ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ

«Διδακτική της Τεχνολογίας & Ψηφιακών Συστημάτων»

Κατεύθυνση: Ηλεκτρονική Μάθηση

Σημασιολογικά Κοινωνικά Δίκτυα

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΤΡΙΚΚΑ ΧΡΗΣΤΟΥ – ΦΩΤΙΟΥ ΜΕ 10049

Επιβλέπων: Γεώργιος Βούρος

Καθηγητής Πανεπιστημίου Πειραιά

Αθήνα, Ιούνιος 2012

Περίληψη

Ο Σημασιολογικός Ιστός στοχεύει στην καλύτερη οργάνωση του διαδικτυακού περιεχομένου και στην απόδοσή του με ρητό και τυπικό τρόπο (formally annotated), κάτι που επιτυγχάνεται κύρια με την χρήστη των οντολογιών. Η δημιουργία και η εξέλιξή των τελευταίων πραγματοποιείται μόνο με τη βοήθεια μηχανικών γνώσης και ειδικών σε ένα γνωστικό αντικείμενο, που διαθέτουν την απαραίτητη τεχνογνωσία για τις απαραίτητες διεργασίες (experts), γεγονός που οδηγεί στο πρόβλημα της άμεσης, ανανεώσιμης πληροφορίας.

Από την άλλη πλευρά, τα τελευταία χρόνια παρατηρείται μια ραγδαία αύξηση του ενδιαφέροντος για τα Κοινωνικά δίκτυα, τα οποία παρέχουν πρόσφορο έδαφος για τη συνεργασία ανάμεσα στα μέλη μιας κοινότητας. Τα μέλη των συγκεκριμένων κοινοτήτων διαμοιράζουν και επισημαίνουν πηγές πληροφορίας με ελεύθερες, χωρίς κανένα περιορισμό, λέξεις-κλειδιά. Αυτές οι αυθαίρετες αναθέσεις ονομάζονται ετικέτες σήμανσης (tags) ενώ πυρήνας της δομής των δεδομένων τους είναι ένα folksonomy. Συγκεκριμένα, ο όρος αυτός αναφέρεται σε ένα είδος ταξινομίας που παράγεται από τον ίδιο τον χρήστη για την κατηγοριοποίηση και ανάκτηση του περιεχομένου στο διαδίκτυο (σύνδεσμοι, φωτογραφίες), με τη χρήση των «ανοικτού τύπου» ετικετών. Παρόλα τα πλεονεκτήματά τους όμως τα folksonomies παρουσιάζουν και ορισμένες αδυναμίες.

Πολλοί ερευνητές θεωρούν ότι ο συνδυασμός των τεχνολογιών του Σημασιολογικού Ιστού και των Κοινωνικών δικτύων δύναται να οδηγήσει στην ολοκληρωμένη λειτουργία του Παγκόσμιου Ιστού. Απόρροια αυτής της άποψης, είναι η δημιουργία και η χρήση του Σημασιολογικού Κοινωνικού Ιστού (Social Semantic Web) και των Σημασιολογικών Κοινωνικών Δικτύων. Με την παρούσα διπλωματική εργασία επιζητείται η λεπτομερής καταγραφή, παράθεση, ομαδοποίηση και κριτική επισκόπηση του συνόλου των προαναφερθεισών προσπαθειών σύνδεσης.

Λέξεις-κλειδιά: Σημασιολογικός Ιστός, Οντολογίες, Κοινωνικά Δίκτυα, Folksonomies, Σημασιολογικά Κοινωνικά Δίκτυα, Μέθοδοι, Εφαρμογές, Σύνδεση.

Περιεχόμενα

Περίληψη 2

Περιεχόμενα 3

Λίστα Εικόνων 5

Λίστα Πινάκων 7

ΚΕΦΑΛΑΙΟ 1 8

1.1 Εισαγωγή 8

1.2 Προσδιορισμός του Προβλήματος 9

1.3 Σκοπός Διπλωματικής Εργασίας 10

1.4 Δομή 10

ΚΕΦΑΛΑΙΟ 2 12

2.1 Σημασιολογικός Ιστός 12

2.1.1 Οντολογίες 12

2.2 Κοινωνικά Δίκτυα (Web 2.0) 15

2.2.1 Folksonomies 16

2.2.3 Σημαντικότερα Κοινωνικά Συστήματα Σήμανσης 20

ΚΕΦΑΛΑΙΟ 3 24

3.1 Σημασιολογικός Κοινωνικός Ιστός 24

3.2 Σημασιολογικά Κοινωνικά Δίκτυα 25

3.3 Μέθοδοι και εφαρμογές σύνδεσης σημασιολογικής πληροφορίας με τα κοινωνικά δίκτυα 25

3.4 Προσπάθειες Σύνδεσης Σημασιολογικής Πληροφορίας με τα Κοινωνικά Δίκτυα Περιληπτικά 26

3.5 Σημαντικότερες Προσεγγίσεις 33

3.5.1 Μοντέλο Υπαγωγής 33

3.5.2 Μέθοδος Εμπλουτισμού των Folksonomies με τη βοήθεια του Σημασιολογικού Ιστού 39

3.5.3 Μοντέλο Σήμανσης Οντολογιών (The TagOntology) 48

3.5.4 Εργαλείο Σημασιολογικού Εμπλουτισμού FLOR 53

3.5.5 Αλγόριθμος SemTagP: Σημασιολογική Ανίχνευση Κοινότητας στα Folksonomies 62

3.5.6 Μέθοδος εξαγωγής έμπειρων προφίλ από τις ετικέτες σήμανσης 74

3.5.7 Το μοντέλο Χρήστης – Έννοια – Στιγμιότυπο των Οντολογιών 84

Κεφάλαιο 4 94

4.1 Συνοπτική Παρουσίαση 94

4.1.1 Είδος Έρευνας 94

4.1.2 Χρήση Συστημάτων Σήμανσης και Εξωτερικών Πηγών 95

4.1.3 Ανθρωποκεντρικές και Αυτοματοποιημένες Διαδικασίες 96

4.1.4 Σχέσεις μεταξύ των Ετικετών Σήμανσης 98

4.2 Χαρακτηριστικά Σημαντικότερων Μεθόδων και Εφαρμογών Σύνδεσης 101

4.3 Χαρακτηριστικά Υπόλοιπων Μεθόδων και Εφαρμογών Σύνδεσης 102

Κεφάλαιο 5 104

5.1 Συμπεράσματα - Επίλογος 104

5.2 Μελλοντικά Σχέδια 105

Κεφάλαιο 6 106

6.1 Βιβλιογραφία 106

Λίστα Εικόνων

Εικόνα 1: Μορφές οντολογιών ανάλογα με την εκφραστικότητα τους 13

Εικόνα 2: Σχέση Παγκόσμιου Ιστού με Σημασιολογικό 15

Εικόνα 3: Μορφή Folksonomy 18

Εικόνα 4: Σημασιολογικός Κοινωνικός Ιστός 24

Εικόνα 5: Αρχιτεκτονική συστήματος των Specia L. και Motta E. 41

Εικόνα 6: Σημασιολογικό στρώμα στο χώρο ετικέτας ενός Folksonomy 55

Εικόνα 7: Ο πυρήνας οντολογίας (The Core Ontology) 56

Εικόνα 8: Οι φάσεις του εργαλείου FLOR 57

Εικόνα 9: Στρατηγική συγχώνευσης με όριο την τιμή 0,5 60

Εικόνα 10: Εμπλουτισμός της ετικέτας "moon" στο εργαλείο FLOR 61

Εικόνα 11: Σημασιολογική διάδοση ετικετών σήμανσης. 63

Εικόνα 12: Μορφή κοινωνικού δικτύου ADEME PhD 71

Εικόνα 13: Διάρθρωση της κατατμημένης (διαρθρωμένης) κοινότητας που παράχθηκε, μετά από κάθε διάδοση διαδρομής με τη χρήση των RAK, TagP, SemTagP και 3 ελεγχόμενων SemTagP αλγορίθμων 73

Εικόνα 14: Κύρια βήματα αλγορίθμου της ερευνήτριας Budura A. 82

Εικόνα 15: Ποσοστό ακρίβειας για συνολικά εκατό χρήστες και από τις τρείς λειτουργίες βαθμολόγησης 84

Εικόνα 16: Η δομή του τριμερούς γραφήματος (Tripartite graph) 86

Εικόνα 17: Ετικέτες σήμανσης από το σύστημα del.icio.us, βάσει των συνδέσεων χρηστών ...87

Εικόνα 18: Η ταξινόμηση σύμφωνα με τους ερευνητές Mika P. και Akkerman. Η ……..92

Λίστα Πινάκων

Πίνακας 1: Αποτελέσματα μοντέλων υπαγωγής 38

Πίνακας 2: Συνολικός αριθμός ετικετών, με τους αντίστοιχους χρήστες και πηγές, για κάθε σύστημα σήμανσης 40

Πίνακας 3: Ποσοστά ακρίβειας από τρία διαφορετικά μοντέλα βαθμολόγησης 83

Πίνακας 4: Αποτελέσματα έρευνας αξιολόγησης 91

Πίνακας 5: Χαρακτηριστικά Σημαντικότερων Μεθόδων και Εφαρμογών Σύνδεσης 101

Πίνακας 6: Χαρακτηριστικά Υπόλοιπων Μεθόδων και Εφαρμογών Σύνδεσης 102

ΚΕΦΑΛΑΙΟ 1

1.1 Εισαγωγή

Το όραμα του Σημασιολογικού Ιστού αφορά στη δυνατότητα καλύτερης οργάνωσης του διαδικτυακού περιεχομένου, και στη βελτίωση σημαντικών τεχνολογιών όπως η αναζήτηση, η περιήγηση και η ευρετηρίαση του συνόλου της διαθέσιμης πληροφορίας. Για να αποδοθεί το διαδικτυακό περιεχόμενό του Παγκόσμιου Ιστού με ρητό και τυπικό τρόπο (formally annotated), χρησιμοποιούνται οι οντολογίες. Παρόλο που ο Σημασιολογικός Ιστός αφορά στην καλύτερη αξιοποίηση της πληροφορίας από τις μηχανές, η δημιουργία και η εξέλιξή των οντολογιών πραγματοποιείται μόνο με την βοήθεια μηχανικών γνώσης και ειδικών σε ένα γνωστικό αντικείμενο.

Τα τελευταία μάλιστα χρόνια, καταγράφεται μια συνεχής όξυνση του ενδιαφέροντος για την ενασχόληση με τα Κοινωνικά δίκτυα, που προωθούν και ευνοούν τη συνεργασία ανάμεσα στα μέλη μιας κοινότητας. Συστήματα σήμανσης όπως το Flickr και το del.icio.us γίνονται όλο και πιο δημοφιλή, καθώς καλύπτουν ένα ευρύ φάσμα πόρων και κοινοτήτων, με έναν τεράστιο αριθμό συμμετεχόντων που διαμοιράζουν και επισημαίνουν τις πηγές πληροφορίας με ελεύθερες, χωρίς κανένα περιορισμό, λέξεις-κλειδιά. Αυτές οι αυθαίρετες αναθέσεις ονομάζονται **ετικέτες σήμανσης** (tags) ενώ ο πυρήνας της δομής των δεδομένων τους είναι ένα **folksonomy**. Πιο συγκεκριμένα, ο όρος αυτός αναφέρεται σε ένα είδος ταξινομίας που παράγεται από τον ίδιο τον χρήστη για την κατηγοριοποίηση και ανάκτηση του περιεχομένου στο διαδίκτυο (σύνδεσμοι, φωτογραφίες), χρησιμοποιώντας αυτές τις «ανοικτού τύπου» ετικέτες.

Κατά καιρούς, οι οντολογίες και τα folksonomies έχουν απασχολήσει αρκετούς ερευνητές, γεγονός που αναπόφευκτα οδήγησε στη σύγκρισή τους. Η πλειοψηφία αυτών, θεωρεί πως ο συνδυασμός των τεχνολογιών του Σημασιολογικού Ιστού, των Κοινωνικών Δικτύων και του περιεχομένου των πόρων πληροφορίας δύναται να οδηγήσει στην πλήρη λειτουργία του Παγκόσμιου Ιστού. Απόρροια της συγκεκριμένης άποψης είναι ο **Σημασιολογικός Κοινωνικός Ιστός** (Social Semantic Web), που στοχεύει στην αντιμετώπιση

των αδυναμιών των folksonomies (πολλαπλά νοήματα ετικετών, αμφισημία, συνώνυμες ετικέτες κοκ) με την αξιοποίηση της ρητής και τυπικής αναπαράστασης που παρέχουν οι οντολογίες για μεγαλύτερη βελτίωση και ακρίβεια στην αναζήτηση, στην περιήγηση και στην ευρετηρίαση της διαθέσιμης πληροφορίας.

1.2 Προσδιορισμός του Προβλήματος

Στον Σημασιολογικό Ιστό, όπως έχει ήδη επισημανθεί, η απόδοση του διαδικτυακού περιεχομένου με επίσημο τρόπο γίνεται μέσω των οντολογιών. Η δημιουργία και η τροποποίησή τους δεν πραγματοποιείται από τους απλούς χρήστες της κοινότητας, αλλά από ειδικούς, γεγονός που καθιστά δύσκολη την άμεση ανανέωση των πληροφοριών. Αντιθέτως, στα Κοινωνικά Δίκτυα, παρά τα πλεονεκτήματα της αυθαίρετης ταξινομίας τους που προκύπτει από την συνεργασία των χρηστών και από τα εύχρηστα εργαλεία τους, παρατηρείται το φαινόμενο της ασάφειας που οδηγεί σε αναποτελεσματικές μεθόδους ευρετηρίασης και ανάκτησης της πληροφορίας.

