Πανεπιστημιο Θεσσαλιας
Τμημα Μηχανικων Η/Υ, Τηλεπικοινωνιων και Δικτυων

Μελέτη, ανάλυση και αξιολόγηση ενός συστήματος πληροφοριακών συστάσεων, που χρησιμοποιεί ετικέτες

Μεταπτυχιακη Διατριβη

του

ΣΤΕΦΑΝΟΥ Α. ΚΟΝΤΟΒΑ

Επιβλέποντες : Εμμανουήλ Βαβαλης
Καθηγητής Τ.Μ.Η/Υ.Τ.Δ
Παναγιώτης Μποζανης
Αναπληρωτής Καθηγητής Τ.Μ.Η/Υ.Τ.Δ
Χρήστος Αντωνοπουλος
Επίκουρος Καθηγητής Τ.Μ.Η/Υ.Τ.Δ

17 Ιουλίου 2012

Πανεπιστημιο Θεσσαλιας
Τμημα Μηχανικων Η/Υ, Τηλεπικοινωνιων και Δικτυων

Μελέτη και ανάλυση/αξιολόγηση ενός συστήματος πληροφοριακών συστάσεων, που χρησιμοποιεί ετικέτες

Μεταπτυχιακη Διατριβη

του

ΣΤΕΦΑΝΟΥ Α. ΚΟΝΤΟΒΑ

Επιβλέπων : Εμμανουήλ Βαβαλης
Καθηγητής Τ.Μ.Η/Υ.Τ.Δ

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Εμμανουήλ Βάβαλης	Παναγιώτης Μποζάνης	Χρήστος Αντωνόπουλος
Καθηγητής	Αναπληρωτής Καθηγητής	Επίκουρος Καθηγητής

17 Ιουλίου 2012

Περίληψη

Η παρούσα διπλωματική εργασία εξετάζει τη λειτουργία των *συστημάτων συστάσεων* (recommender systems) και εστιάζει στη χρήση *ετικετών* (tags) από αυτά. Γίνεται αναφορά στις εφαρμογές των συστημάτων συστάσεων και παρουσιάζονται οι βασικές αρχές της διήθησης της πληροφορίας (information filtering) που τα διέπουν. Εξηγείται ο ρόλος των ετικετών ως μεταδεδομένο και η χρησιμότητά τους στο Web2.0. Τέλος, με βάση δεδομένα από την υπηρεσία του *BibSonomy*, παρουσιάζονται πειραματικά αποτελέσματα και συγκρίνεται η ακρίβεια πρόβλεψης των κυριότερων αλγορίθμων, με αλλά και χωρίς τη χρήση ετικετών.

Λέξεις κλειδιά: *συστήματα συστάσεων, διήθηση πληροφορίας, ετικέτες, BibSonomy*

Ευχαριστίες

Κατ' αρχήν θα ήθελα να ευχαριστήσω τους καθηγητές μου κ.κ Βάβαλη Εμμανουήλ και Νανά Νικόλαο, οι οποίοι πρότειναν το θέμα της διπλωματικής μου εργασίας και προσέφεραν πολύτιμη βοήθεια και καθοδήγηση όλους αυτούς τους μήνες, μέχρι την ολοκλήρωσή της.

Οφείλω, επίσης, ένα "ευχαριστώ" στο κύριο Γιάννη Γιάκα, ο οποίος μας παραχώρησε τον server του εργαστηρίου του για την εκτέλεση των πειραμάτων. Τέλος, πρέπει να ευχαριστήσω τον κ. Χρήστο Αντωνόπουλο και την κ. Βάνα Ντουφεξή για τη καθοριστικότατη συμβολή τους στη βελτιστοποίηση της λειτουργικότητας και της απόδοσης του κώδικα που χρησιμοποιήθηκε στα πειράματα.

Περιεχόμενα

Σχήματα **ix**

1 Εισαγωγή **2**
1.1 Συστήματα συστάσεων . . . 3

2 Φιλτράρισμα Πληροφορίας (Information Filtering) **4**
2.1 Collaborative . . . 4
2.2 Content - based . . . 5
2.3 Υβριδικό ΦΠ . . . 6
2.4 Δομές δεδομένων . . . 7
2.5 Μετρικές Σύγκρισης . . . 7

3 Ετικέτες (tags) **10**
3.1 Ορισμός . . . 10
3.2 Ιστορική Αναδρομή . . . 10
3.3 Folksonomies . . . 11
3.4 Σχετικές Μελέτες . . . 12

4 Πειράματα **16**
4.1 Σύνολα Δεδομένων . . . 16
4.1.1 Στατιστική ανάλυση δεδομένων . . . 17
4.2 Μεθοδολογία . . . 19
4.2.1 Μέθοδος Αξιολόγησης . . . 19
4.3 Πειραματικά Αποτελέσματα . . . 20
4.3.1 User-based filtering . . . 20
4.3.2 Item-based filtering . . . 21
4.3.3 Επικάλυψη ετικετών . . . 22
4.4 Συζήτηση των αποτελεσμάτων . . . 24

5 Σύνοψη και μελλοντικές εργασίες **26**

Βιβλιογραφία **29**

Κατάλογος σχημάτων

2.1 Συνεργατικό Φιλτράρισμα πληροφορίας 5
2.2 Προσαρμοστικό Content-Based Φιλτράρισμα εγγράφων 6
2.3 Αναπαράσταση των συνδέσεων του συνόλου των δεδομένων σε τρισδιάστατο πίνακα. 7

3.1 Χρήση ετικετών στο *Flickr* 11
3.2 Ένα *tag cloud* από ετικέτες που σχετίζονται με το *Web 2.0* 11
3.3 Ταξινόμηση ετικετών σε tag cloud 12

4.1 Δείγμα των εγγραφών στο αρχείο *"tas"* 17
4.2 Δείγμα των εγγραφών στο αρχείο *"bookmark"* 17
4.3 Κατανομή των σελιδοδεικτών ως προς τους χρήστες 18
4.4 Φιλτράρισμα με βάση τους χρήστες της folksonomy 21
4.5 Φιλτράρισμα με βάση τα αντικείμενα της folksonomy 22
4.6 Φιλτράρισμα με βάση την επικάλυψη των ετικετών των χρηστών 23
4.7 Φιλτράρισμα με βάση την επικάλυψη των ετικετών των αντικειμένων . . 23
4.8 Σύγκριση των αποτελεσμάτων όλων των πειραμάτων 24

...στους γονείς μου

1

Εισαγωγή

Το Διαδίκτυο (Internet) κατακλύζεται καθημερινά από πληθώρα πληροφοριών. Πληροφορίες προέρχονται από ειδησεογραφικές ιστοσελίδες, άρθρα σε blogs, πολυμέσα (μουσική, φωτογραφίες, βίντεο), ακαδημαϊκά άρθρα, e-mails, RSS[1]. Εάν προσθέσουμε σε αυτά πληροφορίες που προέρχονται από κοινωνικά δίκτυα (twitter, facebook, pinterest) τότε αντιλαμβανόμαστε πως ένας χρήστης έρχεται σε επαφή με ένα μεγάλο όγκο πληροφοριών καθημερινά.

Ποιες όμως από αυτές τις πληροφορίες ενδιαφέρουν πραγματικά τον χρήστη και ποιες μπορούν να αποδειχθούν χρήσιμες; Συνήθως, ένα μικρό ποσοστό των πληροφοριών, με αποδέκτη το χρήστη, χαρακτηρίζονται ως "ενδιαφέρουσες" από τον ίδιο. Αυτή η δυσκολία να εντοπίσουμε τη χρήσιμη πληροφορία είναι ένα επακόλουθο της υπερ-πληροφόρησης (info-overload) και του info pollution. Η *υπερπληροφόρηση* αναφέρεται στη δυσκολία κατανόησης ενός ζητήματος και λήψης αποφάσεων, που προκαλείται από την παρουσία πολλών πληροφοριών[2]. Αντιστοίχως, ως *info-pollution* ορίζεται η μόλυνση της παροχής (πηγής) πληροφοριών με άσχετες, περιττές, αυτόκλητες και χαμηλής αξίας πληροφορίες[3].

Δημιουργείται, συνεπώς, η ανάγκη για διαχωρισμό της χρήσιμης - ενδιαφέρουσας, σύμφωνα με το χρήστη, πληροφορίας από το σύνολο των διαθέσιμων πληροφοριών. Γίνεται αντιληπτό, πως για να εντοπίσει ο χρήστης "χειροκίνητα" πληροφορίες που σχετίζονται με κάποιο ενδιαφέρον του, απαιτείται τις περισσότερες φορές επίπονη και χρονοβόρα διαδικασία. Γι' αυτό το λόγο, έχουν αναπτυχθεί *συστήματα συστάσεων (recommender systems)*, τα οποία φροντίζουν να εξετάσουν τα διαθέσιμα δεδομένα και να παρουσιάσουν - προ-

[1] *Really Simple Syndication*: μέθοδος ανταλλαγής πληροφοριακού περιεχομένου, στηριγμένη στην πρότυπη γλώσσα σήμανσης XML.

[2] http://en.wikipedia.org/wiki/Information_overload

[3] http://en.wikipedia.org/wiki/Information_pollution

τείνουν στο χρήστη, σχετικές με τα ενδιαφέροντά του, πληροφορίες.

1.1 Συστήματα συστάσεων

Ο όρος *σύστημα συστάσεων* αναφέρεται σε ένα πληροφοριακό σύστημα που προσπαθεί να προβλέψει την προτίμηση ενός χρήστη για κάποιο αντικείμενο (π.χ μουσική, βιβλίο, ταινία), το οποίο δεν έχει λάβει υπόψιν του ή δε γνωρίζει ακόμα ο χρήστης. Οι προβλέψεις αυτές παράγονται από ένα υπολογιστικό μοντέλο, το οποίο κατασκευάζεται είτε από τα χαρακτηριστικά του αντικειμένου (content-based) είτε από το κοινωνικό περιβάλλον του χρήστη (collaborative). Μερικά παραδείγματα τέτοιων συστημάτων είναι:

- το Amazon.com, το οποίο προτείνει επιπλέον προϊόντα με βάση αυτό που επέλεξε ο χρήστης αλλά και με βάση τι άλλο αγόρασαν οι προηγούμενοι αγοραστές αυτού του προϊόντος.
- το Last.fm, το οποίο προτείνει στο χρήστη νέα τραγούδια παρατηρώντας τα τραγούδια που ακούει ο χρήστης και συγκρίνοντας τα ακούσματά του με αυτά άλλων χρηστών.
- το Netflix, το οποίο προτείνει στο χρήστη ταινίες βασιζόμενο στις προηγούμενες αξιολογήσεις του, στις προηγούμενες επιλογές του (σε σύγκριση με τους υπόλοιπους χρήστες), αλλά και στα χαρακτηριστικά της ταινίας (π.χ είδος, σκηνοθέτης).

