

Οικονομικό Πανεπιστήμιο Αθηνών
ΜΠΣ Πληροφοριακών Συστημάτων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

« Ανάπτυξη συστήματος εμπλουτισμού αρχείου καθιερωμένων εγγραφών με κοινωνικές επισημειώσεις. »

Σπύρος Δάγλας

M3100023

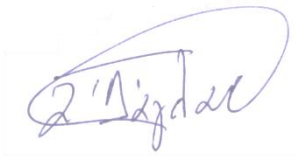
ΑΘΗΝΑ, ΙΑΝΟΥΑΡΙΟΣ 2012

"…Φίλος είναι εκείνος που καταλαβαίνει το παρελθόν σου, πιστεύει στο μέλλον σου και σε δέχεται όπως είσαι…"

-Λαϊκή ρήση-

Στους καλύτερους μου φίλους,
τους γονείς μου, Παναγιώτη και Έλντα
και στις αδελφές μου, Μάιρα- Σπυριδούλα
και Νίκη- Κωνσταντίνα. Σας ευχαριστώ πολύ,
για την στήριξη και τις θυσίες σας.

Σπύρος Π. Δάγλας



Ιδιαίτερες ευχαριστίες στον κύριο **Παπαθεοδώρου Χρήστο**, αναπληρωτή Καθηγητή του τμήματος αρχειονομίας- βιβλιοθηκονομίας του Ιονίου πανεπιστημίου, στην κυρία **Κακάλη Ντίνα**, υποψήφια διδάκτορα του τμήματος αρχειονομίας-βιβλιοθηκονομίας του Ιονίου πανεπιστημίου και στους μεταπτυχιακούς φοιτητές του τμήματος Πληροφοριακών Συστημάτων του Οικονομικού Πανεπιστημίου Αθηνών, **Κουμάκη Μαρίνα** και **Κακαβούλη Διονύσιου**, για την εξαιρετική συνεργασία μας τους τελευταίους 6 μήνες και την βοήθεια, τα σχόλια και τις υποδείξεις, για την συγγραφή της παρούσας διπλωματικής εργασίας.

Σας ευχαριστώ πολύ και ελπίζω και σε άλλες συνεργασίες μας στο μέλλον.

«…Νομίζω ότι οι κοινωνικές επισημειώσεις, θα μεταβάλλουν το τοπίο των πληροφοριακών συστημάτων, με πολύ καθοριστικό τρόπο…»

-- Jon Udell

Πρόλογος. Η συνεχώς αυξανόμενη πρόσβαση σε υπηρεσίες κοινωνικής δικτύωσης (όπως το Facebook, το LinkedIn κ.α.), καθώς και η ανάπτυξη του Web 2.0, βοήθησαν στην «διάσχιση» των συνόρων και την ένταξη των ανθρώπων σε ένα νέο, ψηφιακό τοπίο. Καθώς ο κόσμος κινείται όλο και πιο κοντά στην ενσωμάτωση της τεχνολογίας σε παραδοσιακές κοινωνικές συμπεριφορές, υπάρχει μεγαλύτερη ανάγκη για καινοτόμο έρευνα και ανάπτυξη στις διάφορες πτυχές της κοινωνικής έκφανσης των υπολογιστών.

Το ερώτημα αν εμείς, αλλάξαμε το διαδίκτυο, ή το διαδίκτυο άλλαξε εμάς, είναι δύσκολο να απαντηθεί. Μολονότι η διαδικασία είναι ένα μυστήριο, οι αλλαγές του διαδικτύου είναι γεγονός. Μια πρόσφατη έρευνα μεγάλης κλίμακας, για την χρήση του διαδικτύου, ανέδειξε μια μεγάλη μεταβολή στον τρόπο που προσεγγίζουμε το διαδίκτυο. Αν σκεφτόμαστε το Web ως ένα γιγάντιο πίνακα ανακοινώσεων, τις πρώτες ημέρες του, θα τοποθετούνταν κάποιες σκόρπιες επισημειώσεις, αλλά οι περισσότεροι απλώς θα περνούν από δίπλα του, απρόσεκτα, πλοηγούμενοι από το ένα σημείωμα στο επόμενο. Τώρα πια, η πλοήγηση δεν είναι η ίδια. Το διαδίκτυο έχει γίνει προέκταση του εαυτού μας που μας επιτρέπει να επικοινωνούμε με ανθρώπους από όλων τον κόσμο. Έχουμε μάθει να το χρησιμοποιούμε για να αναζητούμε άλλους ανθρώπους, με αποτέλεσμα να μετατραπεί σε μια τεράστια στρογγυλή τράπεζα[1]. Γύρω της, το «συμβούλιο» θα συζητάει, θα ρωτάει, θα δίνει συμβουλές, θα μοιράζεται τις εμπειρίες και τις ιδέες του και θα οικοδομεί σχέσεις μεταξύ των μελών του.

Το αντικείμενο της παρούσας διπλωματική εργασίας, είναι η ανάπτυξη ενός συστήματος εμπλουτισμού αρχείου καθιερωμένων εγγραφών με κοινωνικές

[1] Στον Αρθουριανό κύκλο, η **Στρογγυλή Τράπεζα** ήταν ένα μυστικό τραπέζι συγκεντρώσεων στο Κάμελοτ, όπου ο Βασιλιάς Αρθούρος και οι Ιππότες του μαζεύονταν για να συζητήσουν ζητήματα υψίστης σημασίας για το βασίλειο. Η ιδέα της Στρογγυλής Τραπέζης χρονολογείται από το έργο του Τζέφρεϊ του Μονμάουθ *Historia Regum Britanniae*, αλλά η πιο δημοφιλής ιστορία προέρχεται από το *Μέρλιν* του Robert de Boron. Σε αυτήν, η Τράπεζα κατασκευάστηκε από το Μέρλιν ως μίμηση του τραπεζιού στο Μυστικό Δείπνο.

επισημειώσεις. Για την αποτελεσματικότητα του συστήματος θα εξετάσουμε τις συνέπειές του, στον ηλεκτρονικό κατάλογο της βιβλιοθήκης του Πάντειου Πανεπιστημίου - Κοινωνικών και Πολιτικών Επιστημών - Αθηνών.

Το πρώτο κεφάλαιο της παρούσας εργασίας, παρουσιάζει αναλυτικά όλους τους όρους και τις διαδικασίες που πρέπει να γνωρίζει ο αναγνώστης για την καλύτερη κατανόηση του συστήματος.

Στην συνέχεια, στο δεύτερο κεφάλαιο, γίνεται μια επιγραμματική παρουσίαση, άλλων ερευνών με παρόμοιο αντικείμενο, που μας βοήθησαν στην εξαγωγή συμπερασμάτων για την σταδιακή δόμηση του δικού μας συστήματος.

Το τρίτο έως και το πέμπτο κεφάλαιο, αποτελούν τα βασικά κεφάλαια της διπλωματικής μας εργασίας, στα οποία παρουσιάζεται αναλυτικά το δικό μας σύστημα, βήμα - βήμα, καθώς και τα προβλήματα που αντιμετωπίσαμε κατά τον σχεδιασμό του. Ο διαχωρισμός των κεφαλαίων αυτών έγινε, για τον καλύτερο διαχωρισμό των βημάτων που ακολουθήθηκαν μέχρι το τέλος της έρευνάς μας.

Στο έκτο κεφάλαιο, αναφέρονται μικρότερα ζητήματα και επιμέρους θέματα που το δικό μας σύστημα δεν προβλέπει και δεν βοηθάει στην λύση τους.

Στο έβδομο κεφάλαιο, γίνεται αναφορά, στα ανοιχτά θέματα του συγκεκριμένου ερευνητικού τομέα και τα οποία αντιμετωπίζει η διεθνής ερευνητική κοινότητα προς εύρεση λύσης.

Στο όγδοο κεφάλαιο, παρουσιάζονται γενικά συμπεράσματα που αποκομίσαμε από την έρευνας μας, καθώς και μια περιληπτική επισκόπηση του συστήματος μας. Τέλος, ο επίλογος της διπλωματικής μας εργασίας, ακολουθείται από την παρουσίαση της αναλυτικής βιβλιογραφίας που χρησιμοποιήσαμε.

Περιεχόμενα

Πρόλογος **iii**

Εισαγωγή **1**

1.1. Web 2.0 1

1.1.1. Ιστορική αναδρομή στο Web 2.0 4

1.2. Κοινωνικά Δίκτυα 5

1.2.1. Ανάλυση κοινωνικών δικτύων: Σύντομη ιστορική ανασκόπηση 7

1.2.2. Χρησιμότητα Ανάλυσης Κοινωνικών Δικτύων 10

1.3 Λογισμικό κοινωνικών δικτύων και δημόσιες βιβλιοθήκες. 11

1.4 Library 2.0 12

1.5 Επισημειώσεις (Tagging) 15

1.5.1 Τομείς δημιουργίας επισημειώσεων. 18

1.6 Ταξονομίες και ελεγχόμενα λεξιλόγια 20

1.6.1. Ελεγχόμενα λεξιλόγια 20

1.6.1.1. Δαχτυλίδια Συνωνύμων. 21

1.6.1.2. Αρχεία Καθιερωμένων όρων. 21

1.6.2. Ταξονομίες. 22

1.7. Όψεις. 23

1.8. Κοινωνική Ταξονομία. 23

Βιβλιογραφική Επισκόπηση. **25**

2.1 Πρόβλεψη επισημειώσεων. 26

2.2 Συστήματα συστάσεων με χρήση επισημειώσεων. 29

2.3 Δημιουργία ταξονομιών επισημειώσεων. 33

Προτεινόμενο Μοντέλο – Βήμα 1 **37**

3.1 Σύντομη περιγραφή του μοντέλου μας. 38

3.2. Διαδικασία εκτίμησης επισημειώσεων. 39

3.2.1. Επεξεργασία αρχείου καθιερωμένων όρων. 41

3.2.2. Διαχωρισμός χρήσιμων επισημειώσεων. 47

Προτεινόμενο Μοντέλο – Βήμα 2 **50**

4.1. Σύντομη περιγραφή του LibraryThing. 52

4.2. Διαδικασία εξόρυξης πληροφοριών στο LibraryThing. 54

4.2.1. Θέματα υλοποίησης της ανάκτησης δεδομένων από το LibraryThing. 61

4.3. Σύντομη περιγραφή του Google Books. 66

4.4. Διαδικασία εξόρυξης πληροφοριών στο Google Books. 67

Προτεινόμενο Μοντέλο – Βήμα 3 **69**

5.1. Εύρεση TF-IDF. 69

5.2. Μετρικές και πειράματα. 72

Ιδιαιτερότητες του μοντέλου μας. **77**

Ανοιχτά Θέματα. **79**

Παράρτημα. **84**

Βιβλιογραφία **86**

"Data mining creates dossiers on people."
-- Larry Frankel

1 Εισαγωγή

Η ανάπτυξη του Web 2.0, καθώς και η εντυπωσιακή ανταπόκριση των χρηστών του διαδικτύου, σε εφαρμογές και υπηρεσίες κοινωνικής δικτύωσης, οδήγησε την ερευνητική κοινότητα στην ανάγκη ανάπτυξης ενός νέου κλάδου έρευνας, αναφορικά με την ανάλυση κοινωνικών δικτύων. Σε αυτό το κεφάλαιο, θα παρουσιάσουμε τα χαρακτηριστικά του Web 2.0 και πως καταλήξαμε σε αυτό, θα αναφέρουμε τι είναι τα κοινωνικά δίκτυα, καθώς επίσης θα κάνουμε και μια μικρή ιστορική ανασκόπηση σημαντικών γεγονότων που δημιούργησαν αυτόν τον κλάδο. Στην συνέχεια, θα παρουσιάσουμε και θα εξηγήσουμε, όλους τους όρους που θα χρησιμοποιήσουμε στην έρευνα μας, προς διευκόλυνση των αναγνωστών.

1.1. Web 2.0

Ο όρος **Web 2.0** (**Ιστός 2.0**), χρησιμοποιείται για να περιγράψει τη νέα γενιά του Παγκόσμιου Ιστού η οποία βασίζεται στην όλο και μεγαλύτερη δυνατότητα των χρηστών του Διαδικτύου να μοιράζονται πληροφορίες και να συνεργάζονται διαδικτυακά. Αυτή η νέα γενιά είναι μια δυναμική διαδικτυακή πλατφόρμα στην

οποία μπορούν να αλληλεπιδρούν χρήστες χωρίς εξειδικευμένες γνώσεις σε θέματα υπολογιστών και δικτύων.

Το Web 2.0 ξεπερνά τα όρια της περιορισμένης πλατφόρμας ενός υπολογιστή. Όταν οι άνθρωποι μιλούν για το Web 2.0, μιλάνε για ισοπέδωση των ιεραρχιών, αποκεντρωμένη λήψη αποφάσεων, απελευθέρωση της δημιουργικότητας, αξιοποίηση της συλλογική γνώσης και δυνατότητα στους χρήστες να δημιουργούν περιεχόμενο. Ο χρήστης μπορεί να δρα στον Παγκόσμιο Ιστό όπως ενεργούσε μέχρι τώρα στον υπολογιστή του. Οι πιο ειδικοί μιλούν για έναν νέο τρόπο σχεδίασης των ιστοσελίδων ο οποίος θα βασίζεται στην διάδραση με τον χρήστη. Θα επιτρέπει στον χρήστη να αλλάξει τόσο το περιβάλλον της σελίδας όσο και να παρέμβει στο περιεχόμενό της. Χαρακτηριστικές εφαρμογές του Web 2.0 είναι τα κοινωνικά μέσα (social media), τα wikis και τα blogs. Πολλές από τις εντολές διάδρασης που χαρακτηρίζουν την λειτουργία του Web 2.0, μας είναι ήδη γνωστές από διάφορες ιστοσελίδες κοινωνικών μέσων, όπως το facebook ή το youtube. Τέτοιες εκφράσεις είναι η αναζήτηση (search), οι επισημάνσεις (tagging), ή η παράθεση (link) ή οι συγγραφές (authoring) που λειτουργούν σε πολλά wikis, όπου οι χρήστες μπορούν να δημιουργήσουν άρθρα αλλά και να ανανεώσουν ή να διαγράψουν τα ήδη υπάρχοντα.

Οι αρχές τις οποίες επιδιώκει το Web 2.0, μέσω της σχεδίασης του, είναι οι παρακάτω [11]:

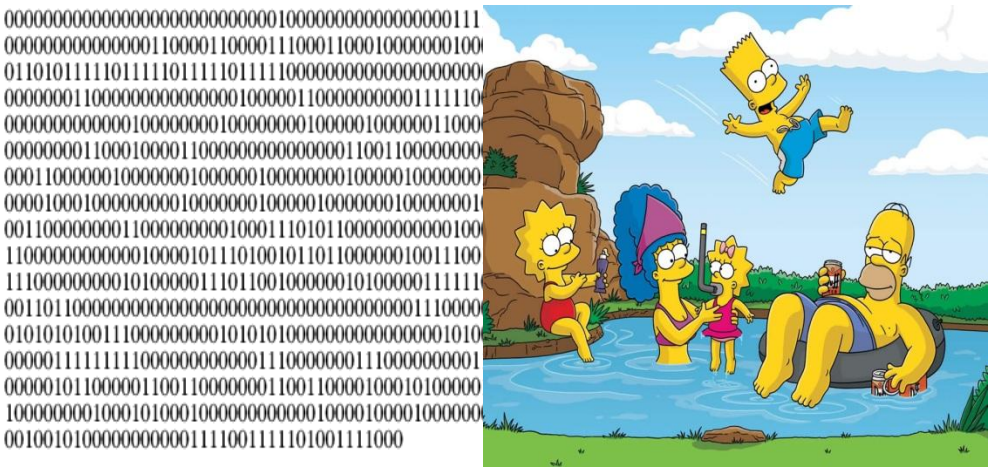
- Συνομιλίες: Συμμετοχή χρηστών. Συζητήσεις και ανατροφοδότηση είναι τα ζητούμενα και ενθαρρύνονται,
- Κοινότητα: Οι συνομιλίες οδηγούν στην αίσθηση ότι οι χρήστες ανήκουν σε μια κοινότητα,
- Συμμετοχή: Νέες πληροφορίες δημιουργούνται μέσω της συνεργασίας μεταξύ των χρηστών,
- Κοινή χρήση: Οι χρήστες μπορούν να συνεισφέρουν στις κοινότητες όσο θέλουν.

Ποιος λόγος όμως μας οδήγησε στο Web 2.0; Παρακάτω, θα δώσουμε για παράδειγμα 3 επερωτήσεις που το Web 1.0 δεν μπορούσε να μας δώσει ικανοποιητικές απαντήσεις.

? Ποιος είναι ο Dennis MacAlistair Ritchie;

- ? Ποιο καινούργιο άλμπουμ μουσικής μπορεί να μου αρέσει;
- ? Δείξε μου φωτογραφίες από το Παρίσι (Paris).

Οι παραπάνω ερωτήσεις προβάλλουν ένα γενικό πρόβλημα όσον αφορά την πρόσβαση σε τεράστιο όγκο πληροφοριών στο Διαδίκτυο. Δηλαδή, και στις τρεις περιπτώσεις έχουμε να κάνουμε με ένα κενό γνώσης. Ότι ο υπολογιστής κατανοεί και είναι σε θέση να επεξεργαστεί, είναι πολύ πιο περιορισμένο από την γνώση του χρήστη. Το μειονέκτημα του υπολογιστή, ως επί το πλείστον, οφείλεται στις τεχνολογικές δυσκολίες που συναντούμε, ώστε να καταφέρουμε τους υπολογιστές μας, να κατανοήσουν την φυσική μας γλώσσα ή να «βλέπει» το περιεχόμενο των εικόνων και άλλων πολυμέσων. Ακόμη και αν οι πληροφορίες είναι εκεί, και είναι φανερές σε κάποιον φυσικό χρήστη, ο υπολογιστής μπορεί να μην είναι σε θέση να δει τίποτα άλλο, εκτός από μια σειρά χαρακτήρων (εικόνα 1.1). Στην περίπτωση αυτή μπορεί να επεξεργαστεί και να συγκρίνει τις ερωτήσεις μας, με τις λέξεις-κλειδιά που παρέχουν οι χρήστες, αλλά χωρίς καμία κατανόηση για το τι αυτές οι λέξεις-κλειδιά σημαίνουν.



Εικόνα 1.0.1 Δεξιά η εικόνα γνωστής τηλεοπτικής επιτυχίας, που ο χρήστης κατανοεί, και αριστερά η "μετάφραση" την οποία κατανοεί ο υπολογιστής.

Η ιδέα του Web 2.0, είναι η εφαρμογή προηγμένων τεχνολογιών γνώσης, προκειμένου να καλυφθεί το κενό γνώσεων μεταξύ ανθρώπου και μηχανής. Αυτό σημαίνει να παρέχετε η γνώση σε μορφές που οι υπολογιστές θα μπορούν εύκολα να καταλάβουν και να επεξεργαστούν. Αυτή η γνώση μπορεί να είναι είτε πληροφορίες

που είναι ήδη περιγράφονται στο περιεχόμενο των ιστοσελίδων, αλλά δύσκολο να εξαχθούν, είτε πρόσθετες γνώσεις που θα προσφέρουν το κατάλληλο υπόβαθρο, ώστε να απαντηθούν οι ερωτήσεις με κάποιο τρόπο [15].

1.1.1. Ιστορική αναδρομή στο Web 2.0

Ο όρος "Web 2.0" επινοήθηκε τον Ιανουάριο του 1999 από τον Darcy DiNucci. Σε άρθρο του, ο DiNucci γράφει:

> *Το Διαδίκτυο όπως το γνωρίζουμε σήμερα, το οποίο φορτώνει σε ένα παράθυρο του προγράμματος περιήγησης, στατικές δομές πληροφοριών, είναι μόνο ένα έμβρυο από το Διαδίκτυο που πρόκειται να έρθει. Τα πρώτα στοιχεία του Web 2.0 έχουν αρχίσει να εμφανίζονται, και εμείς μόλις τώρα μπορούμε να δούμε πώς αυτό έμβρυο πρόκειται να εξελιχθεί. Ο Παγκόσμιος Ιστός θα μετατραπεί σε έναν μηχανισμός μεταφοράς, όπως ο αέρας, μέσω του οποίου θα υπάρχει διαδραστικότητα. [...] Θα εμφανίζεται στην οθόνη του υπολογιστή, [...] στην οθόνη της τηλεόρασής, [...] στο ταμπλό του αυτοκινήτου, [...] το κινητό σας [...] ίσως ακόμη και φούρνο μικροκυμάτων σας.[7]*

Ο όρος Web 2.0 δεν θα επανέλθει μέχρι το 2002. Οι συγγραφείς επικεντρώνονται στις έννοιες που συνδέονται σήμερα με τον όρο, μέχρι που, όπως Scott Dietzen το θέτει, «το Διαδίκτυο γίνεται καθολική, προτυποποιημένη πλατφόρμα». Ο John Robb έγραψε: «Τι είναι το Web 2.0; Είναι ένα σύστημα που διαφοροποιείται με το παλιό συγκεντρωτικό μοντέλο του Διαδικτύου και μετακινεί την δύναμη του στην επιφάνεια εργασίας του χρήστη.»

Το 2003, ο όρος άρχισε την άνοδό τους σε δημοτικότητα όταν τα O'Reilly Media και MediaLive, φιλοξένησαν το πρώτο συνέδριο Web 2.0. Στην ομιλία τους, ο John Battelle και ο Tim O'Reilly περιγράφουν τον ορισμό «Διαδίκτυο ως πλατφόρμα», όπου οι εφαρμογές λογισμικού βασίζονται στο Διαδίκτυο και όχι στην επιφάνεια εργασίας. Προέβαλαν το επιχείρημα ότι οι δραστηριότητες των χρηστών δημιουργίας περιεχομένου (με τη μορφή ιδεών, κείμενο, βίντεο ή εικόνες) θα μπορούσε να "αξιοποιηθεί" για τη δημιουργία αξίας.

Όσον αφορά το μη επαγγελματικό κοινό, ο όρος Web 2.0 σε μεγάλο βαθμό υπερασπίστηκε από bloggers, με αποκορύφωμα το περιοδικό TIME "Πρόσωπο της χρονίας 2006". Αυτό συνέβη γιατί το περιοδικό, επέλεξε σαν πρόσωπο της χρονιάς,

τους χρήστες που συμμετείχαν στη δημιουργία περιεχομένου για τα κοινωνικά δίκτυα, blogs και τα wikis. Σχετικά με την επιλογή, ο Λεβ Γκρόσμαν, εξηγεί:

Είναι μια ιστορία για την κοινότητα και τη συνεργασία σε τέτοια κλίμακα που δεν είχαμε γνωρίσει ποτέ πριν. Έχει να κάνει με την παγκόσμια συλλογή της γνώσης της Wikipedia και του δίκτυο των εκατομμυρίων καναλιών των ανθρώπων του YouTube και του MySpace. Έχει να κάνει με τις πολλές δύναμης από τους λίγους και της βοήθειας που δίνει ο ένας στον άλλο για το τίποτα και πώς αυτό δεν θα αλλάξει μόνο τον κόσμο, αλλά θα μεταβάλλει επίσης και τον τρόπο που ο κόσμος αλλάζει [8].

Από τότε, το Web 2.0 έχει βρει μια θέση στο λεξικό. Το 2009 η λέξη Web 2.0 δηλώθηκε ως η λέξη ένα εκατομμύριο στην αγγλική γλώσσα [9]. Το ακριβές νόημα του όρου, ωστόσο, παραμένει ανοιχτό προς αντιπαράθεση, και μερικοί ειδικοί, συμπεριλαμβανομένου και του Τιμ Μπέρνερς Λι, έχουν αμφισβητήσει εάν ο όρος έχει κάποιο πραγματικό νόημα. Μεταξύ άλλων έχει κατηγορηθεί ότι αποτελεί εφεύρεση του μάρκετινγκ. Αναφέρεται σαν επιχείρημα πως οι τεχνολογίες που είναι σε χρήση στο Διαδίκτυο αναβαθμίζονται συνεχώς και η προηγούμενη μέρα πριν το Web 2.0 δεν απείχε πολύ. Επίσης με την ίδια λογική, μερικούς μήνες μετά την χρήση ενός υποτιθέμενου Web 2.0, το Web 2.1 θα εμφανιζόταν. Είναι όμως αδύνατο να εκφραστεί η τεχνολογική ιδιότητα του διαδικτύου ακριβώς με έναν αριθμό μιας και το αμάλγαμα των τεχνολογιών που είναι αυτή τη στιγμή σε χρήση στο διαδίκτυο είναι αχανώς πολυσύνθετο. Αυτό που μπορεί να ειπωθεί ότι άλλαξε είναι η ευκολία χρήσης, κάτι αναμενόμενο με την πάροδο του χρόνου.

1.2. Κοινωνικά Δίκτυα

Τα κοινωνικά δίκτυα αποτελούν μια προσέγγιση για τη διερεύνηση των κοινωνικών δομών. Αυτός είναι ο λόγος που συχνά αναφέρονται ως δομική δίκτυα. Οι αναλυτές των κοινωνικών δικτύων ενδιαφέρονται για το πώς το άτομο είναι ενσωματωμένο μέσα σε μια δομή και τον τρόπο με τον οποίο μια δομή προκύπτει από τις σχέσεις μεταξύ των επιμέρους τμημάτων της. Κοινωνικό δίκτυο είναι μια ομάδα από συνεργαζόμενα άτομα, ή / και ανταγωνιστικά άτομα ή ακόμα και επιχειρήσεις που σχετίζονται μεταξύ τους [2]. Ένα κοινωνικό δίκτυο μπορεί να παρουσιάζεται ως μια γραφική παράσταση, ή ένα πολύ-γράφημα. Κάθε συμμετέχοντας ονομάζεται

παράγοντας (actor) και απεικονίζεται ως κόμβος στο γράφημα. Οι σχέσεις μεταξύ των παραγόντων απεικονίζονται ως σύνδεσμος, είτε κατευθυνόμενος, είτε μη κατευθυνόμενος, μεταξύ των αντίστοιχων κόμβων. Οι παράγοντες μπορεί να είναι άτομα, οργανώσεις ή οποιοδήποτε σύνολο από συσχετιζόμενες έννοιες. Ως εκ τούτου, η ανάλυση κοινωνικών δικτύων μπορεί να χρησιμοποιηθεί σε διαφορετικά επίπεδα, που κυμαίνονται από άτομα, οικογένειες και μικρές ομάδες, έως μεγάλους οργανισμούς, πολιτικές παρατάξεις ακόμα και έθνη [2].

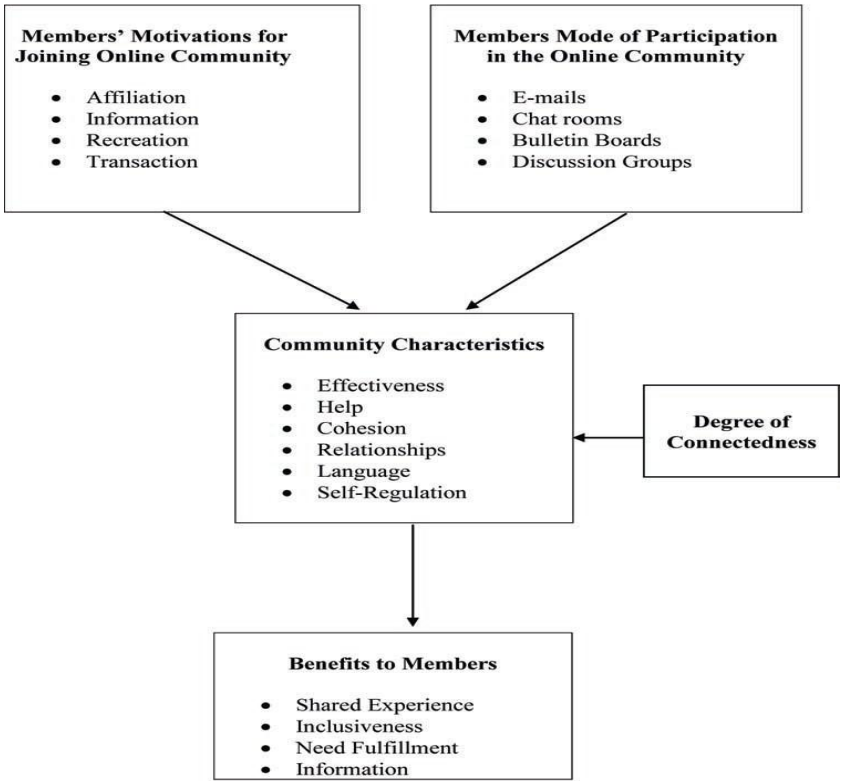
Με μια ιστορία πάνω από 70 χρόνια, η ανάλυση των κοινωνικών δικτύων, ως διεπιστημονική τεχνική, αναπτύχθηκε κάτω από πολλές επιρροές, οι οποίες προέρχονται από διάφορους τομείς, όπως η κοινωνιολογία, τα μαθηματικά και η επιστήμη των υπολογιστών, οι οποίοι αποκτούν ολοένα και μεγαλύτερη σημασία σε πολλούς κλάδους, συμπεριλαμβανομένης της κοινωνιολογίας, της οικονομίας, της επιστήμης της επικοινωνίας, και της ψυχολογίας.

Παραδοσιακά, η κύρια κοινωνική έρευνα επικεντρώνεται αποκλειστικά στη συμπεριφορά των ατόμων. Η προσέγγιση αυτή αγνοεί το κοινωνικό μέρος ή τη δομή της ανθρώπινης συμπεριφοράς, δηλαδή, το τμήμα που ασχολείται με τους τρόπους που τα άτομα αλληλεπιδρούν μεταξύ τους και την επιρροή που έχουν στο μεταξύ τους [3].

Οι σχέσεις που οι αναλυτές κοινωνικών δικτύων εξετάζουν, είναι συνήθως αυτές που συνδέουν μεμονωμένα ανθρώπινα όντα, αφού αυτοί οι κοινωνικοί επιστήμονες πιστεύουν ότι εκτός από τα ατομικά χαρακτηριστικά, τις σχεσιακές συνδέσεις ή την κοινωνική δομή, είναι αναγκαίο και απαραίτητο να κατανοήσουμε πλήρως τα κοινωνικά φαινόμενα [3].