Προκειμένου να αντιμετωπιστούν τα προαναφερθέντα προβλήματα αρκετοί ερευνητές πρότειναν την «αναδυόμενη σημασιολογία» (emergent semantics). Ο συγκεκριμένος όρος αναφέρεται σε ένα σύνολο αρχών και τεχνικών ανάλυσης της εξέλιξης των αποκεντρωμένων σημασιολογικών δομών, που είναι κατανεμημένα σε μεγάλης κλίμακας συστήματα πληροφοριών. Η γενική ιδέα είναι πως κανένας δεν λειτουργεί ως διαχειριστής, με την σημασιολογία να αναδύεται μέσω της αλληλεπίδρασης οντοτήτων (για παράδειγμα πρακτόρων λογισμικού). Γι' αυτόν ακριβώς τον λόγο, παρόλο το αρχικό στάδιο, οι ερευνητές θέλησαν μέσω ποικίλων μεθόδων και εφαρμογών να συνδέσουν τον Σημασιολογικό Ιστό με τα Κοινωνικά δίκτυα ώστε να υπερκαλυφθούν οι αδυναμίες και από τις δύο πλευρές.

Με τον όρο «σύνδεση» αναφερόμαστε είτε στην αξιοποίηση της σημασιολογικής πληροφορίας για καλύτερες υπηρεσίες κοινωνικών δικτύων, είτε στην αξιοποίηση της πληροφορίας από Κοινωνικά δίκτυα για την εύρεση σημασιολογικών συσχετίσεων που αποτυπώνουν την αντίληψη μιας κοινότητας ή και των μελών της για ένα πεδίο γνώσης. Το όραμά τους αφορά σε μια διαδικτυακή αυτόνομη κοινότητα, η οποία θα οργανώνεται από μόνη της και θα περιλαμβάνει περισσότερους εξειδικευμένους πράκτορες λογισμικού που θα

συνεργάζονται δυναμικά σε ανοιχτά περιβάλλοντα ενώ καθένας από αυτούς θα μπορεί να οργανώνει την πληροφορία σύμφωνα με μια αυτοκαθοριζόμενη οντολογία, καθιερώνοντας σχέσεις και τροποποιώντας έννοιες μόνο όταν είναι απαραίτητο για την συνεργασία.

1.3 Σκοπός Διπλωματικής Εργασίας

Σκοπός της παρούσας διπλωματικής εργασίας είναι η καταγραφή των προσπαθειών σύνδεσης των τεχνολογιών του Σημασιολογικού Ιστού με τα Κοινωνικά Δίκτυα, προκειμένου να καταλήξουμε σε έναν «ανοιχτό» και «έξυπνο» Παγκόσμιο Ιστό. Συγκεκριμένα, οι στόχοι που θα καλυφθούν είναι οι εξής:

- Να περιγραφούν οι έννοιες του Σημασιολογικού Ιστού και των Κοινωνικών Δικτύων.
- Να προσδιοριστεί η φύση, καθώς και όλα τα προβλήματα και οι αδυναμίες των οντολογιών και των folksonomies.
- Να παρουσιαστούν τα σημαντικότερα συστήματα σήμανσης, που έπαιξαν καταλυτικό ρόλο για τον προσδιορισμό των αποτελεσμάτων πολλών ερευνητικών προσεγγίσεων.
- Να καταγραφούν περιληπτικά όλες οι προσπάθειες σύνδεσης του Σημασιολογικού Ιστού με τα Κοινωνικά δίκτυα.
- Να περιγραφούν με λεπτομέρειες και να συγκριθούν οι σημαντικότερες προσεγγίσεις, ανάλογα με την πρωτοτυπία ή με την ακρίβεια των αποτελεσμάτων τους.
- Να προσδιοριστούν τα χαρακτηριστικά όλων των προσεγγίσεων της έρευνάς μας, συγκεντρωμένα σε έναν πίνακα.

1.4 Δομή

Η Διπλωματική εργασία αποτελείται από έξη κεφάλαια:

Στο **2ο κεφάλαιο** αναλύονται λεπτομερώς οι όροι Σημασιολογικός Ιστός, Κοινωνικά δίκτυα, Οντολογίες και Folksonomies. Παράλληλα επισημαίνονται όλα τα πλεονεκτήματα και οι αδυναμίες τους, επιχειρώντας επιπροσθέτως μια σύγκριση ανάμεσα στις δύο τελευταίες έννοιες (Οντολογίες, Folksonomies). Στο τελευταίο κομμάτι του κεφαλαίου παρατίθενται

τέσσερα από τα πιο δημοφιλή συστήματα σήμανσης, στοχεύοντας στην παρουσίαση συγκεκριμένων λειτουργιών τους που χρησιμοποιήθηκαν από τους ερευνητές στις προσεγγίσεις τους.

Στο **3ο κεφάλαιο** προσδιορίζονται οι ιδιαιτέρως σημαντικοί και ουσιαστικοί για τη διπλωματική εργασία, όροι «Σημασιολογικός Κοινωνικός Ιστός» και «Σημασιολογικά Κοινωνικά Δίκτυα». Εν συνεχεία, αναφέρονται περιληπτικά όλες οι μέθοδοι και οι εφαρμογές σύνδεσης της Σημασιολογικής πληροφορίας με τα Κοινωνικά δίκτυα και μετέπειτα αναλύονται οι σημαντικότερες από αυτές.

Στο **4ο κεφάλαιο** γίνεται σύγκριση ανάμεσα στις κυριότερες προσεγγίσεις, παραθέτοντας παράλληλα δύο πίνακες με τα κύρια χαρακτηριστικά όλων των μεθόδων και εφαρμογών που ερευνήθηκαν.

Στο **5ο κεφάλαιο** επιχειρείται μια γενική ανασκόπηση με την προβολή των τελικών συμπερασμάτων από την ενασχόλησή μας με τα Σημασιολογικά Κοινωνικά Δίκτυα.

Τέλος, στο **6ο κεφάλαιο** παρατίθεται η βιβλιογραφία που χρησιμοποιήθηκε για την ολοκλήρωση της παρούσας Διπλωματικής εργασίας.

ΚΕΦΑΛΑΙΟ 2

2.1 Σημασιολογικός Ιστός

Το όραμα του Σημασιολογικού Ιστού (Semantic Web), εμφανίστηκε για πρώτη φορά πριν από περίπου δέκα χρόνια από τον Tim Berners-Lee [7]. Συμφώνα με αυτόν ορίζεται ως «η προέκταση του σημερινού Παγκόσμιου Ιστού (*Εικόνα 2*) στον οποίο η πληροφορία έχει καλά καθορισμένο νόημα, πράγμα που διευκολύνει τη συνεργασία ανάμεσα στους υπολογιστές και στους ανθρώπους-χρήστες τους». Επιτρέπει την αυτόματη κατανόηση, την επεξεργασία και την ενοποίηση των πόρων στο διαδίκτυο από «έξυπνα» προγράμματα - πράκτορες, κάτι που δεν απαιτεί επομένως την ανθρώπινη διαμεσολάβηση. Αποστολή του είναι η καλύτερη οργάνωση του περιεχομένου στο διαδίκτυο, έτσι ώστε να βελτιωθεί η αναζήτηση, η περιήγηση και η ένταξη της πληροφορίας σε αυτό.

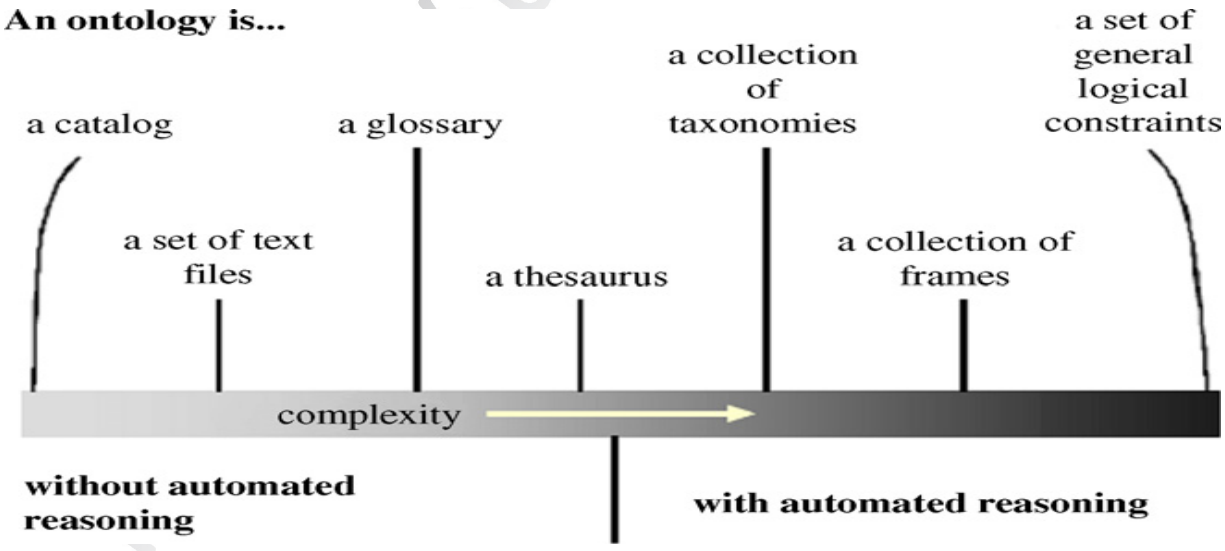
Ο Σημασιολογικός Ιστός, σε αντίθεση με το Web 2.0, απαιτεί ένα σημαντικό επίπεδο γνώσης από τους χρήστες του, ειδικότερα σε γλώσσες και τεχνικές, για την αναπαράσταση της γνώσης. Οι βασικές τεχνολογίες του αντιστοιχούν σε ένα σύνολο τυπικών γλωσσών και εργαλείων καθορισμού εννοιολογικών μοντέλων με συγκεκριμένες προδιαγραφές (formal conceptual models). Οι βασικές γλώσσες που χρησιμοποιούνται για το σκοπό αυτό είναι το RDF (Resource Description Framework), το RDF Schema, η OWL (Web Ontology Language) και η SPARQL, μια γλώσσα επερωτήσεων για την RDF. Όλες αυτές οι γλώσσες δημιουργήθηκαν στα θεμέλια των URIs (Uniform Resource Identifier), XML (Extensible Markup Language) και XML namespaces. Για να αποδοθεί το διαδικτυακό περιεχόμενό του Παγκόσμιου Ιστού με ρητό και τυπικό τρόπο (formally annotated), χρησιμοποιούνται οι οντολογίες, οι οποίες ουσιαστικά περιγράφουν τα αντικείμενα ενός συγκεκριμένου πεδίου γνώσης (domain) και τις μεταξύ τους σχέσεις.

2.1.1 Οντολογίες

Πρόκειται για όρο που χρησιμοποιείται ήδη από την αρχαιότητα και σχετίζεται, όπως άλλωστε μαρτυρά και η ετυμολογική ονομασία του, με τη φύση των όντων, ήτοι καθετί που τα προσδιορίζει, τα περιγράφει και τα διαχωρίζει μεταξύ τους. Η οντολογία έχει διατυπωθεί από

πολλούς ερευνητές με διαφορετικούς ορισμούς, κάτι που αποδεικνύει πως δεν είναι προφανής η σημασία της σε όλους. Μερικοί από αυτούς παρατίθενται ως εξής :

1. «Μια οντολογία ορίζει τους βασικούς όρους και τις σχέσεις που περιλαμβάνει το λεξιλόγιο μιας συγκεκριμένης θεματικής περιοχής (γνωστικού πεδίου- domain), καθώς και τους κανόνες για τον συνδυασμό τους και τις μεταξύ τους σχέσεις» [8].
2. «Είναι ένα ιεραρχικά δομημένο σύνολο όρων, για την περιγραφή μιας θεματικής περιοχής που μπορεί να χρησιμοποιηθεί ως ραχοκοκαλιά μιας βάσης γνώσεων» [9].
3. «Παρέχει τα μέσα για την σαφή περιγραφή εννοιών, πέρα από την ήδη υπάρχουσα γνώση» [10].
4. «Είναι ένα σύνολο μηχανισμών αναπαράστασης, που μπορούν να ταξινομηθούν ανάλογα με την εκφραστικότητα τους» [11], όπως φαίνεται και στην *εικόνα 1*. Οι ερευνητές υποστηρίζουν πως υπάρχει ένα σημείο (ένδειξη με τη μαύρη γραμμή) όπου ο αυτοματοποιημένος συλλογισμός γίνεται χρήσιμος : Αυτό είναι το σημείο όπου η οντολογία έχει τουλάχιστον μια σταθερή ιεραρχία, επιτρέποντας σχέσεις υπαγωγής.



Εικόνα 1: Μορφές οντολογιών ανάλογα με την εκφραστικότητα τους κατά Smith και Welty [11].