Τα συστήματα συστάσεων αποτελούν ένα υποσύνολο - υποκλάση των συστημάτων φιλτραρίσματος (διήθησης) πληροφορίας *(information filtering)*. Αναπόφευκτα, λόγω της φύσης του προβλήματος το οποίο καλούνται να επιλύσουν, χρησιμοποιούν τεχνικές και ιδιότητες του information filtering, βασικά χαρακτηριστικά του οποίου αναφέρονται στο επόμενο κεφάλαιο.

2

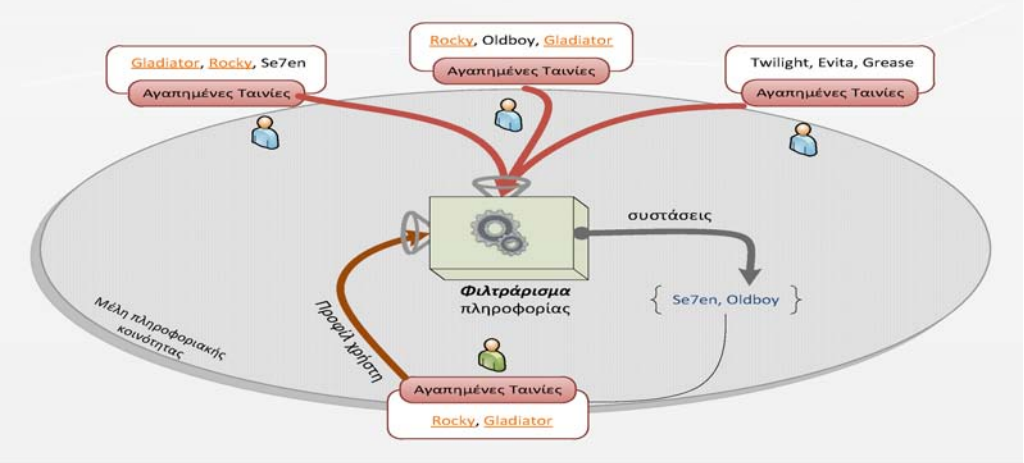
Φιλτράρισμα Πληροφορίας (Information Filtering)

Το Φιλτράρισμα της Πληροφορίας (ΦΠ) βασίζεται στη δημιουργία και διαχείριση, για κάθε χρήστη, ενός προφίλ (profile) το οποίο αναπαριστά τα ενδιαφέροντά του. Κάθε πληροφορία συγκρίνεται με το προφίλ του χρήστη. Αν κριθεί "σχετική", τότε παρουσιάζεται στον χρήστη καταλλήλως· συνήθως, δίνοντάς της προτεραιότητα. Ξεχωρίζουν κυρίως δύο κατηγορίες Φιλτραρίσματος της Πληροφορίας: το *συνεργατικό φιλτράρισμα (collaborative)* και το *φιλτράρισμα βάσει περιεχομένου (content - based)*. Διαφέρουν στον τρόπο με τον οποίο η πληροφορία και το προφίλ του χρήστη αναπαρίστανται και συγκρίνονται. Τέλος, υπάρχει και μια τρίτη κατηγορία, το *υβριδικό φιλτράρισμα (hybrid)*, η οποία συνδυάζει χαρακτηριστικά των προηγούμενων δυο κατηγοριών, προσπαθώντας να ενισχύσει την ακρίβεια των προβλέψεων.

2.1 Collaborative

Η πληροφορία χαρακτηρίζεται από το πως την έχουν αξιολογήσει τα μέλη μιας κοινότητας. Σκοπός είναι να προταθούν στο χρήστη νέες πληροφορίες που πιθανόν τον ενδιαφέρουν. Αυτό επιτυγχάνεται με τη εύρεση ενός συνόλου χρηστών ("γείτονες") που μοιράζονται τα ίδια ενδιαφέροντα με τον εν λόγω χρήστη. Από τη στιγμή που θα οριστεί αυτό το σύνολο, επιλέγονται να παρουσιαστούν στο χρήστη οι πληροφορίες οι οποίες έχουν αξιολογηθεί υψηλά από τα μέλη του παραπάνω συνόλου (Σχήμα 2.1). Με τον τρόπο αυτό, προτείνονται στο χρήστη, βάσει των αξιολογήσεων των "γειτόνων" του, αντικείμενα πληροφορίας τα οποία ο ίδιος δεν είχε προσπελάσει στο παρελθόν ή δε γνώριζε

καν την ύπαρξή τους. Το Collaborative Filtering δεν προϋποθέτει πρόσβαση σε αυτό καθ'



Σχήμα 2.1: *Συνεργατικό Φιλτράρισμα πληροφορίας*

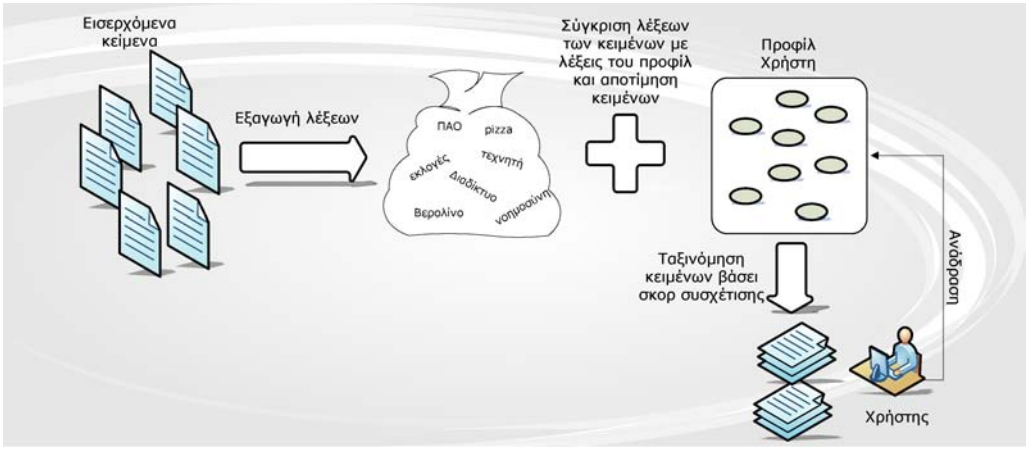
αυτό το περιεχόμενο της πληροφορίας και, συνεπώς, δεν περιορίζεται στην "κειμενική" (textual) πληροφορία. Ένα απλό παράδειγμα εφαρμογής του collaborative filtering είναι η εισήγηση στο χρήστη ταινιών, βιβλίων ή μουσικής, βάσει των προτιμήσεων των "φίλων" του [1].

2.2 Content - based

Η πληροφορία και το προφίλ του χρήστη αναπαρίστανται χρησιμοποιώντας στοιχεία από το περιεχόμενο της πληροφορίας. Λόγω της φύσης του χρησιμοποιείται, κυρίως, σε φιλτράρισμα κειμένων και εγγράφων, από τα οποία μπορούμε να εξάγουμε διακριτά στοιχεία - γνωρίσματα (features). Συνήθως, χρησιμοποιείται ένας κοινός διανυσματικός χώρος (vector space model) όπου τα προφίλ και τα κείμενα, που αναπαρίστανται ως δυαδικά ή σταθμισμένα διανύσματα, προβάλλονται στις ίδιες διαστάσεις[1]. Αυτό επιτρέπει στη συνέχεια τη χρήση τριγωνομετρικών μεθόδων για την σύγκριση της πληροφορίας με το προφίλ του χρήστη (Σχήμα 2.2).

Το διάνυσμα του προφίλ αποτελείται, συνήθως, από τις λέξεις που χαρακτηρίζουν τα ενδιαφέροντα του χρήστη, συνοδευόμενες από βάρη που δηλώνουν τη σημασία της κάθε λέξης στο προφίλ.

[1] Το μέγεθος του διανυσματικού χώρου είναι ίσο με τον αριθμό των features που εξάγονται από το κείμενο των εγγράφων.



Σχήμα 2.2: *Προσαρμοστικό Content-Based Φιλτράρισμα εγγράφων*

2.3 Υβριδικό ΦΠ

Τα τελευταία χρόνια το ενδιαφέρον των ερευνητών έχει επικεντρωθεί στην προσπάθεια τους να συνδυάσουν τεχνικές φιλτραρίσματος πληροφορίας με σκοπό να πετύχουν καλύτερες προβλέψεις αλλά και μεγαλύτερη προσαρμογή στα ενδιαφέροντα των χρηστών. Δημιουργήθηκε, λοιπόν, ένα νέο είδος διήθησης, το Υβριδικό (hybrid) ΦΠ.

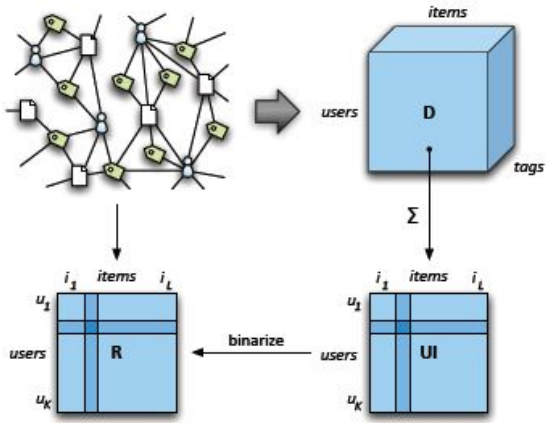
Τα περισσότερα συστήματα που έχουν αναπτυχθεί αφορούν συνδυασμούς collaborative και content - based φιλτραρίσματος [2]. Βασιζόμενοι στις προβλέψεις του συνεργατικού φιλτραρίσματος, χρησιμοποιούν δεδομένα από το περιεχόμενο της πληροφορίας για να ενισχύσουν την ακρίβειά τους [3]. Άλλα συστήματα χρησιμοποιούν κάποια μεταδεδομένα (metadata) όπως ετικέτες, κατηγορίες κ.α για επιτύχουν καλύτερη και αναλυτικότερη περιγραφή του προφίλ ενός χρήστη [4, 5]. Επίσης, υπάρχουν και υβριδικά συστήματα συστάσεων που συνδυάζουν το ΦΠ με τεχνικές νευρωνικών δικτύων αλλά και αλγόριθμων μηχανικής μάθησης [6].

Στα πλαίσια της παρούσης διπλωματικής εργασίας θα επικεντρωθούμε στο *Υβριδικό ΦΠ* με βάση τις *ετικέτες* καθώς παρουσιάζει ενδιαφέροντα χαρακτηριστικά και ιδιότητες. Μερικά από αυτά είναι:

- η αυξανόμενη χρήση τους στο διαδίκτυο
- ο δυναμικός τρόπος ομαδοποίησης και οργάνωσής τους, αλλά και
- το γεγονός ότι με τη χρήση ετικετών μπορούν να ”χαρακτηριστούν” πληροφορίες στο διαδίκτυο ανεξάρτητα από τη μορφή τους (κείμενο, ήχος, εικόνα, βίντεο).