Συγκεκριμένα, ο Wetherell περιγράφει την ανάλυση των κοινωνικών δικτύων ως εξής [4]:

Η ανάλυση των κοινωνικών δικτύων (1) αντιλαμβάνεται εννοιολογικά την κοινωνική δομή ως δίκτυο με τους δεσμούς που συνδέουν τα μέλη και τη διοχέτευση των πόρων, (2) εστιάζεται στα χαρακτηριστικά των δεσμών και όχι στα χαρακτηριστικά των επιμέρους μελών της, και (3) αντιλαμβάνεται της κοινότητες ως «προσωπική κοινότητες», που είναι, όπως τα δίκτυα επιμέρους σχέσεων που οι άνθρωποι διατηρούν κατά την διάρκεια της καθημερινής τους ζωής.



Εικόνα 1.2 Σχεδιάγραμμα σχέσεων και κινήτρων για την συμμετοχή κινήτρων [10].

1.2.1. Ανάλυση κοινωνικών δικτύων: Σύντομη ιστορική ανασκόπηση

Στα τέλη του 1800, οι Emile Durkheim και Ferdinand Tonnies, έγιναν πρόδρομοι της θεωρίας του κοινωνικού διαδικτύου. Ο Tonnies υποστήριξε ότι οι κοινωνικές ομάδες μπορεί να υπάρχουν ως προσωπικοί και άμεσοι κοινωνικοί δεσμοί που συνδέουν τα άτομα που συμμερίζονται τις ίδιες αξίες και πεποιθήσεις. Ο Durkheim έδωσε μια μη ατομικιστική εξήγηση των κοινωνικών γεγονότων υποστηρίζοντας ότι κοινωνικά φαινόμενα προκύπτουν κατά την αλληλεπίδραση των ατόμων, τα οποία αποτελούν μια πραγματικότητα που δεν μπορεί πλέον να λαμβάνεται υπόψη μόνο από τις ιδιότητες των επιμέρους φορέων.

Ο Georg Simmel, γράφοντας στα τέλη του εικοστού αιώνα, ήταν ο πρώτος μελετητής σκεπτόμενος άμεσα σε επίπεδο κοινωνικού δικτύου. Το Κοινωνικό δίκτυο, δημιουργήθηκε ως μια ξεχωριστή ακαδημαϊκή έννοια το 1920-1930 στον τομέα της ανθρωπολογίας στη Βρετανία. Ο ανθρωπολόγος Ρότζερ Μπράουν ήταν ο πρώτος ερευνητής που χρησιμοποίησε τον όρο κοινωνικό δίκτυο[5].

Πρωτοστάτης της ανάλυσης των κοινωνικών δικτύων ήταν ο κοινωνικός ψυχολόγος Jacob Levy Moreno στην δεκαετία του 1930. Ο Moreno πρωτοστάτησε στην συστηματική καταγραφή και ανάλυση της κοινωνικής αλληλεπίδρασης σε μικρές ομάδες και ομάδες εργασίας. Το 1934 στο βιβλίο του, χρησιμοποίησε τον όρο "δίκτυο" με την έννοια που χρησιμοποιείται μέχρι και σήμερα. Το 1940-1960 θεωρείται ως ο Μεσαίωνας στην ιστορία της ανάπτυξης της ανάλυσης κοινωνικών δικτύων. Σε αυτή την περίοδο, δεν υπήρξε καμία ευρέος αναγνωρισμένη προσέγγιση για την δικτυακή κοινωνική έρευνα .

Στη δεκαετία του 1960 έως το 1970, ένας συνεχώς αυξανόμενος αριθμός επιστημόνων εργάστηκε για να συνδυάσει την διαφορετικά κομμάτια των κοινωνικών δικτύων. Μια μεγάλη ομάδα από το Πανεπιστήμιο του Χάρβαρντ υπό την επίβλεψη του Harrison White δημοσίευσε μια σημαντική θεωρία ώστε όλοι οι κοινωνικοί επιστήμονες, από όλων τον κόσμο και ανεξάρτητα από τον επιμέρους τομέα τους, δεν μπορούσαν πλέον να αγνοήσουν την ιδέα. Μέχρι το τέλος της δεκαετίας του 1970, η ανάλυση των κοινωνικών δικτύων άρχισε να αναγνωρίζεται παγκοσμίως μεταξύ των κοινωνικών επιστημόνων.

Ο πίνακας 1.1 παραθέτει ορισμένους ιδρυτές και τους σημαντικότερους ερευνητές της ανάλυσης κοινωνικών δικτύων από το 1940 μέχρι την δεκαετία του 1970.

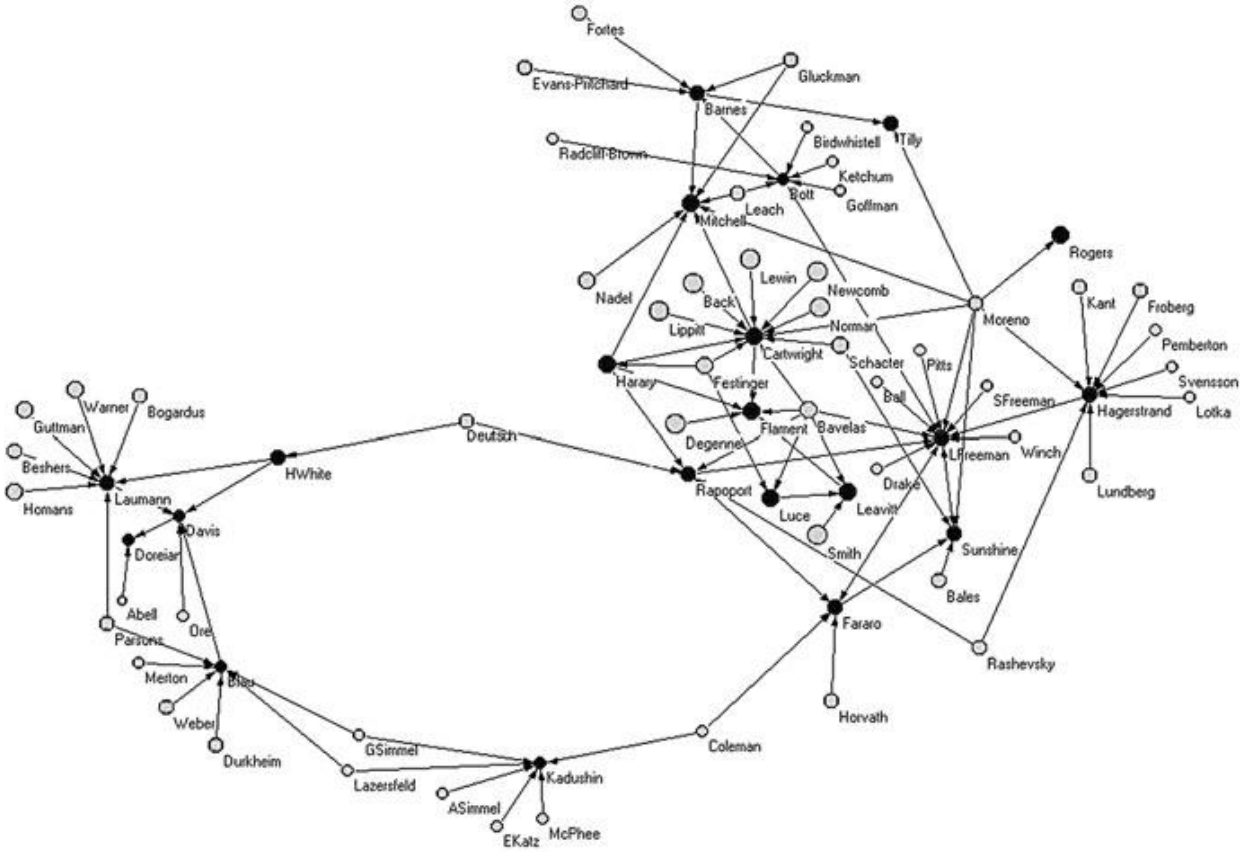
Time period	Group leader	Known researchers	University/institution
1930s–1940s	Kurt Lewin, John French, Alex Bavelas	Dorwin Cartwright, Leon Festinger, Duncan Luce	University of Iowa, MIT, University of Michigan
Mid 1940s	Charles P. Loomis	Leo Katz, Charles Proctor, T. N. Bhargava	Michigan State College
Late 1940s	Lévi–Strauss	–	University of Chicago
Early 1950s	Torsten Hägerstrand	–	Lund University
Early 1950s	Nicolas Rashevsky	Walter Pitts, Herbert D. Landahl, Hyman G. Landau, Anatol Rapoport	University of Chicago
Mid 1950s	Paul Lazarsfeld, Robert K. Merton	James S. Coleman, Elihu Katz, Herbert Menzel, Peter Blau, Charles Kadushin	Columbia University
Mid 1950s	Everett M. Rogers	George Barnett, James Danowski, Richard Farace, Peter Monge, Nan Lin, William Richards, Ronald Rice	Iowa State University, Michigan State University
Mid 1950s	Alfred Reginald Radcliffe–Brown, Max Gluckman	John Barnes, John Barnes, J.Clyde Mitchell, Elizabeth Bott, Sigfried Nadel	Manchester University, London School of Economics
Late 1950s	Karl Wolfgang Deutsch, Ithiel de Sola Pool	Fred Kochen	MIT
Late 1950s	Linton C. Freeman, Morris H. Sunshine	Thomas Fararo, Sue Freeman	Syracuse University
Early 1960s	Claude Flament	–	Paris Sorbonne University

Mid 1960s	Edward O. Laumann	Stephen Berkowitz, Ronald Burt, Joseph Galaskiewicz, Alden Klovdahl, David Knoke, Peter Marsden, Martina Morris, David Prensky, Philip Schumm	University of Michigan
Late 1960s	Peter Blau, James A. Davis	–	University of Chicago
Late 1960s	Robert Mokken	Jac Anthonisse, Frans Stokman	
1960s–1970s	Harrison Colyer White	Peter Bearman, Paul Bernard, Phillip Bonacich, Ronald L. Breiger, Kathleen M. Carley, Ivan Chase, Bonnie Erickson, Claude S. Fischer, Mark Granovetter, Joel Levine, Siegwart M. Lindenberg, Barry Wellman, Christopher Winship	Harvard University

Εικόνα 1.3 Ιδρυτές και σημαντικοί ερευνητές της Ανάλυσης Κοινωνικών Δικτύων (1940-1970).

Στη δεκαετία του 1980, ένας αριθμός κοινωνιολόγων άρχισε να χρησιμοποιεί την ανάλυση κοινωνικών δικτύων για να εξετάσει κοινωνικά και οικονομικά φαινόμενα. Στις αρχές της δεκαετίας του 1980, ο Mark Granovetter διατύπωσε τη θεωρία της εμπορίας του διαδικτύου σε γνωστό άρθρο του "Where Do Markets Come From?" .Στα μέσα της δεκαετίας του 1980, ο Granovetter, υποστήριξε ότι η λειτουργία της οικονομίας είναι ενσωματωμένη στην κοινωνική δομή, ωστόσο, ο πυρήνας της κοινωνικής δομής βρίσκεται σε ατομικά κοινωνικά δίκτυα. Από το 1990 μέχρι και σήμερα, η ανάλυση κοινωνικών δικτύων σταδιακά αντλεί την προσοχή ακαδημαϊκών, από το χώρο της κοινωνιολογίας, της πολιτικής, της οικονομίας, της επικοινωνίας και άλλων κλάδων.

Παρακάτω στο σχήμα 1.1 περιγράφονται οι επιρροές διαφόρων επιστημών στους παραπάνω ιδρυτές και ερευνητές.



Σχήμα 1.1 Επιρροές μεταξύ των ερευνητών σχετικά με την ανάλυση κοινωνικών δικτύων.

1.2.2. Χρησιμότητα Ανάλυσης Κοινωνικών Δικτύων

Η ιεράρχηση και η ταξινόμηση συγκεκριμένων αρχών (authorities), είναι δυο κρίσιμα συστατικά για ένα ευρύ φάσμα κοινωνικών διαδικτυακών εφαρμογών και υπηρεσιών, όπως η θεματική εστίαση ή η σύσταση επαφών. Διαδικτυακές κοινότητες, όπως το Facebook και το Twitter, παρέχουν πολύ περιορισμένα στατιστικά στοιχεία σχετικά με τις σχέσεις μεταξύ των χρηστών τους, όπως ο αριθμός των διαδικτυακών φίλων ή κάποια κοινές ομάδες ενδιαφέροντος [1]. Αν και αυτά τα κριτήρια, αντικατοπτρίζουν κατάλληλα την δημοτικότητα του συνόλου των χρηστών στην διαδικτυακή κοινότητα, η ερμηνεία τους σε ένα πιο εστιασμένο πλαίσιο γίνεται πολύ δύσκολη. Για να γίνουμε πιο ακριβείς θα παρουσιάσουμε ένα ρεαλιστικό παράδειγμα

για αυτό το πρόβλημα. Δύο πολύ δημοφιλείς χρήστες της διαδικτυακής κοινότητας του Twitter είναι ο Tim Berners Lee (με όνομα χρήστη: timberners lee) και είκοσι χιλιάδες (20.000) οπαδοί (followers) και η Paris Hilton (με όνομα χρήστη parishilton) και 1,7 εκατομμύρια οπαδούς. Στις αρχές του 2010, και οι δύο χρήστες σχολίασαν την έναρξη της νέας τεχνολογίας iPad της Apple - ταυτόχρονα - ακολουθώντας αναρτήσεις (tweets), όπως φαίνεται στο σχήμα 1.2.

timberners_lee	*parishilton*
Following: 59	Following: 272
Followers: 20.692	Followers: 1.709.116
RT @janl: Apple:	I Love my new I-Pad.
"Preparing your web content for iPad:	So much fun!
2. Use W3C standard web technologies." #w3c	Technology rocks!
9:46 AM Mar 21st via TweetDeck	about 2 h ago via UberTwitter

Εικόνα 1.4 Αναρτήσεις των χρηστών Timbernerslee και parishilton

Με μια καθαρά αριθμητική οπτική γωνία, θα μπορούσαμε να συμπεράνουμε ότι ο χρήστης parishilton είναι ισχυρότερη αρχή του Twitter, σχετικά με το θέμα «νέες τεχνολογίες» από ότι ο χρήστης timberners Lee.

Ωστόσο, η βαθύτερη ανάλυση των τοποθετήσεων και των δυο χρηστών αποδεικνύει αμέσως το αντίθετο. Ενώ οι αναρτήσεις του timberners Lee, έχουν επικεντρωθεί στις νέες ερευνητικές / τεχνολογικές πτυχές (όπως η Επιστήμη του Διαδικτύου, Semantic Web κτλ.), ο χρήστης parishilton έχει μάλλον επικεντρωθεί σε θέματα όπως «διασημότητες», «τρόπος ζωής» κ.α [1].

Η ανάγκη για καλύτερη και ρητή κατηγοριοποίηση των συνεισφορών σε πληροφορία, οδήγησε σε διάφορους μηχανισμούς, αυτο-οργάνωσης, με την χρήση λεξιλογικών επεκτάσεων. Η πιο γνωστή μορφή κατηγοριοποίησης απλού περιεχόμενου και πληροφορίας είναι οι επισημειώσεις (tagging), δηλαδή, η χρήση σύντομων αποσπασμάτων κειμένου (ή συγκεκριμένους όρους) για το σχολιασμό του περιεχομένου - ευρέως γνωστή από κοινωνικές ιστοσελίδες όπως Flickr, YouTube, Google Books, και άλλα.

1.3 Λογισμικό κοινωνικών δικτύων και δημόσιες βιβλιοθήκες.

Η φιλοσοφία των υπηρεσιών των βιβλιοθηκών, είναι με επίκεντρο τον χρήστη, γεγονός που προϋπάρχει της εμφάνισης του Web 2.0. Ωστόσο, με την έλευση των τεχνολογιών κοινωνικής αλληλεπίδρασης, όπως τα blogs και τα wikis, οι βιβλιοθήκες έχουν δει μια ριζική αλλαγή. Οι βιβλιοθήκες αναγνωρίζουν το γεγονός ότι οι τεχνολογίες του Web 2.0, μπορούν να παρέχουν στους χρήστες τους πολλές υπηρεσίες, καθώς εφαρμόζουν το «κοινωνικό λογισμικό» με καινοτόμους τρόπους.

Οι πιο συχνά χρησιμοποιούμενες τεχνολογίες «κοινωνικού λογισμικού» σε βιβλιοθήκες είναι οι παρακάτω [11]:

- Ιστολόγια (blogs),
- Wikis (συχνά χρησιμοποιούνται μέσα σε βιβλιοθήκες από το προσωπικό των βιβλιοθηκών),
- Ιστοσελίδες κοινωνικών σελιδοδεικτών, όπως το Del.icio.us και το LibraryThing, καθώς και ιστοσελίδες ανταλλαγής φωτογραφιών, που επιτρέπουν στους χρήστες να σχολιάσουν βιβλία και φωτογραφίες,
- Λογισμικό που επιτρέπει την χρήση σχολίων και επισημειώσεων,
- Εργαλεία κοινωνικής δικτύωσης όπως το Facebook και το MySpace (για να δίνεται η δυνατότητα σε συγκεκριμένα σύνολα χρηστών, να ενημερώνονται για νέα προγράμματα και πηγές βιβλιοθηκών),

1.4 Library 2.0

Οι τεχνολογίες «κοινωνικού λογισμικού» που βρήκαν χρήση στα πλαίσια βιβλιοθηκών, δημιούργησαν την ανάγκη ενός νέου όρου, του Βιβλιοθήκη 2.0 (Library 2.0, ή για συντομία απλά L2). Ο όρος Βιβλιοθήκη 2.0 αποδίδεται από τον Michael Casey το 2005 στο blog του "LibraryCrunch" [12]. Οι Casey και Savastinuk (2006) τονίζουν την ανάγκη για αλλαγές με επίκεντρο τον χρήστη, την ενθάρρυνση της συμμετοχής των χρηστών στη δημιουργία των φυσικών και των εικονικών υπηρεσιών μιας βιβλιοθήκης. Ωστόσο, στις βιβλιοθήκες, όπως και σε κάθε άλλο διοικητικό ή πολιτιστικό πλαίσιο, απαιτείται ριζοσπαστική εμπιστοσύνη για να υιοθετηθούν σημαντικές αλλαγές. Στην εποχή, του Web 2.0 και εν συνεπεία του Library 2.0, «ριζοσπαστική εμπιστοσύνη» συνεπάγεται εμπιστοσύνη στο εσωτερικό της Κοινότητας, που επιτρέπει και ενθαρρύνει τους χρήστες να συμμετέχουν στην

ανάπτυξη των υπηρεσιών της βιβλιοθήκης. Οι βιβλιοθηκονόμοι ανησυχούν με την απώλεια των κανόνων και της ακρίβειας, όταν οι χρήστες συνεισφέρουν στο περιεχόμενο, επισημειώσεις, σχόλια ή παρατηρήσεις σε βιβλιοθήκες αρχείων καταλόγου. Παρ' όλα αυτά, οι Abbas και Graham [13], θεωρούν τη συμμετοχή των χρηστών, ως το μέσο για να γεφυρωθεί το χάσμα μεταξύ των χρηστών και των πόρων της βιβλιοθήκης.

Παρακάτω είναι ένα σύντομο χρονοδιάγραμμα-ορόσημα των εκδόσεων «κοινωνικού λογισμικού» για τις βιβλιοθήκες. Το χρονοδιάγραμμα, δεν περιλαμβάνει όλες τις μορφές «κοινωνικού λογισμικού», καθώς ελάχιστα στοιχεία είναι διαθέσιμα για πρώτες χρήσεις των blogs, wikis και άλλων λογισμικών σε βιβλιοθήκες.

Πριν από το 2003 - 2004: Μια από τις πρώτες ιστοσελίδες που επέτρεψε στους χρήστες να συνεισφέρουν συνδέσμους και ιστοσελίδες στα θέματά του, ήταν το Slashdot, το οποίο αναπτύχθηκε το 1997. Άλλες ιστοσελίδες όπως το Del.icio.us και το Furl αναπτύχθηκαν κατά τη διάρκεια του 2003-2004.

Το έτος 2004 χαρακτηρίστηκε, όπως προαναφέραμε, από την πρώτη χρήση του όρου Web 2.0 από τους O'Reilly και Battele.

Το 2005, το LibraryThing, το Flickr, και άλλοι χώροι κοινωνικής χρήσης διατέθηκαν στο διαδίκτυο. Οι χρήστες ενθαρρύνονται να οργανώνουν και να επισημειώνουν τις δικές τους συλλογές από βιβλία, φωτογραφίες, και σελιδοδείκτες, καθώς και να επισημειώσουν και να σχολιάσουν τις συλλογές άλλων χρηστών.

Το Steve.museum.org tagging project, άρχισε να εξερευνά τις επισημειώσεις των χρηστών σε διαδικτυακές συλλογές μουσείων. Το έργο, αυτό, απεικονίζει μια από τις πρώτες μεγάλης κλίμακας ερευνητικές προσπάθειες για να μάθουμε περισσότερα σχετικά με τη σκοπιμότητα και την αξία των προσφορών πληροφορίας των χρηστών σε ιστοσελίδες μουσείων.

Το 2006, το έργο PennTags στο Πανεπιστήμιο της Πενσυλβάνια ξεκίνησε. Το πανεπιστήμιο της Πενσυλβανία, κάλεσε χρήστες να κάνουν επισημειώσεις και να κρατήσουν σχόλια σχετικά με τους πόρους της βιβλιοθήκης του πανεπιστημίου, με την βοήθεια ενός λογισμικού που ανέπτυξε η ίδια η πανεπιστημιακή κοινότητα.

Το 2007, πολλές βιβλιοθήκες αρχίσουν να ζητούν την συμβολή των χρηστών στις ηλεκτρονικές βιβλιοθήκες τους. Για παράδειγμα, τον Μάιο του 2007 ξεκίνησε η λειτουργία του LibraryThing για βιβλιοθήκες. Οι βιβλιοθήκες μπορούσαν να συνδεθούν με το LibraryThing με την χρήση εξειδικευμένου λογισμικού, που

επέτρεπε στους χρήστες να δημιουργούν επισημειώσεις για τους πόρους της βιβλιοθήκης. Η βιβλιοθήκη του Danbury, στο Κονέκτικατ, ήταν η πρώτη βιβλιοθήκη που συνδέθηκε με το LibraryThing. Τον Φεβρουάριο του 2008, το LibraryThing έφτασε τις 40 βιβλιοθήκες μέλη.

Παρακάτω παρουσιάζουμε μια έρευνα που έγινε από το OCLC Online Computer Library Center, το 2007, σχετικά με το «κοινωνικό λογισμικό» που μπορεί να χρησιμοποιηθεί από βιβλιοθήκες. Η έρευνα έγινε σε παγκόσμια κλίμακα[2]. Τα σημαντικότερα αποτελέσματα της έρευνας φαίνονται παρακάτω [14]:

- Το κοινό στην πραγματικότητα δεν προβλέπει κανένα ρόλο για τις βιβλιοθήκες στα κοινωνικά δίκτυα,
- Το 6% του κοινού, απάντησε ότι είναι πολύ πιθανό να περιέγραφαν την συλλογή τους σε ένα κοινωνικό δίκτυο βιβλιοθηκών,
- Το 5% του κοινού, θα μοιραζόταν τα βίντεο και τις φωτογραφίες του σε κοινωνικά δίκτυα βιβλιοθηκών,
- Το 13% του κοινού, θα ήθελε να ενημερώνεται όταν νέα αντικείμενα προστίθενται στην βιβλιοθήκη,
- Το 8% του κοινού, θα μοιραζόταν τις ιδέες του με το προσωπικό των βιβλιοθηκών για την βελτίωση των υπηρεσιών της,
- Το 6% του κοινού, θα έπαιρνε μέρος σε δημόσιες συζητήσεις για να βρει άλλους ανθρώπους για γνωριμία, με τα ίδια ενδιαφέρονται,
- Το 14% των αμερικανικών κοινωνικών συμβουλίων, θα ήθελε να δημιουργήσει μια ιστοσελίδα κοινωνικής δικτύωσης για την κοινότητά του.

Στην ίδια έρευνα, έγινε και η ανάλυση 70 ιστοσελίδων βιβλιοθηκών, ακαδημαικών ιδρυμάτων και μουσείων. Η έρευνα αφορούσε την δομή και το περιεχόμενο των ιστοσελίδων αυτών. Τα αποτελέσματα φαίνεται παρακάτω [14] :

- Το 70% των ιστοσελίδων που εξετάστηκαν περιείχε κάποιον μηχανισμό επισημειώσεων,
- Το 81% των ιστοσελίδων βιβλιοθηκών είχαν τον παραπάνω μηχανισμό επισημειώσεων,

[2] Για περισσότερες πληροφορίες, [14].

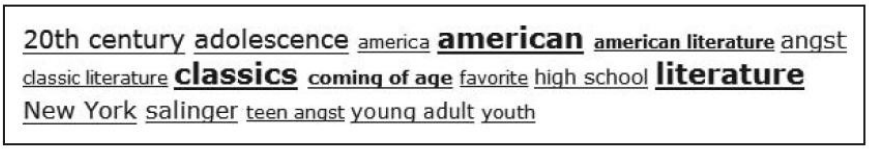
- Το 76% των ακαδημαικών ιστοσελίδων είχαν τέτοιο μηχανισμό,
- Το 37.5% των ιστοσελίδων των μουσείων είχαν τον μηχανισμό επισήμανσης των περιεχομένων τους.

Στο σημείο αυτό θα παρουσιάσουμε κάποιους όρους και έννοιες, που θα μας βοηθήσουν να καταλάβουμε καλύτερα τα επόμενα κεφάλαια της έρευνας. Όροι όπως επισημείωση, αρχείο κανόνων και ταξονομία, εξετάζονται παρακάτω.

1.5 Επισημειώσεις (Tagging)

Περπατώντας στη δημόσια βιβλιοθήκη στο Danbury, στο Connecticut, θα βρείτε το συνηθισμένα ράφια με στοιβαγμένα βιβλία, οργανωμένα σε τακτοποιημένες σειρές. Μυθοπλαστικά έργα ομαδοποιούνται κατά αλφαβητική σειρά με βάση το επώνυμο του συγγραφέα. Οι τίτλοι πεζού λόγου τοποθετούνται στην κατάλληλη θέση τους σύμφωνα με την κατηγορία τους, κατά Dewey Decimal, όπως ακριβώς και χιλιάδες άλλες βιβλιοθήκες ανά τον κόσμο.

Ωστόσο, με μια επίσκεψη στον ηλεκτρονικό κατάλογο της Βιβλιοθήκης Danbury (http://danburylibrary.org), θα βρείτε κάτι που μάλλον, βρίσκεται σε αντίθεση με μια τυπική βιβλιοθήκη. Μια αναζήτηση για το βιβλίο «The Catcher in the Rye», εμφανίζει όχι απλά έναν αριθμό για ταξιθέτησης του βιβλίου για να το παραλάβουμε, αλλά επίσης και μια λίστα σχετικών βιβλίων και ετικετών (λέξεων-κλειδιά), όπως «Literature» «America», «Youth» και «New York», που περιγράφουν το κλασσικό μυθιστόρημα του JD Salinger (Εικόνα 1.5) [16].



Εικόνα 1.5 Επισημειώσεις για τον βιβλίο "The Catcher in the Rye".

Αν επιλέξετε στην ετικέτα "Youth", θα βρείτε μια λίστα τίτλων που έχουν και εκείνοι επισημανθεί με την σειρά τους με την ίδια ετικέτα. Κάθε ετικέτα είναι το εφαλτήριο για ένα νέο σύνολο βιβλίων και ιδεών. Αυτές οι ετικέτες δεν είναι το έργο

των επιμελών βιβλιοθηκονόμων της βιβλιοθήκης του Danbury, αλλά προέρχονται από το LibraryThing (σχήμα 1.2).



Σχήμα 1.2 LibraryThing - The Catcher in the Rye, J.D. Salinger

Κάθε φορά που προσθέτει κάποιος χρήστης μια επισημείωση σε κάποιο τίτλο, πραγματοποιεί πέντε πράγματα ταυτόχρονα [16]:

1. Αρχικά, δημιουργεί έναν τρόπο για να περιηγηθεί σε μια συλλογή τίτλων που ο ίδιος δημιουργεί, σύμφωνα με την γλώσσα του και τις εκφράσεις του,
2. Επίσης, για κάθε επισημείωση που προστίθεται, δημιουργείται και ένας ακόμα τρόπος για να βρεθεί ο συγκεκριμένος τίτλος. Την επόμενη φορά που θα αναζητήσει τον τίτλο ο χρήστης, δεν χρειάζεται να θυμάται τίποτα παραπάνω, παρά την επισημείωση που του πρόσθεσε,
3. Ακόμα, η επισημείωση που προστέθηκε, μπαίνει σε ένα σύνολο επισημειώσεων για τον συγκεκριμένο τίτλο, με επισημειώσεις και από άλλους χρήστες και βοηθά στην δημιουργία μιας γέφυρας μεταξύ της γνώσης του χρήστη και της γνώσης ολόκληρης της κοινότητας,
4. Επιπροσθέτως, οι επισημειώσεις των χρηστών μπορεί να δημιουργήσουν μια διασύνδεση μεταξύ του συγκεκριμένου τίτλου, με

άλλους τίτλους που ο χρήστης δεν γνωρίζει και μπορεί να τον ενδιαφέρουν,

5. Τέλος, οι επισημειώσεις των χρηστών, είναι μικρά άγκιστρα που μπορούν να χρησιμοποιηθούν για τη συγκέντρωση πληροφοριών από άλλους διαδικτυακούς τόπους που χρησιμοποιούν την μέθοδο αυτή. Περιεχόμενο από το LibraryThing, το Flickr, Del.icio.us, και άλλες Web 2.0 ιστοσελίδες, μπορούν να συγκεντρώνεται χρησιμοποιώντας την ετικέτα "Social tagging".



Εικόνα 1.6 Σε ένα σύστημα επισημειώσεων, οι χρήστες προσθέτους επισημειώσεις σε πήγες.

Επειδή η δημιουργία επισημειώσεων είναι ελεύθερη, οι επισημειώσεις μπορεί να είναι κάθε τύπου. Μπορεί να είναι περιγραφή του αντικειμένου του πόρου, μιας τοποθεσίας που περιγράφεται, μια προβλεπόμενη χρήση του πόρου, ή ακόμα και μια υπενθύμιση. Μπορεί να είναι ατομική λέξη όπως "αστείο", ή φράση όπως "δώρο από τη μαμά". Διαφορετικοί άνθρωποι έχουν διαφορετικά μοτίβα χρήσης ετικετών, κάποιοι άνθρωποι είναι πιο εκφραστικοί ενώ άλλοι προτιμούν πιο απλές και περιγραφικές επισημειώσεις. Τα συστήματα τοποθέτησης επισημειώσεων επιτρέπουν και ενθαρρύνουν σε πολλές περιπτώσεις, τις διαφορές αυτές.