Πέρα από αυτούς όμως, ο πιο διαδεδομένος ορισμός στην βιβλιογραφία είναι αυτός που εκφράστηκε από τον ***Gruber T.*** [12]. Σύμφωνα με αυτόν:

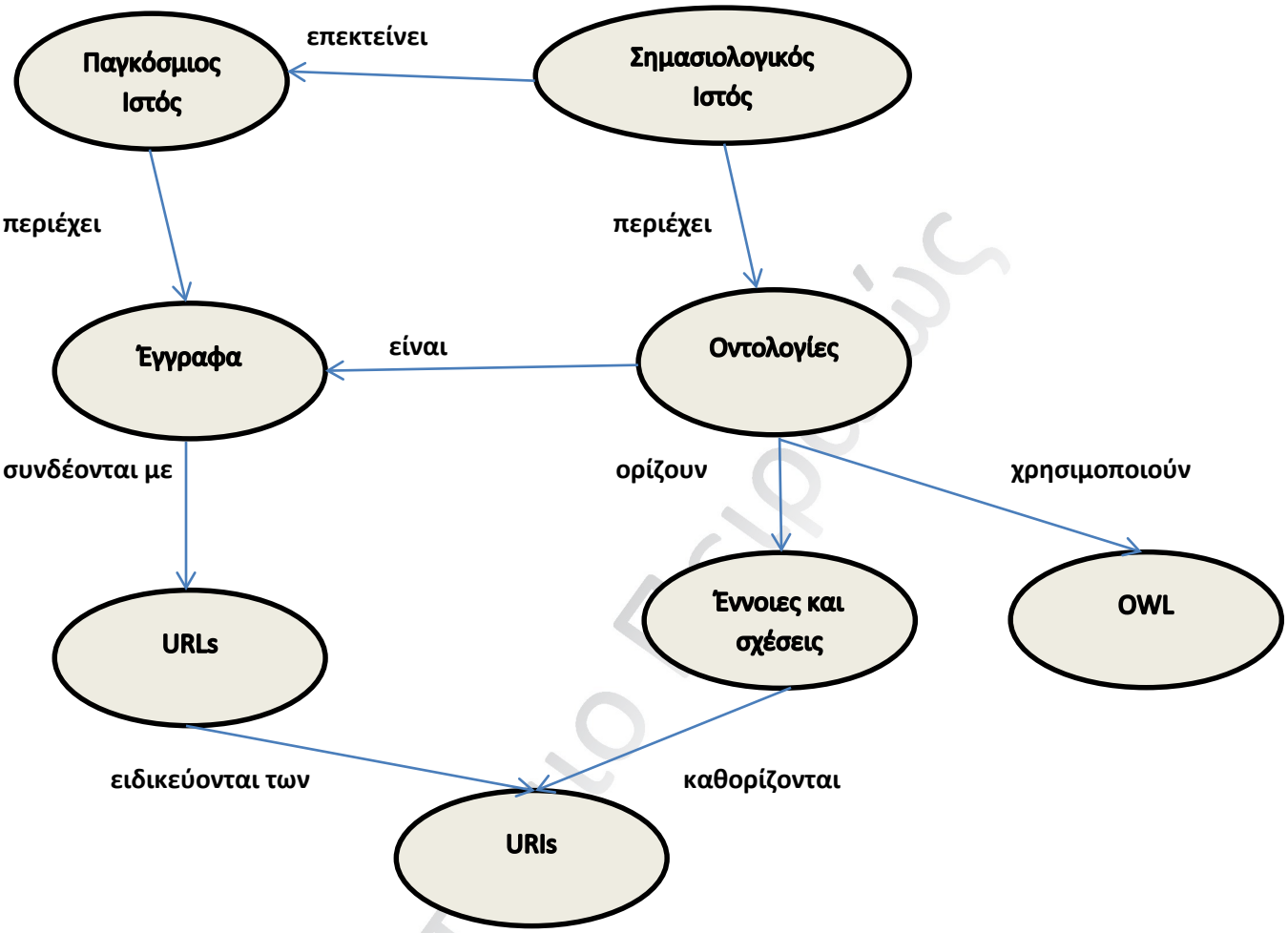
5. «Μία οντολογία είναι η επίσημη και σαφή προδιαγραφή, μιας κοινής εννοιολογικής αναπαράστασης».

Ο όρος «εννοιολογική αναπαράσταση» (conceptualization) αναφέρεται σε ένα αφηρημένο μοντέλο φαινομένων/οντοτήτων του κόσμου, με προσδιορισμένες τις έννοιες που σχετίζονται με αυτά (Consensual Knowledge). Θα πρέπει να είναι τυπική (formal), καθώς θα πρέπει να είναι μηχανικά αναγνώσιμη και σαφής. Τέλος, ο όρος «κοινή» (shared) υποδηλώνει το γεγονός ότι θα πρέπει να παρέχει κοινά αποδεκτή γνώση σε όλα τα μέλη μιας κοινότητας.

Οι οντολογίες διαδραματίζουν κεντρικό ρόλο στο όραμα του Σημασιολογικού Ιστού και θεωρούνται από πολλούς ως η «ραχοκοκαλιά» του. Καθιερώνουν λεξιλόγια και σημασιολογικές ερμηνείες όρων που γίνονται εύκολα κατανοητές από τις μηχανές. Όπως ανέφερε χαρακτηριστικά ο *Gruber T.* [13], χρησιμοποιούνται κυρίως για τον διαμοιρασμό κοινής κατανοητής πληροφορίας ανάμεσα στους απλούς χρήστες και στους ειδικούς (expert agents). Ουσιαστικά προσδίδουν στις εκφράσεις σημασία, κάνοντας δηλαδή το περιεχόμενο του Παγκόσμιου Ιστού να έχει κάποιο νόημα. Παρόλο που ο Σημασιολογικός Ιστός προορίζεται για μηχανές, η δημιουργία και η ανάκτηση περιεχομένου μπορεί να χαρακτηριστεί ως ανθρωποκεντρική (human-incentive).

Στις περισσότερες υπάρχουσες οντολογίες οι χρήστες δεν λαμβάνουν μέρος στην διαδικασία παραγωγής τους, εξαιτίας κάποιων σημαντικών εμποδίων, όπως :

- Η έλλειψη εύχρηστων και έξυπνων εργαλείων.
- Ο περιορισμένος χρόνος για αξιόπιστη γνώση. Η εξέλιξη και τροποποίηση των παραδοσιακών οντολογιών απαιτεί τη συμμετοχή του δημιουργού-ειδικού τους. Ο περιορισμένος χρόνος για την υλοποίηση των απαιτούμενων αλλαγών οδηγεί πολλές φορές σε καθυστέρηση της ανανέωσης της οντολογίας, γεγονός μη αποδεκτό στα συστήματα με δυναμικά πεδία γνώσης (domains).



Εικόνα 2: Σχέση Παγκόσμιου Ιστού με Σημασιολογικό

2.2 Κοινωνικά Δίκτυα (Web 2.0)

Ο όρος Web 2.0 χρησιμοποιήθηκε για πρώτη φορά το 2004 από τον Tim O'Reilly [14] κατά τη διάρκεια ενός συνεδρίου, όπου προτείνονταν ιδέες για την αναβάθμιση του παγκόσμιου ιστού. Ο όρος θεωρείται ακόμα δόκιμος αν και ορισμένοι ειδικοί, όπως ο Tim Berners Lee [15], διατηρούν ισχυρές αμφιβολίες για την αξία του ως όρο. Εξαιτίας της πληθώρας των κοινωνικών διαστάσεών του, όλες οι νέες εφαρμογές και ιστοσελίδες είναι ιδιαιτέρως εύχρηστες και κατανοητές στο ευρύ κοινό, καθώς δεν απαιτούνται ιδιαίτερες γνώσεις και δεξιότητες για την ενασχόληση τους με αυτές. Ουσιαστικά, παρακινεί όλους τους χρήστες, απλούς και ειδήμονες, να χρησιμοποιήσουν ετικέτες για τη σήμανση του πόρου της πληροφορίας που τους ενδιαφέρουν κατά κάποιο τρόπο, με αποτέλεσμα καθένας από αυτούς να επωφελείται από τα εργαλεία αυτά χωρίς ιδιαίτερη επιβάρυνση.

Για τη σήμανση αυτή οι χρήστες έχουν τη δυνατότητα να χρησιμοποιήσουν ελεύθερο λεξιλόγιο, δίχως να υπάρχει κάποιος περιορισμός ή σύσταση. Όπως επισημαίνει χαρακτηριστικά ο Tim O'Reilly [14], «οι Web 2.0 εφαρμογές δεν έχουν κάποιο όριο αλλά ένα δυνατό πυρήνα. Ο καθένας μπορεί να τις απεικονίσει ως ένα σύνολο από αρχές και πρακτικές που ενώνονται σε ένα σύνολο από ιστοσελίδες, κάτι που αποδεικνύει και τις διαφορετικές αποστάσεις τους από τον πυρήνα». Οι ελευθερίες που προσφέρουν έχουν ως αποτέλεσα τη μετατροπή του παραδοσιακού συμβατικού διαδικτύου σε συνεχώς εμπλουτιζόμενο, με αυτό να αποτελεί σαφώς μια εξέλιξη που μπορεί να οδηγήσει σε Web 2.0 ιστοσελίδες με σημασιολογικό εμπλουτισμό.

2.2.1 Folksonomies

Ένα χαρακτηριστικό παράδειγμα Web 2.0 συστήματος είναι το *folksonomy* (*Εικόνα 3*), που αποτελεί ένα είδος ταξινομίας, παραγόμενης από τον ίδιο τον χρήστη, για την κατηγοριοποίηση και την ανάκτηση περιεχομένου στο διαδίκτυο, όπως για παράδειγμα φωτογραφίες (*Flickr*) ή συνδέσμους (*del.icio.us*), χρησιμοποιώντας ετικέτες σήμανσης με ελεύθερο λεξιλόγιο, χωρίς κάποιον περιορισμό. Είναι ένας όρος που πρωτοειπώθηκε από τον *Thomas Vander Wal* [16] και προέρχεται από τις λέξεις Folks (φίλοι) και taxonomy (ταξινομία). Είναι επίσης γνωστός και με άλλες ονομασίες όπως κοινωνική σήμανση (*social tagging*), κοινωνική ευρετηρίαση (*social indexing*), συνεργατική σήμανση (*collaborative tagging*) και κοινωνική ταξινόμηση (*social classification*). Τα μεταδεδομένα του παράγονται όχι μόνο από ειδικούς, αλλά και από δημιουργούς και καταναλωτές του περιεχομένου. Τα *folksonomies* θεωρούνται από πολλούς ως το κλειδί για την ανάπτυξη του Σημασιολογικού Ιστού - Web 3.0, στον οποίο κάθε ιστοσελίδα περιέχει τα αναγνώσιμα (από το σύστημα) μεταδεδομένα που περιγράφουν το περιεχόμενό της. Τέτοια μεταδεδομένα μπορούν να βελτιώσουν σε μεγάλο βαθμό την ακρίβεια (το ποσοστό των σχετικών εγγράφων που ανακτούνται) στις μηχανές αναζήτησης και ανάκτησης περιεχομένου. Γι' αυτόν ακριβώς τον λόγο απασχόλησαν και απασχολούν μέχρι και σήμερα πολλούς ερευνητές, οι οποίοι -εκτός από το γεγονός ότι τα χρησιμοποιούν για την έρευνα τους- εξετάζουν διάφορα συστήματα σήμανσης (όπως το *Flickr* και το *del.icio.us*) για να κατανοήσουν, πέρα από τα χαρακτηριστικά τους, τον τρόπο λειτουργίας και χρήσης τους [1][2].

Ένα folksonomy μπορεί να οριστεί [17] ως μία πλειάδα F= (U,T,R,Y,≺) όπου:

- U,T και R είναι πεπερασμένα σύνολα, των οποίων τα στοιχεία είναι οι χρήστες (users), οι ετικέτες (tags) και οι πόροι (resources),
- Y είναι η τριμερής μεταξύ τους σχέση, δηλαδή Y⊆U×T×R , των οποίων τα στοιχεία είναι γνωστά και ως αναθέσεις ετικετών (tag assignments),
- ≺ είναι ουσιαστικά η σχέση ανάμεσα στον χρήστη και στις ετικέτες σήμανσης, έχοντας ο ίδιος την δυνατότητα να ορίσει τη δυναμική τους, δηλαδή $\prec \subseteq U \times T \times T$.

Συνολικά, παρόλο που ο όρος ταξινομία (taxonomy) συχνά χρησιμοποιείται για συστήματα folksonomies, θα λέγαμε πως γίνεται περισσότερο κατηγοριοποίηση (categorization). Η τελευταία είναι λιγότερο αυστηρή, με τα όριά της να είναι λιγότερο σαφή.

Ένα ιδιαίτερα σημαντικό χαρακτηριστικό των folksonomies σύμφωνα με τον [18], είναι ότι δεν υπάρχει ιεραρχία (parent-child) ανάμεσα στους όρους που χρησιμοποιούνται για τη σήμανση του περιεχομένου. Μπορούν όμως να παραχθούν συσχετίσεις μεταξύ των ετικετών σήμανσης, με την ομαδοποίηση των οποίων να βασίζεται στα κοινά URLs. Αντιθέτως, στα επίσημα συστήματα ταξινόμησης, παρατηρούνται πολλαπλές σαφείς σχέσεις ανάμεσα στους όρους. Αυτές οι σχέσεις περιλαμβάνουν κατά κύριο λόγο ευρεία (*broad*) ή περιορισμένα (*narrow*) folksonomies, δύο κατηγορίες που πρωτοαναφέρθηκαν από τον ερευνητή Thomas Vander Wal [16] με συγκεκριμένες ιδιότητες και χρήσεις.

