation agents,* In: Engineering Applications of Artificial Intelligence 18, pp. 781–789, 2005