Οι επισημειώσεις είναι ουσιαστικά μετά-δεδομένα (metadata) για τον συγκεκριμένο πόρο. Τα μετά-δεδομένα, σύμφωνα με τον NIST[3], είναι «δομημένες

[3] Το Εθνικό Ινστιτούτο Προτύπων και Τεχνολογίας (NIST), είναι ένα Αμερικάνικο ινστιτούτο προτυποποίησης. Επίσημη αποστολή του Ινστιτούτου είναι, η προώθηση της καινοτομίας στις ΗΠΑ και της ανταγωνιστικότητα της βιομηχανίας, ταυτόχρονα με την προώθηση της επιστήμης των μετρήσεων, των προτύπων και της τεχνολογίας με τρόπους που ενισχύουν την οικονομική ασφάλεια και βελτιώνουν την ποιότητα της ζωής μας [17].

πληροφορίες που περιγράφουν, εξηγούν, εντοπίζουν, ή άλλως καθιστούν ευκολότερη για την ανάκτηση, τη χρήση, ή την διαχείριση μια πηγή πληροφοριών». Με άλλα λόγια, οι επισημειώσεις είναι πληροφορίες για τις πληροφορίες.

1.5.1 Τομείς δημιουργίας επισημειώσεων.

Οι επισημειώσεις βρίσκονται στο σημείο τομής τριών κατηγοριών [16] (εικόνα 1.7) :

- Αρχιτεκτονική της Πληροφορίας,
- Κοινωνικό λογισμικό,
- Διαχείριση προσωπικών πληροφοριών (ΔΠΠ).

Αρχιτεκτονική της Πληροφορίας

Το Ινστιτούτο Αρχιτεκτονικής Πληροφοριών ορίζει την αρχιτεκτονική πληροφοριών ως «τον δομικό σχεδιασμό των κοινόχρηστων περιβαλλόντων πληροφοριών» και «η τέχνη και η επιστήμη της οργάνωσης και της υποστήριξης της χρηστικότητας τους Διαδικτύου». Οι ερευνητές σε αυτόν τον τομέα, έχουν επικεντρωθεί σχετικά με τη χρήση ελεγχόμενων λεξιλογίων, συστήματα αναζήτησης και περιήγησης, και συστήματα πλοήγησης για την επίτευξη αυτής της επιτυχίας [16].

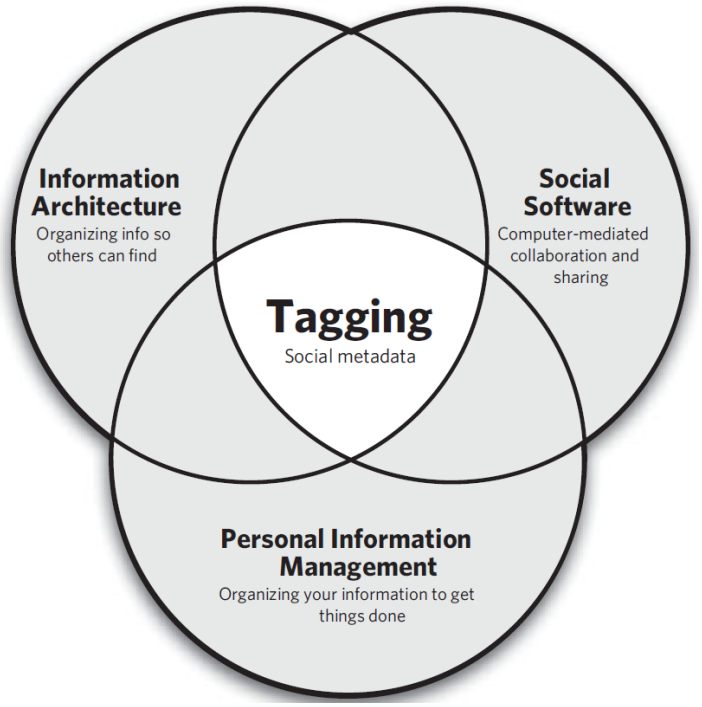
Κοινωνικό λογισμικό

Μπορούμε να σκεφτούμε το κοινωνικό λογισμικό, ως εφαρμογές που οι άνθρωποι χρησιμοποιούν για να επικοινωνούν, να συνεργάζονται, και να μοιράζονται πράγματα διαδικτυακά.

Διαχείριση προσωπικών πληροφοριών

Η διαχείριση προσωπικών πληροφοριών, αναφέρεται τόσο στην πρακτική, όσο και στην μελέτη των δραστηριοτήτων που επιτελούν οι άνθρωποι, ώστε να αποκτούν, να οργανώνουν, να διατηρούν, να ανακτούν και να χρησιμοποιούν

πληροφοριακά στοιχεία, όπως έγγραφα (έντυπες και ψηφιακές), ιστοσελίδες, και μηνύματα ηλεκτρονικού ταχυδρομείου.



Εικόνα 1.7. Οι επισημειώσεις βρίσκονται στο σημείο τομής τριών κατηγοριών.

1.6 Ταξονομίες και ελεγχόμενα λεξιλόγια

Οι ταξονομίες και τα ελεγχόμενα λεξιλόγια είναι δύο ειδών συστήματα ταξινόμησης που καθορίζουν τις σχέσεις μεταξύ των όρων. Αυτές οι σχέσεις μπορεί να είναι σημασιολογικές, όπως η θέσπιση ότι «μαθηματικά» και «αριθμητική» είναι συνώνυμα, ή μπορεί να είναι εννοιολογικές, όπως η σχέση μεταξύ φιλοσοφίας και επιστημολογίας, ένας κλάδος της φιλοσοφίας που αφορά τη φύση της γνώσης. Οι ταξονομίες και τα ελεγχόμενα λεξιλόγια, θα μας βοηθήσουν να κατανοήσουμε και να πλοηγηθούμε σε έννοιες καθιστώντας την γλώσσα λιγότερο αμφίσημη. Βοηθούν στην σύλληψη της σχέσης, μεταξύ αντικειμένων που παρατηρούνται στον πραγματικό κόσμο.

1.6.1. Ελεγχόμενα λεξιλόγια

Θα προσπαθήσουμε να εξηγήσουμε τον όρο, ελεγχόμενο λεξιλόγιο, δίνοντας ένα παράδειγμα. Ποιόν όρο θα χρησιμοποιούσατε αν θέλατε να περιγράψετε ένα ανθρακούχο ποτό[4];

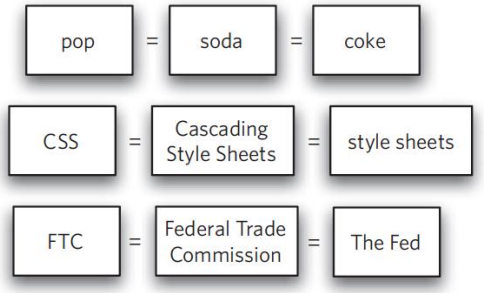
Αν είστε από τις βόρειες Ηνωμένες Πολιτείες ή τον Καναδά, το πιθανότερο είναι ότι θα πείτε "Pop." Αν είστε από το νοτιοδυτικό ή βορειοανατολικό τμήμα των ΗΠΑ, θα πρέπει πιθανώς να πείτε "soda." Αν είστε από τις νότιες ΗΠΑ, ιδιαίτερα τα Κράτη του Κόλπου, τότε μάλλον θα χρησιμοποιήσετε την ονομασία “coke”. Όλες αυτές οι λέξεις αυτές αναφέρονται στην ίδια έννοια, Για οποιαδήποτε συζήτηση σχετικά με τα ανθρακούχα ποτά, οι παραπάνω έννοιες είναι ουσιαστικά συνώνυμες.

Ένα ελεγχόμενο λεξιλόγιο είναι ένα σύστημα διαχείρισης του νοήματος των λέξεων. Αφαιρεί την ασάφεια της γλώσσας και εξασφαλίζει ότι οι άνθρωποι που χρησιμοποιούν τις έννοιες «pop», «soda», ή «coke» αναφέρονται στο ίδιο αντικείμενο, χωρίς να χρειαστεί να εξετάσουν κάποια άλλη ορολογία. Τα δύο πιο κοινά είδη ελεγχόμενα λεξιλόγια είναι τα δαχτυλίδια συνωνύμων (Synonym Rings) και τα αρχεία καθιερωμένων όρων (Authority Files).

[4] Αυτήν την ερώτηση έκανε ο Alan McConchie, στην ιστοσελίδα του http://popvssoda.com.

1.6.1.1. Δαχτυλίδια Συνωνύμων.

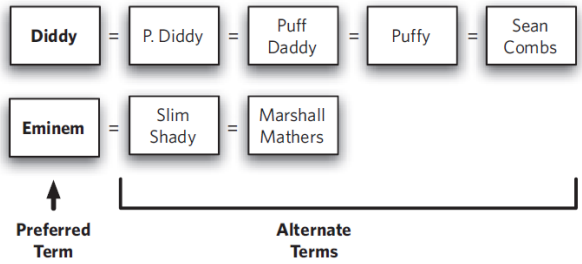
Ένα συνώνυμο δαχτυλίδι, δίνει, σε δύο ή περισσότερες λέξεις ισοδύναμη έννοια. Μπορεί να είναι χρήσιμα για την εξομάλυνση των διαφορών μεταξύ ακρωνυμίων και των πλήρως ανεπτυγμένων τους εννοιών. Η εικόνα 1.7 δείχνει τον τρόπο λειτουργίας των συνώνυμων δαχτυλιδιών [16].



Εικόνα 1.7 Συνώνυμα δαχτυλίδια.

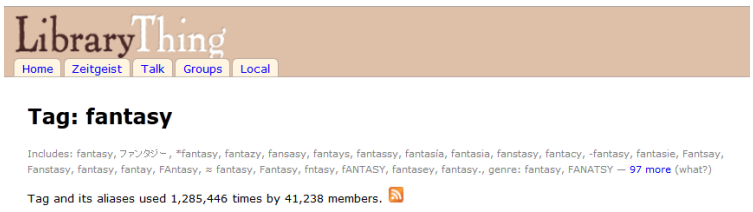
1.6.1.2. Αρχεία Καθιερωμένων όρων.

Ένα αρχείο καθιερωμένων όρων είναι παρόμοιο με ένα δαχτυλίδι συνώνυμων, με την διαφορά πως μία από τις λέξεις προσδιορίζεται ως προτιμητέος όρος. Ο έγκυρος όρος εμφανίζεται στους χρήστες, ενώ οι άλλοι όροι μπορούν να χρησιμοποιηθούν για την παροχή υποδείξεων (παράδειγμα, - για πληροφορίες για "soda", δείτε "pop"-). Τα αρχεία καθιερωμένων όρων είναι χρήσιμα για τη χαρτογράφηση ψευδωνύμων, παραλλαγών και συντμήσεων σε ένα επίσημο όρο [16].



Εικόνα 1.8 Αρχεία καθιερωμένων όρων.

Για άλλη μια φορά θα δούμε το LibraryThing, την καινοτόμα ιστοσελίδα κοινωνικής καταλογογράφησης του διαδικτύου, για ένα παράδειγμα που αναμιγνύει την προσθήκη επισημειώσεων με ελεγχόμενα λεξιλόγια. Η λειτουργία «Συνδυασμός επισημειώσεων» του LibraryThing είναι το χαρακτηριστικό επιτρέπει στους χρήστες να δημιουργήσουν συσχετίσεις μεταξύ επισημειώσεων της ιστοσελίδας, δημιουργώντας ένα ιδιότυπο λεξιλόγιο (εικόνα 1.9).

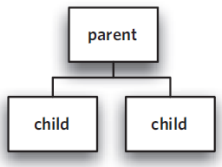


Εικόνα 1.9. Πληροφορίες επισημείωσης στο LibraryThing.

Το LibraryThing χρησιμοποιεί επίσης την δημοτικότητα ενός όρου στην δημιουργία καθιερωμένων όρων. Όταν επισημειώσεις συνδυάζονται, ο πιο δημοφιλής γίνεται ο προτιμητέος όρος.

1.6.2. Ταξονομίες.

Οι ταξονομίες είναι ένα ελεγχόμενο λεξιλόγιο που καθορίζει, μεταξύ όρων, σχέσεις όπως γονέα-παιδιού, ή ευρύτερο-στενότερο, καθώς και άλλες σχέσεις μεταξύ των όρων. Μια ταξονομία είναι συνήθως ιεραρχική (Σχήμα 1.3).



Σχήμα 1.3 Ιεραρχική Ταξονομία.

Η πιο γνωστή ταξονομία είναι ίσως η Dewey Decimal Classification system. Ο Dewey σχεδίασε ένα σύστημα με 10 υψηλού επιπέδου κατηγορίες, ακολουθούμενη

από 10 υποκατηγορίες και αυτές ακολουθούμενες με την σειρά τους από άλλες 10 υποκατηγορίες [16].

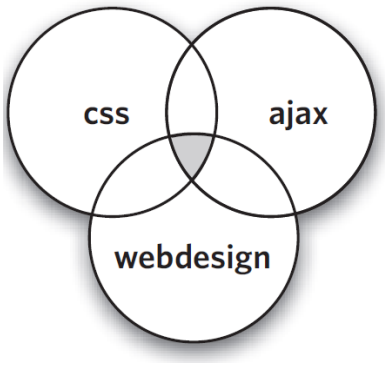
1.7. Όψεις.

Η πολυεδρική ταξινόμηση (facet classification) είναι ένας τρόπος οργάνωσης των πραγμάτων από τις διάφορες ιδιότητές τους. Ένας Ινδός βιβλιοθηκονόμος, ο Ranganathan εφηύρε την πολυεδρική ταξινόμηση στη δεκαετία του 1930. Ο Ranganathan, θεώρησε ότι τα καθαρά ιεραρχική συστήματα ταξινόμησης ήταν περιορισμένα. Αντίθετα, η πολυεδρική ταξινόμηση επιτρέπει στα αντικειμένα να ταξινομούνται σε πολλές διαστάσεις, σύμφωνα με ένα σύνολο χαρακτηριστικών τους, καθώς και να έχουν πολλές διαδρομές εύρεσής τους [16].

1.8. Κοινωνική Ταξονομία.

Η κοινωνική ταξονομία περιγράφει τα κάτω προς τα πάνω συστήματα ταξινόμησης που προκύπτουν από την κοινωνική σήμανση. Όμως τι ακριβώς είδους δομή ταξινόμησης είναι μια κοινωνική ταξονομία; Άλλα συστήματα ταξινόμησης καθορίζουν, όπως προαναφέραμε, τις σχέσεις ευρύτερος, στενότερος, ισοδύναμος, που σχετίζονται μεταξύ των όρων ή των εννοιών που αναφέρονται από τις επισημειώσεις.

Σε μια κοινωνική ταξονομία, οι σχέσεις μεταξύ των ετικετών βασίζονται στην χρήση των όρων και στις συνήθειές τους. Δεν υπάρχουν επίσημες σχέσεις σε μια κοινωνική ταξονομία, εκτός από μια σχέση που βρίσκει τον «βαθμό συγγένειας». Επειδή μια κοινωνική ταξονομία χρησιμοποιεί αλγόριθμους για να βρει τα πρότυπα σήμανσης, δύο επισημάνσεις που δεν έχουν καμία γνωστή σημασιολογική σχέση (όπως το "Blowfish" και "sql") μπορεί να έχουν μια στατιστική σχέση [16].



Σχήμα 1.4 Σελιδοδείκτες με επισημειώσεις, από το Del.icio.us

Στο σημείο αυτό θα δούμε ένα παράδειγμα. Ας υποθέσουμε ότι έχουμε μια συλλογή σελιδοδεικτών από το Del.icio.us. Διακόσια από αυτά είναι με επισήμανση "webdesign," 200 είναι με επισήμανση "Css", και 200 είναι με επισήμανση "ajax". Κάποιοι από τους σελιδοδείκτες περιέχουν έναν, δύο ή και τους τρεις παραπάνω όρους, όπως φαίνεται και στο σχήμα 1.4.

Μπορούμε να πούμε για την σημασιολογική σχέση μεταξύ αυτών των τριών επισημειώσεων, πως με βάση το πόσο πολύ οι κύκλοι επικαλύπτονται, αν οι όροι με κάποιον τρόπο σχετίζονται. Στο συγκεκριμένο παράδειγμα, μπορούμε με ικανοποιητική ασφάλεια να μαντέψουμε ότι ο όρος “Ajax” είναι μια τεχνολογία που έχει κάποια σχέση με το “web design”. Παρόμοιες σχέσεις μπορούμε να μαντέψουμε, φυσικά και για τα υπόλοιπα δυο ζευγάρια που δημιουργούνται.

"You can already see the privacy debate moving to the realm of automated massive data mining. When governments begin to suspect people because of where they were at a certain time, it can get very worrying."

--Jonathan Zittrain

2 Βιβλιογραφική Επισκόπηση

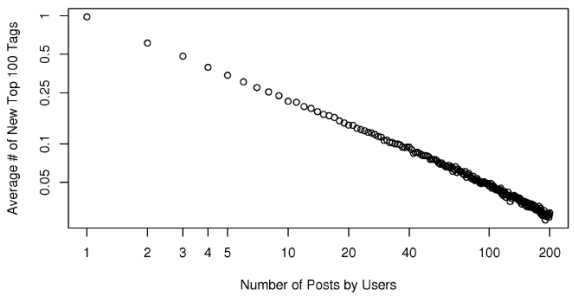
Σημαντική ερευνητική έξαρση παρατηρείται στον χώρο της πληροφορικής, αναφορικά με τις κοινωνικές επισημειώσεις που εμπλουτίζουν αρχεία καθιερωμένων εγγραφών. Πολλές από αυτές τις έρευνες, ασχολούνται με την «πρόβλεψη» συγγενών επισημειώσεων, με σκοπό την διασύνδεση διαφορετικών αρχείων. Η ανάλυση των επισημειώσεων και η αξιοποίησή τους στα συστήματα πληροφόρησης είναι πλούσια και πολυσχιδής. Στη διπλωματική αυτή παρουσιάζουμε 4 ανάλογες ερευνών, που είχαν σημαντικά αποτελέσματα και βοήθησαν στην ανάπτυξη του δικού μας συστήματος εμπλουτισμού αρχείων καθιερωμένων εγγραφών με κοινωνικές επισημειώσεις.

2.1 Πρόβλεψη επισημειώσεων.

Οι Paul Heymann, Daniel Ramage και Hector Garcia-Molina, στην έρευνα τους "Social Tag Prediction", εξετάζουν πόσο αποτελεσματικά μπορεί να είναι διαφορετικά είδους δεδομένα, στην πρόβλεψη επισημειώσεων, σε ένα σύστημα κοινωνικών επισημειώσεων. Πιο συγκεκριμένα, μελετούν κατά πόσο είναι δυνατό να προβλέψουμε επισημειώσεις ιστοσελίδων, βασιζόμενοι στα ίδια τα κείμενα των ιστοσελίδων, καθώς και στις ιστοσελίδες με τις οποίες διασυνδέονται [18].

Για την έρευνά τους αυτή, χρησιμοποίησαν δεδομένα τις διαδικτυακής εφαρμογής del.icio.us[5] καθώς και μια από τις μεγαλύτερες βάσεις δεδομένων επισημειώσεων, την Stanford Tag Crawl 2007.

Ένα σύστημα κοινωνικών επισημειώσεων, αποτελείται από ένα σύνολο χρηστών (users) $u \in U$, από ένα σύνολο επισημειώσεων (tags) $t \in T$ και από ένα σύνολο αντικειμένων (objects) $o \in O$. Για κάθε αντικείμενο o, υπάρχει ένα μεγάλο σύνολο επισημειώσεων (R_n), που λέμε ότι περιγράφουν αρνητικά (negatively describes) το αντικείμενο, καθώς δεν το χαρακτηρίζουν. Ακόμα, υπάρχει ένα μικρότερο σύνολο επισημειώσεων (R_p), οι οποίες χαρακτηρίζουν το αντικείμενο και εν συνεπεία το περιγράφουν θετικά (passively describes). Τέλος, υπάρχει και ένα ακόμα μικρότερο σύνολο επισημειώσεων (R_a), το οποίο χρησιμοποιούν οι χρήστες ενός συστήματος ως είσοδος σε αυτό και το οποίο αποδίδει το νόημα του αντικειμένου (annotates) και επομένως το ταυτοποιεί.



Εικόνα 2.1. Πιθανότητα εισαγωγής των κορυφαίων 100 επισημειώσεων, ανά πλήθος νέων μηνυμάτων.

[5] Το del.icio.us είναι μια διαδικτυακή υπηρεσία κοινωνικών σελιδοδεικτών, για αποθήκευση, ανταλλαγή και αναζήτηση διαδικτυακών σελιδοδεικτών. Η υπηρεσία ιδρύθηκε από τον Joshua Schachter το 2003, ενώ στην συνέχεια αγοράστηκε από την Yahoo! το 2005. Μέχρι το τέλος του 2008, η υπηρεσία είχε δεχθεί περισσότερους από 5.3 εκατομμύρια χρήστες και 180 εκατομμύρια διαφορετικούς σελιδοδείκτες.

Για την ακριβέστερη προσέγγιση των συνόλων R_p και R_n , η έρευνα συγκεντρώνει σαφείς επισημειώσεις, χωρίς προσωπικούς χαρακτηρισμούς, όπως «to read» ή «nice» κ.α. . Για του σκοπούς της έρευνας, ανέπτυξαν κάποιες μετρικές, με τις οποίες βγήκε το συμπέρασμα ότι, ο πρώτος χρήστης που θα εισάγει ένα νέο URL, θα προσθέσει μια (1) επισημείωση από τις εκατό (100) συχνότερα χρησιμοποιημένες, ο δεύτερος χρήστης θα προσθέσει 0,6 επισημειώσεις από τις 100 συχνότερα χρησιμοποιημένες κ.ο.κ. . Ακόμα παρατήρησαν πως για ένα αντικείμενο, η πιθανότητα το $100^{ο}$ μήνυμα, να αναφέρεται σε κάποια από τις 100 συχνότερα χρησιμοποιούμενες επισημειώσεις, είναι μικρότερη του 5% και παραμένει στα ίδια επίπεδο και στα επόμενα μηνύματα (εικόνα 2.1).

Για την πρόβλεψη των επισημειώσεων, ακολουθούνται δύο μέθοδοι. Η πρώτη μέθοδος, αξιοποιεί τα δεδομένα των ιστοσελίδων της βάσης δεδομένων του Stanford, ενώ η δεύτερη μέθοδος αξιοποιεί τις ήδη υπάρχοντες επισημειώσεις ενός αντικειμένου.

Για την πρώτη μέθοδο, αξιοποιήθηκαν τα δεδομένα των σελίδων σχετικά με τις 100 συχνότερα χρησιμοποιούμενες επισημειώσεις (T_{100}). Τα δεδομένα προέκυψαν από τις ιστοσελίδες που περιέχουν το αντικείμενο, από ιστοσελίδες που προωθεί η ιστοσελίδα του αντικειμένου, καθώς και από σελίδες που καταλήγουν στην ιστοσελίδα του αντικειμένου. Απεικόνισαν την πρόβλεψη κάθε επισημείωσης, με μια δυαδική μέθοδο ταξινόμησης. Για κάθε μέθοδο ταξινόμησης, ορίστηκαν δυο νέες μέθοδοι, με τυχαία ή προκαθορισμένη επιλογή επισημειώσεων του συνόλου R_p και R_n. Σε κάθε νέα μέθοδο επιλογής συνόλων, ένα τμήμα των συνόλων αυτών, αξιοποιείται για την εκπαίδευση του συστήματος πρόβλεψης νέων επισημειώσεων, ενώ το υπόλοιπο σύνολο ελέγχει την αξιοπιστία του συστήματος.

Για την υλοποίηση της μεθόδου, αρχικά κατακερματίστηκαν οι χρησιμοποιούμενες ιστοσελίδες και από τις λέξεις που εξήχθησαν, διαγράφηκαν εκείνες με το μικρότερο πλήθος εμφανίσεων. Κάθε λέξη από εκείνες που κρατήθηκαν για την υλοποίηση της έρευνας, αντιστοιχίζονται σε μια τιμή, αναφορικά με το πλήθος των εμφανίσεων της και με το μέρος που εμφανίστηκε. Στην συνέχεια για κάθε λέξη, γίνεται η αξιολόγηση της, σύμφωνα με την κατηγορία από την οποία εξήχθηκε. Η αξιολόγηση γίνεται με τον υπολογισμό της τιμής TFIDF, κανονικοποιημένης προς το πλήθος του συνόλου που εξεταζόταν κάθε φορά.

Στην δεύτερη μέθοδο, χρησιμοποιήθηκαν οι τιμές των λέξεων, με έναν διαφορετικό τρόπο. Αρχικά για κάθε λέξη υπολογίσθηκαν οι κανόνες συσχέτισης, support, confidence και interest (εικόνα 2.2.).

Int.	Conf.	Supp.	Rule
0.81	0.989	1097	open_source & source → opensource
0.55	0.979	1003	downloads & os → software
0.42	0.967	1686	free & webservice → web
0.73	0.964	1134	accessibility & css → webdev
0.84	0.952	1305	app & osx → mac
0.40	0.950	2162	webdesign & websites → web
0.47	0.947	2269	technology & webtools → tools
0.40	0.945	1662	php & resources → web
0.63	0.937	2754	html & tips → webdesign
0.50	0.934	1914	xp → software
0.45	0.928	1332	freeware & system → tools
0.62	0.919	1513	cool & socialsoftware → web2.0
0.61	0.915	1165	business & css → webdesign
0.61	0.912	2231	tips & webdevelopment → development
0.35	0.900	6337	toread & web2.0 → web
0.69	0.897	1010	fotografia & inspiration → art
0.33	0.895	1723	help & useful → reference

Εικόνα 2.2. Τυχαίο δείγμα από τους εξαχθέντες κανόνες συσχέτισης της έρευνας.

Οι τιμές υπολογίσθηκαν για το σύνολο των ιστοσελίδων. Στην συνέχεια, οι ιστοσελίδες χωρίστηκαν σε υποκατηγορίες όπως και στην προηγούμενη μέθοδο, σε κατηγορία προσομοίωσης και κατηγορία δοκιμών. Ύστερα, για ένα υποσύνολο των ιστοσελίδων, εφαρμόστηκαν κάποιοι κανόνες συσχετίσεων ταξινομημένοι με φθίνουσα τιμή confidence. Για κάθε κανόνα πάνω από ένα συγκεκριμένο κατώφλι (threshold), οι όροι του προστίθενται στους προβλεπόμενους όρους.

2.2 Συστήματα συστάσεων με χρήση επισημειώσεων.

Στην έρευνα που θα παρουσιάσουμε παρακάτω, η ερευνητική ομάδα από το Ηνωμένο Βασίλειο και την Ιταλία, προτείνει 3 μεθόδους για αποτελεσματικότερα συστήματα προτάσεων για την κάλυψη των ενδιαφερόντων των χρηστών υπηρεσιών σχετικών με ταινίες [20].

Τα συστήματα συστάσεων έχουν εξελιχθεί τα τελευταία χρόνια για την υποστήριξη των χρηστών στην ανακάλυψη νέων στοιχείων, μέσω κατασκευής προφίλ που εκπροσωπούν τα συμφέροντά τους, και των δικτύων που τους συνδέουν με άλλους χρήστες που μοιράζονται παρόμοια ενδιαφέροντα. Ωστόσο, πολλές από αυτές τις στρατηγικές συστάσεων στηρίζονται στην μοντελοποίηση των εγγενών χαρακτηριστικών για κάθε στοιχείο (π.χ. τις λέξεις-κλειδιά για ένα έγγραφο), έτσι ώστε τα στοιχεία να μπορούν να ταξινομηθούν, με άμεσο αποτέλεσμα το επίπεδο του ενδιαφέροντος που έχει ο χρήστης να μπορεί να εκφραστεί με όρους από αυτά τα χαρακτηριστικά. Αυτές οι συστάσεις βασίζονται εξ ολοκλήρου στο τι συμβαίνει στο εσωτερικό του συστήματος, αγνοώντας οποιαδήποτε εξωτερική γνώση για τα αντικείμενα ή τους ίδιους τους χρήστες.

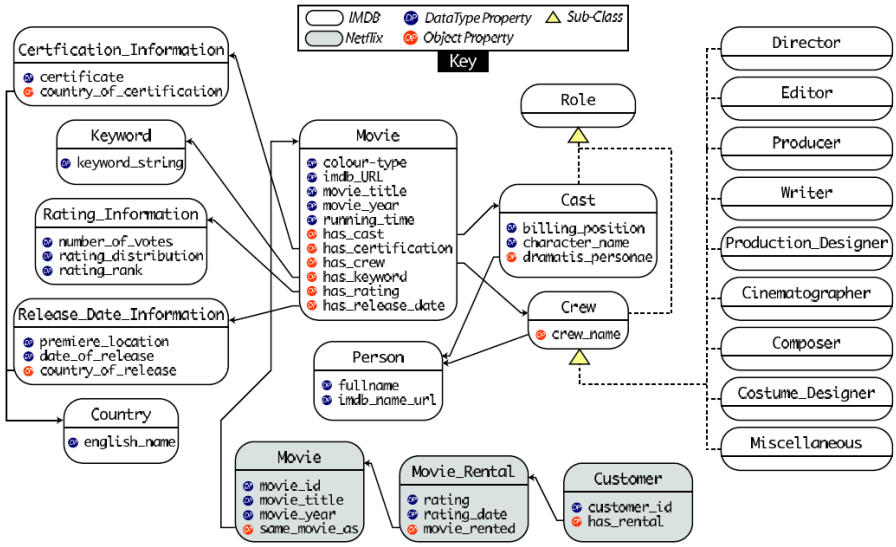
Για να βελτιωθούν τέτοιες τεχνικές σύστασης, πιστεύουμε ότι θα ήταν χρήσιμο να ενσωματωθούν στοιχεία και από πολλές πηγές, για το όσο το δυνατόν την οικοδόμηση πιο πλούσιου προφίλ. Στην παρούσα εργασία, εξετάζετε η υπόθεση πως οι κοινωνικές ταξονομίες (folksonomies) που περιγράφουν τίτλους ταινιών μπορεί να χρησιμοποιηθεί ως βάση για την οικοδόμηση προφίλ συστάσεων με την συμμετοχή του κάθε χρήστη και του σύννεφου επισημειώσεων που εκπροσωπεί τα συμφέροντά του.