Σε ένα **broad folksonomy** (όπως το σύστημα σήμανσης del.icio.us) πολλοί χρήστες επισημειώνουν τον ίδιο πόρο, χωρίς κάποιο περιορισμό στο λεξιλόγιο και την γλώσσα. Η συγκεκριμένη κατηγορία παρέχει εργαλεία για την διερεύνηση των τάσεων σήμανσης από μεγάλες ομάδες ανθρώπων, για την περιγραφή των αντικειμένων. Μπορεί να χρησιμοποιηθεί επίσης για την επιλογή επιθυμητών όρων ή για την εξαγωγή ελεγχόμενου λεξιλογίου. Η πραγματική δύναμη των broad folksonomies αφορά τον πλούτο της μάζας, με τον πλουραλισμό του καθορισμού και της περιγραφής των πραγμάτων να οδηγεί στην άμεση εξέλιξη της ταξινόμησης και ισχύς τους.

Από την άλλη μεριά σε ένα **narrow folksonomy**, όπως το σύστημα σήμανσης Flickr, μόνο ο δημιουργός του εκάστοτε πόρου τον επισημειώνει (χρησιμοποιώντας μία ή περισσότερες ετικέτες), είτε για προσωπική μελλοντική ανάκτηση ή για δική του διευκόλυνση. Η συγκεκριμένη κατηγορία παρέχει οφέλη όσον αφορά τα αντικείμενα σήμανσης που δεν εντοπίζονται εύκολα με τα παραδοσιακά εργαλεία (full-text search, text-related tools) ή δεν μπορούν απλά να περιγραφούν με βάση το υπάρχον λογισμικό του διαδικτύου. Τέλος ένα narrow folksonomy παρέχει σε διάφορες ομάδες-στόχους ένα κοινό ειδικό λεξιλόγιο, με τους χρήστες να προσθέτουν ετικέτες με δική τους γλώσσα, κάνοντας την μελλοντική ανάκτηση γρήγορη και αποδοτική.

Εικόνα 3: Μορφή Folksonomy

2.1.1.1 Αδυναμίες

Το ανεξέλεγκτο λεξιλόγιο που χρησιμοποιείται στα *folksonomies*, είναι λογικό να επιφέρει και ορισμένες αδυναμίες. Η ασάφεια του νοήματος των ετικετών είναι ένα συχνό φαινόμενο, με τις πολλαπλές λέξεις και τα συνώνυμα να οδηγούν πολλούς χρήστες στη σήμανση του ίδιου περιεχομένου με διαφορετικές ετικέτες. Το ζήτημα αυτό δύναται να οδηγήσει στο πρόβλημα της αναποτελεσματικότητας, όσον αφορά στο περιεχόμενο της αναζήτησης και της ευρετηρίασης.

- **Ασάφεια**: Προέρχεται από την τάση έτερων χρηστών να χρησιμοποιούν διαφορετικές ετικέτες για τη σήμανση του ίδιου περιεχομένου. Τα ακρωνύμια (*acronyms*) είναι ένα χαρακτηριστικό παράδειγμα ασάφειας, καθώς χρησιμοποιούνται ευρέως σε τέτοιου είδους συστήματα. Για παράδειγμα σε ορισμένες ιστοσελίδες το "ANT" αναφέρεται στις λέξεις "Actor Network Theory", ενώ στο σύστημα σήμανσης del.icio.us στο εργαλείο δημιουργίας προγραμματισμού Apache Ant.

- **Κενά και Πολλαπλές Λέξεις**: Σε πολλά συστήματα σήμανσης οι χρήστες χρησιμοποιούν πολλαπλές λέξεις, για τον ορισμό μιας ετικέτας, χωρίς κάποιο κενό ανάμεσα σε αυτές. Ένα σχετικό παράδειγμα είναι η σήμανση "elearningvideouniversitypiraeus". Με αυτό τον τρόπο ίσως θέλουν να ιεραρχήσουν τις λέξεις ή απλά να αποδώσουν μια κατηγορία που έχει πολλαπλούς όρους.
- **Συνώνυμα**: Ένα σημαντικό πρόβλημα στα συστήματα *folksonomies*, είναι ότι δεν υπάρχει έλεγχος συνωνύμων. Ως εκ τούτου, πολλές ετικέτες έχουν το ίδιο νόημα. Για παράδειγμα οι ετικέτες "mac", "mackintosh" και "apple", αναφέρονται στην εταιρία υπολογιστών Apple. Επιπροσθέτως, σε ορισμένα συστήματα σήμανσης (κυρίως στο Flickr), ο ενικός και ο πληθυντικός αριθμός οδηγούν σε ορισμένα προβλήματα ασάφειας που αφορούν στην αναζήτηση ορισμένων ετικετών σήμανσης. Χαρακτηριστικό παράδειγμα αποτελεί οι ετικέτες "flower" και "flowers".

2.1.1.2 Δυναμική και Πλεονεκτήματα

Πέρα από τις αδυναμίες, ένα *folksonomy* διακρίνεται για τα ποικίλα πλεονεκτήματά του γεγονός που αποδεικνύει και τον βαθμό διάδοσης και χρησιμότητάς του. Πέρα από αυτά, αναδεικνύονται ορισμένα στοιχεία που φανερώνουν ιδιαίτερα τη δυναμική του όπως :

- **Η τυχαία και απρόσμενη ανακάλυψη** (*serendipity*) περιεχομένου, που προέρχεται από την περιήγηση του χρήστη στο σύστημα. Οποιοσδήποτε θα μπορούσε να αξιολογήσει ένα folksonomy, χρησιμοποιώντας συγκεκριμένα ερωτήματα (queries) από τους χρήστες και αξιολογώντας ποια δεδομένα έχουν σημανθεί με τις αντίστοιχες σχετιζόμενες λέξεις-κλειδιά. Βέβαια, ένα τέτοιο ενδεχόμενο θα αγνοούσε την ευρύτερη περιήγηση στην οποία τέτοιου είδους συστήματα σήμανσης στηρίζουν την δυναμική τους.
- **Αντικατοπτρίζουν τις ανάγκες του κάθε χρήστη** (*desire lines*). Τα παραδοσιακά συστήματα ανάκτησης περιεχομένου, περιλαμβάνουν δύο ή και περισσότερα λεξιλόγια [18]. Πολλές φορές η κατανόηση αυτών (για παράδειγμα τα λεξιλόγια του δημιουργού ή του διαχειριστή) είναι ιδιαίτερα δύσκολη, πράγμα μη συμβατό για τέτοιου είδους

συστήματα. Τα folksonomies παράγονται από τους απλούς χρήστες, και όχι από ειδήμονες μιας γνωστικής περιοχής ή τους δημιουργούς της γνώσης.

- Είναι **"φθηνά" και επεκτάσιμα**, καθώς δημιουργούνται από τους ίδιους τους χρήστες, σε αντίθεση με τις παραδοσιακές μεθόδους πρόσθεσης δεδομένων. Το συνολικό κόστος σε χρόνο και προσπάθεια είναι μακράν χαμηλότερο από ότι στα συστήματα που στηρίζονται σε πολύπλοκες ιεραρχικές ταξινομήσεις.
- Η **ανατροφοδότηση** μέσα σε αυτά είναι άμεση, καθώς, όταν ο χρήστης υποσημειώνει ένα αντικείμενο με μια ετικέτα, του εμφανίζονται ομάδες πόρων με την ίδια σήμανση. Εφόσον ο ίδιος θεωρήσει πως η διαδικασία δεν έχει πραγματοποιηθεί σωστά, μπορεί να αλλάξει την ετικέτα ή να προσθέσει κάποια άλλη. Εάν κάτι τέτοιο δεν αποφέρει τα αναμενόμενα στην αναζήτηση αποτελέσματα, δύναται να προσαρμόσει με κάποιο άλλο πρότυπο τη σήμανσή του, επηρεαζόμενος από άλλους.
- Η ανατροφοδότηση αυτή μπορεί να οδηγήσει σε μία μορφή **ασύμμετρης επικοινωνίας ανάμεσα στους χρήστες**, διαμέσου των μεταδεδομένων. Από τη στιγμή που ένας από αυτούς δημιουργεί μια ετικέτα, συνεισφέροντας με τον τρόπο αυτό στο περιεχόμενο, καθίσταται πιο εύκολα αντιληπτός στους υπόλοιπους χρήστες. Οι ετικέτες σήμανσης, σε τέτοιου είδους συστήματα, ουσιαστικά αποτελούν ένα δίαυλο επικοινωνίας ανάμεσα στα μέλη της κοινότητας.
- Τέλος, ενθαρρύνουν τη **συνεργασία**, καθώς η ετικέτα κάθε χρήστη δεν αποτελεί μόνο προσωπικό υλικό αλλά μπορεί επίσης, εφόσον ο ίδιος επιθυμεί, να διαμοιραστεί και σε άλλα μέλη της κοινότητας. Πιο συγκεκριμένα, υπάρχουν συστήματα σήμανσης που επιτρέπουν στο χρήστη να διακρίνει τις επαφές του σε κατηγορίες φίλων ή οικογένειας (Flickr), καθώς και άλλα που χρησιμοποιούν συνδρομές (subscribes) για διάφορες λίστες σήμανσης άλλων μελών (del.icio.us).

2.2.3 Σημαντικότερα Κοινωνικά Συστήματα Σήμανσης

Το ***Del.icio.us*** [1] είναι ίσως το δημοφιλέστερο εργαλείο οργάνωσης ιστοσελίδων. Σύμφωνα με τον δημιουργό του Schachter J [19]. «είναι ένας κοινωνικός διαχειριστής σελιδοδεικτών (social bookmarks manager) που επιτρέπει την εύκολη πρόσθεση ιστοσελίδων στην προσωπική συλλογή συνδέσμων ενός χρήστη, ανάλογα με την επιθυμία του, με

αποτέλεσμα να έχει την δυνατότητα αυτός με την σειρά του να τις κατηγοριοποιεί με λέξεις-κλειδιά και να τις διαμοιράζει από τον browser στο web, ώστε να είναι προσβάσιμα από οποιονδήποτε χρήστη». Παρότι δε μπορεί να χαρακτηριστεί ως καινούρια ή μοναδική ιδέα πρωτοπορεί στο γεγονός ότι οι λέξεις-κλειδιά αποτελούν ουσιαστικά τη βασική δομή ταξινόμησης στο σύστημα. Με αυτές, ο κάθε χρήστης περιγράφει ή οργανώνει το περιεχόμενο με το λεξιλόγιο που ο ίδιος επιθυμεί, χωρίς να υπάρχει κανένας περιορισμός. Πέρα από το ότι πρόκειται για ένα δωρεάν εργαλείο, είναι ιδιαίτερα κατανοητό και εύκολο στην χρήση του. Για να γίνει συγκεκριμένα κάποιος μέλος χρειάζεται μόνο το όνομα χρήστη (Username), τον κωδικό πρόσβασης (Password) και το προσωπικό του e-mail. Μετά την δημιουργία του λογαριασμού, ο χρήστης έχει τη δυνατότητα να εισάγει το εργαλείο ως πρόσθετη εφαρμογή (add-on) στον browser του, για εύκολη και γρήγορη χρήση.

Κατά την περιήγησή του για παράδειγμα σε μια ιστοσελίδα ενδιαφέροντος (η διαδικασία μπορεί να πραγματοποιηθεί και για πολλές ιστοσελίδες), εφόσον ο χρήστης επιθυμεί να την αποθηκεύσει στο del.icio.us, επιλέγει αρχικά τον σελιδοδείκτη. Στη συνέχεια, εισάγει μια ετικέτα σήμανσης, που τη συνδέει με την ιστοσελίδα και την αποθηκεύει. Εκτός από τη γρήγορη αποθήκευση και διαχείριση των προσωπικών ιστοσελίδων ενός χρήστη όπου κι αν βρίσκεται, το del.icio.us προβάλει και τους δημοφιλέστερους σελιδοδείκτες από πολλούς τομείς ενδιαφέροντος, καθώς και άλλους, ανάλογα με την θεματολογία σήμανσης που ο ίδιος ο χρήστης χρησιμοποιεί. Βέβαια η ευκολία αυτή οδηγεί ορισμένους χρήστες σε αρκετές λανθασμένες υποδείξεις στη σήμανση. Παρατηρείται κυρίως το φαινόμενο της ασάφειας, καθώς η πλειοψηφία επιλέγει γενικούς όρους για να περιγράψει συγκεκριμένα δεδομένα, χρησιμοποιεί μεγάλες φράσεις-κλειδιά (αντί λέξεις-κλειδιά) και σημεία στίξης. Παρόλα τα προβλήματα, το συγκεκριμένο σύστημα σήμανσης αριθμεί πάνω από πέντε εκατομμύρια εγγεγραμμένους στο σύστημα χρήστες με 180 εκατομμύρια περίπου σελιδοδείκτες σήμανσης.