Στις επόμενες ενότητες, παρουσιάζονται οι δομές δεδομένων και τα μέτρα σύγκρισης που θα χρησιμοποιηθούν σε αυτή την υβριδική υλοποίηση.

2.4 Δομές δεδομένων



Σχήμα 2.3: *Αναπαράσταση των συνδέσεων του συνόλου των δεδομένων σε τρισδιάστατο πίνακα.*

Μία δομή δεδομένων που συγκεντρώνει όλες τις απαραίτητες πληροφορίες για την υλοποίηση κάποιας από τις παραπάνω τεχνικές φιλτραρίσματος, είναι αυτή ενός τρισδιάστατου πίνακα. Σε αυτόν τον πίνακα, στον πρώτο από τους άξονες αναπαριστώνται οι χρήστες(users) του συνόλου δεδομένων. Στο δεύτερο άξονα αναπαριστώνται τα urls / items που έχει αξιολογήσει (άμεσα ή έμμεσα) ο κάθε χρήστης και, τέλος, στον τρίτο άξονα υπάρχουν οι ετικέτες που έχει αναθέσει κάθε χρήστης στα items. Επομένως, είναι δυνατόν από αυτόν τον πίνακα να εξαχθούν επιμέρους υποπίνακες που να αναπαριστούν τις σχέσεις users-items, users-tags, items-tags. Συστήματα συστάσεων που εξετάζουν τις ομοιότητες και τις συσχετίσεις μεταξύ των χρηστών (*user based*), χρησιμοποιούν τον user-items πίνακα, ενώ αντιθέτως, συστήματα που εξετάζουν τις ομοιότητες των αντικειμένων (*item based*), κάνουν χρήση του πίνανκα items-users. Οι τιμές που περιέχονται στους πίνακες μπορούν να είναι βεβαρημένες ή να δηλώνουν τον αριθμό εμφανίσεων κάποιου στοιχείου ή απλώς δυαδικές. Στα πλαίσια της παρούσης εργασίας, όλοι οι υποπίνακες είναι δυαδικοί. Περιέχουν, δηλαδή, τιμές 1 ή 0 με τις οποίες δηλώνεται αντίστοιχα η ύπαρξη ή μη, σύνδεσης των δυο υπό εξέταση στοιχείων.

Στο Σχήμα 2.3 παρατηρούμε πως ο δυαδικός πίνακας R που αναπαριστά τις σχέσεις users-items μπορεί να παραχθεί με δύο τρόπους. Ο ένας είναι απευθείας από το γράφο που περιγράφει το σύνολο δεδομένων, και αναπαριστά τις συνδέσεις χρηστών - αντικειμένων. Ο δεύτερος είναι μέσω του τρισδιάστατου πίνακα D. Συνυπολογίζοντας τη διάσταση με τις ετικέτες, παίρνουμε τον πίνακα UI που περιέχει τον αριθμό των ετικετών για κάθε αντικείμενο του χρήστη. Μετατρέποντας σε δυαδικές τις τιμές του UI, παίρνουμε πάλι τον πίνακα R.

2.5 Μετρικές Σύγκρισης

Σημαντική λειτουργία των συστημάτων συστάσεων και κατ' επέκταση των αλγορίθμων φιλτραρίσματος πληροφορίας είναι ο εντοπισμός ομοιότητας μεταξύ χρηστών ή αντικειμένων. Η σύγκριση αυτή γίνεται μεταξύ των προφίλ τους, τα οποία αναπαριστώνται ως

διανύσματα. Το μέγεθος του διανυσματικού χώρου ορίζεται από τον αριθμό των χρηστών, αντικειμένων ή ετικετών που υπάρχουν. Όπως γίνεται αντιληπτό, τα μεγέθη αυτά ισούνται από τις διαστάσεις του τρισδιάστατου πίνακα που περιγράψαμε στην προηγούμενη υποενότητα.

Για τη σύγκριση δύο διανυσμάτων οι πιο δημοφιλείς μέθοδοι σύγκρισης είναι:
το *εσωτερικό γινόμενο* :

$$\langle P, D \rangle = \sum_{i=1}^{N} p_i \cdot d_i \tag{2.1}$$

το *συνημίτονο* της μεταξύ τους γωνίας :

$$\cos(\theta) = \frac{P \cdot D}{\|P\|\|D\|} \tag{2.2}$$

και ο *συντελεστής* ομοιότητας *Jaccard*:

$$sim_{Jaccard}(P, D) = \frac{|P \cap D|}{|P \cup D|} \tag{2.3}$$

όπου: *P, D*[2] είναι ισομεγέθη διανύσματα με δυαδικές ή βεβαρημένες τιμές

Στις επόμενες ενότητες ακολουθεί μία γενική αναφορά στη χρήση ετικετών στο διαδίκτυο και τη σημασία τους και στη συνέχεια θα εστιάσουμε σε εφαρμογές και υλοποιήσεις που τις ενσωματώνουν στη λειτουργιά τους.

[2]P = profile, D = document/item

3

Ετικέτες (tags)

3.1 Ορισμός

Η “ετικέτα”, όταν αναφερόμαστε σε πληροφοριακά συστήματα, είναι ένας όρος ή μια λέξη-κλειδί η οποία αποδίδεται σε κάποιο αντικείμενο πληροφορίας. Αποτελεί ένα είδος μεταδεδομένου (metadata) που βοηθά στην περιγραφή ενός στοιχείου και διευκολύνει τον εντοπισμό του μέσα από εφαρμογές περιήγησης ή αναζήτησης. Οι ετικέτες δεν είναι απαραιτήτως κάποιες προκαθορισμένες λέξεις, αλλά συνήθως επιλέγονται προσωπικά είτε από το δημιουργό-πομπό είτε τον αποδέκτη της πληροφορίας, αναλόγως με τις δυνατότητες που προσφέρει το εκάστοτε πληροφοριακό σύστημα.

3.2 Ιστορική Αναδρομή

Η χρήση των ετικετών βρήκε πρόσφορο έδαφος με την ανάπτυξη του Web2.0 και αποτέλεσε σημαντικό χαρακτηριστικό των υπηρεσιών του. Το 2003, η ιστοσελίδα κοινωνικής διαχείρισης σελιδοδεικτών Delicious παρείχε στους χρήστες της τη δυνατότητα να προσθέσουν ετικέτες στους σελιδοδείκτες τους, ώστε να μπορούν να τους αναζητήσουν εύκολα αργότερα. Παρείχε, επίσης, τη δυνατότητα να παρουσιάζονται συγκεντρωμένοι οι σελιδοδείκτες, όλων των χρηστών, που περιείχαν μία συγκεκριμένη ετικέτα. Το Flickr, με τη σειρά του, επιτρέπει στους χρήστες να προσθέσουν ετικέτες δικής τους επιλογής (Σχήμα 3.1), δημιουργώντας “ευέλικτα” μεταδεδομένα τα οποία καθιστούν ευκολότερη την αναζήτηση μεταξύ των εικόνων. Η επιτυχία του Flickr αλλά και η επιρροή του Delicious διέδωσαν την ιδέα, με αποτέλεσμα και άλλες ιστοσελίδες “κοινωνικών” υπηρεσιών - όπως YouTube, Last.fm, BibSonomy αλλά και Blogger, να ενσωματώσουν στο

Σχήμα 3.1: *Χρήση ετικετών στο* Flickr

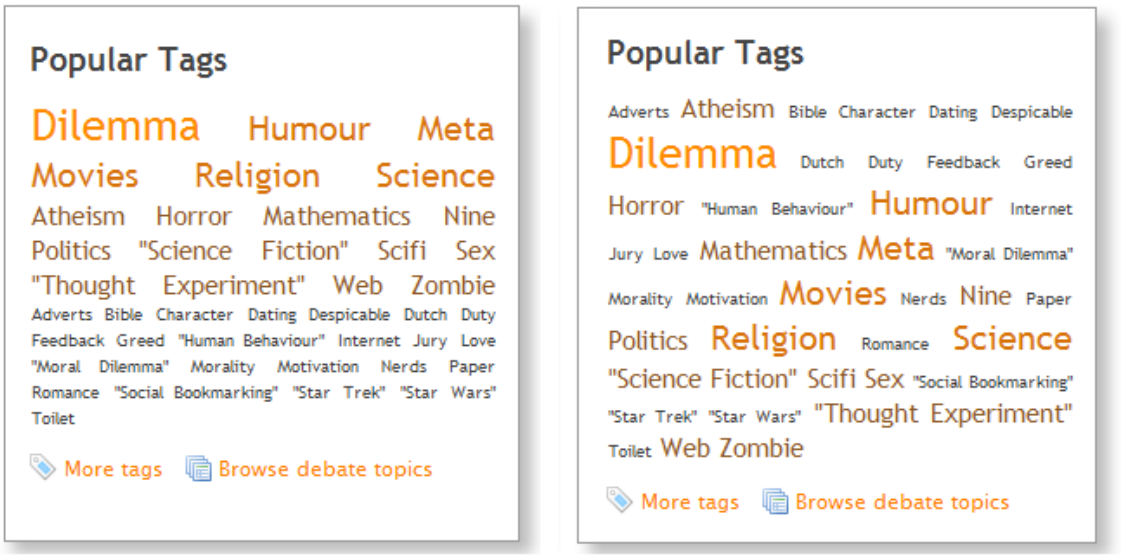
σύστημά τους ετικέτες.

Όπως μπορεί να παρατηρήσει κάποιος, οι ιστοσελίδες που αναφέρονται στην προηγούμενη παράγραφο προσφέρουν διαφορετικές, η καθεμία, υπηρεσίες και διαχειρίζονται διαφορετικής μορφής πληροφορίες. Όλες όμως χρησιμοποιούν τις ετικέτες με παρόμοιο τρόπο. Αυτό είναι και ένα ισχυρό χαρακτηριστικό του μηχανισμού των ετικετών. Μπορούν να προστεθούν σε οποιοδήποτε είδος πληροφορίας (άρθρα, φωτογραφίες, μουσική, βίντεο) και να βοηθήσουν στην περιγραφή και την κατηγοριοποίησή της.

3.3 Folksonomies

Συνεργατικά συστήματα ετικετών, όπως αυτά που προαναφέρθηκαν, είναι γνωστά ως folksonomies. Ο όρος *folksonomy*, επινόηση του Thomas Vander Wal [7], αποτελεί μία συνένωση των λέξεων folks ("ομάδα ατόμων", σε ελεύθερη μετάφραση) και taxonomy ("ταξονομία"). Σε αυτά τα συστήματα οι ετικέτες βοηθούν τους χρήστες να κατηγοριοποιήσουν και να αναζητήσουν πληροφορίες. Σε πολλά από αυτά χρησιμοποιούνται *tag clouds* για την απεικόνιση των ετικετών που περιέχονται σε μια folksonomy (Σχήμα 3.2). Στο tag cloud κάθε ετικέτα αναπαριστάται με μέγεθος γραμματοσειράς ανάλογο με τη σπουδαιότητα και τη συχνότητα εμφάνισής της στη folksonomy. Επιπροσθέτως, είναι σύνηθες οι ετικέτες σε κάθε tag cloud να ταξινομούνται

Σχήμα 3.2: *Ένα* tag cloud *από ετικέτες που σχετίζονται με το* Web 2.0



(α΄) Βάσει δημοτικότητας (β΄) Αλφαβητικά

Σχήμα 3.3: *Ταξινόμηση ετικετών σε tag cloud*

βάσει της δημοτικότητάς τους ή αλφαβητικά (Σχήμα 3.3)[8].