Για να συγκεντρώσουν τις πληροφορίες που απαιτούνται για την κατασκευή του προφίλ που περιγράφει τα είδη των ταινιών, για τις οποίες ένας χρήστης ενδιαφέρεται, συνδυάζονται δεδομένα που συλλέγονται από δύο πηγές.

1. Για τα δεδομένα επισημειώσεων ταινιών, οι ερευνητική ομάδα έκανε χρήση του Internet Movie Database (IMDB), μια ηλεκτρονική βάση δεδομένων που περιέχει εκτενείς πληροφορίες για ταινίες, ηθοποιούς, τηλεοπτικές εκπομπές, και το προσωπικό των παραγωγών.

2. Για τον έλεγχο του συστήματος χρησιμοποιούνται τα δεδομένα που παρέχονται στο Netflix. Το Netflix είναι μια διαδικτυακή υπηρεσία ενοικίασης DVD, που ιδρύθηκε το 1998 και παρέχεται σε πελάτες στις Ηνωμένες Πολιτείες[6].

Για να συνδυάσουν τη βάση δεδομένων της imdb και τα στοιχεία αξιολόγησης της Netflix, εισήγαγαν τα δύο σύνολα δεδομένων σε μια πρότυπη σχεσιακή βάση δεδομένων. Συνδυάζοντας τα παραπάνω δεδομένα, παρείχαν έναν τρόπο για ανάκτησης λέξεων κλειδιών από την IMDB για κάθε τίτλο ταινίας Netflix, με συγκεκριμένη βαθμολογία, καθώς και άλλα στοιχεία, όπως φαίνονται στο παρακάτω σχήμα.



Σχήμα 2.1. Η οντολογία που συνδυάζεις τις πληροφορίες του IMDB και του Netflix.

Για να προβλεφθεί η σχέση μεταξύ του τρόπου που οι χρήστες βαθμολογούν τις ταινίες και των λέξεων-κλειδιά που εκχωρούν σε αυτές, η ερευνητική ομάδα έχει επινοήσει δυο αλγορίθμους πρόβλεψης.

[6] Τον Οκτώβριο του 2006, το Netflix άρχισε ένα διαγωνισμό για να βρει την καλύτερη πρόταση για συστήματα σύστασης, προσφέροντας ένα μεγάλο βραβείο αξίας 1 εκατ. δολάρια σε όποιον καταφέρει να βελτιώσει το δικό τους αλγόριθμο κατά 10%. Για να επιτύχει αυτός ο διαγωνισμός, το Netflix δημοσίευσε ένα μεγάλο σύνολο δεδομένων βαθμολογίας ταινιών από τη βάση δεδομένων που χαρακτηρίζει τους 480.189 πελάτες και 17.770 τίτλους ταινιών.

Average – Based Rating.

Μια πολύ απλή στρατηγική πρόβλεψης αξιολόγησης, μπορεί να εφαρμοστεί από την παραδοχή ότι ένας χρήστης θα βαθμολογήσει μια νέα ταινία, σύμφωνα με τη μέση εκτίμηση που λαμβάνει η ταινία από όλους τους άλλους χρήστες. Η μέθοδος αυτή θα χρησιμοποιηθεί στο τέλος της έρευνας για την αξιολόγηση των αποτελεσμάτων των προτάσεων της ερευνητικής ομάδας.

Απλή Σύγκριση σύννεφου επισημειώσεων.

Στο εν λόγω σύστημα, η ερευνητική ομάδα θα μαντέψει τη βαθμολογία του χρήστη που θα έδινε σε μια ταινία, συνδυάζοντας το σύνολο των λέξεων-κλειδιών που συνδέονται με την ταινία με την βαθμολογία που έδωσε ο χρήστης στις ταινίες που αντιστοιχούν στο σύννεφο-επισημειώσεων που σχετίζονται με την ταινία. Προβλέπουν πως η βαθμολογία, που αντιστοιχεί στην ταινία που εξετάζεται, είναι εκείνη που ταιριάζει περισσότερο στο σύνολο λέξεων-κλειδιών της ταινίας. Και μετράται με την βοήθεια των λέξεων-κλειδιών που μοιράζεται η ταινία με το σύννεφο επισημειώσεων του χρήστη για διαφορετικές βαθμολογίες. Πιο συγκεκριμένα εξετάζουν, ποιες επισημειώσεις έχει δώσει ο συγκεκριμένος χρήστης και ταιριάζουν με τις επισημειώσεις της ταινίας που εξετάζεται και εξάγουν την βαθμολογία της ταινίας, χρησιμοποιώντας την βαθμολογία που έχει δώσει ο χρήστης στις ταινίες των επισημειώσεων αυτών.

Σταθμισμένη Σύγκριση σύννεφου επισημειώσεων.

Σε αυτό το υβριδικό σύστημα προσπαθούν να λάβουν υπόψη τόσο τα βάρη σε επίπεδο λέξεων-κλειδιών (μέσω των συχνοτήτων τους), όσο και στο επίπεδο σύννεφου επισημειώσεων, μέσω ενός μέτρου ομοιότητας. Δεδομένης μιας νέας ταινίας, θεωρούν το σύνολο των λέξεων-κλειδιών της και εισάγουν την έννοια της «ομοιότητας» μεταξύ των επισημειώσεων της ταινίας και ένας συγκεκριμένου σύννεφου επισημειώσεων του χρήστη.

Στο πείραμά τους αυτό, αποφάσισαν να χρησιμοποιήσουν την βαθμολογία, που υπολογίζεται από την ομοιότητα του σύννεφου επισημειώσεων, για να βελτιωθεί

η βαθμολογία με απλή εκτίμηση. Συνδυάζουν τις δύο εκτιμήσεις, υπολογίζοντας σταθμισμένο μέσο όρο τους.

Για τον έλεγχο των αλγορίθμων που παρουσιάστηκαν παραπάνω, η ερευνητική ομάδα έχει καταρτίσει έναν πίνακα με τα αποτελέσματα από 500 τυχαία επιλεγμένες ταινίας του Netflix (σχήμα 2.2).

	Average Rating	Unweighted	Weighted
Correct	36.12%	44.15%	42.47%
Incorrect	63.99%	55.85%	57.53%
RMSE	1,131	1.074	0.961

Σχήμα 2.2. Αποτελέσματα μεθόδων της ερευνητικής ομάδας.

Η μη σταθμισμένη σύγκριση του σύννεφου επισημειώσεων έχει καλύτερες επιδόσεις από την αφελή μέθοδο μέσης βαθμολογίας, με μια μέτρια αύξηση του ποσοστού των σωστά βαθμολογημένων ταινιών. Χρησιμοποιώντας την σταθμισμένη σύγκριση επισημειώσεων βελτιώνετε η μέση τετραγωνική απόκλιση (RMSE), αλλά με μια μικρή πτώση στο κλάσμα των σωστά βαθμολογημένων ταινιών.

2.3 Δημιουργία ταξονομιών επισημειώσεων.

Η συγκεκριμένη έρευνα πραγματοποιήθηκε από τους Paul Heymann και Hector Garcia – Molina, από το τμήμα επιστήμης των υπολογιστών του πανεπιστημίου του Stanford [19].

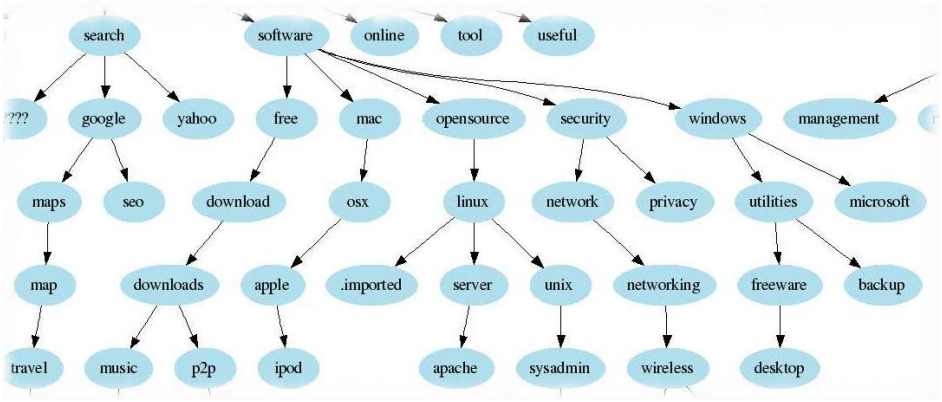
Καθώς το μέγεθος των πληροφοριακών συστημάτων μας, έχει επεκταθεί, δημιουργείται μια σταδιακή τάση αποφυγής συστημάτων με κεντρική διαχείριση, τα οποία βασίζονται σε ελεγχόμενα λεξιλόγια (δηλαδή, ένα πρότυπο βιβλιοθήκης). Η αλλαγή παρατηρείτε καθώς μεταφερόμαστε σε χαοτικά κατανεμημένα συστήματα, στα οποία όλοι οι χρήστες συνεργάζονται.

Μέχρι τώρα, οι χρήστες μπορούν να περιηγηθούν σε αντικείμενα χρησιμοποιώντας ένα σύστημα τοποθέτησης ετικετών με τρεις κύριες ιδιότητες:

1. Μια λίστα όλων των αντικειμένων που έχουν επισημανθεί με μια δεδομένη και συγκεκριμένη ετικέτα,
2. Μια λίστα από τις πιο δημοφιλείς λέξεις-κλειδιά στο σύστημα.
3. Μια λίστα των επισημειώσεων που έχουν υψηλό βαθμό επικάλυψης με μια επισημείωση την οποία ο χρήστης ερευνά εκείνη την στιγμή.

Ωστόσο, οι περιορισμοί αυτοί καθιστούν δύσκολη την αναζήτηση ευρύτερων ή στενότερων επισημειώσεων, οι οποίες μπορεί να αντιπροσωπεύουν καλύτερα τα τρέχοντα ενδιαφέροντα του χρήστη.

Οι Molina και Heymann, σχεδίασαν έναν αλγόριθμο που βοηθά στην επίλυση αυτού του προβλήματος με την αυτόματη δημιουργία μιας ιεραρχίας επισημειώσεων από τα δεδομένα ενός συστήματος τοποθέτησης επισημειώσεων (εικόνα 2.3). Ο αλγόριθμος αξιοποιεί τις έννοιες της ομοιότητας και της γενικότητας που είναι συνυφασμένες με τα δεδομένα που δημιουργούνται από τους χρήστες, δεδομένου ότι σχολιάζουν αντικείμενα. Ο αλγόριθμός τους αξιοποιεί τις εισφορές σε πληροφορία, του τεράστιου αριθμού των χρηστών, επιτρέποντας τους ταυτόχρονα να χρησιμοποιούν μια πιο φυσική περιήγηση.



Εικόνα 2.3. Όψη του ιεραρχικού μοντέλου που προτείνουν οι Molina και Heymann.

Πριν εξηγήσουμε, περιληπτικά, το μοντέλο των Molina και Heymann, θα αναφέρουμε κάποια στοιχεία που πρέπει να γνωρίζει ο αναγνώστης, όπως τα θεωρούν οι ερευνητές.

Υποθέτουμε ότι σύστημα ένα επισημειώσειων αποτελείται από αντικείμενα (O_1, O_2, …), χρήστες (u_1,U_2,…), και επισημειώσεις (t_1,t_2,…). Τα δεδομένα σε ένα σύστημα τοποθέτησης επισημειώσεων αποτελούνται από σχολιασμούς των αντικειμένων (annotations) (a_1,a_2,…), καθένα από τα οποία περιέχει ένα χρήστη (U_i), ένα αντικείμενο (O_j), και μία ή περισσότερες επισημειώσεις (t_k, …). Οι επισημειώσεις αποτυπώνονται σε διανύσματα επισημειώσεων, τα οποία αναδεικνύουν το πλήθος των φορών που η επισημείωση t_l «σχολιάζει» το αντικείμενο o_m.

Η ερευνητική ομάδα, υπολογίζει την ομοιότητα μεταξύ των επισημειώσεων, χρησιμοποιώντας την ομοιότητα συνημίτονου μεταξύ των διανυσμάτων τους, αν και άλλες μετρήσεις ομοιότητας μπορεί να είναι αποδεκτές εξίσου. Η γραφική παράσταση ομοιότητας επισημειώσεων για ένα συγκεκριμένο σύνολο δεδομένων είναι ένα μη σταθμισμένο γράφημα όπου κάθε επισημείωση αντιπροσωπεύεται από μια κορυφή, ενώ δύο κορυφές συνδέονται με μια ακμή αν η ομοιότητα των κόμβων που αντιπροσωπεύουν βρίσκεται πάνω από ένα συγκεκριμένο κατώφλι (threshold).

Μετά την περιορισμένη επιτυχία που παράγουν οι ιεραρχικές ταξινομίες με ιεραρχική ομαδοποίηση, αναπτύχθηκε ο αλγόριθμος της εικόνας 2.4. Ο αλγόριθμος ξεκινά με ένα δέντρο του οποίου ο μόνος κόμβος είναι η ρίζα του, που αντιπροσωπεύει την κορυφή του δέντρου (γραμμή 1). Στη συνέχεια, προσθέτει κάθε ετικέτα στο σύστημα σήμανσης στο δέντρο κατά φθίνουσα σειρά (γραμμές 2-4). Ο αλγόριθμος αποφασίζει πού θα τοποθετηθεί η κάθε υποψήφια επισήμανση με τον

υπολογισμό της ομοιότητάς του με κάθε κόμβο που υπάρχει ήδη στο δέντρο (γραμμές 5-10).

Require: $L_{generality}$ is a list of tags $t_i, ...t_j$ in descending order of their centrality in the similarity graph.
Require: Several functions are assumed: $s(t_i, t_j)$ computes the similarity (using cosine similarity, for example) between t_i and t_j. $getVertices(G)$ returns all vertices in the given graph, G.
Require: *taxThreshold* is a parameter for the threshold at which a tag becomes a child of a related parent rather than of the root.

1: $G_{taxonomy} \leftarrow \langle \emptyset, root \rangle$
2: **for** $i = 1...|L_{generality}|$ **do**
3: $t_i \leftarrow L_{generality}[i]$
4: $maxCandidateVal \leftarrow 0$
5: **for all** $t_j \in getVertices(G_{taxonomy})$ **do**
6: **if** $s(t_i, t_j) > maxCandidateVal$ **then**
7: $maxCandidateVal \leftarrow s(t_i, t_j)$
8: $maxCandidate \leftarrow t_j$
9: **end if**
10: **end for**
11: **if** $maxCandidateVal > taxThreshold$ **then**
12: $G_{taxonomy} \leftarrow G_{taxonomy} \cup \langle maxCandidate, t_i \rangle$
13: **else**
14: $G_{taxonomy} \leftarrow G_{taxonomy} \cup \langle root, t_i \rangle$
15: **end if**
16: **end for**

Εικόνα 2.4. Αλγόριθμος δημιουργίας ιεραρχική ταξινόμησης.

Η υποψήφια ετικέτα στη συνέχεια, είτε προστίθεται ως παιδί στον κοντινότερο κόμβο αν η ομοιότητά τους είναι μεγαλύτερη από κάποιο όριο, είτε προστίθεται στον κόμβο-ρίζα, αν δεν υπάρχει μέχρι εκείνη την στιγμή μια καλή μητρική (γραμμές 11-16).

Ο αλγόριθμος, των Molina και Heymann, μπορεί να χρησιμοποιηθεί σε συστήματα όπως αυτό που θα εξετάσουμε παρακάτω και να «προτείνει» ένα σύνολο όρων σχετικών με τις επισημειώσεις ενός τεκμηρίου, για να προστεθούν στο αρχείο καθιερωμένων όρων.

Ο αλγόριθμός που ανέπτυξαν οι Molina και Heymann, χαρακτηρίζεται από την ταχύτητα και την επεκτασιμότητα του.

"You can already see the privacy debate moving to the realm of automated massive data mining. When governments begin to suspect people because of where they were at a certain time, it can get very worrying."
--Jonathan Zittrain

3 Προτεινόμενο Μοντέλο – Βήμα 1

Όπως έχουμε αναφέρει στην περίληψη της εργασίας μας, σε αυτό το κεφάλαιο καθώς και στα επόμενα δυο που ακολουθούν, θα παρουσιάσουμε λεπτομερώς βήμα προς βήμα, όλες τις ενέργειες που κάναμε για να καταλήξουμε στο μοντέλο μας. Πιο συγκεκριμένα, παρακάτω, θα κάνουμε μια σύντομη επισκόπηση του προβλήματος που έχουμε να αντιμετωπίσουμε και της λύσης με την οποία το προσεγγίζουμε. Στην συνέχεια θα καταγράψουμε τα πρώτα βήματα που κάναμε για την πλήρη κατανόηση του προβλήματος μας, αλλά και την συλλογή των επιθυμητών δεδομένων προς επεξεργασία. Ύστερα, στα επόμενα δυο κεφάλαια, θα αναλύσουμε τον κορμό των εργασιών μας, ενώ στην συνέχεια θα καταγράψουμε το τελευταίο βήμα με τα πειράματα που εκτελέσαμε.

3.1 Σύντομη περιγραφή του μοντέλου μας.

Στην παρούσα διπλωματική εργασία, ασχοληθήκαμε με την σχεδίαση ενός μοντέλου, για τον εμπλουτισμό ενός αρχείου καθιερωμένων εγγραφών με κοινωνικές επισημειώσεις. Πιο συγκεκριμένα η εργασία επικεντρώνεται στην ανάπτυξη ενός εργαλείου σύστασης κοινωνικών επισημειώσεων, για τον εμπλουτισμό της θεματικής περιγραφής, των εγγραφών ενός καταλόγου μιας βιβλιοθήκης. Τα δεδομένα της εργασίας μας είναι τα εξής:

1. Το αρχείο καθιερωμένων εγγραφών (authority file), το οποίο πήραμε από τον κοινωνικό κατάλογο (OPACIAL[7]), του Πάντειου Πανεπιστημίου – Κοινωνικών και Πολιτικών Επιστημών –Αθηνών[8],
2. Μια κοινωνική ταξονομία (folksonomy), η οποία αποτελείται από κοινωνικές επισημειώσεις χρηστών και προέρχεται από το LibraryThing,
3. Το αρχείο καταγραφής των επερωτήσεων (query log) στον κατάλογο της βιβλιοθήκης για την αναζήτηση πόρων-τεκμηρίων (οι αναζητήσεις βασίζονται στα ακόλουθα ευρετήρια: θέμα, συγγραφέας, τίτλος, έτος έκδοσης, ISBN, γλώσσα, σημειώσεις, εκδότης, τίτλος σειράς καθώς και σε άλλα πεδία).

[7] Η βιβλιοθήκη του Παντείου Πανεπιστημίου προχώρησε στην ανάπτυξη ενός νέου «κοινωνικού καταλόγου» (OPACIAL) με τον εμπλουτισμό του από τις τεχνολογίες του Ιστού 2.0. Εκτός όμως από τις δυνατότητες αυτές ο νέος κατάλογος έχει δυνατότητες ενοποιημένης θεματικής αναζήτησης από την ψηφιακή βιβλιοθήκη ΠΑΝΔΗΜΟΣ, καλύτερης προβολής των εγγραφών σε συνοπτική και εξαντλητική μορφή, προβολή των θεματικών επικεφαλίδων καθώς και των νέων εγγραφών, προβολή σε xml, δημιουργία βιβλιογραφικών παραπομπών κ.α.

[8] Το Πάντειο Πανεπιστήμιο Αθηνών, οφείλει την ύπαρξη του στους Γεώργιο Σ. Φραγκούδη (1869-1939) και Αλέξανδρο Ι. Πάντο (1888-1930). Σήμερα περιλαμβάνει δέκα (10) Τμήματα: Α) Τμήμα Πολιτικής Επιστήμης και Ιστορίας, Β) Τμήμα Διεθνών και Ευρωπαϊκών Σπουδών, Γ) Τμήμα Δημόσιας Διοίκησης, Δ) Τμήμα Κοινωνιολογίας, Ε) Τμήμα Οικονομικής και Περιφερειακής Ανάπτυξης, ΣΤ) Τμήμα Κοινωνικής Πολιτικής, Ζ) Τμήμα Κοινωνικής Ανθρωπολογίας, Η) Τμήμα Επικοινωνίας, Μέσων και Πολιτισμού, Θ) Τμήμα Ψυχολογίας, Ι) Γενικό Τμήμα Δικαίου.

3.2. Διαδικασία εκτίμησης επισημειώσεων.

Σε αυτό το πρώτο βήμα της εργασίας, αναζητούμε την σχέση των επισημειώσεων με την θεματική περιγραφή, ή κάποιο σημείο πρόσβασης των αντικειμένων μας.

Ωστόσο, πριν προχωρήσουμε σε αυτό το βήμα, θα πρέπει να διερευνήσουμε το ποσοστό κάλυψης – σε λεξικογραφικό επίπεδο – ανάμεσα στα κύρια σύνολα των δεδομένων μας, όπως τα αναφέραμε προηγουμένως. Πιο συγκεκριμένα, αναζητήσαμε το ποσοστό κάλυψης των εξής συνόλων:

Σχέση F ∩ A : (Κοινωνικής ταξονομία ∩ Αρχείο καθιερωμένων εγγραφών - Folksonomy ∩ Authority file)

Σχέση A ∩ Q : (Αρχείο καθιερωμένων εγγραφών ∩ Αρχείο καταγραφής επερωτήσεων - Authority File ∩ Queries)

Σχέση F ∩ Q : (Κοινωνικής ταξονομία ∩ Αρχείο καταγραφής επερωτήσεων -Folksonomy ∩ Queries)

Τα αποτελέσματα μας έδωσαν τα πρώτα δείγματα για να ξεκινήσει η σχεδίαση της μεθόδου μας. Έτσι λοιπόν, είδαμε τα παρακάτω αποτελέσματα:

- ✓ Η κάλυψη του συνόλου του αρχείου καταγραφής επερωτήσεων, από το σύνολο του αρχείου καθιερωμένων εγγραφών, δηλαδή, $\frac{Q \cap A}{Q}$, είναι ίση με το 73,61%. Ενώ, αντίθετα, η κάλυψη του συνόλου του αρχείου των καθιερωμένων όρων, από το σύνολο του αρχείου καταγραφής επερωτήσεων, δηλαδή, $\frac{A \cap Q}{A}$, είναι ίση με το 81,3%.
- ✓ Η κάλυψη του συνόλου της κοινωνικής ταξονομίας, από το σύνολο του αρχείου καθιερωμένων εγγραφών, δηλαδή, $\frac{F \cap A}{F}$, είναι ίση με το 48,35%.
- ✓ Το σύνολο των επερωτήσεων, καλύπτεται από το σύνολο της κοινωνικής ταξονομίας, δηλαδή, $\frac{Q \cap F}{Q}$, σε ποσοστό 13%. Ενώ,

αντίθετα, το σύνολο της κοινωνικής ταξονομίας, καλύπτεται από το σύνολο των επερωτήσεων, δηλαδή, $\frac{F \cap Q}{F}$, ίσο με 43%.

- ✓ Τέλος, θα διαχωρίσουμε το σύνολο των επερωτήσεων, σε εκείνες που έγιναν με σημείο πρόσβασης «Κοινωνικές Επισημειώσεις». Το σύνολο που προκύπτει Q_0, καλύπτεται από το σύνολο της κοινωνικής ταξονομίας, δηλαδή, $\frac{Q_0 \cap F}{Q_0}$, σε ποσοστό 91%. Αντίθετα, το σύνολο της κοινωνικής ταξονομίας, καλύπτεται από το σύνολο Q_0, $\frac{F \cap Q_o}{F}$, σε ποσοστό ίσο με 40%.

Παρατηρούμε, λοιπόν, πως η γλώσσα που χρησιμοποιούν οι χρήστες προς αναζήτηση τεκμηρίων, καλύπτει σε σημαντικό βαθμό το αρχείο καθιερωμένων εγγραφών της βιβλιοθήκης. Ωστόσο, παρατηρούμε επίσης, πως και το αρχείο καθιερωμένων εγγραφών καλύπτεται σε σημαντικό βαθμό από την γλώσσα που χρησιμοποιούν οι χρήστες. Τα γεγονότα αυτά, μας δίνουν τα περιθώρια, να προσθέσουμε νέες επισημειώσεις στο αρχείο καθιερωμένων όρων για να καλύψουμε σε μεγαλύτερο βαθμό την γλώσσα των χρηστών, ενώ παράλληλα να αφαιρέσουμε επισημειώσεις που δεν προσφέρουν χρήσιμη πληροφορία στους χρήστες.

Παρατηρούμε επίσης, την τάση των χρηστών να προσθέτουν επισημειώσεις στα τεκμήρια, με βάση τους θεματικούς όρους που παρέχει η βιβλιοθήκη για αυτά. Σχεδόν οι μισές επισημειώσεις των χρηστών, λοιπόν, δεν προσφέρουν νέες πληροφορίες για την βιβλιοθήκη.

Ακόμα, παρατηρούμε πως λίγες φορές οι χρήστες χρησιμοποιούν όρους επισημειώσεων για να αναζητήσουν τεκμήρια. Θεωρούμε λοιπόν, πως οι χρήστες δεν έχουν γνώση των επισημειώσεων και συνήθως αναζητούν τα τεκμήρια τους, με βάση άλλα στοιχεία, όπως συγγραφέας, εκδοτικός οίκος κ.α. . Ωστόσο, βλέπουμε πως όταν οι χρήστες αναζητούν τεκμήρια, με βάση μόνο τις επισημειώσεις που έχουν, τότε καλύπτονται σε σημαντικό βαθμό από τα αποτελέσματά. Τέλος, παρατηρούμε πως λιγότερες από τις μισές επισημειώσεις της βιβλιοθήκης, καλύπτονται από την γλώσσα των χρηστών. Θεωρούμε πως πρέπει να βρεθεί ένας τρόπος να γεφυρωθεί το χάσμα, μεταξύ της γλώσσας των χρηστών και των επισημειώσεων, ώστε να

αντιπροσωπεύουν περισσότερο τους χρήστες και να αποφέρουν καλύτερα αποτελέσματα στην αναζήτηση τεκμηρίων.

Για τα παραπάνω ποσοστά κάλυψης, εφαρμόσαμε έναν απλό αλγόριθμο αναζήτησης, με σταθερό σύνολο, εκείνο που εξετάζεται η κάλυψη του, κάθε φορά και μεταβλητό το άλλο σύνολο. Πριν το έλεγχο, ωστόσο, για κάθε αρχείο – σύνολο, αφαιρέσαμε (από) τους όρους, όλα τις περιττές λέξεις (stop-words[9]), καθώς επίσης κάθε όρο τον αντιστοιχήσαμε στην κατάλληλη ρίζα του. Για παράδειγμα, το όρο-φράση «Ο θαλασσινός πελαργός» την μετατρέψαμε κατά την αναζήτηση σε «Θαλασσ πελαργ», αφαιρώντας το stop-word «ο» και τις καταλήξεις. Η κάλυψη υπολογίζεται όπως στον αλγόριθμό 1.

Ο αλγόριθμος 1, έχει υλοποιηθεί σε γλώσσα java και βρίσκεται διαθέσιμος στο συνοδευτικό CD, της παρούσας εργασίας. Περισσότερα για το περιβάλλον εκτέλεσης του προγράμματος, θα δούμε στο παράρτημα του παρόντος εγγράφου.

3.2.1. Επεξεργασία αρχείου καθιερωμένων όρων.

Το αρχείο καθιερωμένων όρων του Πάντειου Πανεπιστημίου Αθηνών, που παραλάβαμε για την έρευνά μας, περιείχε δεδομένα που δεν χρειαζόμασταν στο σύνολο της έρευνάς μας. Έπρεπε, λοιπόν, να επεξεργαστούμε το αρχείο και να το φτιάξουμε σε μια μορφή που εξυπηρετούσε τους σκοπούς

```
Για τα δύο σύνολα που εξετάζουμε, αφαίρεσε όλα τα
stop-words και βρες την ρίζα όλων των όρων τους;

While (υπάρχουν όροι του πρώτου συνόλου){
        Έστω όρος1;
        While(υπάρχουν όροι του δεύτερου
        συνόλου) {
                Έστω όρος2;
                Έλεγξε αν ο όρος2 περιέχεται στον
                όρο1 με οποιονδήποτε τρόπο;
                If(περιέχεται){
                        Αύξησε τον μετρητή κατά  1;
                }Else προχώρα στον επόμενο όρο2;
        }
}
```

Αλγόριθμος 1.

[9] Στην επιστήμη των υπολογιστών, τα stop-words, είναι λέξεις οι οποίες φιλτράρονται πριν, ή μετά, την επεξεργασία της φυσικής γλώσσας δεδομένων (κείμενο). Δεν υπάρχει μία οριστική λίστα των stop-words, που χρησιμοποιούν όλα τα εργαλεία. Μερικά εργαλεία αποφεύγουν τη χρήση τους για την υποστήριξη αναζήτηση φράσης. Κάθε ομάδα λέξεων μπορεί να επιλεγεί ως stop-word για ένα συγκεκριμένο σκοπό. Για μερικές μηχανές αναζήτησης, αυτά είναι μερικά από τα πιο κοινά, σύντομες λέξεις, όπως το, «The», «Το» κ.α. . Σε αυτή την περίπτωση, τα stop-words, μπορεί να προκαλέσουν προβλήματα κατά την έρευνα για τις φράσεις που τα περιλαμβάνουν, ιδιαίτερα σε ονόματα όπως «The Who», «Το Ο», ή «Take That».
Ο Hans Peter Luhn, ένας από τους πρωτοπόρους στην ανάκτηση πληροφοριών, πιστώνεται την επινόηση της φράσης stop-words.

μας και βοήθαγε στο έργο μας. Η αρχική μορφή του αρχείου ήταν όπως παρουσιάζεται παρακάτω, σύμφωνα με το πρότυπο MARC:

```
100| a20080516cgrey0103    ga-
250| aΠαιδαγωγικ΄ηxΙστορ΄ια-
*
100| a20060621cgrey0103    ba-
250| aΜν΄ημηxΚοινωνικ΄ες απ΄οψειςyΕυρ΄ωπη-
450| aMemoryxSocial aspectsyEurope0750-
*
100| a20080516cgrey0103    ga-
250| aΕγκληματικ΄οτητα ανηλ΄ικωνyΕλλ΄αδα-
xΙστορ΄ια-
450| aJuvenile delinquencyyGreecexHistory0750-
*
100| a19920107agrey0105    ga-
250| aΠρολεταρι΄ατο, Δικτατορ΄ια του-
300|1 aLC Control Number: sh 85037698  -
450| aDictatorship of the proletariat0750-
450| aΔικτατορ΄ια του προλεταρι΄ατου-
550| aΚομμουνισμ΄ος5g-
550| aΠρολεταρι΄ατο5g-
550| aΟλοκληρωτισμ΄ος5g-
550| aΚομμουνιστικ΄ο κρ΄ατος-
801| 0aGRbΠ΄αντειο Πανεπιστ΄ημιο-
*
```

Με την παραπάνω μορφή, δηλαδή με την μορφή MARC εγγραφών, εισάγονται οι επικεφαλίδες στο πληροφοριακό σύστημα των βιβλιοθηκών.