Το ***Flickr*** [2] είναι ένα σύστημα αποθήκευσης, διαχείρισης και διαμοιρασμού φωτογραφιών, στα πλαίσια των εφαρμογών κοινωνικής δικτύωσης. Παρά το γεγονός ότι είναι σύστημα ελεύθερης πρόσβασης, μπορεί κάποιος να πληρώσει για έναν λογαριασμό με περισσότερες δυνατότητες (μεγαλύτερος αποθηκευτικός χώρος). Μέσω αυτού ο εκάστοτε

χρήστης ουσιαστικά κοινοποιεί τις εικόνες που θέλει και τις οργανώνει, δίνοντας παράλληλα την ίδια δυνατότητα στους εγγεγραμμένους στο σύστημα χρήστες να παρέχουν ετικέτες σήμανσης. Παρά την ύπαρξη της ελευθερίας, οποιοσδήποτε δύναται να ρυθμίσει την ιδιωτικότητα των δεδομένων του. Με τις ετικέτες (tags) να αποτελούν βασικό χαρακτηριστικό του ιστοχώρου, ο εκάστοτε χρήστης βρίσκει σχετικές εικόνες με παρόμοιο περιεχόμενο, πατώντας επάνω στην ετικέτα που ο ίδιος έχει βάλει. Επιπροσθέτως, χαρακτηριστική είναι η χρήση των tag clouds, τα οποία ουσιαστικά αποτελούνται από ετικέτες που έχουν σημανθεί με τις δημοφιλέστερες λέξεις-κλειδιά. Γι' αυτό ακριβώς, το Flickr θεωρείται ίσως το αντιπροσωπευτικότερο παράδειγμα χρήσης της κοινωνικής σήμανσης (Folksonomy, social tagging), καθώς μέσω της συνεργασίας των χρηστών πραγματοποιείται η σήμανση του περιεχομένου. Τέλος, ο σχολιασμός είναι ελεύθερος, χωρίς κάποια σύσταση ή περιορισμό από το ίδιο το σύστημα, γεγονός που προκαλεί ορισμένα προβλήματα κατά την αναζήτηση. Παρόλα αυτά, η βάση δεδομένων του ιστοχώρου περιλαμβάνει εκατομμύρια φωτογραφίες που οργανώνονται και σημαίνονται από περίπου οκτώμισι εκατομμύρια εγγεγραμμένους χρήστες. Είναι τόσο δημοφιλές που υπολογίζεται πως περίπου δώδεκα χιλιάδες φωτογραφίες εισάγονται στο σύστημα το δευτερόλεπτο, με δύο εκατομμύρια να αποτελούν το ρεκόρ σε μία μέρα.

Το ***BibSonomy*** [3] είναι μια κοινωνική πλατφόρμα σελιδοσήμανσης και διαμοιρασμού βιβλιογραφίας. Μέσα σε αυτήν οι χρήστες έχουν τη δυνατότητα να οργανώσουν τους προσωπικούς τους σελιδοδείκτες και να δημοσιεύσουν τις καταχωρήσεις τους. Επιπλέον, μπορούν να χρησιμοποιήσουν αυτές για σήμανση, γεγονός που συμβάλει στη δομή και στην εύρεση της νέας πληροφορίας. Στο BibSonomy, η ταξινομία folksonomy εξελίσσεται ουσιαστικά από τη συμμετοχή των ερευνητικών ομάδων, των κοινοτήτων μάθησης και των μεμονωμένων χρηστών, οργανώνοντας τις πληροφορίες ανάλογα με τις ανάγκες τους.

Το ***CiteULike*** [4] είναι μια πλατφόρμα κοινωνικής σελιδοσήμανσης, η οποία δημιουργήθηκε από τον *Cameron R.* το 2004. Μέσα σε αυτόν τον ιστόχωρο, οι ερευνητές έχουν τη δυνατότητα να ανταλλάξουν πληροφορίες και να διαμοιράσουν τις επιστημονικές εργασίες τους. Κατά την περιήγηση ενός χρήστη, περισυλλέγεται από scripts η βιβλιογραφία

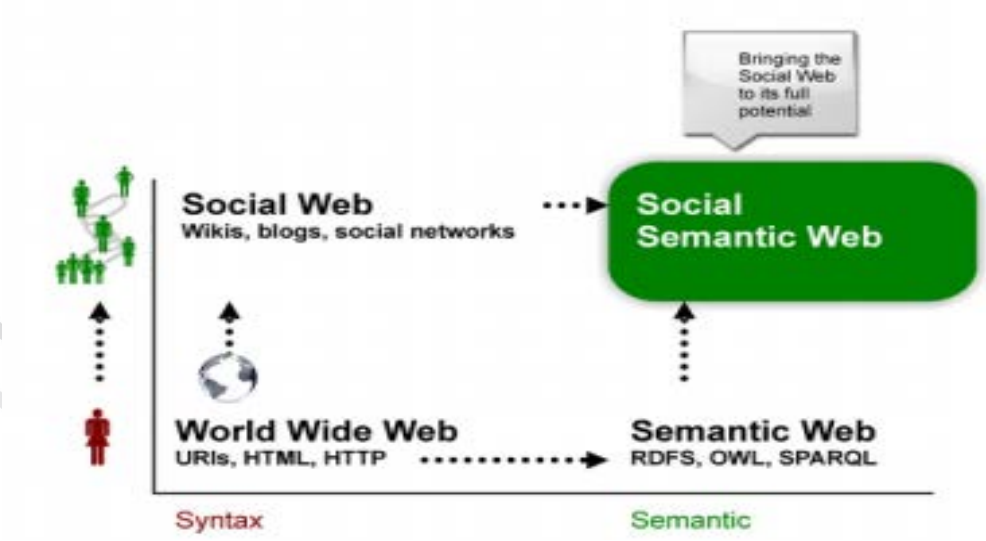
σε σελιδοδείκτες, κάτι που επιτρέπει την εισαγωγή των άρθρων από τις διάφορες βάσεις δεδομένων στο συγκεκριμένο σύστημα σήμανσης. Στη συνέχεια, το σύστημα προσπαθεί να καθορίσει από μόνο του τα μεταδεδομένα κάθε άρθρου (τίτλος, συγγραφείς κλπ.) αυτόματα. Η οργάνωση της κάθε προσωπικής βιβλιοθήκης επιτυγχάνεται από τις ελεύθερες, χωρίς περιορισμούς, ετικέτες σήμανσης, γεγονός που παράγει ένα folksonomy ταξινομημένο ανάλογα με τις ακαδημαϊκές ανάγκες των ερευνητών.

ΚΕΦΑΛΑΙΟ 3

3.1 Σημασιολογικός Κοινωνικός Ιστός

Ο Σημασιολογικός Ιστός είναι ένα ιδανικό μέρος για τη δημιουργία κοινωνικών ιστοσελίδων, με συγκεκριμένους κανόνες, παρέχοντας κατάλληλα πρότυπα για τον διαμοιρασμό δεδομένων και την κατανεμημένη πληροφορία, από τη συμμετοχή και τη συνεργασία των μελών των διάφορων κοινοτήτων. Η εφαρμογή του στο Κοινωνικό Δίκτυο αποφέρει τον Σημασιολογικό Κοινωνικό Ιστό (*Εικόνα 4*), η έννοια του οποίου καλύπτει τα οποιαδήποτε προβλήματα στις κοινωνικές αλληλεπιδράσεις στο διαδίκτυο και οδηγεί στη δημιουργία σαφούς και ρητής , με σημασιολογικά πλούσιες αναπαραστάσεις γνώσης.

Προκειμένου να επιτευχθεί αυτή η σημασιολογική αναβάθμιση των κοινωνικών δικτύων, δίνεται έμφαση στην ανάδειξη της ταυτότητας του κάθε χρήστη και της ηλεκτρονικής προσωποποίησής του (ετικέτες σήμανσης). Συνεπώς, ο Σημασιολογικός Κοινωνικός Ιστός μπορεί να θεωρηθεί ως ένα δίκτυο συστημάτων συλλογικής γνώσης (collective knowledge systems), ικανό να παρέχει χρήσιμες πληροφορίες βασισμένες στις συνεισφορών των μελών του. Συνδυάζει τεχνολογίες, στρατηγικές και μεθοδολογίες από τον Σημασιολογικό Ιστό, από τα κοινωνικά δίκτυα και από τις τεχνολογίες Web 2.0.



Εικόνα 4: Σημασιολογικός Κοινωνικός Ιστός [20]

3.2 Σημασιολογικά Κοινωνικά Δίκτυα

Οι ετικέτες, στα παραδοσιακά κοινωνικά συστήματα σήμανσης, περιλαμβάνουν σημασιολογικές πτυχές του καθενός που δεν ορίζονται με κατανοητό τρόπο. Αναγνωρίζοντας τα τυπικά και ρητά στοιχεία που αντιστοιχούν στις ετικέτες και στις μεταξύ τους σχέσεις, υπάρχει πιθανότητα να γίνει σαφές ένα σημαντικό μέρος της υποκειμενικής γνώσης. Αυτό ακριβώς επιθυμούν να επιτύχουν τα **Σημασιολογικά Κοινωνικά Δίκτυα**, τα οποία στοχεύουν στο να ξεπεράσουν τις αδυναμίες των folksonomies (πολλαπλά νοήματα, συνώνυμα κτλ) με την αξιοποίηση των Σημασιολογικών τεχνολογιών στην αναζήτηση, στην περιήγηση και στην ευρετηρίαση του περιεχομένου των πόρων πληροφορίας. Το κύριο χαρακτηριστικό τους είναι ότι περιλαμβάνουν ετικέτες σήμανσης που συντελούν στη δημιουργία οντολογιών, με έννοιες που είναι κοινωνικά αντιληπτές και "συμφωνημένες" μεταξύ των μελών κοινοτήτων. Επομένως αντί να στηρίζονται αποκλειστικά στην σημασιολογία που παρέχεται από μια προδιαγεγραμμένη οντολογία, οι χρήστες διαμορφώνουν διαφανώς την οντολογία, μέσω των αντίστοιχων συστημάτων, με την αλληλεπίδρασή τους.

3.3 Μέθοδοι και εφαρμογές σύνδεσης σημασιολογικής πληροφορίας με τα κοινωνικά δίκτυα

Στην βιβλιογραφία η πλειοψηφία των ερευνητών υποστηρίζει το ίδιο επιχείρημα: Ο συνδυασμός των Σημασιολογικών και Web 2.0 τεχνολογιών, μπορεί να οδηγήσει στην πλήρη λειτουργία του διαδικτύου. Οι περισσότεροι από αυτούς θεωρούν πως ο Σημασιολογικός Ιστός έχει ωριμάσει αρκετά για να υποστηρίξει τα ανοιχτά δεδομένα και τη διαλειτουργικότητα του περιεχομένου. Επιπροσθέτως, είναι της άποψης ότι οι διάφορες αδυναμίες των Κοινωνικών Δικτύων, όπως αυτές έχουν ήδη αναφερθεί παραπάνω, δύναται να υπερκαλυφθούν με τις τεχνολογίες του Σημασιολογικού Ιστού.

Ορισμένοι ερευνητές από αυτούς, μέσω των μεθόδων τους, έχουν ως στόχο να διευρύνουν τα folksonomies σε οντολογίες (folksonomy enrichment), χρησιμοποιώντας είτε αυτοματοποιημένες διαδικασίες, είτε με την παρέμβαση του χρήστη. Κάποιοι, σε αντίθεση με άλλους ερευνητές, χρησιμοποίησαν στις προσεγγίσεις τους αλγόριθμους ενώ αρκετοί είχαν ως στόχο απλά να προσδιορίσουν τα προβλήματα των folksonomies [18], ενώ ειπώθηκε επιπλέον

ότι η συνύπαρξη ανάμεσα στον Σημασιολογικό Ιστό και στα Κοινωνικά Δίκτυα υπάρχει ήδη [21].

Παρακάτω παρατίθενται περιληπτικά όλες οι μέθοδοι και οι εφαρμογές σύνδεσης της σημασιολογικής πληροφορίας με τα κοινωνικά δίκτυα, με τις οποίες ασχοληθήκαμε για την ολοκλήρωση της διπλωματικής εργασίας. Όλες αυτές οι προσεγγίσεις χωρίστηκαν σε δύο κατηγορίες: στις **σημαντικότερες**, οι οποίες αποτελούν και το κύριο μέρος της έρευνάς και στις **υπόλοιπες**, με την αναφορά μας μόνο στα κύρια χαρακτηριστικά τους. Οι σημαντικότερες προσεγγίσεις (κεφάλαιο 3.5) επιλέχθηκαν με γνώμονα την αναγνωρισιμότητα, την πρωτοπορία και την αποτελεσματικότητά τους. Η σύνοψη των κύριων χαρακτηριστικών καθεμίας από το σύνολο τους, που θα μας απασχολήσει σε επόμενα κεφάλαια της εργασίας μας, πραγματοποιήθηκε με την δημιουργία δύο πινάκων, ενός για κάθε κατηγορία (κεφάλαια 4.2, 4.3).

3.4 Προσπάθειες Σύνδεσης Σημασιολογικής Πληροφορίας με τα Κοινωνικά Δίκτυα Περιληπτικά

Ο ***Mika P.*** [21] θέλησε να προσδώσει κοινωνική διάσταση στο παραδοσιακό διμερές μοντέλο των οντολογιών (όροι / ετικέτες σήμανσης, πόροι), προσθέτοντας τους χρήστες. Χρησιμοποιώντας το υποσύνολο ετικετών σήμανσης του del.icio.us, δημιούργησε γραφήματα σχετικότητας ανάμεσα στις ετικέτες με τους χρήστες, καθώς και στις ετικέτες με τους πόρους. Χρησιμοποίησε τεχνικές ανάλυσης δικτύου τις οποίες εφάρμοσε σε αυτά για να ανακαλύψει τις παραγόμενες οντολογίες.

Οι ***Specia L.*** και ***Motta E.*** [22] με την προσέγγισή τους είχαν ως στόχο να καταστήσουν σαφή τη σημασιολογία στα κοινωνικά συστήματα σήμανσης. Γι' αυτό ακριβώς τον λόγο, ασχολήθηκαν τόσο με τις σχέσεις ανάμεσα στις ίδιες τις ετικέτες και όσο και με το πως αυτές μπορούν να οδηγήσουν στην ανάπτυξη των οντολογιών, εξετάζοντάς τις σε συνδιασμό με τη χρήση πηγών πληροφοριών, όπως το Google και η Wikipedia.