Ενδιαφέρον έχει μια εμπειρική ανάλυση της δυναμικής των συστημάτων σήμανσης [9], που δημοσιεύθηκε το 2007, και έδειξε ότι εμφανίζεται μία "σύμπνοια" και μία σταθερή κατανομή στις λέξεις που χρησιμοποιούνται ως ετικέτες, ακόμη και εν απουσία ενός, κεντρικά, ελεγχόμενου λεξιλογίου. Η αιτία αυτής της "σύμπνοιας" είναι το, εκ των πραγμάτων, κοινό λεξιλόγιο που χρησιμοποιείται από τα μέλη μιας folksonomy για να περιγραφούν έννοιες αλλά και τομείς ενδιαφέροντος.

Παρόλο που υπήρχε η αντίληψη ότι τα σύνολα των λέξεων που θα χρησιμοποιούνταν ως ετικέτες, έπρεπε να είναι προκαθορισμένα, για να είναι το περιεχόμενο της πληροφορίας κατηγοριοποιημένο και ευκόλως αναζητήσιμο, πρόσφατη έρευνα έδειξε πως σε μεγάλες folksonomies, δημιουργούνται, αυθόρμητα, κοινές μορφές κατηγοριοποιήσεων [10]. Ως εκ τούτου, είναι δυνατόν να αναπτυχθούν μαθηματικά μοντέλα που επιτρέπουν τη μετάφραση ενός προσωπικού λεξιλογίου ετικετών (personomies) σε ένα λεξιλόγιο που διαμοιράζονται περισσότεροι χρήστες.

3.4 Σχετικές Μελέτες

Λόγω της αυξανόμενης ενσωμάτωσης των ετικετών σε διαδικτυακές εφαρμογές και κοινωνικά δίκτυα, υπάρχει έντονο ενδιαφέρον και έχουν γίνει αρκετές μελέτες από ομάδες

ερευνητών. Για τους σκοπούς της εργασίας μας, εστιάσαμε σε μελέτες που χρησιμοποιούν τις ετικέτες σε συνδυασμό με συστήματα συστάσεων. Στις επόμενες παραγράφους αναφέρουμε τις πιο αντιπροσωπευτικές.

Οι Tso-Sutter et al.[2] κάνουν ένα συνδυασμό των αποτελεσμάτων του user-based και item-based φιλτραρίσματος και, επιπροσθέτως, εκμεταλλεύονται την πληροφορία των ετικετών. Κατά την εκτέλεση του user-based φιλτραρίσματος οι ετικέτες λογίζονται ως επιπλέον αντικείμενα, ενώ στο item-based ως επιπλέον χρήστες. Υπολογίζουν τις εν δυνάμει συστάσεις και με τους δύο τρόπους και έπειτα αθροίζουν τα αποτελέσματα, χρησιμοποιώντας μια παράμετρο λ, και παράγουν τις τελικές συστάσεις.

Στο [5] οι συγγραφείς, αρχικά, ομαδοποιούν τις ετικέτες από ένα σύνολο αντικειμένων και τις κατατάσσουν σε topics. Στη συνέχεια, αντιστοιχούν τα υπό εξέταση αντικείμενα σε αυτά τα topics, σύμφωνα με τις ετικέτες που τους έχει αναθέσει ο χρήστης. Ανάλογα με τα topics στα οποία ενδιαφέρεται ο κάθε χρήστης, αναθέτουν και μεγαλύτερο βάρος στις ετικέτες που τα περιγράφουν.

Ένας άλλος τομέας όπου εντοπίζονται μελέτες γύρω από ετικέτες, είναι αυτός της αυτόματης σύστασης ετικετών για αντικείμενα που πρωτοσυναντά ο χρήστης. Συνήθως οι εργασίες σε αυτόν τον τομέα χωρίζονται σε τρεις κατηγορίες: content-based, collaborative και graph-based. Οι συστάσεις ετικετών βασίζονται τόσο στην συμπεριφορά της folksonomy όσο και σε αυτή του χρήστη. Μία τέτοια μελέτη περιγράφεται στο [11] όπου οι ετικέτες, οι χρήστες και τα αντικείμενα αναπαριστώνται ως κόμβοι σε ένα γράφο και οι ακμές μεταξύ τους δείχνουν το βαθμό διασύνδεσης τους. Η σύσταση σχετικών ετικετών σε ένα νέο χρήστη ή αντικείμενο, προκύπτει εμμέσως από το γράφο, αναλύοντάς τον από την προοπτική του χρήστη ή του αντικειμένου αντίστοιχα.

Οι Hotho et al. στο [12] παρουσιάζουν ένα τυπικό μοντέλο και έναν νέο αλγόριθμο αναζήτησης για folksonomies, τον *FolkRank*, ο οποίος προσπαθεί να εκμεταλλευτεί τη δομή της folksonomy. Το σύστημα κατάταξης του *FolkRank* έχει χρησιμοποιηθεί στη συγκεκριμένη δημοσίευση για τη δημιουργία εξατομικευμένης κατάταξης των αντικειμένων σε μια folksonomy, και για να συστήσει τους χρήστες, τις κατάλληλες ετικέτες. Ο αλγόριθμος αυτός εφαρμόζεται επίσης για να εντοπίσει μικρές "κοινότητες" εντός της folksonomy και χρησιμοποιείται για να δομήσει τα αποτελέσματα αναζήτησης.

Στο [3] προτείνουν μια λύση στο πρόβλημα της σύστασης κατάλληλων ετικετών (κυρίως για έγγραφα), που περιλαμβάνει την ομαδοποίηση των υφιστάμενων εγγράφων με σκοπό τον εντοπισμό σύνολων με παρόμοια έγγραφα. Τα σύνολα αυτά, με τη σειρά τους, προσδιορίζουν το σύνολο των χρηστών των οποίων οι ετικέτες μπορούν να ανατεθούν στο τρέχον, υπό εξέταση, έγγραφο. Η λίστα με τις πιθανές ετικέτες που μπορεί να αναθέσει

ένας συγκεκριμένος χρήστης στο έγγραφο προκύπτει από τις ετικέτες των παρεμφερών εγγράφων. Στη συνέχεια, για κάθε πιθανή ετικέτα, υπολογίζεται μία τιμή που αποτελείται από τον σταθμισμένο συνδυασμό της ομοιότητα κάθε εγγράφου στο οποίο έχει ανατεθεί και της ομοιότητας του χρήστη που την έχει αναθέσει, διαιρεμένο με τον αριθμό των εγγράφων στα οποία εμφανίζεται.

Στο [13] υπολογίζεται η ομοιότητα μεταξύ ενός προφίλ χρήστη και ενός προφίλ αντικειμένου σε επίπεδο περιεχομένου. Οι ετικέτες θεωρούνται χαρακτηριστικά (features) του περιεχομένου, τα οποία περιγράφουν τόσο τους χρήστες όσο και τα αντικείμενα. Επιπλέον, παρουσιάζονται αρκετά συστήματα στάθμισης για τη αποτίμηση της "σημασίας" μιας συγκεκριμένης ετικέτας για κάθε χρήστη και αντικείμενο. Μερικά από αυτά τα συστήματα στάθμισης βασίζονται σε μεμονωμένα προφίλ, ενώ άλλα σε ολόκληρη τη folksonomy.

Ακόμη, υπάρχουν περιπτώσεις όπου γίνεται προσπάθεια να εκμεταλλευθούν οι σημασιολογικές έννοιες των ετικετών (semantics). Για παράδειγμα στο [4], διερευνάται η ενσωμάτωση της folksonomy για ταινίες με μια σημασιολογική - γνωστική βάση σχετικά με τις ενοικιάσεις ταινιών από τους χρήστες. Η folksonomy χρησιμοποιείται για τον εμπλουτισμό της γνωστικής βάσης[1] με περιγραφές και κατηγοριοποιήσεις των ταινιών, αλλά και με τα ενδιαφέροντα και τις απόψεις των χρηστών. Χρησιμοποιώντας ετικέτες που συλλέγονται από το IMDB[2], και στοιχεία αξιολόγησης ταινιών από το Netflix, διεξάγονται πειράματα για να διερευνηθεί κατά πόσον ένα *tag-cloud* μιας folksonomy μπορει να χρησιμοποιηθεί για την κατασκευή ενός καλύτερου προφίλ, που θα αντανακλά τα ενδιαφέροντα ενός χρήστη σε διαφορετικά είδη ταινιών, και ως εκ τούτου, θα παρέχει μια βάση για την πρόβλεψη της βαθμολογίας - αξιολόγησής του για μια ταινία που δεν έχει, προηγουμένως, παρακολουθήσει.

Τέλος, οι Bogers και Van Den Bosch στην εργασία τους [14] εξετάζουν τις εγγραφές των χρηστών στις υπηρεσίες κοινωνικής διαχείρισης σελιδοδεικτών, BibSonomy[3] και CiteULike[4]. Επιδιώκουν να ενσωματώσουν τις ετικέτες αλλά και άλλα μεταδεδομένα, όπως τίτλο, περιγραφή κ.λ.π, στο μοντέλο του collaborative filtering εκμεταλλευόμενοι νέες μετρικές σύγκρισης, όπως η "επικάλυψη ετικετών". Εξετάζουν, επίσης, μία υλοποίηση συστήματος συστάσεων που συνδυάζει τα παραπάνω μεταδεδομένα με τεχνικές content-based filtering.

Η παραπάνω δημοσίευση αποτέλεσε τον "οδηγό" μας για τα πειράματα που εκτελέσαμε και τα οποία περιγράφονται στο επόμενο κεφάλαιο. Αποφασίσαμε να επιλέξουμε την εν

[1] Knowledge base
[2] Internet Movie Database: www.imdb.com
[3] www.bibsonomy.org
[4] www.citeulike.org

λόγω εργασία ως σημείο αναφοράς για τους παρακάτω, κυρίως, λόγους:

- Ενσωματώνουν σε συστήματα συστάσεων, ετικέτες.
- Παραθέτουν και συγκρίνουν τα αποτελέσματα αρκετών πειραμάτων στα οποία χρησιμοποιούν ποικίλους αλγορίθμους και μετρικές σύγκρισης.
- Σε αντίθεση με άλλες δημοσιεύσεις που μελετήσαμε, έχουν αναλυτική περιγραφή των αλγορίθμων αλλά και των συνόλων δεδομένων που χρησιμοποιούν.