Σε αυτό το σημείο θα κάνουμε μια ανασκόπηση στις εγγραφές MARC, όπως αυτή παραπάνω, για να εξηγήσουμε την σημασία των πεδίων και των συμβόλων που παρατηρούμε, ενώ στην συνέχεια θα δούμε αναλυτικά την επεξεργασία που κάναμε

στο αρχείο καθιερωμένων όρων. Τα γράμματα και τα σύμβολα που έχουν επισημανθεί, είναι όλα όσα πρέπει να διαγράψουμε ή να επεξεργαστούμε.

Κάθε εγγραφή έχει μια συγκεκριμένη μορφή επικεφαλίδας. Για ονόματα, ένα από τα πιο ευρέως χρησιμοποιούμενα αρχεία καθιερωμένων όρων, είναι τι αρχείο καθιερωμένων όρων της Βιβλιοθήκης του Κογκρέσου (ή LCNAF). Το αρχείο αυτό είναι διαθέσιμο από την Υπηρεσία διανομής και καταλογογράφησης (CDS) του Κογκρέσου ή από την διεύθυνση: http://authorities.loc.gov/. Όλες οι κατηγορίες είναι ένα από τα παρακάτω είδη: ονόματα, όνομα / τίτλος συνδυασμούς, ομοιόμορφη τίτλους, ή τα θέματα. Παραδείγματα των κλάσεων όνομα περιλαμβάνουν[32]:

Η **επικεφαλίδα – όνομα** (name heading) μιας εγγραφής μπορεί να έχει σχέση με πρόσωπο, τοποθεσία κτλ.

```
100 1#  $a Woolf, Virginia, $d 1882-1941
            (personal name heading)
110 2#  $a Association for Childhood Education International
            (corporate name heading)
111 2#  $a La Crosse Health and Sports Science Symposium
            (meeting name heading)
151 ##  $a Mexico
            (geographic name heading)
```

Η επικεφαλίδα – όνομα/τίτλος (name/title heading) μιας εγγραφής αποτελείται τόσο από ένα όνομα, όσο και από έναν τίτλο ενός έργου. Το όνομα μπορεί να είναι όπως προηγουμένως, ενός το τμήμα τίτλος περιέχει τον τίτλο με τον οποίο ένα στοιχείο ή μια σειρά πρόκειται να προσδιοριστεί στην καταλογογράφηση. Χρησιμοποιείται κυρίως για κλασικά έργα που έχουν δημοσιευθεί με διαφορετικούς τίτλους κατά την πάροδο του χρόνου. Ένα παράδειγμα μιας επικεφαλίδας όνομα / τίτλος[32]:

```
100 1#  $a Dostoyevsky, Fyodor,
        $d 1821-1881.
        $t Crime and punishment
```

Μια επικεφαλίδα θέμα μπορεί να αποτελείται από έναν απλό όρο, μια φράση, ή μια ομάδα όρων. Η ετικέτα για το θέμα βασίζεται στον πρώτο όρο ή φράση της επικεφαλίδας. Οι λοιποί όροι της εγγραφής είναι σε συγκεκριμένα υποπεδία που καλούνται υποδιαιρέσεις[32].

```
100 1#  $a Shakespeare, William, $d 1564-1616
        $x Criticism and interpretation $x History
        $y 18th century
150 ##  $a Books and reading $z Argentina
```

Κάθε MARC εγγραφή, αποτελείται από ένα σύνολο γραμμών, ενώ διαχωρίζονται μεταξύ τους με μια νέα γραμμή με τον χαρακτήρα « * ». Κάθε γραμμή της εγγραφής με την σειρά της, αποτελείται από ακριβώς ένα πεδίο (με το οποίο και ξεκινάει), ένα ή περισσότερα υποπεδία και κανέναν, έναν ή περισσότερους δείκτες.

Το **πεδίο** περιέχει δύο δείκτες ενός και δύο χαρακτήρων, αντίστοιχα, που δείχνουν την κατηγορία και την υποκατηγορία που ανήκει η εγγραφή.

Ο πρώτος χαρακτήρα της ετικέτας, που δηλώνει την κατηγορία της εγγραφής, προσδιορίζει τη λειτουργία των δεδομένων εντός του πεδίου. Ο τύπος των πληροφοριών εντός του πεδίου αναγνωρίζεται από το υπόλοιπο της ετικέτας. Για παράδειγμα:

- Η ετικέτα 0XX περιλαμβάνει πρότυπους αριθμούς, αριθμούς ταξινόμησης και κωδικούς αριθμούς,
- Η ετικέτα 1XX περιλαμβάνει επικεφαλίδες,
- Η ετικέτα 6XX περιλαμβάνει σημειώσεις.

Επίσης, με ορισμένες εξαιρέσεις, παρέχονται στους τελικούς δύο χαρακτήρες της ετικέτας των πεδίων, οι παρακάτω υποκατηγορίες:

- X00 Προσωπικά ονόματα
- X10 Εταιρικά ονόματα
- X48 Χρονολογικοί όροι
- X51 Γεωγραφικά ονόματα
- X55 Όροι σχετικοί με είδος/στυλ

Το **υποπεδίο** είναι πληροφορία που περιλαμβάνεται μέσα σε ένα πεδίο και σκοπό έχει να προσδιορίζει συγκεκριμένα κομμάτια δεδομένων στην γραμμή. Κάθε υποπεδίο αποτελείται από 2 χαρακτήρες. Ο πρώτος είναι ένας ειδικός χαρακτήρας, ο οποίος σηματοδοτεί την αρχή του υποπεδίου και είναι ίδιος για όλα τα υποπεδία, αλλά διαφέρει από αρχείο σε αρχείο. Στο αρχείο καθιερωμένων όρων του OPACIAL, ο χαρακτήρας που δείχνει την αρχή ενός υποπεδίου είναι ο χαρακτήρας US (unit separator). Για λόγους καλύτερης κατανόησης από τον αναγνώστη, ο χαρακτήρας US, έχει αντικατασταθεί με τον χαρακτήρα '$': Ενώ, ο δεύτερος χαρακτήρας είναι ένας εκ των : a, b, f, x, y, z, d, e, c, 4, g, 0, 2, 3, 5, 6, 7, 8. Κάθε χαρακτήρας δίνει μια διαφορετική σημασία στα δεδομένα, επίσης αξίζει να σημειωθεί πως, ακόμα και ο ίδιος χαρακτήρας σε διαφορετικό πεδίο, μπορεί να δίνει διαφορετική σημασία. Για παράδειγμα, το υποπεδίο $a στο πεδίο 102 δείχνει χώρα προέλευσης, ενώ το ίδιο υποπεδίο στο πεδίο 220 δείχνει το όνομα οικογενείας. Το ίδιο υποπεδίο μπορεί να υπάρχει παραπάνω της μιας φοράς σε μια γραμμή

Ο **δείκτης** είναι ένας χαρακτήρας που τοποθετείται προαιρετικά πριν από το πρώτο υποπεδίο. Σκοπός του είναι να παρέχει επιπρόσθετη πληροφορία σχετικά με τα περιεχόμενα του πεδίου, την σχέση μεταξύ του πεδίου και άλλων πεδίων ή τις δράσεις, που απαιτούνται σχετικά με κάποιες διαδικασίες διαχείρισης δεδομένων. Για παράδειγμα, ο δείκτης 1 στο πεδίο 200 δείχνει ότι ονοματεπώνυμο έχει καταχωρηθεί με πρώτα το επίθετο και μετά το όνομα.

Στο σημείο αυτό θα αναλύσουμε την επεξεργασία που κάναμε στο αρχείο καθιερωμένων όρων, με σκοπό να το μετατρέψουμε σε μια μορφή που να μπορούμε να επεξεργαστούμε και να πάρουμε το περιεχόμενό του χωρίς περιττές πληροφορίες.

Αρχικά, οι αριθμοί που βρίσκονται πριν από το σύμβολο « | », οι οποίοι όπως αναφέραμε εκφράζουν το πεδίο - σημείο πρόσβασης της επικεφαλίδας, αφαιρούνται. Ταυτόχρονα αφαιρείται και το σύμβολο « | », καθώς και όποιος αριθμός το ακολουθεί εάν υπάρχει, με εξαίρεση το πρώτο υποπεδίο κάθε γραμμής και το τελευταίο, αν μετά από αυτό δεν ακολουθούν δεδομένα, τα οποία απλά διαγράφονται, χωρίς να αντικατασταθούν με 2 παύλες. Ακόμα, εξαίρεση αποτελούν τα πεδία 200 και 400, στα οποία πλην του υποπεδίου X (όπου και σ' αυτό μπαίνουν 2 παύλες), αντί για 2 παύλες μπαίνει κόμμα « , » .

Οι γραμμές που ξεκινούν από 1XX, 3XX και 8XX, όπου XX ε ℕ, π.χ. 100, 101, 300, 330, 801) δεν χρειάζονται και διαγράφονται μαζί με το περιεχόμενο τους. Το ίδιο συμβαίνει και με τα σύμβολα υποπεδίων 2, 3, 4, 5, 6, 7, 8, 0 και g μαζί με το περιεχόμενο τους, αλλά και για τον χαρακτήρα « - » αν εμφανίζεται στο τέλος κάποιας γραμμής.

Ακόμα, διαγράφεται και το σύμβολο «$», όπου υπάρχει, καθώς δηλώνει την έναρξη κάποιου υποπεδίου.

Τέλος, όπως φαίνεται και στο δείγμα του αρχείου καθιερωμένων όρων, παραπάνω, οι επικεφαλίδες που είναι γραμμένες στα ελληνικά, έχουν τους τόνους τους μετά το γράμμα που τονίζεται, με αποτέλεσμα να πρέπει να αντικατασταθούν τα γράμματα που τονίζονται με τους κατάλληλους χαρακτήρες.

Μετά την επεξεργασία που περιγράψαμε παραπάνω, το αρχείο καθιερωμένων όρων είχε την παρακάτω μορφή:

Παιδαγωγική Ιστορία

Μνήμη--Κοινωνικές απόψεις--Ευρώπη

Memory--Social aspects--Europe—

Εγκληματικότητα ανηλίκων--Ελλάδα Ιστορία

Juvenile delinquency----Greece—History

Προλεταριάτο, Δικτατορία του

Δικτατορία του προλεταριάτου

Κομμουνιστικό κράτος

3.2.2. Διαχωρισμός χρήσιμων επισημειώσεων.

Μετά την εύρεση της επικάλυψης των παραπάνω συνόλων και την επεξεργασία του αρχείου καθιερωμένων όρων, εξετάσαμε την σχέση των επισημειώσεων που έχουμε στην κατοχή μας, με τα σημεία πρόσβασης των εγγραφών μας. Πιο συγκεκριμένα, παρατηρήσαμε ποιες επισημειώσεις, υπάρχουν αυτούσιες, σαν όροι, στα θέματα, στον συγγραφέα, στον τίτλο ή σε κάποιο άλλο πεδίο μιας εγγραφής.

Μία τέτοια επισημείωση, όπου βρίσκεται σε κάποιο από τα πεδία αναζήτησης μιας εγγραφής, βρίσκονται ήδη στο αρχείο καθιερωμένων όρων που θέλουμε να εμπλουτίσουμε και ως εκ τούτου, δεν τα λαμβάνουμε υπ' όψιν μας για τα επόμενα στάδια της έρευνάς μας, απλά μπορούμε να πούμε ότι καλύπτουν μια ανάγκη του χρήστη για να διαχειριστεί με αυτό τον τρόπο την πληροφορία.

Στην κατοχή μας είχαμε ένα αρχείο, που περιείχε όλες τις εγγραφές της βιβλιοθήκης του Πάντειου Πανεπιστημίου, που περιλάμβαναν επισημειώσεις. Το αρχείο έχει την μορφή:

10022404	0-226-46812-7	poetry
10022404	0-226-46812-7	poetry criticism
10022404	0-226-46812-7	Poetique
10022404	0-226-46812-7	rhetoric
10022404	0-226-46812-7	semiotics
10022404	0-226-46812-7	social commentary

Οι στήλες του αρχείου είναι :

1. ο μοναδικός αριθμός εγγραφής (MAE) του τεκμηρίου στον κατάλογο της βιβλιοθήκης,
2. το ISBN του τεκμηρίου και
3. η επισημείωση που έχει εισαχθεί για το τεκμήριο.

Με βάση το αρχείο αυτό, διαχωρίσαμε ποιες επισημειώσεις ήταν χρήσιμες για εμάς και θα τις επεξεργαζόμασταν στην συνέχεια της έρευνάς μας.

Το παραπάνω αρχείο σε συνδυασμό με το αρχείο όλων των πληροφοριών των τεκμηρίων της βιβλιοθήκης, θα μας βοηθήσουν στην εξαγωγή των χρήσιμων (worthy) και μη επισημειώσεων.

Παρακάτω θα δώσουμε ένα παράδειγμα για την καλύτερη κατανόηση των λεγόμενών μας, σε ένα πραγματικό αντικείμενο της βιβλιοθήκης του Πάντειου Πανεπιστημίου.

Παράδειγμα 1. Η επισημείωση “Foucault”.

Η επισημείωση “Foucault” υπάρχει σαν θέμα στο παρακάτω τεκμήριο, http://library.panteion.gr/opacial/details.php?catalog=1&view=1&language=el&numhits=10&query=Foucault&search_field=0&bool_field1=&query2=&search_field2=1&bool_field2=&query3=&search_field3=1&page=4&show=normal&rec_id=10024425

Επομένως, αναμένεται να υπάρχει ένας μεγάλος αριθμός από επισημειώσεις που να ταυτίζονται με κάποιο από τα σημεία πρόσβασης της εγγραφής μας. Έτσι λοιπόν, η αναζήτηση του τεκμηρίου αυτού με τον όρο “Foucault”, θα είναι επιτυχής, ωστόσο, ο όρος αυτός δεν προσθέτει καμία νέα πληροφορία που αξίζει να προστεθεί στο αρχείο καθιερωμένων εγγραφών. Εμείς επιθυμούμε, την αναζήτηση του τεκμηρίου, με κάποιον όρο που δεν περιέχετε σε κανένα σημείο πρόσβασης και θα προσθέσει νέα πληροφορία στο αρχείο καθιερωμένων όρων.

Μετά το τέλος και της παραπάνω διαδικασίας, καταφέραμε και δημιουργήσαμε ένα νέο αρχείο με όλες τις επισημειώσεις που θα ενδεχομένως θα προσφέρουν νέα πληροφορία στο αρχείο καθιερωμένων εγγραφών μας. Στο επόμενο κεφάλαιο, αναφερόμαστε στην διαδικασία κατά την οποία λαμβάνουμε και εξετάζουμε την αξία των επισημειώσεων που κρατήσαμε, στο LibraryThing και στο Google Books.

"I have had my results for a long time:
but I do not yet know how I am to
arrive at them."
--Carl Friedrich Gauss

4 Προτεινόμενο Μοντέλο – Βήμα 2

Σε αυτό το κεφάλαιο, θα αναλύσουμε τις μεθόδους με τις οποίες εξορύξαμε τα δεδομένα μας. Πιο συγκεκριμένα, θα δούμε αναλυτικά όλες τις διαδικασίες με τις οποίες καταφέραμε να πάρουμε πληροφορίες σχετικά με τις επισημειώσεις που ξεχωρίσαμε από το προηγούμενο κεφάλαιο.

Τελειώνοντας τις προηγούμενες διαδικασίες, πιθανότατα, για κάποιες επισημειώσεις να ανακαλύψουμε ότι η επισημείωση δεν υπάρχει ως όρος, δηλαδή δεν ταυτίζεται λεξικογραφικά με κανένα σημείο πρόσβασης της υπό εξέταση εγγραφής. Έτσι λοιπόν διακρίνουμε δύο πιθανές περιπτώσεις:

1. Η επισημείωση είναι προσωπική ενός τυχαίου χρήστη και δεν είναι σημαντική για το σύνολο των χρηστών μας ή ακόμα για την πληροφορία της εγγραφής.

2. Η επισημείωση είναι σημαντική και το ενδεχόμενο αυτό εξετάζεται στο επόμενο στάδιο, αυτό, της αξιολόγησης της επισημείωσης.

Πριν προχωρήσουμε στα επόμενα βήματα της εργασίας μας, θα ορίσουμε, ποια είναι η «αξία» μιας επισημείωσης και πως θα την αξιολογήσουμε. Η διαδικασία αξιολόγησης μιας επισημείωσης, χωρίζεται σε δύο στάδια. Το πρώτο είναι η εξόρυξη από το LibraryThing και το Google Books, της πληροφορίας που έχει να μας δώσει κάθε επισημείωση, ενώ το δεύτερο στάδιο είναι ο συνδυασμός των αποτελεσμάτων που πήραμε και το συμπέρασμα που βγάλαμε για αυτές.

Η αξία των επισημειώσεων που θεωρούμε χρήσιμες για το σύστημά μας είναι ένας συνδυασμός των πληροφοριών που λαμβάνουμε από το LibraryThing και το Google Books. Για την ακρίβεια, στο LibraryThing, θα δούμε το πλήθος των εμφανίσεων που κάθε επισημείωση παρατηρείται στο τεκμήριο που αναφερόμαστε. Ενώ, στο Google Books, θα δούμε αν η συγκεκριμένη επισημείωση είναι σημαντική στο τεκμήριο, καθώς κοιτάμε το πλήθος των σελίδων που εμφανίζεται σε αυτό. Περισσότερα για τις πληροφορίες των επισημειώσεων που λαμβάνουμε, θα αναφέρουμε στο αντίστοιχο υποκεφάλαιο.

Με τις πληροφορίες αυτές που λαμβάνουμε, θα αξιολογήσουμε κάθε επισημείωση, όχι μόνο σύμφωνα με την προτίμηση που έχουν οι χρήστες σε αυτήν για το συγκεκριμένο τεκμήριο, αλλά παράλληλα θα προσμετράτε και η βαρύτητα της επισημείωσης στο τεκμήριο, σύμφωνα με τα γραφόμενα του εκάστοτε συγγραφέα.

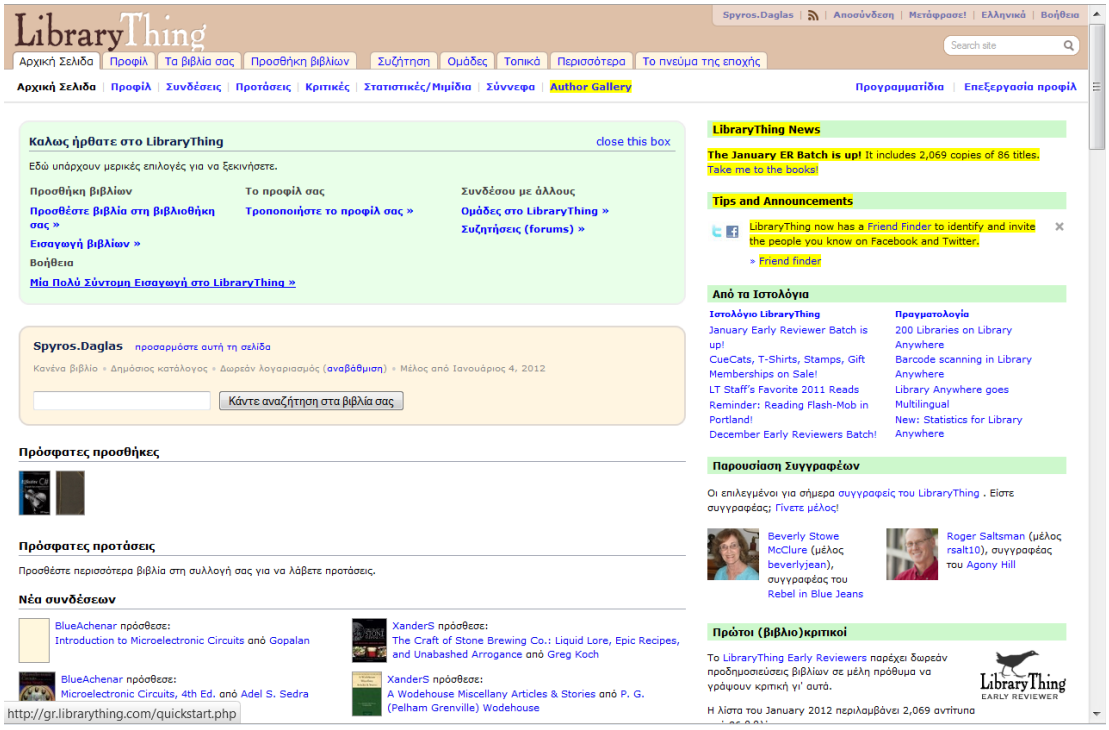
4.1. Σύντομη περιγραφή του LibraryThing.

"LibraryThing can also connect likeminded readers — a sort of MySpace for bookworms."

--The Wall Street Journal

Το LibraryThing είναι μια διαδικτυακή υπηρεσία που βοηθάει τους χρήστες της, να καταλογογραφήσουν τα βιβλία τους εύκολα. Το LibraryThing συνδέει, επίσης, τα άτομα με τα ίδια ενδιαφέροντα σε βιβλία και προτείνει βιβλία που ενδεχομένως κάποιος χρήστης να ενδιαφερόταν να διαβάσει στο μέλλον.

Αποτελεί μια πλήρη εφαρμογή καταλογογράφησης, οποία «περιλαμβάνει» την Βιβλιοθήκη του Κογκρέσου, και τους πέντε εθνικούς δικτυακούς τόπους του Amazon, και πάνω από 690 βιβλιοθήκες παγκοσμίως.



Εικόνα 4.1. Η κεντρική σελίδα του LibraryThing.

Δίνει την δυνατότητα στους χρήστες του, να επεξεργαστούν τις πληροφορίες του, να αναζητήσουν και να ταξινομήσουν τα βιβλία τους, να προσθέσουν επισημειώσεις σε βιβλία κατά το δοκούν, ή να χρησιμοποιήσουν την Βιβλιοθήκη του Κογκρέσου για να οργανώσουν τη δική τους συλλογή.

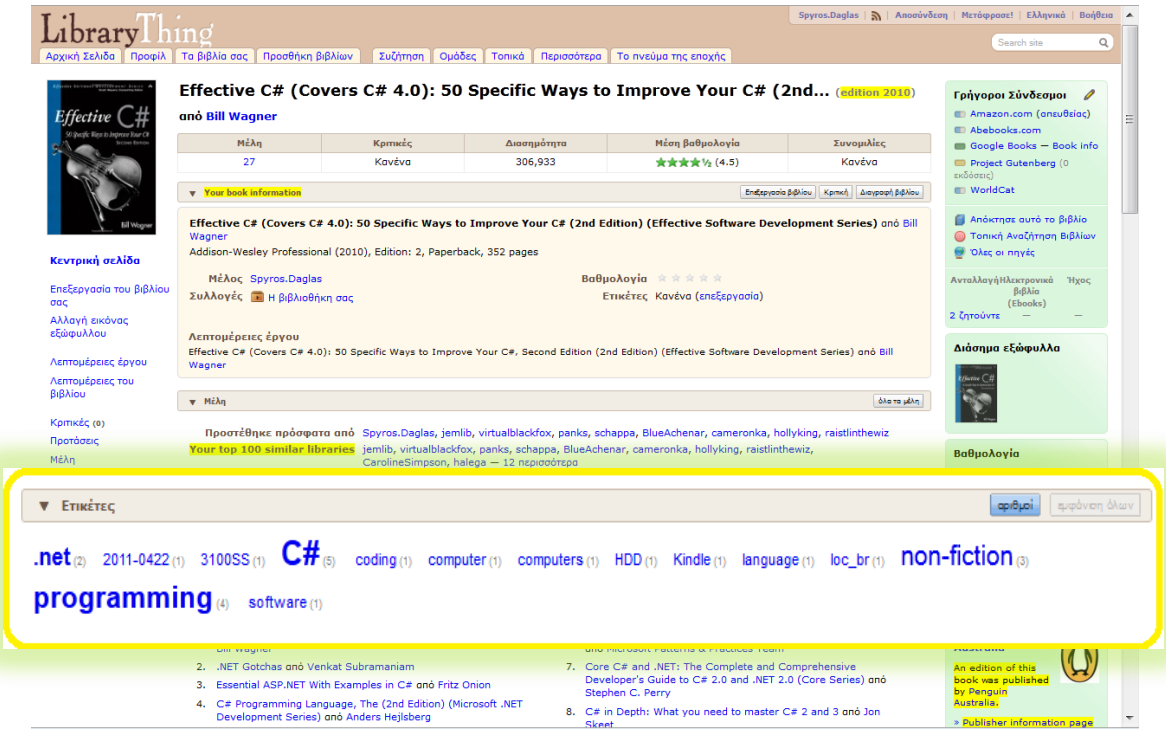
Ακόμα, το LibraryThing είναι ένας χώρος κοινωνικής δικτύωσης, που συχνά περιγράφεται ως "MySpace για τα βιβλία" ή "το Facebook για τα βιβλία". Το LibraryThing διατυπώνει επίσης συστάσεις βιβλίων, βασισμένες στις πληροφορίες και άλλων βιβλιοθηκών.

Δημιουργήθηκε από τον Tim Spalding, έναν προγραμματιστή και εκδότης εφαρμογών του Ιστού από το Πόρτλαντ. Από τον Μάιο του 2006, όπου και έγινε πραγματική επιχείρηση, το LibraryThing απασχολεί τώρα δεκάδες ανθρώπους.

Το LibraryThing, παίρνει πολλές από τις πληροφορίες του από το Amazon και από πολλές βιβλιοθήκες που παρέχουν ελεύθερη πρόσβαση στις συλλογές τους, σύμφωνα με το πρωτόκολλο Z39.50. Το πρωτόκολλο που χρησιμοποιείται από μια ποικιλία προγραμμάτων γραφείου, κυρίως βιβλιογραφικού λογισμικού όπως το EndNote.

4.2. Διαδικασία εξόρυξης πληροφοριών στο LibraryThing.

Η εφαρμογή που σχεδιάσαμε για την εξόρυξη πληροφοριών από το LibraryThing, στοχεύει να μας δώσει για κάθε τεκμήριο – βιβλίο που επιθυμούμε, όλες τις επισημειώσεις που οι χρήστες του έχουν προσθέσει, καθώς και το πλήθος των εμφανίσεων τους. Στη παρακάτω εικόνα, μέσα στο κίτρινο πλαίσιο, φαίνονται οι επισημειώσεις ενός βιβλίου μαζί με το πλήθος τους σε παρένθεση.



Εικόνα 4.2. Σε μεγέθυνση οι επισημειώσεις του τεκμηρίου με το πλήθος εμφανίσεων τους σε παρένθεση.

Η εφαρμογή, γράφτηκε με την γλώσσα προγραμματισμού Java[10], στην έκδοση 6 και σε περιβάλλον Netbeans[11]. Για την εξόρυξη των δεδομένων μας,

[10] Η **Java** είναι μια αντικειμενοστρεφής γλώσσα προγραμματισμού που σχεδιάστηκε από την εταιρεία πληροφορικής Sun Microsystems [21]. Ένα από τα βασικά πλεονεκτήματα της Java έναντι των περισσότερων άλλων γλωσσών είναι η ανεξαρτησία του λειτουργικού συστήματος και πλατφόρμας. Τα προγράμματα που είναι γραμμένα σε *Java* τρέχουνε ακριβώς το ίδιο σε Windows, Linux, Unix και Macintosh.

[11] Το Netbeans δημιουργήθηκε το 1996 με αρχικό όνομα το Xelfi. Είναι μια πλατφόρμα σχεδίασης προγραμμάτων, με χρήση Java, Javascript, PHP, Python, C, C++ καθώς και άλλες γλώσσες προγραμματισμού. Έχει σχεδιαστεί σε περιβάλλον Java και είναι συμβατό σε Windows, Linus και MacOS[22].

χρησιμοποιήσαμε μια βιβλιοθήκης της Java, την Jsoup. Παρακάτω ακολουθεί μια σύντομη περιγραφή της βιβλιοθήκης.

Για περισσότερες πληροφορίες σχετικά με την εκτέλεση των προγραμμάτων της εργασίας μας, ανατρέξτε στο παράρτημα του παρόντος εγγράφου.

JSOUP: Java HTML Parser

Η jsoup είναι μια βιβλιοθήκη της Java, που εφαρμόζεται πάνω σε κώδικα HTML[12]. Παρέχει μια πολύ βολική διασύνδεση προγραμματισμού εφαρμογών (API) για την εξαγωγή και το χειρισμό των δεδομένων, χρησιμοποιώντας την λογική των DOM[13] και CSS[14] μεθόδων. Η Jsoup δεν έχει εξαρτήσεις με καμία άλλη βιβλιοθήκη και μπορεί να εκτελεστεί σε περιβάλλον με java έκδοσης 1.5 και νεότερο, Scala, Android, OSGi, καθώς και Google App Engine.