Η ***Budura A.*** [23], εμπνευσμένη από το χώρο των επιχειρήσεων, πρότεινε μια νέα προσέγγιση για το πρόβλημα της εξόρυξης της τεχνογνωσίας. Ανέπτυξε μια μέθοδο

πιθανοτήτων για την οικοδόμηση προφίλ για κάθε χρήστη (έμπειρων σε κάποιους τομείς) βασισμένο στις ετικέτες σήμανσής του, πειραματιζόμενη με τα κοινωνικά επιχειρησιακά δίκτυα σήμανσης Dogear και IBMr.

Ο ***Schmitz P.*** [24] επιδίωξε να κατασκευάσει ένα μοντέλο το οποίο αξιοποιεί τις στατιστικές τεχνικές επεξεργασίας γλωσσικών πόρων για να δημιουργήσει οντολογίες μέσα από τη βάση δεδομένων του Flickr. Για τον λόγο αυτό, διερεύνησε ένα πιθανολογικό - υπαγωγικό μοντέλο (subsumption – based model) , προσαρμόζοντας τα στατιστικά όρια για την αυθαίρετη χρήση των ετικετών σήμανσης και φιλτράροντας το εξαιρετικά ιδιότυπο λεξιλόγιο της εκάστοτε ιστοσελίδας σήμανσης που χρησιμοποίησε για την προσέγγιση του. Τα παραγόμενα υποδέντρα (subtrees) του μοντέλου του αξιολογούνται με μη αυτοματοποιημένο τρόπο.

Οι ***Heymann P.*** και ***Garcia-Molina H.*** [25] χρησιμοποίησαν τον αλγόριθμό τους προκειμένου να μετατρέψουν ένα μεγάλο σύνολο ετικετών σε ιεραρχική ταξινομία πλοήγησης. Οι ετικέτες συγκεντρώνονται σε διανύσματα που δηλώνουν τον αριθμό των φορών που έχει χρησιμοποιηθεί μια ετικέτα για κάθε σημαινόμενο πόρο. Η συνάρτηση ομοιότητας τους υπολογίζεται μέσω του συνημίτονου μεταξύ των διανυσμάτων και στη συνέχεια θεσπίζεται ένα όριο για να ελαττωθούν οι μη σχετικές τιμές. Δημιούργησαν το διάγραμμα ομοιότητας για ένα συγκεκριμένο σύνολο δεδομένων και αξιοποίησαν την έννοια της κεντρικότητας του κοινωνικού δικτύου για να δείξουν ότι η κοινωνική της έννοια (graph centrality) μπορεί να είναι καθοριστικής σημασίας σε συνεργατικά συστήματα σήμανσης. Τέλος, εφάρμοσαν τον αλγόριθμό τους σε δύο διαφορετικά σύνολα δεδομένων (del.icio.us και CiteULike), αλλά μόνο για το πρώτο σύνολο δεδομένων τα αποτελέσματά τους θεωρήθηκαν ενθαρρυντικά.

Ο ***Wu Z.*** [26] προτείνει στατιστικές τεχνικές για την εξόρυξη της σιωπηρής σημασιολογίας (implicit semantics), η οποία είναι ενσωματωμένη στη συνύπαρξη μεταξύ των χρηστών, των πόρων και των ετικετών στα Folksonomies. Χρησιμοποιεί ένα πιθανοτικό – παραγωγικό μοντέλο (probabilistic generative model), για να αναπαραστήσει τη συμπεριφορά σήμανσης του εκάστοτε χρήστη στο del.icio.us και να αντλήσει αυτόματα την αναδυόμενη σημασιολογία των ετικετών. Κατά την προσέγγιση ομαδοποιούνται συνώνυμες ετικέτες με

πολλά νοήματα εντοπίζονται και διαχωρίζονται. Επιπλέον, η παραγόμενη σημασιολογία αξιοποιείται για την ανακάλυψη και την αναζήτηση σελιδοδεικτών σήμανσης.

Ο ***Jaschke R.*** [17] χρησιμοποίησε έναν αλγόριθμο (TRIAS) για την εύρεση των χρηστών, που εμμέσως συμφωνούν σε μια κοινή έννοια (στα σύνολα των πόρων). Εφόσον γίνει η διαδικασία της εύρεσης, επιστρέφει ένα σύνολο από τριπλέτες (A,B,C), που κάθε μια από αυτές αποτελείται από ένα σύνολο από A χρήστες, ένα άλλο από B ετικέτες και ένα από C πόρους. Κάθε τριπλέτα έχει την ακόλουθη ιδιότητα : Κάθε χρήστης από το A, έχει κάνει σήμανση για κάθε πόρο στο C, με όλες τις ετικέτες από το B. Επιπροσθέτως, κανένα από τα παραπάνω σύνολα δε μπορεί να αυξηθεί εάν δεν μειωθεί ταυτόχρονα και ένα από αυτά. Επομένως, το σύνολο A χρηστών μοιράζεται έμμεσα μια έννοια, το σύνολο B των ετικετών είναι η πρόθεση της νοηματοδότησης και το σύνολο C των πόρων η πραγμάτωση της. Ουσιαστικά η προσέγγισή του αφορά στην εξόρυξη δεδομένων, με τη δουλειά του να επεκτείνεται στην ανακάλυψη όλων των άγνωστων στοιχειοσυνόλων των χρηστών, των ετικετών σήμανσης και των πηγών, που βρίσκονται μέσα στα Folskonomies. Θέλοντας να αποδείξει την αποδοτικότητά του, ο ερευνητής εφάρμοσε τον αλγόριθμο σε τρία διαφορετικά συστήματα σήμανσης (del.icio.us, IT Baseline Security Manual, BibSonomy).

Οι ***Ereteo G.*** και ***Gandon F.*** [27] έχοντας ως στόχο την ανίχνευση της διαδικτυακής κοινότητας, δημιούργησαν τον αλγόριθμο SemTagP. Επωφελούμενος από τα σημασιολογικά δεδομένα κατά την δόμηση των RDF σχημάτων από τα κοινωνικά δίκτυα, ο SemTagP όχι μόνο εντοπίζει αλλά και σημαίνει κοινότητες, αξιοποιώντας τις ετικέτες και τις σημασιολογικές μεταξύ τους σχέσεις. Εν τέλει, για να αξιολογήσουν τα αποτελέσματα της μεθοδολογίας τους τον εφάρμοσαν στο κοινωνικό δίκτυο ADEME Agency.

Ο ***Begelman G.*** [28] πρότεινε μια αυτοματοποιημένη τεχνική πιο αποτελεσματική στην εγγραφή, στην αναζήτηση και στην εξερεύνηση ετικετών σήμανσης. Για όσες ετικέτες χρησιμοποιούνται για τον ίδιο πόρο, ο αλγόριθμος υπολογίζει τον αριθμό συνύπαρξης κάθε ζεύγους εξ αυτών. Όλες οι σχετικές ή μη ετικέτες αναπαριστώνται σε ένα γράφημα, στο οποίο ο προτεινόμενος αλγόριθμός τις ομαδοποιεί, μετρώντας τη συνεκτικότητα αρθρωμάτων

(modularity) του γραφήματος. Ο ερευνητής, θέλοντας να παρατηρήσει την αποδοτικότητα της προσέγγισής του, εφάρμοσε τον αλγόριθμό του στη βάση δεδομένων RawSugar.

Οι ερευνητές ***Xu Z.*** και ***Fu Y.*** [29] καθόρισαν γενικά κριτήρια για τη δημιουργία ενός επιθυμητού συστήματος σήμανσης για την αναγνώριση των καταλληλότερων ετικετών, μειώνοντας με τον τρόπο αυτό τα ανεπιθύμητα αποτελέσματα (noisy, spam). Αυτά τα κριτήρια, τα οποία προσδιορίστηκαν χάρη σε μια μελέτη τους μέσω πραγματικών χρηστών στο σύστημα My Web 2.0, περιλαμβάνουν εξασφάλιση καλής ανάκλησης, λιγότερη προσπάθεια για να μειωθεί το κόστος πλοήγησης και υψηλή δημοτικότητα ετικετών σήμανσης για να διασφαλιστεί η ποιότητά τους. Ο προτεινόμενος αλγόριθμός, χρησιμοποιεί ένα μέτρο καταλληλότητας, το οποίο έχει ως κύριο στόχο την ελάττωση των ανεπιθύμητων ετικετών. Επιπλέον εξασφαλίζει ότι οι προτεινόμενες ετικέτες διατηρούν καλή ισορροπία μεταξύ της κάλυψης και τη δημοτικότητάς της (coverage, popularity). Τα πρώτα αποτελέσματα, φανέρωσαν πως ένας τέτοιος αλγόριθμος είναι αποτελεσματικός για να προτείνει κατάλληλες ετικέτες που ταιριάζουν με τις αναμενόμενες ιδιότητες.

Ο ***Gruber T.*** [13] πρότεινε τη θέσπιση ενός κοινού προτύπου σήμανσης για να πετύχει την διαλειτουργικότητα ανάμεσα σε τέτοιου είδους συστήματα. Θεώρησε τις ετικέτες ως μορφή ψήφου και τη σήμανση από τους χρήστες, με καινοτομία ενσωμάτωσης υπερσυνδέσεων (hyperlinks), ως μέτρο αναγνωρισημότητας. Ο ερευνητής προβάλλει την TagOntology, μια οντολογία που από τη μια αναγνωρίζει και επισημοποιεί μια κοινή αντίληψη της δραστηριότητας της σήμανσης και από την άλλη αναπτύσσει την τεχνολογία που οδηγεί τα folksonomies, μέσω της οντολογίας, σε σημασιολογικό επίπεδο.

Οι ***Golder S.*** και ***Huberman B.*** [30] για να κατανοήσουν ευκολότερα την δομή των κοινωνικών δικτύων σήμανσης ανέλυσαν τα δεδομένα ενός δημοφιλούς Folksonomy, του del.icio.us. Η έρευνά τους εξετάζει την πορεία σήμανσης από τους χρήστες, καθώς και το πως η διαδικασία σήμανσης φτάνει σε μια κατάσταση "ισορροπίας". Επιπλέον, οι ερευνητές ασχολούνται με την τακτικότητα σήμανσης από τον χρήστη, με τη συχνότητα των ετικετών και τα είδη αυτών που είναι περισσότερο δημοφιλή, προτείνοντας ένα δυναμικό μοντέλο για την πρόβλεψη σταθερών μοντέλων σήμανσης. Τέλος, συζήτησαν για τις δυσκολίες των

Folksonomies, που σχετίζονται με τα σημασιολογικά θέματα των ετικετών, όπως η συνωνυμία και η πολυσημία.

Ο ***Mathes A.*** [18] εξετάζοντας δύο γενικές προσεγγίσεις για τη δημιουργία των μεταδεδομένων από επαγγελματίες (professional creation) και από συντάκτες (author creation), παρατήρησε ότι οι χρήστες που συμβάλλουν στη δημιουργία των πληροφοριών ουσιαστικά δεν συμμετέχουν στη διαδικασία. Γι' αυτό ακριβώς τον λόγο και υποστήριξε ότι οι ετικέτες, ως μεταδεδομένα που δημιουργούνται από τον ίδιο τον χρήστη (user generated metadata), συνήθως διευκολύνουν κάποια οργάνωση και πρόσβαση στις πληροφορίες. Η έρευνά του πραγματοποιήθηκε με τη βοήθεια δύο κοινωνικών δικτύων σήμανσης, του Flickr και του del.icio.us, για την καλύτερη κατανόηση της βάσης ταξινόμησης.

Ο ***Φωντόπουλος Γ.*** [5] με την διδακτορική διατριβή του "RichTags: A Social Semantic Tagging System" ερεύνησε τα παραδοσιακά Κοινωνικά δίκτυα σήμανσης, παρατηρώντας πολλές ασάφειες και αδυναμίες στην ακρίβειά των μεθόδων ανάκτησης. Εξέτασε παρόμοιες προσεγγίσεις έτερων ερευνητών, προτείνοντας εν συνεχεία το «Rich Tags». Πρόκειται για ένα εργαλείο που φαίνεται να υπερκαλύπτει τις αδυναμίες των Folksonomies σε ορισμένο βαθμό, με τη χρήση Σημασιολογικών τεχνολογιών. Το κύριο χαρακτηριστικό του είναι ότι οι ετικέτες αποτελούν μια κατανοητή οντολογία, την οποία διαχειρίζονται συλλογικά οι χρήστες.

Η ***Αγγελέτου Σ.*** [6] ασχολήθηκε με τη μελέτη των πιθανών συνδυασμών των ετερογενών τεχνολογιών του Σημασιολογικού ιστού και του Web2.0, με σκοπό να συνεισφέρει στη δημιουργία ενός «ανοιχτού» και «έξυπνου» διαδικτυακού περιβάλλοντος. Επιθυμώντας να εμπλουτίσει τα Folksonomies, δημιούργησε ένα σημασιολογικό στρώμα σαν ενδιάμεση διεπαφή μεταξύ των ερωτημάτων των χρηστών και των ετικετών σήμανσης. Κατάφερε, στο τέλος της μεθοδολογίας της, να καθορίσει την σημασιολογία των ετικετών αυτοματοποιημένα, χρησιμοποιώντας το εργαλείο FLOR.