4

Πειράματα

4.1 Σύνολα Δεδομένων

Για την εκτέλεση των πειραμάτων χρησιμοποιήθηκε μία συλλογή δεδομένων από την υπηρεσία του BibSonomy[1] [15]. Το *BibSonomy* είναι ένα σύστημα κοινωνικής διαχείρισης σελιδοδεικτών και διαμοιρασμού δημοσιεύσεων[2]. Προσφέρει στους χρήστες τη δυνατότητα να αποθηκεύσουν και να οργανώσουν σελιδοδείκτες και καταχωρήσεις δημοσιεύσεων. Υποστηρίζει την ενσωμάτωση μεμονωμένων αλλά και ομάδων χρηστών, προσφέροντας, με αυτόν τον τρόπο, μια κοινωνική πλατφόρμα για την ανταλλαγή βιβλιογραφίας. Σε κάθε σελιδοδείκτη ή δημοσίευση μπορούν να προστεθούν *ετικέτες*, βοηθώντας στην οργάνωση και την αναζήτηση πληροφορίας. Η υπηρεσία αναπτύχθηκε από μια ομάδα φοιτητών και επιστημόνων που εργάζονται στο Institute of Knowledge and Data Engineering του Πανεπιστημίου του Kassel, στη Γερμανία.

Η συγκεκριμένη συλλογή επιλέχθηκε ως η καταλληλότερη για τη διενέργεια των πειραμάτων καθώς, εκτός από τα στοιχεία που έχει κάνει bookmark κάθε χρήστης, περιλαμβάνει χρονικές επισημάνσεις (timestamps) για κάθε ανάθεση του χρήστη και ένα ικανοποιητικό πλήθος ετικετών. Πιο συγκεκριμένα, χρησιμοποιήθηκε η συλλογή ***2008-30-06*** η οποία χωρίζεται σε τρία αρχεία - βάσεις δεδομένων[3]. Το *πρώτο* από αυτά (αρχείο *tas*) περιέχει χρονολογημένες καταγραφές για κάθε σελιδοδείκτη ή δημοσίευση που προστέθηκε στο σύστημα. Περιλαμβάνει πληροφορίες για το χρήστη, το είδος του στοιχείου που προστέθηκε (bookmark ή bibtex) και τις ετικέτες που του ανέθεσε ο χρήστης (Σχ. 4.1).

[1] Η πρόσβαση στα στοιχειά παραχωρήθηκε μετά από αίτηση και η χρήση τους αφορά αποκλειστικά ερευνητικούς σκοπούς

[2] Στα αγγλικά αποδίδεται ως “social bookmarking and publication-sharing system”.

[3] Περισσότερες πληροφορίες για τα περιεχόμενα της συλλογής δεδομένων υπάρχουν στη διεύθυνση

```
------------------------------------------
 ⋮
11 webservices 688970 1 2005-12-09 23:09:13
11 government  688970 1 2005-12-09 23:09:13
11 owl         688970 1 2005-12-09 23:09:13
11 engineering 688971 1 2005-12-09 23:02:15
11 ontology    688971 1 2005-12-09 23:02:15
 ⋮
------------------------------------------
```

Σχήμα 4.1: *Δείγμα των εγγραφών στο αρχείο* "tas"

Το *δεύτερο* (αρχείο *bookmark*) και *τρίτο* αρχείο της συλλογής περιέχουν επιπλέον πληροφορίες για τα στοιχεία που αφορούν τους *σελιδοδείκτες* και τις *δημοσιεύσεις* αντίστοιχα. Αυτές οι πληροφορίες, μεταξύ άλλων, περιλαμβάνουν ένα μοναδικό URL (ως MD5 hash) και μια μικρή περιγραφή, για κάθε στοιχείο (Σχ. 4.2).

```
-------------------------------------------------------------------------------
 ⋮
688968 89b23812c69176cf6782a17b9dbf000a [url] [description] 2005-12-12 14:25:42
688969 2579bed76238dee1e745040e66d97ea5 [url] [description] 2005-12-12 14:23:46
688970 a44441112d9524ac7b7cd9b5e5c1eee4 [url] [description] 2005-12-09 23:09:13
688971 1292468dfefb8169d9cba95f2df7c4eb [url] [description] 2005-12-09 23:02:15
 ⋮
-------------------------------------------------------------------------------
```

Σχήμα 4.2: *Δείγμα των εγγραφών στο αρχείο* "bookmark"

Στην παρούσα μεταπτυχιακή διατριβή, αξιοποιήθηκαν και επεξεργάστηκαν τα δεδομένα από το πρώτο και δεύτερο αρχείο της συλλογής, καθώς αντικείμενο μελέτης αποτέλεσαν αποκλειστικά οι πληροφορίες που αφορούσαν του σελιδοδείκτες (bookmarks) των χρηστών.

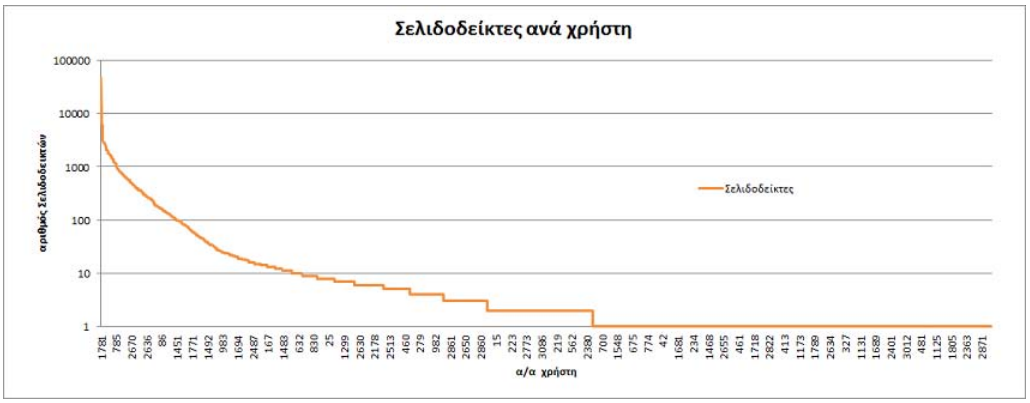
4.1.1 Στατιστική ανάλυση δεδομένων

Τα δεδομένα που περιέχονταν στα αρχεία *tas* και *bookmark*, υπέστησαν μια προεργασία πριν χρησιμοποιηθούν στα πειράματά μας. Οι προεργασίες αφορούσαν, κυρίως, τις ετικέτες και ήταν η αφαίρεση των σημείων στίξεως στην αρχή και στο τέλος των λέξεων, την αφαίρεση μερικών μη αναγνωρίσιμων χαρακτήρων και, τέλος, η διαγραφή των όσων ετικετών αποτελούνταν από μόνο έναν χαρακτήρα.

Μετά πέρας των προεργασιών το σύνολο των διαθέσιμων δεδομένων περιέχει *2.288* **χρήστες** οι οποίοι έχουν αναθέσει *668.878* **ετικέτες** σε *204.784* **σελιδοδείκτες**. Αυτός ο αριθμός των σελιδοδεικτών αντιστοιχεί σε *182.029 μοναδικά* **url hashes**, ενώ οι *μοναδικές* **ετικέτες** είναι *49.749*.

www.kde.cs.uni-kassel.de/bibsonomy/dumps.

Στο Σχήμα 4.3 παρουσιάζεται, σε φθίνουσα σειρά, η κατανομή των σελιδοδεικτών ανάμεσα στους χρήστες. Πόσοι, δηλαδή, σελιδοδείκτες προστέθηκαν από τον κάθε χρήστη. Παρατηρούμε πως σχεδόν οι μισοί χρήστες έχουν προσθέσει, το πολύ, μέχρι 2 σελιδοδείκτες. Αντιθέτως, είναι μερικοί μεμονωμένοι χρήστες που έχουν προσθέσει έναν υπερβολικά μεγάλο αριθμό σελιδοδεικτών. Αυτοί οι τελευταίοι, μετά από περαιτέρω εξέταση, φαίνεται να είναι κάποιοι "αυτοματοποιημένοι" λογαριασμοί χρηστών του Bibsonomy οι οποίοι προσθέτουν συγκεκριμένες λέξεις ως ετικέτες[4]. Λαμβάνοντας υπόψη τα παρα-



Σχήμα 4.3: *Κατανομή των σελιδοδεικτών ως προς τους χρήστες*

πάνω, αποφασίσαμε να φιλτράρουμε τους χρήστες βάσει των σελιδοδεικτών τους, έτσι ώστε από τη μία, να αποφύγουμε ακραίες συμπεριφορές και από την άλλη, να εξασφαλίσουμε ότι κάθε χρήστης που θα συμμετέχει στο πείραμα θα διαθέτει έναν ικανοποιητικό αριθμό σελιδοδεικτών. Επομένως, συμπεριλάβαμε στα πειράματά μας τους χρήστες που είχαν στο προφίλ τους *από 20 έως 3000* σελιδοδείκτες. Το σύνολο των *τελικών* χρηστών ήταν *346*.

	αρχικά	μετά το φιλτράρισμα
χρήστες	2.288	346
σελιδοδείκτες	204.784	123.739
ετικέτες	668.878	436.071
μον. url hash	182.029	107.240
μον. ετικέτες	49.749	38.946

Πίνακας 4.1: *Στατιστικά στοιχεία του συνόλου δεδομένων*

[4]Σχεδόν όλοι τους ανέθεταν στους σελιδοδείκτες τους την ετικέτα "imported".

4.2 Μεθοδολογία

Τα πειράματα σε αυτή την εργασία είχαν ως σκοπό τη σύγκριση της απόδοσης διαφόρων τεχνικών που χρησιμοποιούνται για την παραγωγή συστάσεων. Συνολικά εξετάστηκαν τέσσερις διαφορετικές τεχνικές.

Για την αποτίμηση και σύγκριση της απόδοσης καθενός από τους αλγόριθμους συστάσεων απαιτείται η δημιουργία ενός πειραματικού πλαισίου αναφοράς. Μελετήθηκαν αρκετές μέθοδοι εκτέλεσης και αξιολόγησης πειραματικών διαδικασιών. Αποφασίστηκε να επιλέξουμε αυτές που περιγράφονται και στο [14] έτσι ώστε εκτός από ποιοτική να έχουμε και τη δυνατότητα ποσοτικής σύγκρισης με τα αποτελέσματα της συγκεκριμένης δημοσίευσης.