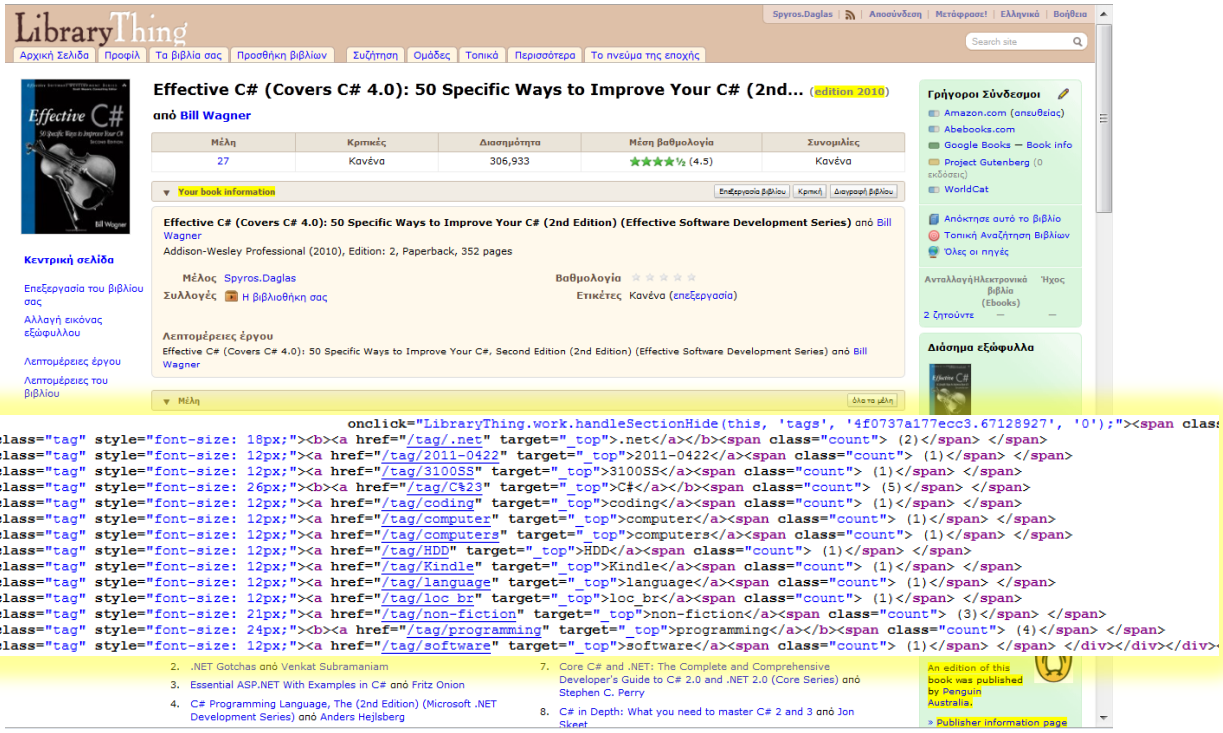
Η jsoup υλοποιεί μια προδιαγραφή της HTML, και αναλύει το έγγραφο HTML σε αρχείο DOM, όπως κάνουν και οι σύγχρονοι φυλλομετρητές.

[12] Η **HTML** (ακρωνύμιο του αγγλικού **H**yper**T**ext **M**arkup **L**anguage, ελλ. Γλώσσα Σήμανσης Υπερκειμένου) είναι η κύρια γλώσσα σήμανσης για τις ιστοσελίδες, και τα στοιχεία της είναι τα βασικά δομικά στοιχεία των ιστοσελίδων. Δημιουργήθηκε από τον Τιμ Μπέρνερς Λι. Τα στοιχεία της HTML χρησιμοποιούνται για να κτίσουν όλους του ιστότοπους. Παρέχει τις μεθόδους δημιουργίας δομημένων εγγράφων (δηλαδή εγγράφων που αποτελούνται από το περιεχόμενο που μεταφέρουν και από τον κώδικα μορφοποίησης του περιεχομένου) καθορίζοντας δομικά σημαντικά στοιχεία για το κείμενο, όπως κεφαλίδες, παραγράφους, λίστες, συνδέσμους, παραθέσεις και άλλα. Μπορούν επίσης να ενσωματώνονται σενάρια εντολών σε γλώσσες όπως η JavaScript, τα οποία επηρεάζουν τη συμπεριφορά των ιστοσελίδων HTML [25].

[13] Το Μοντέλο Αντικειμένου Εγγράφου (DOM) είναι ένα μοντέλο ανεξάρτητο από γλώσσες προγραμματισμού και πλατφόρμας χρήσης, που χρησιμοποιείται στην αντιστοίχηση και την αλληλεπίδραση με αντικείμενα γραμμένα σε HTML, XHTML και XML [26].

[14] Η **CSS** *(Cascading Style Sheets-Διαδοχικά Φύλλα Στυλ)* ή (αλληλουχία φύλλων στύλ) είναι μια γλώσσα υπολογιστή που ανήκει στην κατηγορία των γλωσσών φύλλων στυλ που χρησιμοποιείται για τον έλεγχο της εμφάνισης ενός εγγράφου που έχει γραφτεί με μια γλώσσα σήμανσης. Χρησιμοποιείται δηλαδή για τον έλεγχο της εμφάνισης μιας ιστοσελίδας και γενικότερα ενός ιστοτόπου που γράφτηκε στις γλώσσες HTML και XHTML [27].

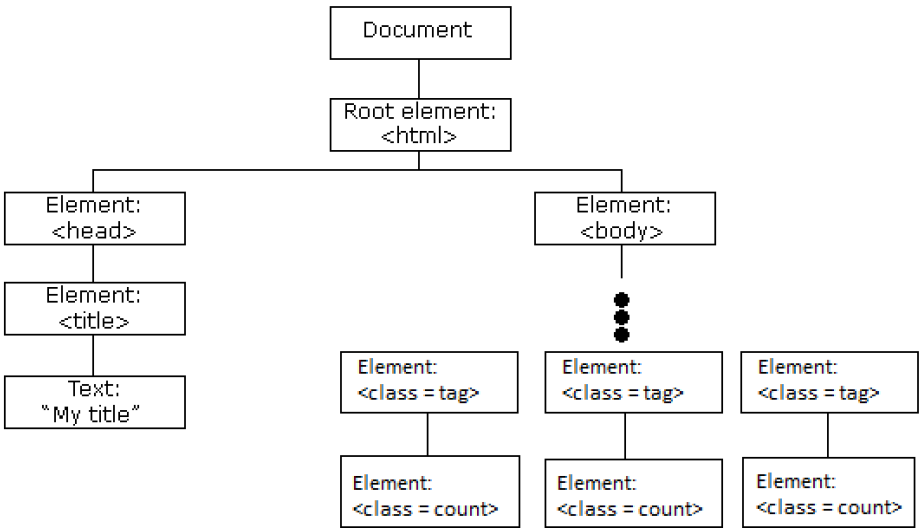
Παρακάτω και με την βοήθεια εικόνων, θα παρουσιάσουμε πως το πρόγραμμά μας, «μεταφράζει» τις ιστοσελίδες για την εξαγωγή των πληροφοριών τους. Στην εικόνα 4.3. βλέπουμε πως ο υπολογιστής, με την χρήση του προγράμματός μας, αντιλαμβάνεται το κομμάτι των πληροφοριών για κάποιο τεκμήριο που έχουμε επιλέξει.



Εικόνα 4.3. Σε μεγέθυνση είναι ο κώδικας HTML των επισημειώσεων του τεκμηρίου και το πλήθος των εμφανίσεων τους.

Στο τμήμα που έχει μεγεθυνθεί, βλέπουμε πως σχεδιάζονται οι πληροφορίες σχετικές με τις επισημειώσεις και του πλήθους εμφανίσεων τους, για το τεκμήριο “Effective C#”. Με τον ίδιο τρόπο εμφανίζονται οι ίδιες πληροφορίες για κάθε τεκμήριο που περιλαμβάνεται στο LibraryThing.

Οι παραπάνω πληροφορίες «μεταφράζονται» από την βιβλιοθήκη Jsoup, στο εξής DOM δέντρο:



Εικόνα 4.4. Διαμόρφωση του DOM tree για τις επισημειώσεις ενός τεκμηρίου.

Κάθε αντικείμενο τάξης επισημείωσης (δηλ. Element: <class = tag>), αντιστοιχίζεται σε μια επισημείωση για το συγκεκριμένο τεκμήριο και συνδέεται με ένα αντικείμενο τάξης αρίθμηση (δηλ. Element: <class = count>). Για παράδειγμα για το τεκμήριο "Effective C#", μια επισημείωση όπως η «C#» είναι ένα αντικείμενο τάξης επισημείωσης και αντιστοιχίζεται με τον αριθμό 5 που είναι ένα αντικείμενο τάξης αρίθμηση. Το ίδιο δέντρο ισχύει για κάθε τεκμήρια αν περιέχει επισημειώσεις.

Με την βοήθεια της βιβλιοθήκης Jsoup, λοιπόν, εμείς προσπαθήσαμε να ανακαλύψουμε το DOM δέντρο, που αναδεικνύει τις επισημειώσεις ενός τεκμηρίου, με σκοπό να τις πάρουμε τοπικά σε ένα αρχείο και να τις επεξεργαστούμε. Το αποτέλεσμα που παίρναμε για κάθε τεκμήριο ήταν το εξής:

110037373	Losing control? : Sovereignty in an age of globalization	Box	17
110037373	Losing control? : Sovereignty in an age of globalization	Economics	1
110037373	Losing control? : Sovereignty in an age of globalization	economy	1
110037373	Losing control? : Sovereignty in an age of globalization	Globalization	3

Στον παραπάνω πίνακα, βλέπουμε τα κάποια από τα αποτελέσματα, όπως τα λάβαμε από το LibraryThing. Στην πρώτη στήλη βρίσκεται το rec_id του τεκμηρίου. Το rec_id, είναι ένας μοναδικός αριθμός για κάθε τεκμήριο, που μας είχε δοθεί από την βιβλιοθήκη του Πάντειου Πανεπιστημίου. Ο αριθμός αυτός ήταν το βασικό κλειδί των πινάκων μας, για να επεξεργαστούμε και να συνδυάσουμε τα αποτελέσματά μας από το LibraryThing και το Google Books. Στην δεύτερη στήλη βρίσκετε ο τίτλος του τεκμηρίου που αναζητούμε, ενώ στην Τρίτη και την τέταρτη βρίσκονται μια επισημείωση που έχει χρησιμοποιηθεί για το τεκμήριο και το πλήθος των χρήσεών της, αντίστοιχα.

Στην αναζήτηση των τεκμηρίων μας, βρεθήκαμε αντιμέτωποι με διάφορα ζητήματα, που σχετίζονταν με τον τρόπο σχεδίασης του LibraryThing και της δομής των πληροφοριών που είχαμε από το Πάντειο Πανεπιστήμιο.

Η αναζήτηση των τεκμηρίων μας από το LibraryThing, έγινε σύμφωνα με τον αριθμό isbn[15] του τεκμηρίου. Ωστόσο, παρατηρήσαμε μια ασυμφωνία σχετικά με τα isbn που είχαμε για τα τεκμήρια από το Πάντειο Πανεπιστήμιο και στα isbn που είχαν τα ίδια τεκμήρια στο LibraryThing και κυρίως στο Google Books. Αποτέλεσμα της παραπάνω διαφοράς των στοιχείων, ήταν να εξετάζουμε τον τίτλο του τεκμηρίου που κάθε φορά μας επέστρεφε το LibraryThing και να τον συγκρίνουμε με τον τίτλο του τεκμηρίου που είχαμε από το Πάντειο Πανεπιστήμιο. Μόνο αν οι τίτλοι ήταν ίδιοι θα προχωρούσαμε στην εξόρυξη των επισημειώσεων τους. Ωστόσο σε κάποιες περιπτώσεις ακόμα και αυτή η σύγκριση δεν ήταν αρκετή. Κάποια τεκμήρια είχαν εισαχθεί στο LibraryThing μαζί με κάποιον υπότιτλο σε παρένθεση ή με κάποια άλλη διαφοροποίηση ανεπαίσθητη στους χρήστες, όχι όμως και στους υπολογιστές. Καταλήξαμε να συγκρίνουμε τους δυο τίτλους, του Πάντειου Πανεπιστημίου και του LibraryThing, σαν σύνολα λέξεων και αν κάποιο σύνολο ήταν υποσύνολο του άλλου

[15] Ο **Διεθνής Πρότυπος Αριθμός Βιβλίου** (αγγλικά: *ISBN - International Standard Book Number*) είναι ένας μηχαναγνώσιμος ραβδοκώδικας ή αλλιώς ένας μηχαναγνώσιμος αριθμός. Ο ραβδοκώδικας (barcode) απαρτίζεται από ψηφία, συγκεκριμένα από αριθμούς και από τον ρωμαϊκό αριθμό X. Σύμφωνα με το πρότυπο ISO-2108 το ISBN είναι ένας 10-ψήφιος αριθμός ο οποίος αποτελείται από τέσσερα τμήματα, τα οποία χωρίζονται μεταξύ τους με ενωτική παύλα ή διάστημα. Τα ψηφία των τριών πρώτων τμημάτων του ISBN μπορούν να είναι ένας αριθμός από το 0 έως το 9, ενώ το τέταρτο τμήμα του ISBN μπορεί να πάρει τιμές από 0 έως 9 καθώς και την τιμή X (X=10). Το ISBN χρησιμοποιείται από εκδοτικούς οίκους, βιβλιοπωλεία, εμπόρους, βιβλιοθήκες, εκπαιδευτικά ιδρύματα, κ.ά [24].

τότε συνεχίζαμε την διαδικασία εξόρυξης των δεδομένων μας, αλλιώς συνεχίζαμε στο επόμενο τεκμήριο.

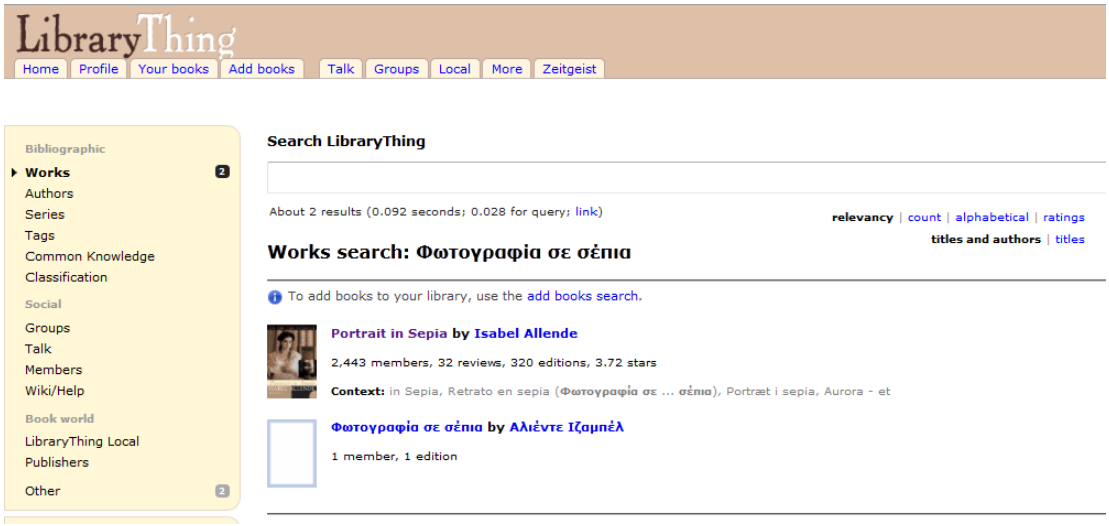
Ακολουθεί ένα παράδειγμα, ενός τεκμηρίου που εντοπίσαμε στο LibraryThing, ενώ ο τίτλος του είχε μια μικρή διαφορά με τον αντίστοιχο τίτλο που είχαμε από το Πάντειο Πανεπιστήμιο. Το τεκμήριο με isbn 1-86064-066-4, είχε τίτλο από το Πάντειο Πανεπιστήμιο τον «An economic history of Russia, 1856-1914», ενώ από το LibraryThing «An economic history of Russia 1856-1914».

Η επιβεβαίωση των τεκμηρίων που θέλαμε να εξετάσουμε, με εκείνα που λαμβάναμε ως απάντηση από το LibraryThing, ήταν ένα μόνο πρόβλημα που αντιμετωπίσαμε κατά την διάρκεια της εργασίας μας.

Μια ακόμα περίπτωση, που έπρεπε να εξετάσουμε κατά τον σχεδιασμό του προγράμματός μας, ήταν η ανακατεύθυνση των διευθύνσεων των τεκμηρίων που ζητούσαμε από το LibraryThing, στις κατάλληλες εσωτερικές διευθύνσεις που το LibraryThing επέλεγε. Πιο συγκεκριμένα, το LibraryThing, για κάθε τεκμήριο έχει μια διεύθυνση με κωδικό πεδίο “work”, ενώ επιτρέπει την αναζήτηση των τεκμηρίων του με κωδικά πεδία τα “isbn” και “title”. Για παράδειγμα, στο βιβλίο που αναφέραμε και προηγουμένως, «An economic history of Russia 1856-1914», το LibraryThing έχει δώσει την εσωτερική διεύθυνση με κωδικό πεδίο το “work” http://www.librarything.com/work/8639292, ενώ ανακατευθύνει τους χρήστες στην ίδια διεύθυνση, αν πληκτρολογήσουν την διεύθυνση με κωδικό πεδίο το “isbn” http://www.librarything.com/isbn/1-86064-066-4 και την διεύθυνση με κωδικό πεδίο το “title” http://www.librarything.com/title/An%20economic%20history%20of%20Russia%201856-1914.

Ωστόσο, σε πολλές περιπτώσεις, για τον ίδιο τίτλο ή για το ίδιο isbn, το LibraryThing, έχει περισσότερες από μια πιθανές ανακατευθύνσεις. Το γεγονός αυτό συμβαίνει καθώς για αρκετά τεκμήρια, έχει δοθεί το ίδιο isbn και για τις μεταφράσεις του τεκμηρίου σε άλλες γλώσσες και για τις νεότερες εκδόσεις του. Βεβαίως υπάρχουν και κάποια τεκμήρια που έχουν 2 ή και περισσότερα αντίτυπα στο LibraryThing. Για όλες αυτές τις περιπτώσεις, το πρόγραμμά μας δεν μπορούσε να γνωρίζει σε πιο τεκμήριο γινόταν η ανακατεύθυνση και κάθε φορά δεχόταν σαν σωστή επιλογή την επιλογή που επέστρεφε το LibraryThing. Και σε αυτήν την περίπτωση, ο έλεγχος του τίτλου του τεκμηρίου, όπως αναφέρθηκε προηγουμένως,

ήταν η λύση που ελαχιστοποιούσε τα σφάλματα. Στην πραγματικότητα, δεν ήταν σφάλμα η εξόρυξη των πληροφοριών από ένα τεκμήριο διαφορετικής έκδοσης από το αντίστοιχο της βιβλιοθήκης του Πάντειου Πανεπιστημίου, ωστόσο ήταν σφάλμα η εξόρυξη των πληροφοριών μας από τεκμήριο μεταφρασμένο από το αντίστοιχο του Πάντειου Πανεπιστημίου. Στην παρακάτω εικόνα, παρουσιάζουμε ένα χαρακτηριστικό παράδειγμα, του παραπάνω προβλήματος. Ενώ εμείς δίνουμε σαν είσοδο την διεύθυνση http://www.librarything.com/isbn/960-410-212-5 και σύμφωνα με την βιβλιοθήκη του Πάντειου Πανεπιστημίου, περιμένουμε το τεκμήριο «Φωτογραφία σε σέπια» από την Ιζαμπέλ Αλιέντε, το LibraryThing μας ανακατευθύνει στην αρχική έκδοση του βιβλίου στην αγγλική γλώσσα, "Portrait in Sepia by Isabel Allende", καθώς την έκδοση αυτή έχει καταχωρημένη ως πρώτη στον κατάλογό του.



Εικόνα 4.6. Παράδειγμα λανθασμένης ανακατεύθυνσης από το LibraryThing.

Παρακάτω, παρουσιάζουμε τον αλγόριθμο που ακολουθήσαμε για την αναζήτηση των τεκμηρίων μας, στο LibraryThing και την εξόρυξη των κατάλληλων δεδομένων (αλγόριθμος 2).

4.2.1. Θέματα υλοποίησης της ανάκτησης δεδομένων από το LibraryThing.

Καθότι στην κατοχή μας είχαμε περίπου 8000 τεκμήρια προς επεξεργασία και εξόρυξη των δεδομένων τους, ήταν φανερή η αναγκαιότητα δημιουργίας ενός προγράμματος που θα έκανε αυτόματα όλες τις διαδικασίες. Ωστόσο, το LibraryThing, όπως και πολλές άλλες ιστοσελίδες, έχουν ειδικούς μηχανισμούς ασφαλείας, με σκοπό την αποφυγή χρήσης των υπηρεσιών τους, από λογισμικού αυτόματης λειτουργίας[16]. Οι μηχανισμοί αυτοί, εντόπιζαν το πρόγραμμά που είχαμε σχεδιάσει, και σταματούσαν την επικοινωνία μαζί μας, μέσω της ηλεκτρονική μας διεύθυνσης IP[17].

```
ψάξε με το ISBN{
        πάρε τον τίτλο του βιβλίου που επιστρέφει
        το LibraryThing = LTtitle;
        δες αν αυτός ο τίτλος είναι ο ίδιος με εκείνο
        του OPACIAL = OPACtitle;

        if(LTtitle == OPACtitle ) {
                Πάρε τα κατάλληλα δεδομένα;
        }
        else {
                πάρε τον τίτλο που λέει οτι ειναι το
                isbn στο έγγραφο του OPACIAL;
                κάνε αναζήτηση με βάση τον τίτλο;
                αν (το βιβλίο με τον τιτλο υπάρχει)
                {
                    Πάρε τα κατάλληλα δεδομένα;
                }
        αλλιώς {
                τύπωσε ότι το βιβλίο δεν υπάρχει
                }

        }
```

Αλγόριθμος 2.

Στην πραγματικότητα, το LibraryThing, επιτρέπει την αυτόματη εξόρυξη κάποιον πληροφοριών του, μέσω συγκεκριμένων βιβλιοθηκών που έχουν αναπτυχθεί για αυτό, ωστόσο καμία βιβλιοθήκη δεν ήταν εξ ολοκλήρου σχεδιασμένη στις δικές μας απαιτήσεις. Με την χρήση περισσοτέρων των 2 βιβλιοθηκών από το LibraryThing, ενδεχομένως να είχαμε το αποτέλεσμα που θέλαμε, αλλά αυτή η

[16] Οι μηχανισμοί αυτοί δημιουργήθηκαν για να αποκλείσουν κακόβουλο λογισμικό να χρησιμοποιεί πόρους ενός συστήματος και ενός εξυπηρετητή άδικα, καθώς επίσης και να αποτρέψουν την «κλοπή» δεδομένων μεγάλης κλίμακας μιας ιστοσελίδας.

[17] Μία **διεύθυνση IP** (***Ip address*** - *Internet Protocol address*), είναι ένας μοναδικός αριθμός που χρησιμοποιείται από συσκευές για τη μεταξύ τους αναγνώριση και συνεννόηση σε ένα δίκτυο υπολογιστών που χρησιμοποιεί το Internet Protocol standard. Κάθε συσκευή που ανήκει στο δίκτυο - όπως επίσης δρομολογητές (routers), υπολογιστές, time-servers, εκτυπωτές, μηχανές για fax μέσω Internet, και ορισμένα τηλέφωνα - πρέπει να έχει τη δική της μοναδική διεύθυνση. Μία διεύθυνση IP μπορεί να θεωρηθεί το αντίστοιχο μιας διεύθυνσης κατοικίας ή ενός αριθμού τηλεφώνου (σύγκριση με VoIP) για έναν υπολογιστή ή άλλη συσκευή δικτύου στο Διαδίκτυο. Όπως κάθε διεύθυνση κατοικίας και αριθμός τηλεφώνου αντιστοιχούν σε ένα και μοναδικό κτίριο ή τηλέφωνο, μια IP address χρησιμοποιείται για τη μοναδική αναγνώριση ενός υπολογιστή ή άλλης συσκευής που συνδέεται στο δίκτυο [23].

κίνηση θα μας στοίχιζε πολύτιμο χρόνο, τον οποίο δεν διαθέταμε στα πλαίσια της διπλωματικής μας εργασίας. Στην εικόνα 4.5 παρουσιάζουμε κάποιες από τις βιβλιοθήκες που προτείνει προς χρήση το LibraryThing, ενώ στην συνέχεια θα αναλύσουμε τον σχεδιασμό του προγράμματός μας, με σκοπό να ξεπεράσουμε το παραπάνω πρόβλημα.



APIs (Application Programming Interfaces)

1. JSON Books API

See LibraryThing JSON Books API

- License: Must be run as Javascript on user's browser, not fetched by a server; cannot be stored, except for browser caching. Commercial use requires a hard (non-JS) link to LibraryThing on every page that returns results.
- Developer key: Not needed.

2. JSON Works API

See http://www.librarything.com/thingology/2008/03/first-cut-works-json-api.php

- License: Must be run as Javascript on user's browser, not fetched by a server; cannot be stored, except for browser caching. Commercial use requires a hard (non-JS) link to LibraryThing on every page that returns results.
- Developer key: Not needed.

3. ThingISBN

ThingISBN takes an ISBN and returns a list of "related" ISBNs—ISBNs from the same "work" (i.e. other editions and translations).

- Format is identical to OCLC's xISBN.
- Sample URL: http://www.librarything.com/api/thingISBN/0441172717
- Read the announcement and a follow-up
- License: Non-commercial use (see below)
- Developer key: Not needed.
- See also Projects currently using ThingISBN

4. What Work?

"What work?" takes an ISBN and/or a title-author and returns the LibraryThing work number.

- See the blog post
- License: Free to all with attribution (see below)

Εικόνα 4.5. Μερικές βιβλιοθήκες για εξόρυξη δεδομένων από το LibraryThing (βλ. http://www.librarything.com/wiki/index.php/LibraryThing_APIs).

Όπως αναφέραμε και προηγουμένως, το LibraryThing έχει μηχανισμούς αποτροπής αυτόματης εκτέλεσης λογισμικού για την εκμετάλλευση του bandwidth του διακομιστή, καθώς και της μαζικής λήψης των πληροφοριών που περιέχει. Οι μηχανισμοί αυτοί έχουν κάποια χρονικά όρια μέσα στα οποία χαρακτηρίζουν κάποια διεύθυνση IP στις παρακάτω κατηγορίες:

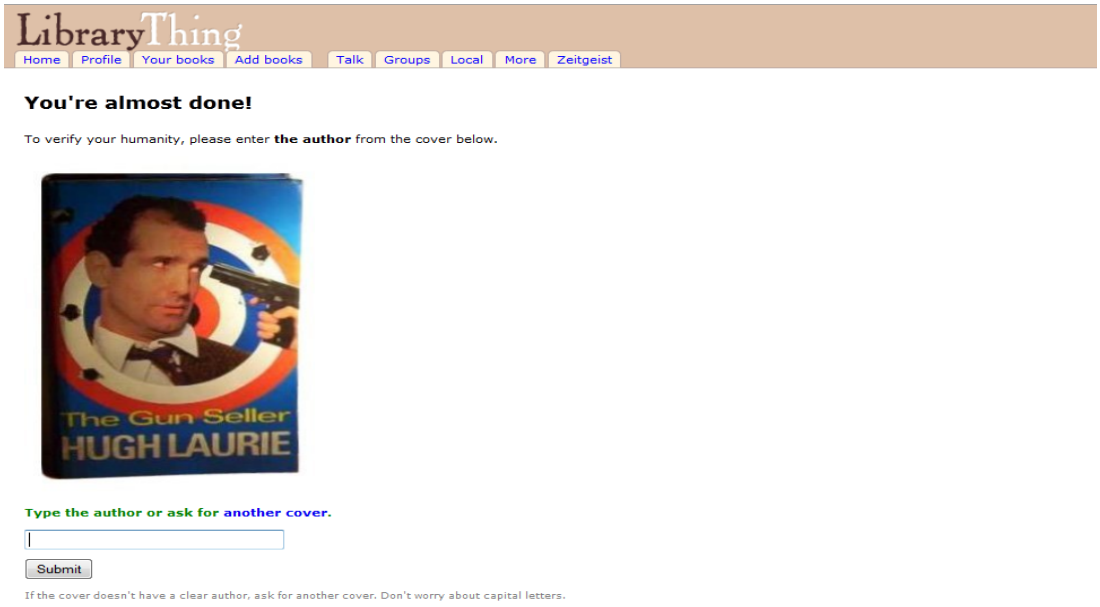
1. Αν χρησιμοποιείται από ενδεχομένως κακόβουλο λογισμικό,
2. αν υπάρχει η πιθανότητα να χρησιμοποιείται από ενδεχομένως κακόβουλο λογισμικό, ή
3. αν χρησιμοποιείται απευθείας από κάποιον άνθρωπο.

Γνωρίζοντας αυτήν την παράμετρο επιλέξαμε μεταξύ των αναζητήσεων κάθε τεκμηρίου να περιμένουμε μερικά δευτερόλεπτα, για να μην αποκλειστούμε από το LibraryThing. Πειραματιζόμενοι με την μέθοδο χρονικής καθυστέρησης,

παρατηρήσαμε και άλλες παραμέτρους που κατατάσσουν τους χρήστες σε μια από τις παραπάνω 3 κατηγορίες. Ωστόσο καμία από αυτές τις παραμέτρους, δεν ήταν αρκετή για να μας αναγνωρίσει το LibraryThing ως κακόβουλο λογισμικό.

Στην πραγματικότητα το LibraryThing, κάθε φορά που εκτελούσαμε το πρόγραμμά μας, μας κατέτασσε στην δεύτερη κατηγορία των εν δυνάμει χρηστών με κακόβουλο λογισμικό. Έτσι λοιπόν, αποφασίσαμε πως η πιο αποδοτική χρονική καθυστέρηση, ανάλογη με τον χρόνο εκτέλεσης του προγράμματος για το σύνολο των 8000 τεκμηρίων, ήταν τα 3 δευτερόλεπτα. Οποιαδήποτε καθυστέρηση μικρότερη των 3 δευτερολέπτων θα μας κατέτασσε στην κατηγορία του κακόβουλου λογισμικού και θα μας απέκλειε από την χρήση των υπηρεσιών του LibraryThing, ενώ οποιαδήποτε μεγαλύτερη χρονική καθυστέρηση θα μεγάλωνε τον χρόνο ολοκλήρωσης του προγράμματος χωρίς ιδιαίτερο όφελος.

Η επιλογή της χρονικής καθυστέρησης εκτέλεσης, μεταξύ των αναζητήσεων των τεκμηρίων μας, ήταν το πρώτο βήμα για την σχεδίαση του προγράμματός μας. Στην συνέχεια έπρεπε να βρούμε έναν τρόπο, ώστε κάθε φορά που το LibraryThing, υποπτευόταν πως ήμασταν κακόβουλο λογισμικό, εμείς να αποδεικνύαμε το αντίθετο. Όπως οι περισσότερες ιστοσελίδες, έτσι και το LibraryThing, έχει για τον σκοπό αυτό έναν μηχανισμό «ανθρώπινης επιβεβαίωσης» (Human Verification). Το μήνυμα αυτό εμφανιζόταν σε εμάς κάθε 17 αναζητήσεις, ενώ σε σύνολο 8000 τεκμηρίων, συνολικά το αντικρίσαμε 500 φορές. Ο μηχανισμός «ανθρώπινης επιβεβαίωσης» του LibraryThing, είχε την παρακάτω μορφή:



Εικόνα 4.7. Μηχανισμός «ανθρώπινης επιβεβαίωσης» του LibraryThing.