Ο ***Cattuto C.*** [31] πρότεινε την άποψη ότι η μέτρηση κάθε σχέσης ανάμεσα στις ετικέτες σήμανσης θα πρέπει να εξαρτάται από τον σημασιολογικό βαθμό τους (semantic relation). Επιπροσθέτως, θεώρησε πως οι διαφορετικοί τύποι ομοιότητας μεταξύ των ετικετών (συνύπαρξη, αλγόριθμος FolkRank, tag context) θα πρέπει να μετρώνται ανάλογα με τον τύπο

των σημασιολογικών σχέσεων στις οποίες αντιστοιχούν. Γι' αυτό ακριβώς τον λόγο εφάρμοσε τη μέθοδό του στο σύστημα σήμανσης del.icio.us και με την βοήθεια του σημασιολογικού λεξικού WordNet για να βρει ετικέτες σχετικές μεταξύ τους, οι οποίες μοιράζονται μια σχέση υπαγωγής με μια δεδομένη ετικέτα t. Με μία σειρά πειραματισμών κατέληξε στο συμπέρασμα πως η μελέτη του είναι καλύτερη στην ανακάλυψη συνωνύμων (μέσω WordNet), στην ιεράρχηση των εννοιών (μέσω FolkRank και συνύπαρξης), στις συστάσεις ετικετών (tag recommendation, μέσω FolkRank και συνύπαρξης) και στην επέκταση ερωτημάτων (query expansion, μέσω ομοιοτήτων πλαισίου πόρου και ετικέτας).

Ο **Damme C.** [32] προσπάθησε να τροποποιήσει τις ετικέτες σήμανσης με την ενσωμάτωση πολλαπλών πόρων και τεχνικών, όπως για παράδειγμα την στατιστική ανάλυση των folksonomies, online λεξιλογικούς πόρους (Google, Wikipedia), οντολογίες (μηχανή αναζήτησης Swoogle) και σημασιολογικούς πόρους (WordNet). Πιο συγκεκριμένα, χρησιμοποίησε σημασιολογικές πηγές πληροφοριών για να εμπλουτίσει το νόημα των ετικετών σήμανσης, συνδυάζοντας αυτή την τεχνική με πραγματικές οντολογίες που απορρέουν από τα folksonomies. Τέλος, ο ερευνητής συνέστησε, για την ομαλή λειτουργία της προσέγγισής του, την παρέμβαση της ανθρώπινης νοημοσύνης στην έγκριση της αυτοματοποιημένης λαμβανόμενης σημασιολογίας των ετικετών.

Οι ***Marihno L.B*** και ***Buza K.*** [33] ασχολήθηκαν με μια μέθοδο που εμπλουτίζει αυτόματα το Folksonomy, με την εισαγωγή ενός αλγορίθμου που βασίζεται στις διαδεδομένες τεχνικές εξόρυξης στοιχειοσυνόλων. Προκειμένου να εκτιμηθεί ποσοτικά η έρευνα τους πρότειναν ένα νέο σημείο αναφοράς, την αξιολόγηση της οντολογίας (task-based) με την ποιότητα της να μετράται με βάση την ευκολία εύρεσης των ατομικών πληροφοριών κάθε χρήστη. Για να αξιολογήσουν την αποτελεσματικότητα της προσέγγισης τους, οι ερευνητές διεξήγαγαν ορισμένα πειράματα σε πραγματικά δεδομένα από τα συστήματα σήμανσης Last.fm και Musicmoz.

Ο ***Knerr T.*** [34] χρησιμοποίησε τις τεχνολογίες του Σημασιολογικού Ιστού για την ανάπτυξη μιας οντολογίας για folksonomies, καθιστώντας εφικτή τη διαλειτουργικότητα και την αυτοματοποιημένη επεξεργασία. Για να αξιολογήσει την μέθοδο του ο ερευνητής

χρησιμοποίησε δεδομένα από το κοινωνικό σύστημα σήμανσης del.icio.us, καθώς και ορισμένα ερωτήματα SPARQL με τα δεδομένα αυτά, για να αποδείξει ότι οι πληροφορίες του Folksonomy μπορεί να είναι εύκολα προσβάσιμες μέσω της οντολογίας του.

Ο ***Markines B.*** [35] ασχολήθηκε με την σύγκριση διαφόρων γενικών θεωρητικών, στατιστικών και πρακτικών μετρήσεων ομοιότητας. Εφόσον εστίασε την προσοχή κυρίως στις σχέσεις ανάμεσα στις ετικέτες σήμανσης, τους πόρους και τους χρήστες, πρότεινε ένα εναλλακτικό τρόπο μέτρησης ομοιότητας βασιζόμενο στην αξιολόγηση από τον χρήστη μέσω του σημασιολογικό λεξικού WordNet και του πολυγλωσσικού καταλόγου συνδέσμων Open Directory Project.

Οι ***Alves H.*** και ***Santanche A.*** [36] προσπάθησαν με την προσέγγιση τους να δημιουργήσουν μια οντολογία με χαρακτηριστικά Folksonomy. Αρχικά διερεύνησαν την φύση των δύο προαναφερθέντων όρων για την ενδυνάμωση των διαδικασιών αναζήτησης και ταξινόμησης, καθώς για την βελτίωση των σχέσεων βαρύτητας (relationship weighting) και για τις λειτουργίες συμπερασμού (inference operations) με εννοιολογικά δεδομένα. Όλες οι παραπάνω ενέργειες πραγματοποιήθηκαν για την δημιουργία ενός εργαλείου (“folksonomized” ontology), στο οποίο εφάρμοσαν σύνολα ετικετών από τα συστήματα σήμανσης del.icio.us και Flickr.

Τέλος, ο ***Hotho A.*** [37] αντιλαμβανόμενος την περιορισμένη υποστήριξη για την ανάκτηση περιεχομένου πρότεινε έναν νέο αλγόριθμο αναζήτησης, τον FolkRank, ο οποίος αξιοποιεί την δομή ενός Folksonomy αναζητώντας κοινότητες μέσα σε αυτό. Για να αξιολογήσει την τεχνική του ο ερευνητής χρησιμοποίησε το σύστημα σήμανσης del.icio.us, με τον αλγόριθμό του να παράγει ως τελικό αποτέλεσμα ένα σύνολο σχετιζόμενων χρηστών και πόρων για μια δεδομένη ετικέτα. Επομένως βασιζόμενος στα προαναφερθέντα στοιχεία, ο FolkRank έχει την δυνατότητα να κάνει συστάσεις χρηστών και πόρων ενδιαφέροντος, οι οποίες παρουσιάζονται στο χρήστη κατά την διάρκεια της χρήσης ενός συστήματος folksonomy.

3.5 Σημαντικότερες Προσεγγίσεις

3.5.1 Μοντέλο Υπαγωγής

3.5.1.1 Γενικά

Ο ***Schmitz P.*** [24] θέλοντας να υπολογίσει πολύπλευρες οντολογίες μέσω ενός συστήματος σήμανσης (faceted ontologies), πρότεινε το μοντέλο υπαγωγής (subsumption-based model) για την εύρεση πιθανών σχέσεων εξάρτησης μεταξύ των ετικετών. Μία πολύπλευρη οντολογία αποτελείται από ένα σύνολο όψεων (facets), όπου κάθε μια από αυτές περιλαμβάνει ένα προκαθορισμένο σύνολο όρων δομημένων από μια υπαγωγική σχέση. Επιπλέον ο ερευνητής θεώρησε πως το μοντέλο του μπορεί να τροποποιηθεί για να ενσωματώνει κοινότητες σήμανσης. Κύριος στόχος του ήταν η δημιουργία ενός συστήματος, το οποίο διατηρεί την ευελιξία σήμανσης για τον σχολιασμό και που επωφελείται από την δύναμη και την χρησιμότητα των πολύπλευρων οντολογιών στην αναζήτηση και στην περιήγηση. Για να αξιολογήσει τα αποτελέσματα και να αποδείξει τις δυνατότητες μιας τέτοιας τεχνικής, χρησιμοποίησε το Flickr.

3.5.1.2 Επιρροές

Αρχικά κρίνεται σκόπιμο να αναφερθούν περιληπτικά τα κύρια χαρακτηριστικά του μοντέλου των *Sanderson M.* και *Croft B.* [38], το οποίο ο ερευνητής υποστήριξε σε αρκετά σημεία και χρησιμοποίησε μέρος του για την υλοποίηση της δικής του προσέγγισης. Με μια σειρά πειραμάτων ουσιαστικά είχε ως στόχο να δώσει λύση, στην βελτίωση της επίδοσης του μοντέλου αυτού. Στην έρευνα τους οι *Sanderson M*. και *Croft B.* περιέγραψαν την διάκριση μεταξύ ομάδων, στις οποίες τα μέλη τους μοιράζονται ίδιες αναλογίες από ένα σύνολο χαρακτηριστικών (*polythetic clusters*), και άλλων στις οποίες όλα τα μέλη μοιράζονται ένα μόνο κοινό χαρακτηριστικό (*monothetic clusters*). Για παράδειγμα η ταξινόμηση των ανθρώπων με βάση το φύλο τους αποτελεί μια monothetic ομαδοποίηση, αλλά μια άλλη με βάση το φύλλο και την δεξιότητα του χεριού για γράψιμο (δεξί ή αριστερό) είναι μία polythetic ομαδοποίηση. Υποστήριξαν πως οι χρήστες μπορούν να κατανοήσουν ευκολότερα την πρώτη (monothetic clusters) προαναφερθείσα κατηγορία ομάδων. Επιπλέον σε αντίθεση με την δεύτερη (polythetic clusters), η πρώτη κατηγορία θεωρείται πως υποστηρίζει πιο εύκολα την

σήμανση προσφέροντας διάφορα κοινά πρότυπα διεπαφής (interface paradigms), όπως για παράδειγμα η καθοδηγούμενη πλοήγηση (guided navigation).

Περιέγραψαν ένα απλό **στατιστικό μοντέλο υπαγωγής**, στο οποίο η ετικέτα X υπάγεται στην Y εάν :

$P(x|y >= 0.8)$ *και* $P(y|x < 1)$

Εφάρμοσαν το συγκεκριμένο μοντέλο για την εννοιολόγηση όρων, οι οποίοι εξάγονται από έγγραφα που επιστρέφονται για ένα ορισμένο ερώτημα. Η χρησιμοποίηση των αποτελεσμάτων ενός ερωτήματος οδηγεί στον περιορισμό του πεδίου των όρων.

3.5.1.3 Μέθοδος και Πειραματισμοί

Ο *Schmitz P.* με την σειρά του διερεύνησε το παραπάνω μοντέλο, εφαρμόζοντάς το σε ένα σύνολο ετικετών σήμανσης, που ήδη υπάρχουν, στη βάση δεδομένων του *Flickr*. Επιπλέον προσάρμοσε τα στατιστικά όρια για να φανερώσουν τη χρηστικότητα των ετικετών σήμανσης (ad hoc usage), προσθέτοντας παράλληλα φίλτρα για έλεγχο του εξαιρετικά ιδιότυπου λεξιλογίου.

Ως εκ τούτου, το X υπάγεται στο Y εφόσον:

$P(x|y >= t)$*και* $P(y|x < t)$,

$D_x >= D_{min}, D_y >= D_{min}$,

$U_x >= U_{min}, U_y >= U_{min}$

Όπου οι μεταβλητές :

- *x, y είναι οι ετικέτες,*
- *t είναι το όριο συνύπαρξης,*
- D_x *είναι το # (πλήθος) των εγγράφων στα οποία ο όρος x εμφανίζεται και πρέπει να είναι μεγαλύτερο από την μικρότερη τιμή* D_{min} *και*
- U_x *είναι το # των χρηστών που χρησιμοποιούν το x για ένα τουλάχιστον σχολιασμό φωτογραφίας και πρέπει να είναι μεγαλύτερος από την μικρότερη τιμή* U_{min}.

Αρχικά φίλτραρε τις φωτογραφίες στο *Flickr*, με σκοπό τον ορισμό της συνύπαρξης των ετικετών σήμανσης, απαιτώντας κατά ελάχιστο δύο όρους εξ αυτών. Στη συνέχεια πραγματοποίησε μια σειρά πειραμάτων μεταβάλλοντας τις μεταβλητές t, D_{min}, U_{min}, έχοντας ως στόχο να βρει μία ισορροπία που θα μειώνει από την μία στο ελάχιστο το ποσοστό λάθους συνύπαρξης και θα αυξάνει από την άλλη κατά πολύ τον αριθμό των προτεινόμενων ζευγών υπαγωγής. Χρησιμοποίησε επιπλέον "αυστηρές" τιμές (που πλησιάζει το 0,9) για τα όρια συνύπαρξης, με σκοπό την μείωση του ποσοστού λάθους ως κάποιο βαθμό αλλά δίχως τη μείωση των προτεινόμενων ζευγών. Οι προσδοκώμενες τιμές ήταν φανερά μειωμένες σε σχέση με τις αντίστοιχες των *Sanderson M. και Croft B.* (κυμαίνονταν μεταξύ 0,7 και 0,8), κάτι που δείχνει πως το μοντέλο του είναι πιο "ευαίσθητο" στις μεταβολές της μεταβλητής U_{min}, παρά της D_{min}. Παρατήρησε πως ρυθμίζοντας την U_{min} κάτω από το 5 παράγονταν πολλοί ιδιότυποι όροι στα ανεπιθύμητα ζεύγη υπαγωγής, κάτι που φανέρωνε πως οι επιθυμητές τιμές κυμαίνονταν ανάμεσα στο 5 και στο 10. Αντιθέτως οι τιμές της μεταβλητής D_{min} κυμαίνονταν ανάμεσα στο 5 και στο 40, κάτι που αποδείκνυε την χρησιμότητά της ως μέσο ρυθμίσεων. Ένα κοινό σημείο στο οποίο κατέληξε, ήταν πως η αύξηση του αριθμού των εγγράφων είναι ανάλογη με αυτήν των δύο παραπάνω μεταβλητών. Με τον αριθμό των φωτογραφιών στο Flickr να πλησιάζει το ένα εκατομμύριο, το λεξιλόγιο παρουσίαζε κάποια αστάθεια κάτι που φανέρωνε την ευαισθησία του μοντέλου σε αυτές τις παραμέτρους.