Τα αρχικά δεδομένα της συλλογής χωρίζονται σε δυο σύνολα. Επιλέγεται τυχαία το 10% των χρηστών που θα αποτελέσει το *σύνολο ελέγχου* και το υπόλοιπο 90% αποτελεί το *σύνολο εκπαίδευσης* του συστήματος. Η απόδοση του συστήματος υπολογίζεται σε αυτό το 10% παρακρατώντας από κάθε χρήστη του συγκεκριμένου συνόλου 10 αντικείμενα. Έπειτα, χρησιμοποιώντας τα εναπομείναντα αντικείμενα μαζί με αυτά του συνόλου εκπαίδευσης παράγονται συστάσεις για αυτό το 10% των χρηστών. Αν τα αντικείμενα που είχαν παρακρατηθεί, εμφανιστούν, μετά το πέρας της διαδικασίας, στην κορυφή της λίστας με τα προτεινόμενα, για τον χρήστη, αντικείμενα τότε ο αλγόριθμος θεωρείται επιτυχής και αποδοτικός. Για να επιτύχουμε μια καλύτερη και "αντικειμενικότερη" εκτίμηση κάνουμε χρήση του *10-fold-cross-validation*. Χωρίζουμε, δηλαδή, το σύνολο των δεδομένων μας σε 10 τμήματα και επιλέγουμε κάθε φορά ένα από αυτό, τυχαία, να αποτελέσει το σύνολο ελέγχου και τα υπόλοιπα εννέα το σύνολο εκπαίδευσης. Επαναλαμβάνουμε τη διαδικασία 10 φορές και υπολογίζουμε το μέσο όρο των επιμέρους εκτιμήσεων.

4.2.1 Μέθοδος Αξιολόγησης

Στο [Herlocker et al.][16] οι συγγραφείς κάνουν εκτιμήσεις για τη χρηστικότητα διαφορετικών μέτρων αξιολόγησης για διαφορετικού τύπου συστήματα συστάσεων. Για συστήματα που παράγουν συστάσεις βάσει της τεχνικής *top-N*[5], η οποία εφαρμόζεται στα πειράματά μας, εκτιμούν πως μέτρα αξιολόγησης που λαμβάνουν υπόψη την κατάταξη των προτεινόμενων αντικειμένων είναι τα πλέον κατάλληλα. Για αυτό το λόγο, για την αξιολόγηση των αλγορίθμων που εξετάζονται στην εργασία χρησιμοποιείται η μέθοδος *Mean Average Precision*. Δηλαδή, ο μέσος όρος των επιμέρους μέσων όρων της *ακρίβειας*

[5] ***top-N***: Η μεθοδολογία κατά την οποία ταξινομούνται τα αποτελέσματα των συγκρίσεων δυο αντικειμένων (ή χρηστών) και επιλέγονται τα Ν πρώτα.

που υπολογίστηκε για κάθε ένα αντικείμενο της λίστας. Όπου ως *ακρίβεια* ορίζεται το ποσοστό των "προτεινόμενων" αντικειμένων που ανήκουν στη λίστα με τα αντικείμενα που είχαν παρακρατηθεί αρχικά.

4.3 Πειραματικά Αποτελέσματα

Συνολικά εκτελέστηκαν τέσσερα πειράματα. Στα πειράματα αυτά εξετάστηκαν τεχνικές συνεργατικού φιλτραρίσματος και χρησιμοποιήθηκε, για τον υπολογισμό των συστάσεων, μια παραλλαγή του αλγόριθμου των *k Πλησιέστερων Γειτόνων* [6]. Τα πειράματα χωρίστηκαν σε δυο σκέλη.

Στο πρώτο σκέλος μελετήθηκαν οι δύο κυριότερες εκδοχές των αλγορίθμων συστάσεων: το *user-based filtering* και *item-based filtering*. Στην πρώτη εκδοχή, όπως προδίδει και η ονομασία του, εντοπίζουμε παρόμοιους χρήστες με αυτούς του *συνόλου ελέγχου* και βρίσκουμε αντικείμενα που θα μπορούσαν να προταθούν σε αυτούς. Στο *item-based filtering* εντοπίζουμε όμοια αντικείμενα με αυτά που έχουν προσθέσει οι χρήστες του *συνόλου ελέγχου*, τα συγκεντρώνουμε σε λίστα και προτείνονται στους χρήστες τα καταλληλότερα.

Στο δεύτερο σκέλος των πειραμάτων ο αλγόριθμος συστάσεων προσπαθεί να εκμεταλλευτεί την πληροφορία που περιέχεται στις ετικέτες των αντικειμένων. Προσμετράται στο υπολογισμό των επιμέρους συστάσεων, και η επικάλυψη των ετικετών, που εντοπίζεται είτε μεταξύ των χρηστών είτε μεταξύ των αντικειμένων του συνόλου δεδομένων.

Σε όλες τις παραπάνω παραλλαγές των πειραμάτων ο αριθμός των *πλησιέστερων γειτόνων* αλλά των *top-N* παρόμοιων χρηστών ή αντικειμένων κυμαίνεται από *N=2* έως *N=20*. Η επιλογή αυτή έγινε διότι, μετά από ανάλυση των αποτελεσμάτων, δεν παρατηρείται κάποια βελτίωση στην ακρίβεια των πειραμάτων για *N>20*.

4.3.1 User-based filtering

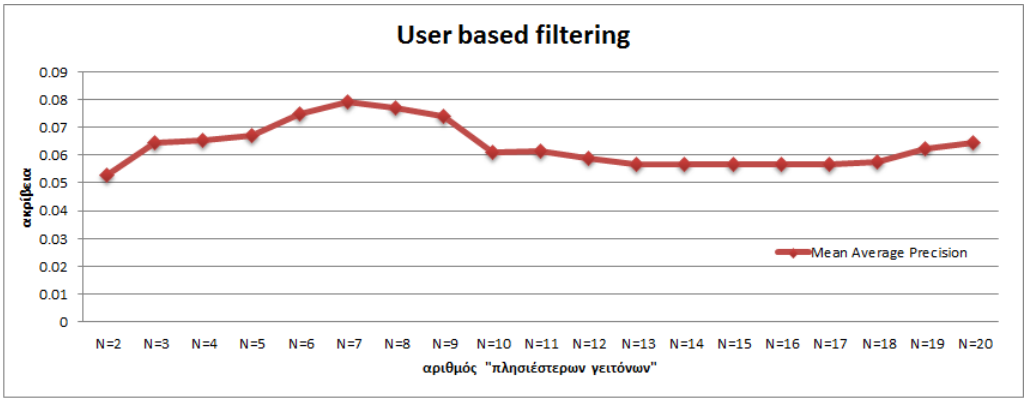
Σε αυτό το πείραμα χρησιμοποιούμε ως δομή δεδομένων τον πίνακα **UI** που απεικονίζεται στο Σχήμα 2.3. Αρχικά, συγκρίνουμε κάθε χρήστη του *συνόλου ελέγχου* με όλους τους υπόλοιπους για να εντοπίσουμε με ποιους έχει τις περισσότερες ομοιότητες. Για τη σύγκριση αυτή χρησιμοποιούμε το μέτρο του *συνημιτόνου* της μεταξύ τους γωνίας (Εξ. 2.2).

Έπειτα, συγκεντρώνουμε τους top N παρόμοιους χρήστες με κάποιον χρήστη u_k του συ-

[6] k-Nearest Neighbor (k-NN)

νόλου ελέγχου. Αυτοί αποτελούν μία λίστα ταξινομημένη σε φθίνουσα σειρά ως προς την ομοιότητά τους. Για κάθε ένα χρήστη (u_a) σε αυτή τη λίστα, εξετάζουμε τα αντικείμενα που έχει προσθέσει. Όσα από αυτά τα αντικείμενα δεν τα έχει προσθέσει ο χρήστης u_k, είναι υποψήφια για να προταθούν σε αυτόν και υπολογίζουμε μια *τιμή πρόβλεψης* για καθένα από αυτά. Η τιμή πρόβλεψης ενός αντικειμένου είναι το άθροισμα των τιμών ομοιότητας των *N κοντινότερων* χρηστών (u_a) που έχουν προσθέσει το αντικείμενο αυτό.

Παρακάτω παρουσιάζεται η γραφική παράσταση με τα αποτελέσματα του πειράματος. Στον *x*-άξονα απεικονίζεται ο αριθμός των "πλησιέστερων γειτόνων" που ελήφθησαν υπόψη και στον *y*-άξονα η τιμή της ακρίβειας (precision) του αποτελέσματος. Να σημειωθεί πως τα ίδια χαρακτηριστικά απεικονίζονται σε όλες τις γραφικές παραστάσεις που παρατίθενται στη συνέχεια.



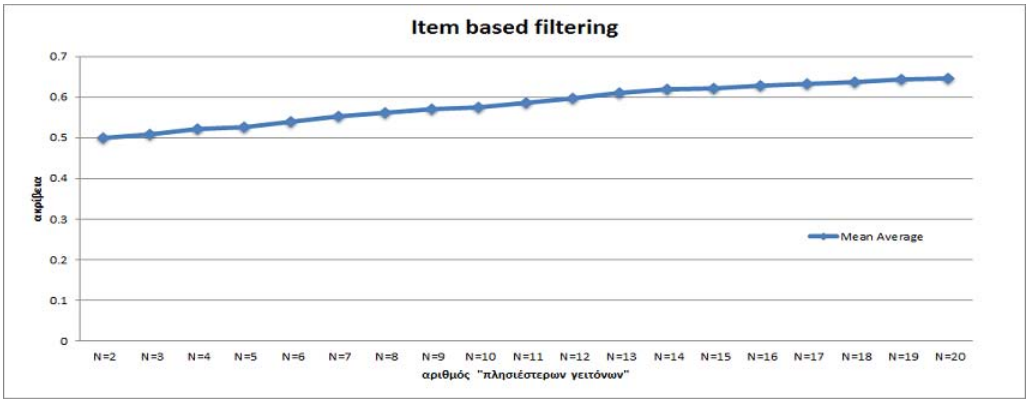
Σχήμα 4.4: *Φιλτράρισμα με βάση τους χρήστες της folksonomy*

Στην παραπάνω γραφική παράσταση παρατηρούμε πως η μέγιστη ακρίβεια του πειράματος επιτυγχάνεται για *N=7*. Μεγαλύτερος αριθμός "γειτόνων" φαίνεται πως προσθέτει περισσότερο "θόρυβο" και οδηγεί σε λανθασμένες συστάσεις. Στο αντίστοιχο πείραμα του [14] η υψηλότερη ακρίβεια επιτυγχάνεται με *N=15* και τιμή *0,0277*.

4.3.2 Item-based filtering

Ο αλγόριθμος που χρησιμοποιήθηκε σε αυτό το πείραμα λειτουργεί με τρόπο ανάλογο του user-based filtering. Μόνο που τώρα η σύγκριση δεν γίνεται σε επίπεδο χρηστών, αλλά αφορά τα αντικείμενα που έχει προσθέσει κάθε χρήστης του συνόλου ελέγχου. Δηλαδή, υπολογίζουμε την ομοιότητα των αντικειμένων ενός χρήστη με τα υπόλοιπα αντικείμενα που δεν έχει ακόμα προσθέσει (δει). Για τον υπολογισμό της ομοιότητας χρησιμοποιούμε κι εδώ το μέτρο του *συνημιτόνου* της μεταξύ τους γωνίας (Εξ. 2.2). Η δομή δεδομένων που χρησιμοποιείται είναι πάλι ο πίνακας **UI** αλλά αυτή τη φορά η επεξεργασία και η προσπέλασή του γίνεται κατά **στήλες**.