Για να επιβεβαιώσει κάποιος χρήστης την ταυτότητά του πρέπει να γράψει τον συγγραφέα του βιβλίου που φαίνεται στην εικόνα. Στην περίπτωση που βλέπουμε πρέπει ο χρήστης να πληκτρολογήσει το όνομα "Hugh Laurie" και να επιλέξει την επιλογή "Submit". Φυσικά κάθε φορά το εξώφυλλο του βιβλίου για την επιβεβαίωση αλλάζει.

Για να χρησιμοποιήσουμε αυτήν την μέθοδο κατά την διάρκεια των αναζητήσεων των τεκμηρίων που θέλουμε, δημιουργήσαμε μια μέθοδο η οποία για κάθε σελίδα που μας ανακατευθύνει το LibraryThing, διαβάζει τον τίτλο της[18]. Αν ο τίτλος που μας επιστρέφεται είναι ο "Human Verification LibraryThing", καταλαβαίνουμε ότι το σύστημα μας υποπτεύεται ως κακόβουλο λογισμικό, με αποτέλεσμα να απαιτείται από εμάς σύνδεση με το LibraryThing και η πληκτρολόγηση του συγγραφέα της μεθόδου «ανθρώπινης επιβεβαίωσης» με το χέρι. Στην συνέχεια, αφού γίνει η επιβεβαίωση και ενημερώσουμε το πρόγραμμά μας για αυτό, τότε εκείνο προσπαθεί και πάλι να συνδεθεί στο ίδιο τεκμήριο και να συνεχίσει την λειτουργία του. Παράδειγμα της εκτέλεσης του προγράμματός μας, όταν απαιτείται η μέθοδος της «ανθρώπινης επιβεβαίωσης» φαίνεται παρακάτω:

```
4)   http://www.librarything.com/isbn/0-415-05218-1.html
HumanVerificationLibraryThing
VERIFY THAT YOU ARE A HUMAN BEING...

StardomIndustryofDesirebyCGledhillLibraryThing
0-415-05218-1
Stardom : industry of desire  FROM OPACIAL
school 1
{}
```

Στην πρώτη γραμμή βλέπουμε την σελίδα που προσπαθεί να συνδεθεί το πρόγραμμά μας, με κωδικό πεδίο τον isbn του τεκμηρίου, όπως μας δίνεται από το Πάντειο Πανεπιστήμιο. Ο αριθμός αριστερά είναι ένας μετρητής για το πόσα τεκμήρια έχουμε ήδη αναζητήσει. Στην δεύτερη σειρά βλέπουμε τον τίτλο της σελίδας που μας επιστρέφετε, χωρίς κενά μεταξύ των λέξεων. Στην συνέχεια το σύστημα μας ενημερώνει πως πρέπει να επιβεβαιώσουμε την ανθρώπινη φύση μας και στην συνέχεια αφού ολοκληρώσουμε την επιβεβαίωση και το ενημερώσουμε

[18] Ο τίτλος της σελίδας διαφέρει από τον τίτλο του τεκμηρίου, για την μέθοδο που ορίσαμε σε προηγούμενο σημείο. Ο τίτλος της σελίδας «ανθρώπινης επιβεβαίωσης» είναι σταθερός.

συνεχίζει την αναζήτηση του τεκμηρίου της διεύθυνσης που του είχαμε δώσει. Επιστέφει, δηλαδή, τον τίτλο της σελίδας που ανακατευθυνθήκαμε, στην συνέχεια επιστρέφει για επιβεβαίωση το isbn του τεκμηρίου, καθώς και τον τίτλο του τεκμηρίου από το Πάντειο Πανεπιστήμιο. Αν ο τίτλος είναι ο επιθυμητός, τότε προχωράμε στην εξόρυξη των δεδομένων που επιθυμούμε. Στην προκειμένη περίπτωση, είδαμε πως για το τεκμήριο "Stardom : industry of desire" πήραμε την επισημείωση "school" η οποία χρησιμοποιήθηκε 1 φορά.

Τέλος για την εφαρμογή μας, προβλέπουμε την συνέχεια της εκτέλεσης της, ακόμα και αν για οποιονδήποτε δικτυακό λόγο, κάποιος από τους κόμβους σύνδεσης, είτε ενδιάμεσος είτε ακριανός, δεν είναι σε λειτουργία.

Το πρόγραμμά μας παράγει ένα αρχείο κειμένου, για διευκόλυνση των επόμενων διεργασιών που θα εκτελέσουμε. Παράδειγμα του παραγόμενου αρχείου σας παρουσιάζουμε παρακάτω. Περισσότερα για την εκτέλεση των προγραμμάτων μας, παρουσιάζονται στο παράρτημα του παρόντος εγγράφου.

10037486	The making of the modern body : sexuality and society in the nineteenth century	19th century	3
10037486	The making of the modern body : sexuality and society in the nineteenth century	biology	2

Η διαδικασία εξόρυξης πληροφοριών από το Google Books, είναι παρόμοια με την διαδικασία για το LibraryThing. Στην παρούσα έρευνα θα παρουσιάσουμε, περιληπτικά κάποια στοιχεία της διαδικασίας, ενώ για περισσότερες πληροφορίες θα σας προτείνουμε τις διπλωματικές εργασίες των μεταπτυχιακών φοιτητών Κουμάκη Μαρίνας και Κακαβούλη Διονύσιου.

4.3. Σύντομη περιγραφή του Google Books.

"Πρόκειται για ένα φανταστικό, και μάλιστα δωρεάν, εργαλείο μάρκετινγκ, που διασφαλίζει τη μεγαλύτερη δυνατή προβολή των βιβλίων μας."

--Eduardo Mirón

Το Google Books (παλαιότερα γνωστό ως Google Book Search και Google Print) είναι μια υπηρεσία από την Google Inc που αναζητά το πλήρες κείμενο των βιβλίων που η Google έχει σαρώσει, έχει μετατρέψει σε κείμενο με τη χρήση οπτικής αναγνώρισης χαρακτήρων, και έχει αποθηκεύσει στην ψηφιακή βάση δεδομένων του [28].

Τα αποτελέσματα από το Google Books εμφανίζονται στην ιστοσελίδα Google Books (books.google.com). Εγγεγραμμένοι χρήστες μπορούν με ένα κλικ σε ένα αποτέλεσμα από το Google Books να δουν σελίδες από το βιβλίο, εάν το βιβλίο δεν προστατεύεται από πνευματικά δικαιώματα ή αν ο κάτοχος των πνευματικών δικαιωμάτων έχει δώσει την άδεια. Βιβλία στο δημόσιο τομέα είναι διαθέσιμα σε "πλήρη επισκόπηση" και είναι δωρεάν για κατέβασμα. Για τα άλλα Βιβλία, η Google περιορίζει τον αριθμό των σελίδων που μπορούν να προβληθούν μέσα από μια ποικιλία μέτρων περιορισμών πρόσβασης και μέτρων ασφάλειας, κάποια εκ των οποίων είναι με βάση τις συνήθειες του χρήστη [28].

Τα περισσότερα σαρωμένα έργα δεν υπάρχουν πλέον σε έντυπη μορφή και δεν διατίθενται στο εμπόριο. Για εκείνα που είναι, η ιστοσελίδα παρέχει διασυνδέσεις με το δικτυακό τόπο του εκδότη και με βιβλιοπωλεία.

Η πρωτοβουλία αυτή έχει αναγνωριστεί για τη δυνατότητά της να προσφέρει άνευ προηγουμένου πρόσβαση σε κάτι που μπορεί να γίνει η μεγαλύτερη διαδικτυακή βάση δεδομένων ανθρώπινης γνώσης και να προωθήσει του εκδημοκρατισμό της γνώσης.

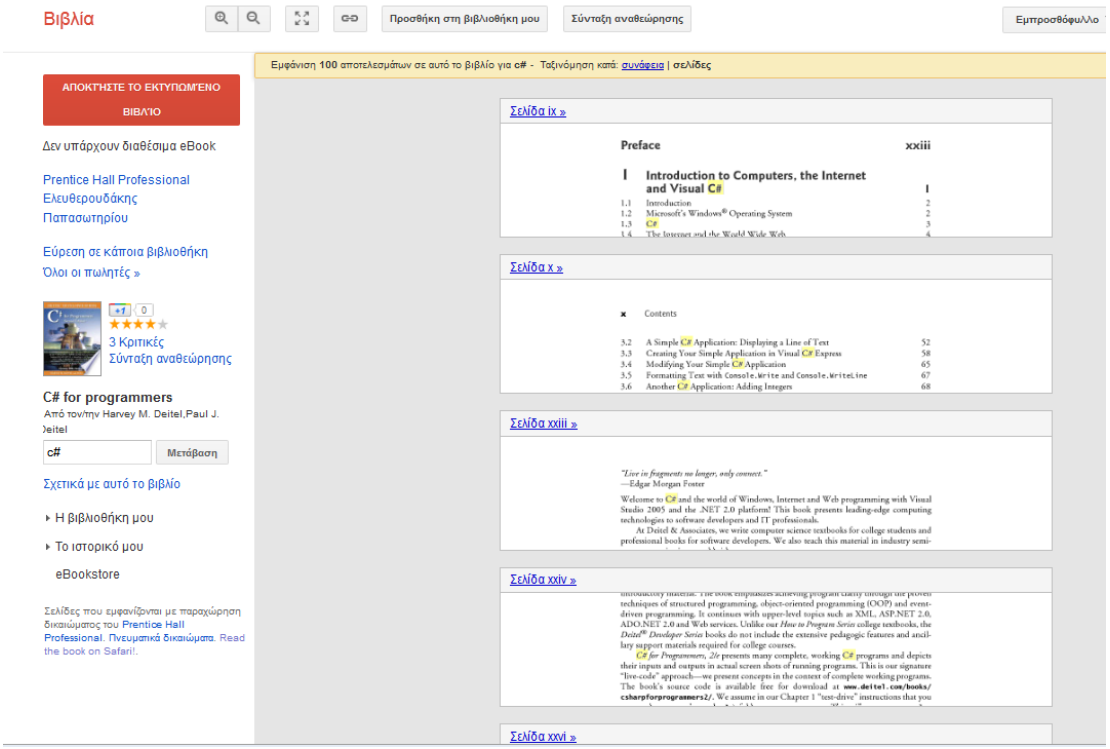
Από το 2010, ο αριθμός των σαρωμένων βιβλίων είναι πάνω από 15 εκατομμύρια. Η Google εκτιμά ότι το 2010 υπήρχαν περίπου 130 εκατομμύρια μοναδικά βιβλία στον κόσμο, και δήλωσε πως έχει την πρόθεση να τα σαρώσει όλα μέχρι το τέλος της δεκαετίας [28].

4.4. Διαδικασία εξόρυξης πληροφοριών στο Google Books.

Η διαδικασία εξόρυξης πληροφοριών στο Google Books, δεν διαφέρει στην βάση της, με την αντίστοιχη διαδικασία του LibraryThing. Ωστόσο υπάρχουν κάποια σημαντικά σημεία που αξίζει να αναφέρουμε.

- Οι επισημειώσεις των τεκμηρίων στο Google Books, δεν φαίνονται σε κάποιο συγκεκριμένο σημείο, που θα μπορούσε να προσδιοριστεί με ένα DOM δέντρο. Αντίθετα, η εφαρμογή περιλαμβάνει ένα κουτί διαλόγου, στο οποίο ο χρήστης πληκτρολογεί όποιον όρο–επισημείωση επιθυμεί και η εφαρμογή αποκρίνεται αν η επισημείωση υπάρχει και πόσες φορές.
- Το πλήθος με το οποίο χαρακτηρίζεται με επισημείωση στο Google Books, δεν δείχνει πόσες φορές οι χρήστες την έχουν χρησιμοποιήσει, όπως το LibraryThing, αλλά δείχνει το πλήθος των σελίδων, που η συγκεκριμένη επισημείωση εμφανίζεται.
- Η αναζήτηση των τεκμηρίων, γίνεται με την χρήση του isbn, ενώ αν δεν βρεθεί, τότε επαναλαμβάνεται με τον συγγραφέα και τον τίτλο του τεκμηρίου. Το LibraryThing, στην σελίδα του τεκμηρίου, έχει διασύνδεση στην αντίστοιχη σελίδα του Google Books, μαζί με το isbn όπως δίνεται από το Google Books. Έτσι λοιπόν, για αρκετά τεκμήρια που δεν βρέθηκαν στο Google Books με το isbn που πήραμε από το Πάντειο Πανεπιστήμιο, χρησιμοποιήσαμε το isbn που έδινε το LibraryThing.

Η διεπαφή του Google Books, φαίνεται στην παρακάτω εικόνα.



Εικόνα 4.8. Διεπαφή Google Books.

Για περισσότερες πληροφορίες μπορείτε να απευθυνθείτε στις διπλωματικές εργασίες των Κουμάκη Μαρίνα και Κακαβούλη Διονύσιου.

Μετά το πέρας των παραπάνω διεργασιών, καταφέραμε να εξορύξουμε 41470 επισημειώσεις από 4018 τεκμήρια. Στο κεφάλαιο που ακολουθεί, ολοκληρώνουμε την έρευνάς μας, με τα πειράματα που εκτελέσαμε για την ολοκλήρωση του συστήματός μας.

"Search engines compile a lot of data on what we search for and much of what we search for is personal. Many people fear that the government wants to do data mining. It'll be interesting to see how much of all this ultimately becomes a political issue."
--Danny Sullivan

5 Προτεινόμενο Μοντέλο – Βήμα 3

Η συλλογή των δεδομένων μας, από το LibraryThing και το Google Books, ολοκληρώθηκε με τις διεργασίες που εξετάσαμε μέχρι τώρα. Στο κεφάλαιο αυτό θα παρουσιάσουμε 3 μετρικές με τις οποίες πειραματιστήκαμε στην συλλογή των δεδομένων μας και τα αποτελέσματα που αυτές είχαν.

5.1. Εύρεση TF-IDF.

Το πρώτο βήμα στην εφαρμογή των πειραμάτων μας, ήταν ο υπολογισμός του σχήματος στάθμισης όρων tf-idf. Το tf-idf μέτρο βάρους, χρησιμοποιείται συχνά στην ανάκτηση πληροφοριών. Το βάρος αυτό είναι ένα στατιστικό μέτρο που χρησιμοποιείται για να εκτιμηθεί πόσο σημαντικός είναι ένας όρος σε ένα τεκμήριο μιας συλλογής.

Το μέτρο εκτίμησης του πόσο σημαντικός είναι ένας όρος αυξάνει αναλογικά με το πλήθος που ο όρος εμφανίζεται στο τεκμήριο, αλλά αντισταθμίζεται από τη συχνότητά του στη συλλογή. Παραλλαγές του σχήματος tf-idf χρησιμοποιούνται συχνά από τις μηχανές αναζήτησης ως ένα κεντρικό εργαλείο για την βαθμολόγηση και την κατάταξη της σχετικότητας ενός εγγράφου που δίνεται για κάποιο ερώτημα χρήστη. Το σχήμα tf-idf μπορεί να χρησιμοποιηθεί επιτυχώς για φιλτράρει stop-words σε διάφορα θεματικά πεδία, συμπεριλαμβανομένων περιλήψεων κειμένων και ταξινομήσεων [29].

Στο σχήμα tf-idf, ο όρος tf, ορίζει την βαρύτητα μιας επισημείωσης σε ένα συγκεκριμένο τεκμήριο. Ισοδυναμεί με το πηλίκο των εμφανίσεων του όρου στο τεκμήριο, δια του πλήθους των εμφανίσεων όλων των όρων στο ίδιο τεκμήριο. Πιο συγκεκριμένα, έστω f_{ij} = η συχνότητα της επισημείωσης i στο τεκμήριο j, τότε η κανονικοποίηση της συχνότητα όρου (term frequency, tf), δίνεται ως,

$tf_{ij} = f_{ij} / \max\{f_{ij}\}$.

Οι όροι που εμφανίζονται σε πολλά διαφορετικά κείμενα είναι λιγότερο ενδεικτικοί του θέματος. Θεωρούμε df_i τον αριθμό των τεκμηρίων που περιλαμβάνουν την επισημείωση i, δηλαδή, τον αριθμό των τεκμηρίων που περιλαμβάνουν τον όρο i. Ακόμα, θεωρούμε idf_i την αντίστροφη συχνότητα τεκμηρίου του όρου i, η οποία ισούται με $\log_2(N/df_i)$, όπου N ο συνολικός αριθμός τεκμηρίων. Ο λογάριθμος χρησιμοποιείται για να μετριαστεί η επίδραση του όρου tf [30].

Έτσι λοιπόν, συνοπτικά, θεωρούμε πως το σχήμα στάθμισης όρων tf-idf ισούται με: $W_{ij} = tf_{ij}idf_i = tf_{ij} \log_2(N/df_i)$. Θεωρητική απόδειξη του σχήματος υπάρχει από τον Panineni[31]. Παρακάτω δίνουμε και ένα παράδειγμα υπολογισμού του σχήματος tf-idf:

Παράδειγμα: Υπολογισμός TF-IDF

- Ένα κείμενο περιέχει τους ακόλουθους όρους με συχνότητα:
  - A(3), B(2), C(1)
- Υποθέτουμε ότι η συλλογή περιλαμβάνει 10.000 κείμενα και ότι οι συχνότητες κειμένων αυτών των όρων είναι:
  - A(50), B(1300), C(250)
- Τότε:
  - A: tf = 3/3; idf = log(10000/50) = 5.3; tf-idf = 5.3
  - B: tf = 2/3; idf = log(10000/1300) = 2.0; tf-idf = 1.3
  - C: tf = 1/3; idf = log(10000/250) = 3.7; tf-idf = 1.2

Εικόνα 4.9. Παράδειγμα tf-idf [30].

Έτσι λοιπόν, υπολογίσαμε το σχήμα tf-idf, για τις επισημειώσεις, ξεχωριστά για εκείνες που πήραμε από το LibraryThing και από το Google Books.

Κατόπιν, δημιουργήσαμε έναν πίνακα στον οποίο συνδυάσαμε τα παραπάνω αποτελέσματα. Πιο συγκεκριμένα υπολογίσαμε μια τιμή tf-idf, η οποία προέρχεται από τις δυο επιμέρους τιμές που περιγράψαμε προηγουμένως, με συγκεκριμένο βάρος κάθε φορά, δηλαδή, $tfidf = (a * tfidf_{LibraryThing}) + (b * tfidf_{Google\ Books})$. Οι τιμές που δώσαμε στα a και b, είναι από το 0 έως το 100, αντίστροφες μεταξύ τους. Παρακάτω, παρουσιάζουμε τον πίνακα που δημιουργήσαμε για μερικές από τις επισημειώσεις που υπολογίσαμε.

	REC ID	TAG	GB tfidf	LT tfidf	100%GB-	90%GB -	80%-20%	70%-30%	60%-40%	50%-50%	40%-60%	30%-70%	20%-80%	10%-90%	0%-100%
2	10011095	textbook - sacred texts	0	1,69689	0	0,16969	0,33938	0,50907	0,67876	0,84845	1,01814	1,18782	1,35751	1,5272	1,69689
3	10037815	advocacy*	7,386	7,38585	7,386	7,38599	7,38597	7,38596	7,38594	7,38593	7,38591	7,3859	7,38588	7,38587	7,38585
4	10035721	advocacy*	7,386	1,84646	7,386	6,83205	6,27809	5,72414	5,17019	4,61623	4,06228	3,50832	2,95437	2,40042	1,84646
5	10048574	gender inequality	0	1,94783	0	0,19478	0,38957	0,58435	0,77913	0,97391	1,1687	1,36348	1,55826	1,75305	1,94783
6	10013628	gender inequality	0	1,94783	0	0,19478	0,38957	0,58435	0,77913	0,97391	1,1687	1,36348	1,55826	1,75305	1,94783
7	10061635	dystopia	0	1,55826	0	0,15583	0,31165	0,46748	0,62331	0,77913	0,93496	1,09078	1,24661	1,40244	1,55826
8	10064230	Customer Service / Loyalty	0	0	0	0	0	0	0	0	0	0	0	0	0
9	10055365	Sandra*	0	4,24223	0	0,42422	0,84845	1,27267	1,69689	2,12112	2,54534	2,96956	3,39379	3,81801	4,24223
10	10056885	philosophy of mind*	0	0	0	0	0	0	0	0	0	0	0	0	0
11	10058229	philosophy of mind*	0	1,52164	0	0,15216	0,30433	0,45649	0,60866	0,76082	0,91299	1,06515	1,21731	1,36948	1,52164
12	10019919	philosophy of mind*	0	0	0	0	0	0	0	0	0	0	0	0	0
13	10032365	philosophy of mind*	0,11565	0	0,11565	0,10409	0,09252	0,08096	0,06939	0,05783	0,04626	0,0347	0,02313	0,01157	0
14	10060795	philosophy of mind*	0	0,14579	0	0,01458	0,02916	0,04374	0,05831	0,07289	0,08747	0,10205	0,11663	0,13121	0,14579
15	10043741	philosophy of mind*	0,42609	0	0,42609	0,38348	0,34087	0,29826	0,25565	0,21305	0,17044	0,12783	0,08522	0,04261	0
16	10046145	philosophy of mind*	0	2,02886	0	0,20289	0,40577	0,60866	0,81154	1,01443	1,21731	1,4202	1,62308	1,82597	2,02886
17	10043431	philosophy of mind*	0,18261	3,04328	0,18261	0,46868	0,75474	1,04081	1,32688	1,61295	1,89901	2,18508	2,47115	2,75722	3,04328
18	10043436	philosophy of mind*	0,48696	0	0,48696	0,43826	0,38957	0,34087	0,29218	0,24348	0,19478	0,14609	0,09739	0,0487	0
19	10062504	disabilities*	0,11878	2,82815	0,11878	0,38971	0,66065	0,93159	1,20253	1,47347	1,7444	2,01534	2,28628	2,55722	2,82815
20	10046346	Staatsrecht*	2,79972	8,48446	2,79972	3,36819	3,93667	4,50514	5,07362	5,64209	6,21057	6,77904	7,34751	7,91599	8,48446

Πίνακας 1. tfidf =(a*tfidf $_{LibraryThing}$) +(b* tfidf $_{Google\ Books}$).

Με τον παραπάνω πίνακα, οι βιβλιοθηκονόμοι έχουν την δυνατότητα να επιλέξουν την τιμή του βάρους a (0<=a<=1) που επιθυμούν να δώσουν στα

αποτελέσματα tfidf του LibraryThing και την τιμή του βάρους b (0<=b<=1) που επιθυμούν να δώσουν στα αποτελέσματα tfidf του Google Books. Είναι στην κρίση κάθε βιβλιοθηκονόμου να επιλέξει αν θεωρεί σημαντικότερες τις επισημειώσεις από τους χρήστες του LibraryThing, ή την χρήση του ίδιου όρου, από τον συγγραφέα του τεκμηρίου.

5.2. Μετρικές και πειράματα.

Για την συνέχεια της έρευνάς μας, επιλέξαμε τα αποτελέσματα των στηλών 70%GB-30%LT έως το 30%GB-70%LT, στα οποία εφαρμόσαμε 3 διαφορετικές μετρικές, που θα μας εμφάνιζαν εναλλακτικά την προσθήκη επισημειώσεων στα περιεχόμενα του αρχείου καθιερωμένων όρων. Οι στήλες που επιλέξαμε με τα αντίστοιχα ποσοστά, θεωρήσαμε πως θα αντιπροσωπεύουν την επιλογή των περισσότερων βιβλιοθηκονόμων, χωρίς ωστόσο να αποκλείεται μια διαφορετική επιλογή. Όποια επιλογή και να κάνει κάποιος βιβλιοθηκονόμος οι διαδικασίες που θα περιγράψουμε παρακάτω, μπορούν να εφαρμοστούν με τον ίδιο ακριβώς τρόπο.

Αρχικά θεωρήσαμε 3 διαφορετικά κατώφλια (threshold) στη συνολική τιμή tfidf - για κάθε συνδυασμό a και b, όπως αναφέρθηκε παραπάνω - με τιμές 1.6 , 2.2 , 2.6 . Κάθε βιβλιοθηκονόμος μπορεί να επιλέξει όποιο κατώφλι θέλει, ωστόσο θα πρέπει να βρεθεί μια τιμή που να εξισορροπεί τον εμπλουτισμό του αρχείου καθιερωμένων όρων και της ποιότητας των επισημάνσεων με τις οποίες εμπλουτίζεται το αρχείο.

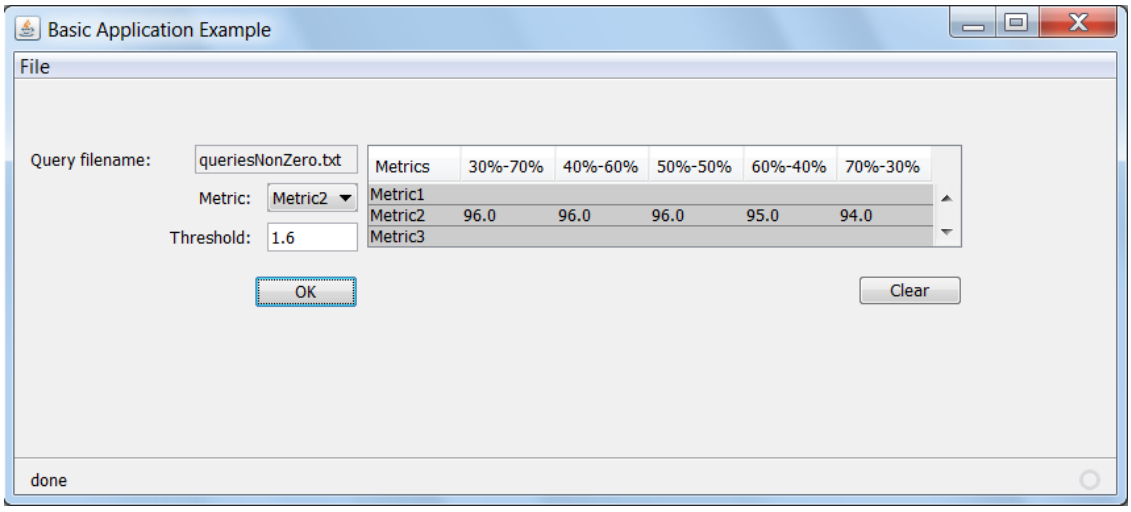
Για κάθε κατώφλι υπολογίσαμε τους ακόλουθους δείκτες:

1) Στο πρώτο μας πείραμα, θελήσαμε να δούμε πόσο αυξήθηκε η κάλυψη του συνόλου του αρχείου καταγραφής επερωτήσεων, από το σύνολο του αρχείου καθιερωμένων εγγραφών.
2) Στο δεύτερο πείραμα, εξετάσαμε το ποσοστό των τεκμηρίων για τα οποία εμπλουτίσαμε τις γνώσεις μας. Είδαμε σε πόσα τεκμήρια, από το σύνολο των 4018 τεκμηρίων που διαθέταμε, καταφέραμε να αυξήσουμε την εγγραφή του στο αρχείο καθιερωμένων όρων. Ποιο

συγκεκριμένα, είδαμε ποιες επισημειώσεις έχουν τιμή tf-idf, μεγαλύτερη από το κατώφλι και κρατήσαμε το τεκμήριο στο οποίο ανήκουν. Το σύνολο των μοναδικών τεκμηρίων που κρατήσαμε προς το σύνολο όλων των τεκμηρίων που είχαμε στην διάθεσή μας (δηλαδή 4018), ήταν το δεύτερο μέτρο που εξετάσαμε.

3) Στο τρίτο και τελευταίο μας πείραμα, βρήκαμε το ποσοστό των επισημειώσεων που προστέθηκαν στο αρχείο καθιερωμένων όρων, συγκριτικά με το σύνολο των εν δυνάμει επισημειώσεων προς εμπλουτισμό του αρχείου. Πιο συγκεκριμένα, γνωρίζοντας το σύνολο των επισημειώσεων που εξορύξαμε από το LibraryThing και το Google Books, για ένα συγκεκριμένο τεκμήριο, παρατηρήσαμε πόσες από αυτές τις επισημειώσεις προστέθηκαν εν τέλει, στο αρχείο καθιερωμένων όρων.

Παρακάτω παρουσιάζουμε τα αποτελέσματα του δεύτερου μέτρου, για κατώφλι με τιμή 1,6 .

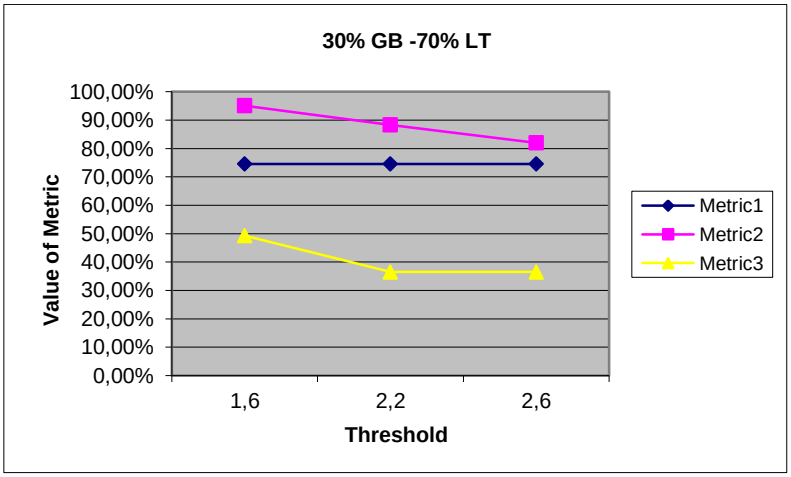


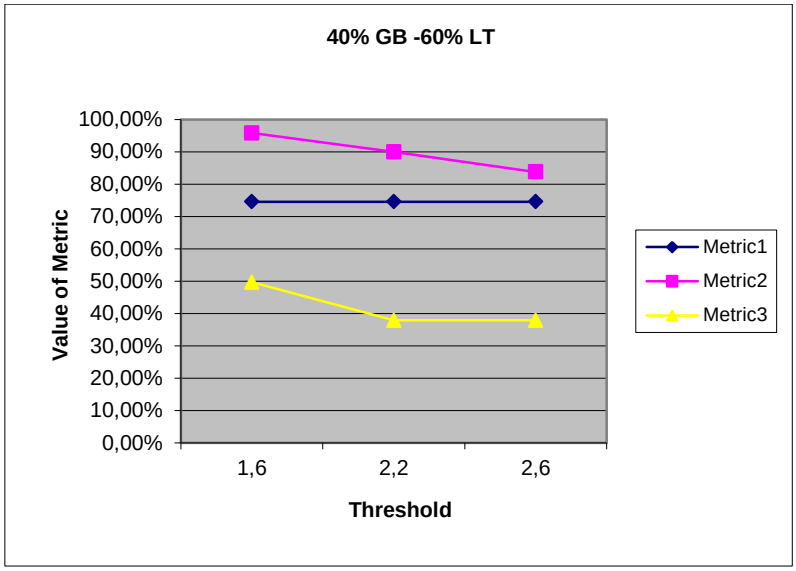
Εικόνα 5.1. Πρόγραμμα εκτέλεσης με τα αποτελέσματα του δεύτερου μέτρου με κατώφλι 1,6.