Για κάθε ένα αντικείμενο του χρήστη επιλέγουμε τα top N παρόμοια με αυτό αντικείμενα, που δεν έχουν ήδη προστεθεί από τον ίδιο. Δημιουργείται πάλι μια ταξινομημένη λίστα που περιέχει τα αντικείμενα αυτά. Για κάθε ένα "άγνωστο" προς το χρήστη αντικείμενο υπολογίζεται μία τιμή πρόβλεψης. Η συγκεκριμένη τιμή αποτελείται από το άθροισμα των τιμών ομοιότητας των N *κοντινότερων* αντικειμένων που έχει προσθέσει ο χρήστης.



Σχήμα 4.5: *Φιλτράρισμα με βάση τα αντικείμενα της folksonomy*

Η ακρίβεια που επιτυγχάνεται μέσα από αυτό το πείραμα είναι αρκετά αυξημένη σε σχέση με το προηγούμενο. Παρατηρείστε ότι ο x-άξονας είναι διαφορετικής κλίμακας με τιμές από 0 έως 0.7, αυτή τη φορά. Η μεγαλύτερη ακρίβεια εμφανίζεται για *N=20* και παρουσιάζει αυξητική τάση για αριθμό "γειτόνων" μεγαλύτερο του 20. Παρ 'όλα αυτά, για λόγους συνέπειας και ομοιομορφίας με τα υπόλοιπα πειράματα περιοριστήκαμε σε αυτό τον αριθμό.

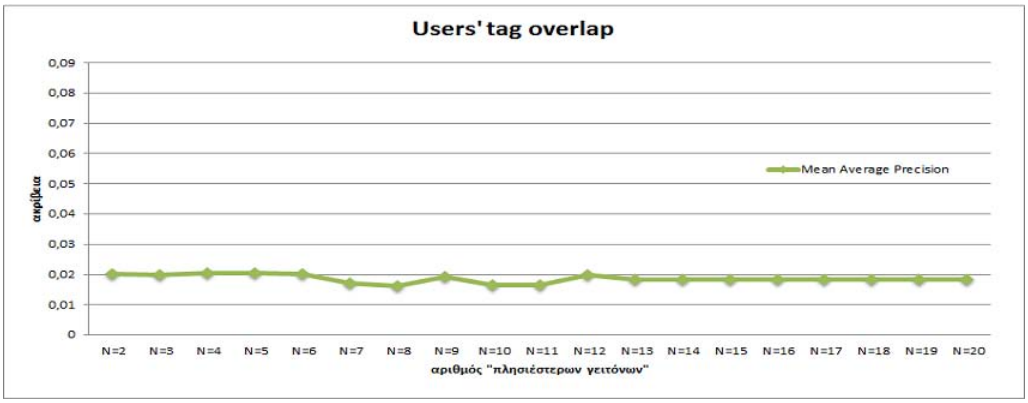
4.3.3 Επικάλυψη ετικετών

Στη folksonomy που περιγράφεται από το σύνολο των δεδομένων που εξετάζουμε, έχουμε και ένα επιπλέον επίπεδο διασύνδεσης μεταξύ των χρηστών και των αντικειμένων. Τις *ετικέτες*. Μπορούμε να εκμεταλλευτούμε αυτή την επιπλέον πληροφορία έτσι ώστε να προσδιορίσουμε και μία άλλου τύπου ομοιότητα μεταξύ χρηστών ή αντικειμένων. Για παράδειγμα, χρήστες που αναθέτουν παρόμοιες ετικέτες μπορούν να θεωρηθούν όμοιοι, ή αντιστοίχως αντικείμενα στα οποία ανατίθενται συχνά οι ίδιες ετικέτες μπορεί να θεωρηθεί ότι έχουν κάποια ομοιότητα.

Για την εκτέλεση των παρακάτω πειραμάτων βασιζόμαστε στους πίνακες **users-tags (UT)** και **items-tags (IT)** (Σχήμα 2.3). Ως μέτρο σύγκρισης ομοιότητας χρησιμοποιούμε, αυτή τη φορά, το *συντελεστή Jaccard* (Εξ. 2.3).

- Μεταξύ *χρηστών*

Ο αλγόριθμος ακολουθεί την ίδια διαδικασία με αυτή του *user-based* πειράματος, με τη διαφορά ότι αυτή τη φορά, τα προφίλ των χρηστών που συγκρίνονται δεν περιέχουν τα αντικείμενα που έχουν προσθέσει, αλλά τις ετικέτες που έχουν αναθέσει στα αντικείμενά τους.

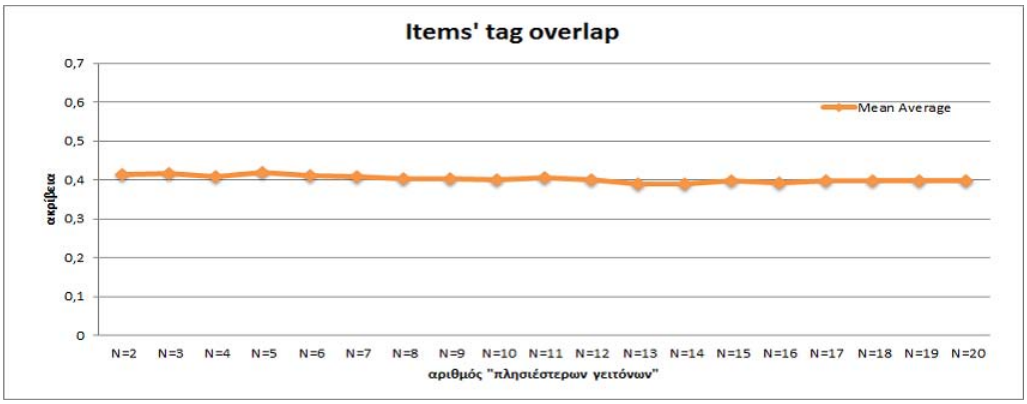


Σχήμα 4.6: *Φιλτράρισμα με βάση την επικάλυψη των ετικετών των χρηστών*

Τα αποτελέσματα του πειράματος είναι στα ίδια επίπεδα με αυτά του user-based. Η μεγαλύτερη ακρίβεια επιτυγχάνεται για *N=4* και *N=5*, ενώ με περισσότερους "γείτονες" έχουμε ελαφρώς χειρότερα αποτελέσματα. Αξιοσημείωτο είναι το γεγονός πως για $13 \leq N \leq 20$ η ακρίβεια παραμένει ανεπηρέαστη.

- Μεταξύ *αντικειμένων*

Ομοίως όπως στο παραπάνω πείραμα, μόνο που αυτή τη φορά χρησιμοποιείται ο **IT** πίνακας και η σύγκριση γίνεται μεταξύ των ετικετών που έχουν ανατεθεί στα συγκεκριμένα αντικείμενα.

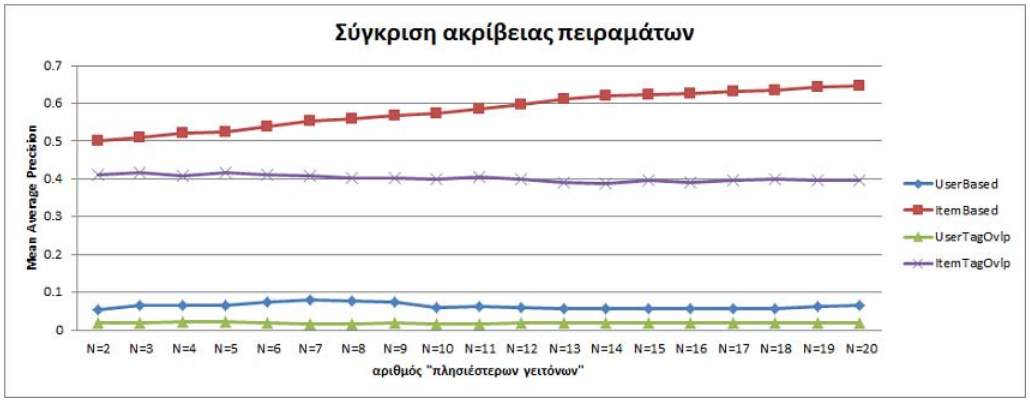


Σχήμα 4.7: *Φιλτράρισμα με βάση την επικάλυψη των ετικετών των αντικειμένων*

Στην παραπάνω γραφική παράσταση είναι εμφανές ότι έχουμε βελτιωμένα αποτελέσματα σε σχέση με το προηγούμενο πείραμα και γι' αυτό άλλωστε χρησιμοποιείται πάλι διαφορετική κλίμακα στον x-άξονα. Η μεγαλύτερη ακρίβεια επιτυγχάνεται για *N=5*, ενώ οι περισσότερους "γείτονες" φαίνεται πως μειώνουν, έστω και σε μικρό βαθμό, την ακρίβεια του συστήματος συστάσεων.

4.4 Συζήτηση των αποτελεσμάτων

Στο παρακάτω Σχήμα 4.8 παρατίθενται συγκεντρωτικά όλες οι γραφικές παραστάσεις των προηγούμενων πειραμάτων.



Σχήμα 4.8: *Σύγκριση των αποτελεσμάτων όλων των πειραμάτων*

Παρατηρούμε πως τα πειράματα με σημείο αναφοράς τα αντικείμενα (items) επιτυγχάνουν εμφανώς καλύτερα αποτελέσματα. Ιδιαίτερα το item-based πείραμα εμφανίζει σημαντικά μεγάλη ακρίβεια, η οποία παρουσιάζει αυξητική τάση όσο ο αριθμός των "πλησιέστερων γειτόνων" μεγαλώνει. Ενδιαφέρον αποτελεί το γεγονός ότι η μέγιστη ακρίβεια του item-based είναι *17x* φορές υψηλότερη κατά από την μέγιστη ακρίβεια του αντίστοιχου πειράματος στο [14]. Μία αιτία για αυτή τη μεγάλη διαφορά, θεωρούμε πως είναι μία μικρή διαφοροποίηση στην υλοποίησή μας. Ενώ εμείς συγκρίνουμε ένα αντικείμενο του χρήστη με όλα τα υπόλοιπα της συλλογής, οι Bogers και Van Den Bosch δεν διευκρινίζουν αν ακολουθούν την ίδια διαδικασία ή αν συγκρίνουν ένα αντικείμενο του χρήστη μόνο με όσα αντικείμενα έχουν προσθέσει οι χρήστες του συνόλου εκπαίδευσης.