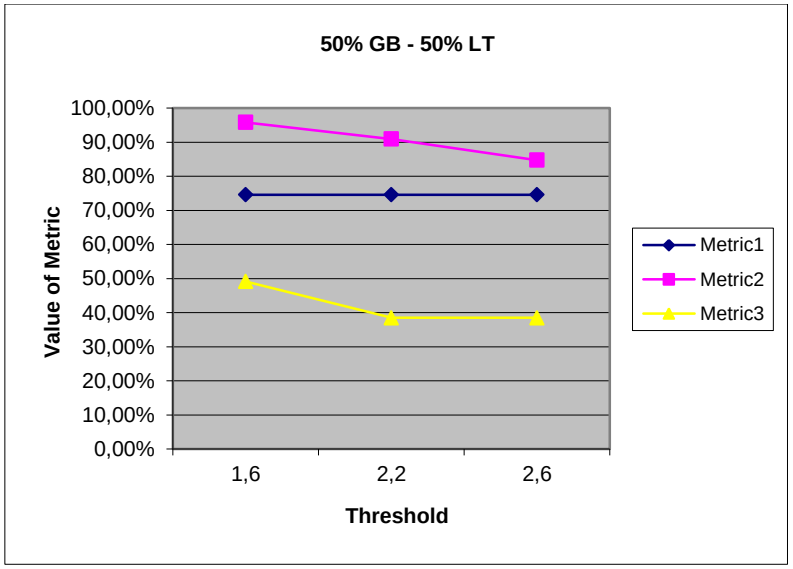
Στην παραπάνω εικόνα, φαίνεται ένα στιγμιότυπο εκτέλεσης ενός προγράμματος που σχεδιάστηκε στα πλαίσια της διπλωματικής μας εργασίας. Οι χρήστες του προγράμματος μπορούν να επιλέξουν οι ίδιοι το κατώφλι που επιθυμούν, όπως επίσης και το αρχείο καταγραφής των επερωτήσεων της δικής τους βιβλιοθήκης.

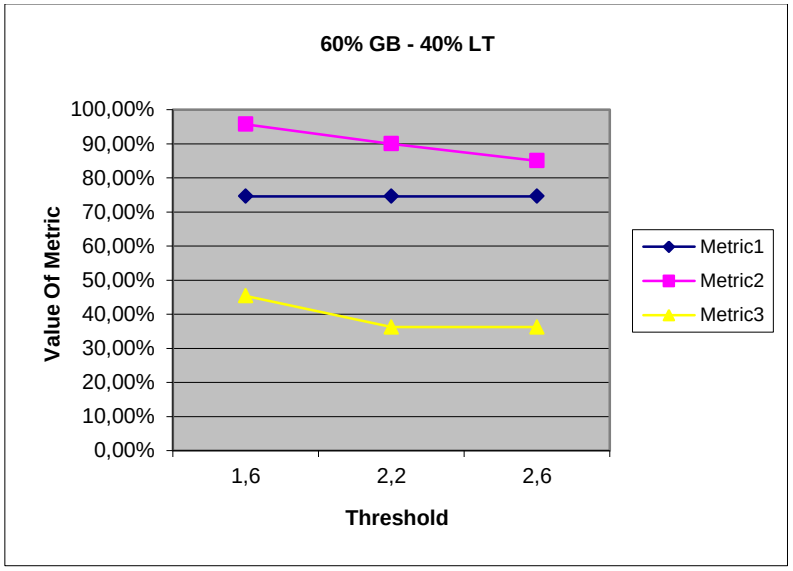
Συνοπτικά τα αποτελέσματα των μέτρων που εξετάσαμε και για τα 3 κατώφλια, παρουσιάζονται στα παρακάτω διαγράμματα. Όπως θα δούμε λοιπόν, με το δεύτερο μέτρο προσθέτουμε πληροφορίες – επισημειώσεις σε ένα μεγάλο ποσοστό τεκμηρίων. Ενώ, όπως είναι φυσιολογικό, όσο το κατώφλι που επιλέγουμε μεγαλώνει, τόσο μειώνονται τα τεκμήρια στα οποία προστίθενται πληροφορίες. Πιο αναλυτικά, παρατηρούμε πως για ένα ποσοστό κοντά στο 90% του συνόλου των τεκμηρίων που είχαμε στην διάθεσή μας, αυξήθηκαν οι εγγραφές τους στο αρχείο καθιερωμένων όρων. Το ποσοστό αυτό είναι πολύ υψηλό και καταδεικνύει την αξία της έρευνάς μας, καθώς αποδεικνύει την ορθότητα των μεθόδων που ακολουθήσαμε. Ασφαλώς, όσο μειώνεται το κατώφλι, τόσο περισσότερες επισημειώσεις θεωρούνται χρήσιμες και προστίθενται στο αρχείο καθιερωμένων όρων, επομένως τόσο περισσότερα είναι και τα τεκμήρια που επηρεάζονται. Ωστόσο, πρέπει να λάβουμε υπ' όψιν μας, πως η σωστή επιλογή του κατωφλίου, είναι ένας πολύ σημαντικός παράγοντας αξιολόγησης των αποτελεσμάτων που σας παρουσιάζουμε, καθώς μικρό κατώφλι σημαίνει πως στο αρχείο προστίθενται περισσότερες επισημειώσεις αλλά με μικρότερη σχέση και αξία για το συγκεκριμένο τεκμήριο. Αντίθετα, υψηλότερο κατώφλι, σημαίνει πως στο αρχείο προστίθενται λιγότερες επισημειώσεις με μεγαλύτερη σχέση και αξία για το συγκεκριμένο τεκμήριο, ωστόσο χάνονται ενδεχομένως κάποιες επισημειώσεις χρήσιμες για το τεκμήριο.

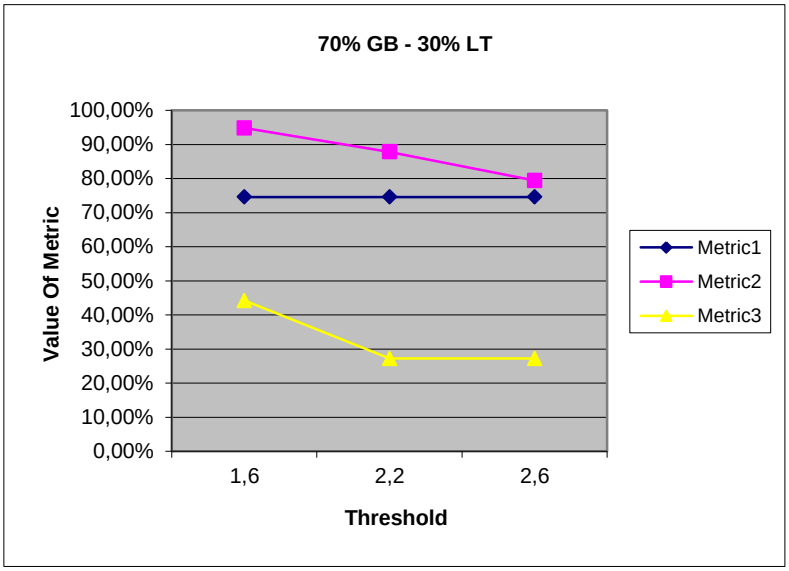
Περισσότερες πληροφορίες σχετικά με τα υπόλοιπα δύο μέτρα που εκτελέσαμε πειραματικά για την έρευνά μας, μπορείτε να βρείτε στις διπλωματικές εργασίες των Μαρίνα Κουμάκη (μέτρο 3) και Κακαβούλη Διονύσιου (μέτρο 1).




40% GB -60% LT
100,00%
90,00%
80,00%
70,00%
60,00%
50,00%
40,00%
30,00%
20,00%
10,00%
0,00%
Value of Metric
1,6
2,2
2,6
Threshold
Metric1
Metric2
Metric3


50% GB - 50% LT
100,00%
90,00%
80,00%
70,00%
60,00%
50,00%
40,00%
30,00%
20,00%
10,00%
0,00%
Value of Metric
1,6
2,2
2,6
Threshold
Metric1
Metric2
Metric3


60% GB - 40% LT
100,00%
90,00%
80,00%
70,00%
60,00%
50,00%
40,00%
30,00%
20,00%
10,00%
0,00%
Value Of Metric
1,6
2,2
2,6
Threshold
Metric1
Metric2
Metric3


70% GB - 30% LT
100,00%
90,00%
80,00%
70,00%
60,00%
50,00%
40,00%
30,00%
20,00%
10,00%
0,00%
Value Of Metric
1,6
2,2
2,6
Threshold
Metric1
Metric2
Metric3

"Life is also a mixture of unsolved problems, ambiguous victories and vague defeats-with very few moments of clear peace."

-- Hugh Prather

6 Ιδιαιτερότητες του μοντέλου μας.

Στο κεφάλαιο αυτό, θα αναφερθούμε σε κάποιες ιδιαιτερότητες του μοντέλου μας, καθώς και σε ζητήματα που δεν καλύπτουμε. Όσων αφορά την αρχική επεξεργασία των δεδομένων μας, εμείς χρησιμοποιούμε μεθόδους αποκοπής της κατάληξης μιας λέξης (stemming), ώστε να μην ασχολούμαστε με το αν η λέξη είναι σε ενικό ή πληθυντικό αριθμό, σε τι γένος (αρσενικό, θηλυκό, ουδέτερο), σε τι πτώση, πρόσωπο ή έγκλιση (αν πρόκειται για ρήμα). Ωστόσο, για τα δεδομένα μας, δεν εφαρμόζουμε μεθόδους προσδιορισμού των μερών του λόγου των λέξεων που απαρτίζουν τα δεδομένα μας. Με την μέθοδο αυτή, ενδεχομένως να είχαμε καλύτερη κάλυψη μεταξύ των κοινών δεδομένων και υψηλότερα ποσοστά στα σύνολα που επεξεργαστήκαμε στο κεφάλαιο 3 (σελίδα 38). Η μέθοδος αυτή έχει την αγγλική ονομασία part of speech tagging – POS tagging, ωστόσο δεν βρήκαμε κάποια εφαρμογή της στην ελληνική γλώσσα και για αυτό δεν την εφαρμόσαμε.

Στην συνέχεια, πρέπει να υπενθυμίσουμε, πως κάποιες από τις επισημειώσεις που πήραμε από τις ιστοσελίδες Google Books και LibraryThing, προέρχονταν πιθανών από διαφορετικές εκδόσεις των τεκμηρίων που αναζητούσαμε ή από μεταφράσεις αυτών, σε άλλες γλώσσες. Σε κάποιες περιπτώσεις, λοιπόν, όπου νέες εκδόσεις τεκμηρίων περιέχουν καινούργια κεφάλαια και νέα δεδομένα, ίσως υπάρξει ένα μικρό ποσοστό επισημειώσεων που προστέθηκαν εκ παραδρομής στο αρχείο καθιερωμένων όρων του Πάντειου Πανεπιστημίου Αθηνών.

Σχετικά με το προγραμματιστικό κομμάτι της διαδικασίας εξόρυξης δεδομένων από το LibraryThing, και όσων αφορά την μέθοδο «ανθρώπινης επιβεβαίωσης» πρέπει να αναφέρουμε μία πιθανή λύση, που θα μας απέτρεπε από την συχνή εισαγωγή του συγγραφέα ενός τεκμηρίου και θα συνέχιζε την εκτέλεση του προγράμματός μας, χωρίς ενδιάμεση διάδραση με τον χρήστη.

Η προτεινόμενη λύση θα ήταν να συνδεόμαστε στην αντίστοιχη διεύθυνση του τεκμηρίου στο LibraryThing, μέσω ενός πληρεξούσιου προγράμματος ή αλλιώς ενός proxy server[19]. Με την χρήση ενός τέτοιου προγράμματος, θα μπορούσαμε να κρύψουμε την ηλεκτρονική μας διεύθυνση από τον διακομιστή του LibraryThing και στην συνέχεια όταν η υποτιθέμενη διεύθυνση μας, τιθόταν εν αμφιβόλω, να αλλάζαμε πληρεξούσιο και να συνεχίζαμε την εκτέλεση του προγράμματός μας.

Πρέπει, ακόμα, να αναφέρουμε πως τα κατώφλια που ο εκάστοτε βιβλιοθηκονόμος θα επιλέξει, για να πραγματοποιήσει ο ίδιος τα πειράματα που περιγράψαμε και εν συνεχεία να εμπλουτίσει το αρχείο καθιερωμένων όρων της βιβλιοθήκης του, χρίζουν ιδιαίτερης προσοχής και σίγουρα αποτελούν αντικείμενο μελέτης από εκείνον. Ο βιβλιοθηκονόμος θα πρέπει να πειραματιστεί με διάφορα κατώφλια και εν συνεχεία να αξιολογήσει τα δεδομένα που πρόσθεσε στο αρχείο καθιερωμένων όρων, καθώς η αλόγιστη μείωση του κατωφλίου με σκοπό της σημαντική αύξηση των όρων στο αρχείο, θα έχει ως συνέπεια, την προσθήκη όρων με μικρή σημασία, μικρό κοινό αποδοχής (τοπικές διαλέκτους) και μικρή ουσιαστική προσφορά στο σύνολο.

[19] Τα πληρεξούσια προγράμματα, είναι μια υπηρεσία κοινής ωφέλειας που μπορεί να χρησιμοποιηθεί από οποιονδήποτε σε όλο τον κόσμο για να πλοηγείται στο διαδίκτυο ανώνυμα. Χρησιμοποιεί ένα διακομιστή-είδωλο με τη βοήθεια του οποίου η ηλεκτρονική διεύθυνση του χρήστη καθώς και άλλες λεπτομέρειες είναι πάντα κρυμμένες στον ενδιάμεσο φορέα που παρέχει την υπηρεσία αυτή. Με μια τέτοια υπηρεσία, μπορείτε να χρησιμοποιήσετε ένα ανώνυμο διακομιστή μεσολάβησης για να επισκεφθείτε τοποθεσίες στο διαδίκτυο χωρίς να αφήνετε κάποιο ίχνος.

7

"Accept the challenges so that you may feel the exhilaration of victory."
--General George S. Patton

Ανοιχτά Θέματα

Στον τομέα των κοινωνικών δικτύων και της ανάλυσης αυτών, όπως επίσης και για «κόσμο» της Βιβλιοθήκης 2.0, υπάρχουν κάποια ανοιχτά θέματα που χρίζουν έρευνας και προσοχής από το σύνολο της ερευνητικής κοινότητας. Στο κεφάλαιο αυτό, θα κάνουμε μια σύντομη επισκόπηση στα θέματα αυτά, που πιθανών να αποτελέσουν μελλοντικό ερευνητικό αντικείμενο των του αναγνώστη.

Μικρή έρευνα έχει επικεντρωθεί σε θέματα σχετικά με τις εφαρμογές του Web 2.0. και τις βιβλιοθήκες. Έρευνες των Bejune και Abbas, όπως:

- Smoothing the information seeking path: Removing representational obstacles in a middle school digital library environment.
- Out of the mouths of middle school children: I. Using student generated keywords for subject access in a digital library.
- Who is tagging information?
- Wikis in libraries.

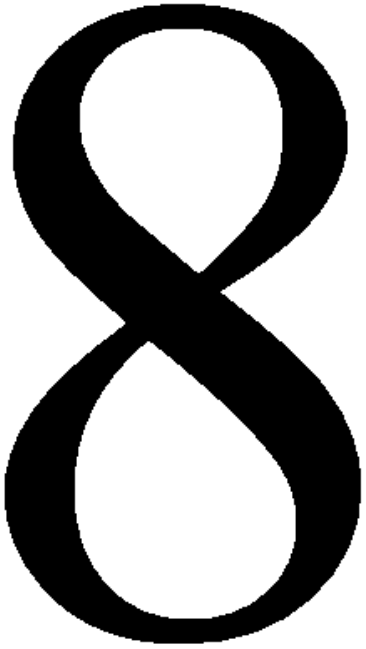
Παρέχουν χρήσιμες πληροφορίες και υλικό για να μπορέσει να μελετήσει κάποιος σε βάθος τον συγκεκριμένο τομέα, ωστόσο δεν υποστηρίζουν υπερβολικά την άποψη της υιοθέτησης των κοινωνικών διαδραστικών τεχνολογιών στις ηλεκτρονικές βιβλιοθήκες, αν και τα ευρήματα ενθαρρύνουν μια τέτοια κίνηση. Αναμφίβολα, απαιτείται περισσότερη έρευνα για την χρήση, τα κίνητρα και την σκοπιμότητα χρήσης τέτοιων τεχνολογιών.

Υπάρχουν ανησυχίες όπου επαγγελματίες του χώρου των βιβλιοθηκών καταδεικνύουν και θα πρέπει να εξεταστούν. Οι ανησυχίες αυτές αφορούν ζητήματα σχετικά με την εμπιστοσύνη την ακρίβεια, την υπευθυνότητα, την αξιοπιστία και την δεοντολογία σχετικά με την χρήση εφαρμογών του Web 2.0. Ζητήματα εγείρονται από τους επαγγελματίες, όπου ενδεχομένως αισθάνονται πως περιθωριοποιούνται, όταν οι βιβλιοθήκες ζητούν την συνδρομή των διαδικτυακών χρηστών της βιβλιοθήκης . Τα ζητήματα αυτά αφορούν κυρίως τις εξής περιπτώσεις [11]:

i. Τη μακροπρόθεσμη επιτυχία του κοινωνικού λογισμικού και της τοποθέτησης επισημειώσεων,
ii. τις περιπτώσεις ειδικών χρήσεων των κοινωνικών ταξονομιών σε ακαδημαϊκά περιβάλλοντα,
iii. την επίδραση των επισημειώσεων, της καταλογογράφησης και της ευρετηρίασης στην βιβλιοθηκονομία και στις επιστήμες πληροφορίας,
iv. την επεκτασιμότητα μεταξύ των διαφόρων συστημάτων και
v. την συνεχή στήριξη και συμμετοχή των χρηστών.

Σημαντικότερο ζήτημα όλων, ωστόσο, δείχνει πως είναι η διασφάλιση της προστασίας των ιδιωτικών πληροφοριών που μπορεί να περιέχονται σε μια βιβλιοθήκη. Σοβαρό θέμα αντίστοιχο της ιδιωτικής ζωής μπορεί να αποδειχθεί η ανάγκη της παραπάνω διασφάλισης. Ένα από τα χαρακτηριστικά του Web 2.0, είναι να υπάρχουν εφαρμογές όπου όλα εξαρτώνται από τους χρήστες σχετικά με το να μοιράζονται και να επεξεργάζονται προσωπικές πληροφορίες με άλλους χρήστες. Αυτό όμως είναι πρόβλημα για τους βιβλιοθηκονόμους οι οποίοι θα ήθελαν να προσφέρουν υπηρεσίες Library 2.0, ωστόσο, πρέπει να εκτιμήσουν και το κόστος των πληροφοριών που βάζουν σε κίνδυνο, αν γίνουν δημόσιες προς επεξεργασία. Θα

πρέπει, επομένως, η ερευνητική κοινότητα να εξετάσει λύσεις σχετικές με το πρόβλημα και τρόπους ώστε όσο το δυνατόν περισσότερες πληροφορίες να βρίσκονται δημόσια στο διαδίκτυο, χωρίς ωστόσο να καταπατώνται δικαιώματα συγγραφέων, εκδοτικών οίκων, βιβλιοθηκών και άλλων φορέων.



*"I came to the **conclusion** that the optimist thought everything good except the pessimist, and that the pessimist thought everything bad, except himself."*

--G. K. Chesterton

Επίλογος

Στην παρούσα διπλωματική εργασία, ασχοληθήκαμε με την εύρεση μεθόδων εμπλουτισμού με κοινωνικές επισημειώσεις, αρχεία καθιερωμένων όρων. Οι κοινωνικές επισημειώσεις, εμφανίστηκαν με τις πρώτες υπηρεσίες εφαρμογών Web 2.0. Το συγκεκριμένο «πρότυπο», επηρέασε και τις εφαρμογές σχετικές με ηλεκτρονικές βιβλιοθήκες και όπως κάθε καινοτομία, έφερε μαζί της νέα προβλήματα και νέες εφαρμογές για την καλύτερη και πιο ολοκληρωμένη εξυπηρέτηση του κοινού.

Με την έρευνά μας, υποστηρίζουμε την χρήση κοινωνικών επισημειώσεων και κατ' επέκταση υπηρεσιών κοινωνικής δικτύωσης, καθώς όπως δείχνουν τα αποτελέσματά μας, δίνουν την δυνατότητα να εμπλουτίσουν σε σημαντικό βαθμό τα αρχεία καθιερωμένων όρων των βιβλιοθηκών και να κάνουν πιο φιλικές προς τους χρήστες τις υπηρεσίες τους.

Τα προβλήματα που εγείρονται από την χρήση υπηρεσιών Library 2.0, είναι αναμφίβολα θέματα που απαιτούν λύση, ωστόσο κατά την γνώμη μας δεν θα πρέπει να αποτελέσουν τροχοπέδη στην εξάπλωσή τους.

Παράρτημα.

Για την ολοκλήρωση της εργασίας μας, σχεδιάσαμε ένα σύνολο εκτελέσιμων προγραμμάτων, καθένα από τα οποία εκτελούσε και διαφορετική διεργασία. Οι γλώσσες προγραμματισμού που χρησιμοποιήσαμε για την σχεδίασή τους, ήταν η Java και η C#[20].

Τα προγράμματα που είναι γραμμένα σε Java εκτελούνται ως εξής:

- Αρχικά πηγαίνουμε στον φάκελο dist, που περιέχεται σε κάθε ένα πρόγραμμα,
- και βρίσκουμε το αρχείο .jar , δηλαδή ένα εκτελέσιμο αρχείο με τις πληροφορίες του προγράμματός μας,
- στην συνέχεια, αν το λειτουργικό μας σύστημα είναι windows, ο χρήστης αρκεί να πληκτρολογήσει στο cmd, την εντολή

  ```
  java -jar program_name.jar,
  ```

  ωστόσο, σε περίπτωση που το πρόγραμμα δεν εκτελείται, τότε θα πρέπει να ακολουθηθεί η παρακάτω διαδικασία:
  - Ανοίγουμε τον windows explorer, και επιλέγουμε «Εργαλεία» και στην συνέχεια «Επιλογές φακέλων»,
  - Κατόπιν, πλοηγούμαστε στην ετικέτα «Τύπος Αρχείων» και ψάχνουμε μέχρι την επιλογή «Αρχεία JAR»,
  - Στην συνέχεια στην επιλογή «Επεξεργασία τύπου αρχείων» επιλέγουμε την «επεξεργασία» και πλοηγούμαστε μέχρι να βρούμε τον φάκελο που περιέχει το αρχείο javaw.exe.
  - Επιλέγοντας «αποδοχή», μπορούμε πλέον να εκτελούμε τα προγράμματά μας, κάνοντας διπλό κλικ στο αρχείο .jar.
- αν το λειτουργικό σύστημα του χρήστη είναι linux, ο χρήστης αρκεί να πληκτρολογήσει την εντολή

  ```
  java -jar program_name.jar,
  ```

[20] Η C# είναι μια σχετικά νέα αντικειμενοστραφής γλώσσα προγραμματισμού η οποία δημιουργήθηκε από την Microsoft. Δανείζεται πολλά στοιχεία, και έχει παρόμοια σύνταξη, με την C++ και την Java, κάνοντας την εκμάθηση της σχετικά εύκολη. Είναι γλώσσα ειδικά σχεδιασμένη για να υποστηρίζει το .NET framework της ίδιας εταιρείας. Βασικό χαρακτηριστικό της είναι ότι δεν παράγει απευθείας κώδικα μηχανής όπως η C++, άλλα ένα ενδιάμεσο κώδικα που στοχεύει το .NET.

- τέλος αν το λειτουργικό σύστημα του χρήστη είναι MacOS, ο χρήστης αρκείται να κάνει διπλό κλικ πάνω στο αρχείο .jar.

Τα προγράμματα σε C#, μπορούν να εκτελεστούν σε περιβάλλον Visual Studio αν ο χρήστης έχει λειτουργικό Windows, Mono αν ο χρήστης έχει λειτουργικό linux, ή Parallels αν το λειτουργικό σύστημα είναι MacOS.

Βιβλιογραφία

[1] S. Sizov, S. Staab και T. Franz, Analysis of Social Networks by Tensor Decomposition, University of Koblenz-Landau, WeST – Institute for Web Science and Technologies, 2010.

[2] M. Zhang, Social Network Analysis: History, Concepts and Research, School of Journalism and Communication, Wuhan University, 2010.

[3] L. C. Freeman, The Development of Social Network Analysis: A Study in The Sociology of Science, Empirical Press Vancouver, 2004.

[4] C. Wetherell, A. Plakans και B. Wellman, Social Networks, Kinship, and Community in Eastern Europe, Journal of Interdisciplinary History,, 1994.

[5] J. Scott, Social Network Analysis: A Handbook, Sage Publications, 1987.

[6] «Web 2.0 - Wikipedia,» [Ηλεκτρονικό]. Available: http://el.wikipedia.org/wiki/Web_2.0.

[7] «Wikipedia (En) - Web 2.0,» [Ηλεκτρονικό]. Available: http://en.wikipedia.org/wiki/Web_2.0.

[8] TIME archive, «Person of the Year: You,» 2006. [Ηλεκτρονικό]. Available: http://www.time.com/time/covers/0,16641,20061225,00.html.

[9] BBC news, «'Millionth English word' declared,» [Ηλεκτρονικό]. Available: http://news.bbc.co.uk/2/hi/americas/8092549.stm.

[10] A. Roy, Online Communities and Social Networking, University of Scranton, USA, 2010.

[11] J. Abbas, Social Software Use in Public Libraries, University of Oklahoma, USA, 2010.

[12] M. Casey και L. Savastinuk,, Library 2.0: Service for the next generation library, Library Journal, 2008.

[13] J. Abbas και J. Graham, So, let’s talk about tagging, user-defined/supplied descriptors. A research and curricular agenda., Annual Meeting of the American Society for Information Science and Technology, 2006.

[14] OCLC, Sharing, privacy, and trust in our networked world: A report to the OCLC membership., OCLC Online Computer Library Center, 2007.

[15] P. Mika, Social Networks and the Semantic Web, Springer, 2007.

[16] G. Smith, Tagging: People-Powered Metadata for the Social Web, New Riders, 2008.

[17] «Wikipedia (En) - National Institute of Standards and Technology,» [Ηλεκτρονικό]. Available: http://en.wikipedia.org/wiki/National_Institute_of_Standards_and_Technology.

[18] P. Heymann, D. Ramage και H. Garcia-Molina, Social Tag Prediction, Department of Computer Science, Stanford University, USA, 2008.

[19] P. Heymann και H. Garcia-Molina, Collaborative Creation of Communal Hierarchical Taxonomies in Social Tagging Systems, Computer Science Department, Stanford University, USA, 2006.

[20] M. Szomszor, C. Cattuto, H. Alani, K. O'Hara, A. Baldassarri, V. Loreto και V. D. Servedio, Folksonomies, the Semantic Web and the Movie Recommendation, School of Electronics and Computer Science, University of Southampton, UK; Dipartimento di Fisica, Universita di Roma "La Sapienza", Italy; Museo Storico della fisica e Centro Studi e Ricerche Enrico Fermi Compendio Viminale, Italy, 2006.

[21] «Wikipedia - Java,» [Ηλεκτρονικό]. Available: http://el.wikipedia.org/wiki/Java.

[22] «Wikipedia - Netbeans {en},» [Ηλεκτρονικό]. Available: http://en.wikipedia.org/wiki/NetBeans.

[23] «Wiipedia {Gr} - IP,» [Ηλεκτρονικό]. Available: http://el.wikipedia.org/wiki/%CE%94%CE%B9%CE%B5%CF%8D%CE%B8%CF%85%CE%BD%CF%83%CE%B7_IP.

[24] «Wikipedia {Gr} - Διεθνής πρότυπος αριθμός βιβλίου,» [Ηλεκτρονικό]. Available: http://el.wikipedia.org/wiki/%CE%94%CE%B9%CE%B5%CE%B8%CE%BD%CE%AE%CF%82_%CF%80%CF%81%CF%8C%CF%84%CF%85%CF%80%CE%BF%CF%82_%CE%B1%CF%81%CE%B9%CE%B8%CE%BC%CF%8C%CF%82_%CE%B2%CE%B9%CE%B2%CE%BB%CE%AF%CE%BF%CF%85.

[25] «Wikipedia {Gr} - HTML,» [Ηλεκτρονικό]. Available: http://el.wikipedia.org/wiki/HTML.

[26] «Wikipedia {En} - Document Object Model,» [Ηλεκτρονικό]. Available: http://en.wikipedia.org/wiki/Document_Object_Model.

[27] «Wikipedia {Gr} - CSS,» [Ηλεκτρονικό]. Available: http://el.wikipedia.org/wiki/CSS.

[28] «Wikipedia {En} - Google Books,» [Ηλεκτρονικό]. Available: http://en.wikipedia.org/wiki/Google_Books.

[29] «Wikipedia {En} - tf–idf,» [Ηλεκτρονικό]. Available:

http://en.wikipedia.org/wiki/Tf%E2%80%93idf.

[30] Ε. Σταματάτος, «Ανάκτηση Πληροφορίας - Κλασσικά Μοντέλα,» Πανεπιστήμιο Αιγαίου, 2011.

[31] K. Panineni, Why inverse document frequency?, New York, USA: IBM T.J. Watson Research Center, 2001.

[32] L. o. C. Cataloging Distribution Service, «What is a MARC record and why is it important,» Library of Congress, 2004. [Ηλεκτρονικό]. Available: http://www.loc.gov/marc/uma/pt1-7.html#pt4.