Επίσης, ένας άλλος λόγος που τα αποτελέσματα των πειραμάτων με σημείο αναφοράς τα αντικείμενα (items) είναι καλύτερα από αυτά με σημείο αναφοράς τους χρήστες (users), είναι η στατιστική αναλογία αντικειμένων ως προς τους χρηστες. Οι γραμμές του πίνακα User-Items, που αναπαριστούν τους χρήστες, είναι αρκετά πιο αραιές (sparse) από τις γραμμές που αναπαριστούν τα αντικείμενα στον πίνακα Item-Users. Επομένως, είναι μι-

κρή η πιθανότητα να εντοπιστεί ικανοποιητικός αριθμός χρηστών που μοιράζονται πολλά κοινά αντικείμενα και αυτό με τη σειρά του μειώνει την ποιότητα των συστάσεων του συστήματος.

Όσον αφορά τα αποτελέσματα μετά την ενσωμάτωση των ετικετών, παρουσιάζονται κατώτερα από τα αντίστοιχα πειράματα στα οποία δε λάβαμε υπόψη τις ετικέτες. Μπορούμε να συμπεράνουμε πως οι αλγόριθμοι που χρησιμοποιούν την επικάλυψη των ετικετών δεν παράγουν καλύτερες συστάσεις από τις "παραδοσιακές" προσεγγίσεις των user-based και item-based πειραμάτων. Αυτό, πιθανότατα, οφείλεται στη χρήση, εκ μέρους των χρηστών, διαφορετικών ετικετών για να περιγράψουν παρεμφερή ή, ακόμη και, ίδια αντικείμενα. Η μέθοδος της επικάλυψης των ετικετών δεν εξετάζει *σημασιολογικά* τις ετικέτες, παρά μόνο ελέγχει την ύπαρξή τους στα προφίλ των χρηστών. Επομένως, παρόμοιες εννοιολογικά ή και συντακτικά ετικέτες θεωρούνται διαφορετικές μεταξύ τους και δεν σημειώνεται κάποια επικάλυψη.

5

Σύνοψη και μελλοντικές εργασίες

Συνοψίζοντας, στην παρούσα εργασία εστιάσαμε στη λειτουργία των πληροφοριακών συστημάτων συστάσεων (recommender systems) και εξετάσαμε τους κυριότερους αλγορίθμους που χρησιμοποιούνται για την παραγωγή αυτόματων συστάσεων. Αναφερθήκαμε στις τεχνικές διήθησης της πληροφορίας που χρησιμοποιούνται από τα συστήματα συστάσεων, δηλαδή, το collaborative filtering, το content based filtering αλλά και σε τεχνικές υβριδικού φιλτραρίσματος.

Στα πλαίσια του υβριδικού φιλτραρίσματος, επικεντρωθήκαμε στο συνδυασμό collaborative filtering με τη χρήση ετικετών (tags). Αναφέραμε, στο Κεφάλαιο 3, τα πλεονεκτήματα και τη χρηστικότητα των ετικετών ως μεταδεδομένα (metadata) σε ένα σύστημα συστάσεων και ορίσαμε την έννοια της *folksonomy*.

Μελετήσαμε και αναφέρουμε αρκετές σχετικές εργασίες που χρησιμοποιούν ετικέτες για την παραγωγή συστάσεων. Πολλές από τις συστάσεις αυτές αφορούν αντικείμενα που προτείνονται στους χρήστες, ενώ κάποιες άλλες αφορούν προτεινόμενες ετικέτες σε αντικείμενα που προσθέτουν, για πρώτη φορά, οι χρήστες στο πληροφοριακό σύστημα. Από τις σχετικές εργασίες ξεχωρίσαμε αυτή των Bogers και Van Den Bosch [14] οι οποίοι εξετάζουν μια πληθώρα αλγορίθμων και μέτρων σύγκρισης για συστήματα συστάσεων και προσπαθήσαμε να αναπαράγουμε και να βελτιώσουμε κάποιες από τις υλοποιήσεις τους.

Αναπαράγαμε τέσσερα διαφορετικά πειράματα. Υλοποιήσαμε αρχικά ένα σύστημα συστάσεων βασισμένο στις αλληλεπιδράσεις των χρηστών (user-based) και ένα στις αλληλεπιδράσεις μεταξύ των αντικειμένων (item-based). Στη συνέχεια, εξελίξαμε τα δύο αυτά συστήματα χρησιμοποιώντας ετικέτες και υπολογίζοντας την επικάλυψη των ετικετών είτε μεταξύ χρηστών (UserTagOverlap) είτε αντικειμένων (ItemTagOverlap). Από τα αποτελέσματα των πειραμάτων προέκυψαν:

- Το item-based πείραμα είχε τα καλύτερα αποτελέσματα. Μάλιστα είχε πολύ καλύτερα αποτελέσματα από από το πείραμα αναφοράς στο [14].
- Τα πειράματα με σημειο αναφοράς τα αντικείμενα είχαν καλύτερα αποτελέσματα από αυτά με σημείο αναφοράς τους χρήστες.
- Η ενσωμάτωση ετικετών στα πειράματα είχε ως αποτέλεσμα, ελαφρώς, μειωμένη ακρίβεια σε σχέση με τα αντίστοιχα πειράματα χωρίς ετικέτες.

Η απόπειρα να χρησιμοποιήσουμε ετικέτες για την παραγωγή πληροφοριακών συστάσεων, παρ' ότι δεν πέτυχε τη μεγαλύτερη ακρίβεια, είχε ενθαρρυντικά αποτελέσματα. Μια πιο εξειδικευμένη και βελτιωμένη υλοποίηση είναι πιθανόν να παράγει ακόμα καλύτερα αποτελέσματα και να ξεπεράσει το item-based. Οι Bogers και Van Den Bosch προτείνουν στην εργασία μία υλοποίηση όπου προσμετράνε, εκτός των ετικετών, και άλλα μεταδεδομένα, όπως τίτλος, περιγραφή, url, ημερομηνία κ.α, και εμφανίζεται να επιτυγχάνει καλύτερες συστάσεις. Επίσης, τεχνικές οι οποίες επεξεργάζονται σημασιολογικά (semantically) τις ετικέτες μπορούν να ενισχύσουν το αποδοτικό φιλτράρισμα των πληροφοριακών αντικειμένων. Επιπροσθέτως, σε όσα αντικείμενα περιέχεται κειμενική πληροφορία[1], θα μπορούσαν να συνδυαστούν οι ετικέτες τους με τα γνωρίσματα (features) που θα προέκυπταν από το, βάσει περιεχομένου, φιλτράρισμά τους. Τέλος, μία a priori επεξεργασία των ετικετών σε μια folksonomy, τόσο λεξικογραφικά όσο και εννοιολογικά, θα επέτρεπε μία προ - κατηγοριοποίηση των ετικετών που θα έκανε πιο ουσιαστική την "επικάλυψη" των ετικετών (ιδιαίτερα στο ItemTagOverlap).

Θεωρούμε πως τα συστήματα πληροφοριακών συστάσεων με χρήση ετικετών χρήζουν περισσότερης έρευνα και μελέτης μιας και οι ετικέτες αποδεικνύονται χρησιμότατο εργαλείο του Web2.0. Επιφυλασσόμαστε, να βελτιώσουμε την ενσωμάτωση των ετικετών στα συστήματα συστάσεων και να επιτύχουμε καλύτερα πειραματικά αποτελέσματα κινούμενοι προς τις κατευθύνσεις που περιγράψαμε στην παραπάνω παράγραφο.

[1] Άρθρα, περιλήψεις(abstract) δημοσιεύσεων, περιγραφές πολυμέσων

Βιβλιογραφία

[1] N. Nanas and A. De Roeck, "A review of evolutionary and immune inspired information filtering," *Natural Computing*, 2007.

[2] K. H. L. Tso-Sutter, L. B. Marinho, and L. Schmidt-Thieme, "Tag-aware recommender systems by fusion of collaborative filtering algorithms," in *Proceedings of the 2008 ACM symposium on Applied computing*, SAC '08, (New York, NY, USA), pp. 1995–1999, ACM, 2008.

[3] N. Landia and S. Anand, "Personalised Tag Recommendation," *Recommender Systems & the Social Web*, 2009.

[4] M. Szomszor, C. Cattuto, H. Alani, K. O'Hara, A. Baldassarri, V. Loreto, and V. D. Servedio, "Folksonomies, the semantic web, and movie recommendation," in *4th European Semantic Web Conference, Bridging the Gap between Semantic Web and Web 2.0*, 2007. Event Dates: 3-7th, June 2007.

[5] H. Liang, Y. Xu, Y. Li, R. Nayak, and G. Shaw, "A hybrid recommender systems based on weighted tags." May 2010.

[6] C. Christakou and A. Stafylopatis, "A Hybrid Movie Recommender System Based on Neural Networks," in *ISDA '05: Proceedings of the 5th International Conference on Intelligent Systems Design and Applications*, (Washington, DC, USA), pp. 500–505, IEEE Computer Society, 2005.

[7] T. V. Wal, "Folksonomy coinage and definition." Website, Februar 2007. http://vanderwal.net/folksonomy.html.

[8] W. McGugan, "Tag clouds look better sorted!." http://www.willmcgugan.com/blog/tech/2007/10/31/tag-clouds-look-better-sorted/, 2007. [Online; accessed 09-April-2012].

[9] H. Halpin, V. Robu, and H. Shepherd, “The complex dynamics of collaborative tagging,” in *Proceedings of the 16th international conference on World Wide Web*, WWW ’07, (New York, NY, USA), pp. 211–220, ACM, 2007.

[10] V. Robu, H. Halpin, and H. Shepherd, “Emergence of consensus and shared vocabularies in collaborative tagging systems,” *ACM Trans. Web*, vol. 3, pp. 14:1–14:34, Sept. 2009.

[11] Y. Song, L. Zhang, and C. L. Giles, “Automatic tag recommendation algorithms for social recommender systems,” *ACM Trans. Web*, vol. 5, pp. 4:1–4:31, Feb. 2011.

[12] A. Hotho, R. Jaschke, C. Schmitz, and G. Stumme, “Information retrieval in folksonomies: Search and ranking,” in *The Semantic Web: Research and Applications* (Y. Sure and J. Domingue, eds.), vol. 4011 of *Lecture Notes in Computer Science*, (Berlin/Heidelberg), pp. 411–426, Springer, June 2006.

[13] I. Cantador, A. Bellogín, and D. Vallet, “Content-based recommendation in social tagging systems,” in *Proceedings of the fourth ACM conference on Recommender systems*, RecSys ’10, (New York, NY, USA), pp. 237–240, ACM, 2010.

[14] T. Bogers and A. Van Den Bosch, “Collaborative and content-based filtering for item recommendation on social bookmarking websites,” *ACM RecSys □09 Workshop on Recommender Systems and the Social Web*, vol. 9, pp. 9–16, 2009.

[15] D. Benz, A. Hotho, R. Jäschke, B. Krause, F. Mitzlaff, C. Schmitz, and G. Stumme, “The social bookmark and publication management system bibsonomy,” *The VLDB Journal*, vol. 19, pp. 849–875, Dec. 2010.

[16] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl, “Evaluating collaborative filtering recommender systems,” *ACM Trans. Inf. Syst.*, vol. 22, pp. 5–53, Jan. 2004.