Εφόσον υπολογιστούν τα στατιστικά συνύπαρξης, επιλέγονται υποψήφια ζεύγη ετικετών με την χρησιμοποίηση καθορισμένων περιορισμών. Έπειτα μπορεί να δημιουργηθεί ένα "δέντρο" με πιθανές σχέσεις παιδιού-γονέα, αποτελούμενο από ορισμένα υποψήφια ζεύγη.

Για *παράδειγμα*, για κάθε όρο x, και δύο πιθανούς όρους γονέα (parent terms) P_{xi} και P_{xj}, εάν ο P_{xi} έχει πιθανή σχέση γονέα με τον όρο P_{xj}, τότε διαγράφεται ο P_{xi} από την λίστα με τους αντίστοιχους πιθανούς όρους για τον x. Ταυτόχρονα η συνύπαρξη των όρων x, P_{xi} και P_{xj} φανερώνει ότι η σχέση x-> P_{xj} είναι πιο πιθανή από μία απλή συνύπαρξη που μπορεί να υποδεικνύεται, και ομοίως η σχέση P_{xi}->P_{xj} πρέπει να ενισχυθεί με κάποιο τρόπο. Αυτό

πραγματοποιείται με την αύξηση της βαρύτητας συνύπαρξης (co-occurrence weight) κάθε όρου αναλογικά.

Για κάθε φύλλο του δέντρου θεωρείται ότι επιλέγεται η καλύτερη διαδρομή για την ρίζα, καθώς υποδεικνύεται το ποσοστό συνύπαρξης με πιθανούς γονείς για κάθε κόμβο και ενώνονται τα πιθανά μονοπάτια στα δέντρα. Ένα σημαντικό όμως πρόβλημα που παρατήρησε ο ερευνητής αφορά το μέγεθος των παραγόμενων δέντρων. Όσο μεγάλο είναι δηλαδή το σύνολο των εγγράφων, τόσο ευρεία είναι τα παραγόμενα δέντρα. Αυτό έχει ως αποτέλεσμα την εμφάνιση πολλών λανθασμένων μονοπατιών, κάτι που αντιμετώπισε μόνο με το φιλτράρισμα του συνόλου των δέντρων. Αυτό μπορεί να δικαιολογηθεί, καθώς ο αριθμός τους θεωρήθηκε υπέρογκος.

Επόμενο βήμα της έρευνάς του, ήταν η εφαρμογή των παρατηρήσεων του από τις προηγούμενες διαδικασίες στο ***Flickr***. Χρησιμοποίησε ένα στιγμιότυπο της βάσης δεδομένων από τον Ιούλιο του 2005. Μέχρι εκείνη την περίοδο είχαν αναρτηθεί περίπου 25 εκατομμύρια εικόνες με 65 εκατομμύρια σχόλια. Από αυτές 5 εκατομμύρια αποκλείστηκαν από την πειραματική του διαδικασία, καθώς δεν ήταν δημόσιες σε όλους τους χρήστες. Η τροποποίηση του συνόλου δεδομένων, με τέτοιο τρόπο ώστε τα προσωπικά στοιχεία των χρηστών (προσωπικά προφίλ, ID ανεβασμένων στοιχείων) και των φωτογραφιών (όλες αυτές με λιγότερους από 2 όρους φιλτράρονται) να γίνουν ανώνυμα, απέφερε ένα σύνολο από περίπου 9 εκατομμύρια φωτογραφίες. Το συνδεδεμένο με αυτά λεξιλόγιο είχε πάνω από 200.000 όρους και συνολικά πάνω από 8 εκατομμύρια ζεύγη. Όσο αφορά τη μορφή σήμανσης στο Flickr ο *Schmitz P.* παρατήρησε μια ασυμβατότητα στο λεξιλόγιο, που σχετιζόταν με τα όρια των λέξεων της. Για παράδειγμα η λέξη "san Francisco" συχνά εμφανιζόταν στο πείραμα του ως δύο όροι "san" και "Francisco". Επιπλέον πολλές φορές κάποιες ετικέτες προκάλεσαν σύγχυση (idiosyncratic annotation terms) για το μοντέλο σήμανσης, καθώς πολλές διαφορετικές λέξεις γράφονταν σαν μία χωρίς κενό ανάμεσά τους. Ένα τέτοιο παράδειγμα είναι η ετικέτα "johnandmaryswedding".

Τα δέντρα που προέκυψαν από την παραπάνω διαδικασία αξιολογήθηκαν με μη αυτοματοποιημένο τρόπο. Απόρροια αυτού, κάθε προτεινόμενο ζεύγος υπαγωγής μπορεί να

θεωρηθεί σωστό, σχετικό, συνώνυμο (περιλαμβάνοντας παραλλαγές της γλώσσας για κοινούς όρους όπως για παράδειγμα "Argos"/ "Argolis") ή λανθασμένο (noisy). Για παράδειγμα, ένα δέντρο μπορεί να περιέχει ως ετικέτα το "Argos" (η οποία αναφέρεται στην ομώνυμη πόλη), καθώς και ένα σύνολο άλλων ετικετών που συνδέονται με αυτήν με σχέση γονέα-παιδιού. Με άλλα λόγια, ετικέτες όπως οι "kastrolarisis", "anthrwpos", "plateiaargous", "Argolis" αποτελούν μερικά από τα παιδιά του γονέα Argos. Μερικά από αυτά με την σειρά τους, μπορούν να θεωρηθούν σχετικά ("Argolis"), λανθασμένα ("anthrwpos") κτλ.

Στηριζόμενος στην εμπειρία του, ο ερευνητής υπέθεσε πως οι εικόνες έχουν την δυνατότητα να σχολιαστούν και να ανακτηθούν ευκολότερα, βάση κάποιων παραγόντων, όπως το *μέρος-περιοχή (place)*, η *δραστηριότητα (activity)* και οι *αναπαραστάσεις (depictions).* Επιπλέον η κοινότητα του Flickr έχει από μόνη της έναν ακόμη παράγοντα, που μπορεί να θεωρηθεί ως συναίσθημα ή αντίδραση. Στα αποτελέσματα του πειράματός του το μεγαλύτερο ποσοστό του διαμοιρασμένου λεξιλογίου συνδέεται με τοπωνύμια.

Για τον παράγοντα *περιοχές* θεώρησε μόνο τα γεωγραφικά τοπωνύμια, όπως και κάποια μέρη ενδιαφέροντος που οριοθετούν το μέρος. Για παράδειγμα το "Argos" είναι πατέρας της ετικέτας "plateiaargous". Υπό την έννοια των τυπικών σχέσεων η παραπάνω συσχέτιση δεν μπορεί να χαρακτηρισθεί σωστή. Παρόλα αυτά μπορεί να θεωρηθεί απολύτως κατανοητή για την χρησιμότητα της τοποθεσίας μιας εικόνας. Με την ίδια λογική, όπως αναφέρθηκε και πιο πάνω, το "Argos" μπορεί να σχετίζεται, αλλά δεν είναι γονέας της ετικέτας "anthrwpos". Όσον αφορά τους γενικούς όρους τα στιγμιότυπα (instances) θεωρούνται παιδιά τους. Τέλος θεωρεί πως σε ένα μεγάλο περιβάλλον διαμοιρασμού φωτογραφιών όπως είναι το Flickr, οι προσωπικές σχέσεις των χρηστών δεν χρησιμεύουν σε τόσο μεγάλο βαθμό στην αναζήτηση, θεωρώντας για κάθε ζεύγος περιεχομένου όλα σχεδόν τα προσωπικά ονόματα ως ανεπιθύμητους όρους (noisy).

Στον παρακάτω πίνακα παρατίθενται συγκεντρωμένα τα αποτελέσματα των δύο σχετικών μοντέλων υπαγωγής των *Schmitz P.*, *Sanderson M.* και *Croft B*.

Μοντέλο	# σχέσεις	σωστές	σχετικές	ίδιες	Λανθασμένες
Sanderson και Croft	?	23%	49%	8%	19%
Schmitz	1200+	51%	21%	5%	23%

Πίνακας 1: Αποτελέσματα μοντέλων υπαγωγής

Όπως αναφύεται από τον παραπάνω πίνακα οι σωστές σχέσεις υπαγωγής στο μοντέλο του Schmitz καταλήγουν στο 51%, αποτέλεσμα σαφώς καλύτερο συγκριτικά με το 23% του μοντέλου των Sanderson M. και Croft B. Ο ερευνητής παρατήρησε επίσης το μεγάλο ποσοστό του σχετικού μοντέλου των *Sanderson M.* και *Croft B.* στις σχετικές ετικέτες, θεωρώντας πως προέρχεται εν μέρει από την ελλειπή αναζήτηση του λεξιλογίου στα συστήματα σήμανσης.

3.5.1.4 Σύνοψη και Μελλοντικά Σχέδια

Το μοντέλο υπαγωγής του Schmitz δημιουργεί ιεραρχίες υπαγωγής μεταξύ όρων, οι οποίες εκφράζουν διαφορετικούς παράγοντες (για παράδειγμα μέρος-περιοχή, δραστηριότητα, αναπαράσταση) αλλά δεν μπορούν να κατηγοριοποιηθούν σε αυτούς έννοιες. Για να αντιμετωπιστεί αυτό το πρόβλημα ο ερευνητής πρότεινε μια σειρά από βελτιώσεις οι οποίες ενδέχεται να πραγματοποιηθούν μελλοντικά. Αυτές αφορούν:

- *Στην μετατροπή του μοντέλου σε καθαρά πιθανοτικό*,
- *την αντιμετώπιση των περιττών και ανορθόγραφων πληροφοριών* (καταγράφοντας την παραγόμενη οντολογία σε γραφική αναπαράσταση εννοιών, στην οποία μπορούν να συσχετιστούν παραλλαγές των ετικετών),
- *στην χρησιμοποίηση μορφολογικών εργαλείων* (τα οποία μπορούν να εφαρμοστούν στην αρχική ανάλυση των ετικετών),
- στην ενασχόληση με πολύπλευρες οντολογίες,
- στην υποστήριξη χρήσης συντονιστών στη κοινότητα (θεωρώντας ότι είναι καλύτερο οι διαδικασίες των εφαρμογών να μην γίνονται εξ ολοκλήρου με αυτοματοποιημένο τρόπο, αλλά να χρησιμοποιούνται συντονιστές για να διατηρούν μια ισορροπία για τις παραγόμενες σχέσεις).

3.5.2 Μέθοδος Εμπλουτισμού των Folksonomies με τη βοήθεια του Σημασιολογικού Ιστού

3.5.2.1 Γενικά

Οι ***Specia L.*** και ***Motta E.*** [22], έχοντας ως στόχο να καταστήσουν σαφή τη σημασιολογία στα κοινωνικά συστήματα σήμανσης, ασχολήθηκαν με τις σχέσεις ανάμεσα στις ίδιες τις ετικέτες και πως αυτές μπορούν να οδηγήσουν στην ανάπτυξη των οντολογιών.

Οι ερευνητές ενδιαφέρθηκαν πρωτίστως για το σκοπό της συλλογής των ετικετών που ανατίθενται σε πόρους. Για αυτό ακριβώς θεώρησαν πως ένα από τα μεγαλύτερα πλεονεκτήματα των κοινωνικών συστημάτων σήμανσης, είναι το μη προκαθορισμένο λεξιλόγιο, που όμως οδηγεί σε μία σειρά από περιορισμούς όσον αφορά τη χρήση των ετικετών για την ανάκτηση περιεχομένου. Επιπλέον αναφέρθηκαν στην άποψη των *Golder S.* και *Huberman B.* [30], πως τα κύρια προβλήματα τέτοιων συστημάτων περιλαμβάνουν την ασάφεια, την έλλειψη συνωνύμων και την έλλειψη συγκεκριμενοποίησης του περιεχομένου.

Η προσέγγιση τους έχει ως στόχο να μειώσει όσο το δυνατόν περισσότερο τέτοια προβλήματα με την βοήθεια του σημασιολογικού ιστού. Αυτό επιτυγχάνεται με μία σειρά από διαδικασίες πού περιλαμβάνουν: τον διαχωρισμό των ετικετών σήμανσης, την ανάλυση συνύπαρξης μεταξύ τους, την ομαδοποίηση τους (που βασίζεται στις πληροφορίες από την ενδεχόμενη συνύπαρξη) και τέλος την αντιστοίχηση των ετικετών σήμανσης σε ομάδες στοιχείων (*concepts, properties, instances*) στις προκαθορισμένες οντολογίες, και τις παραγόμενες σημασιολογικές σχέσεις μεταξύ τους με την χρήση πηγών πληροφοριών (όπως το *Google* και η *Wikipedia*).

3.5.2.2 Μεθοδολογία

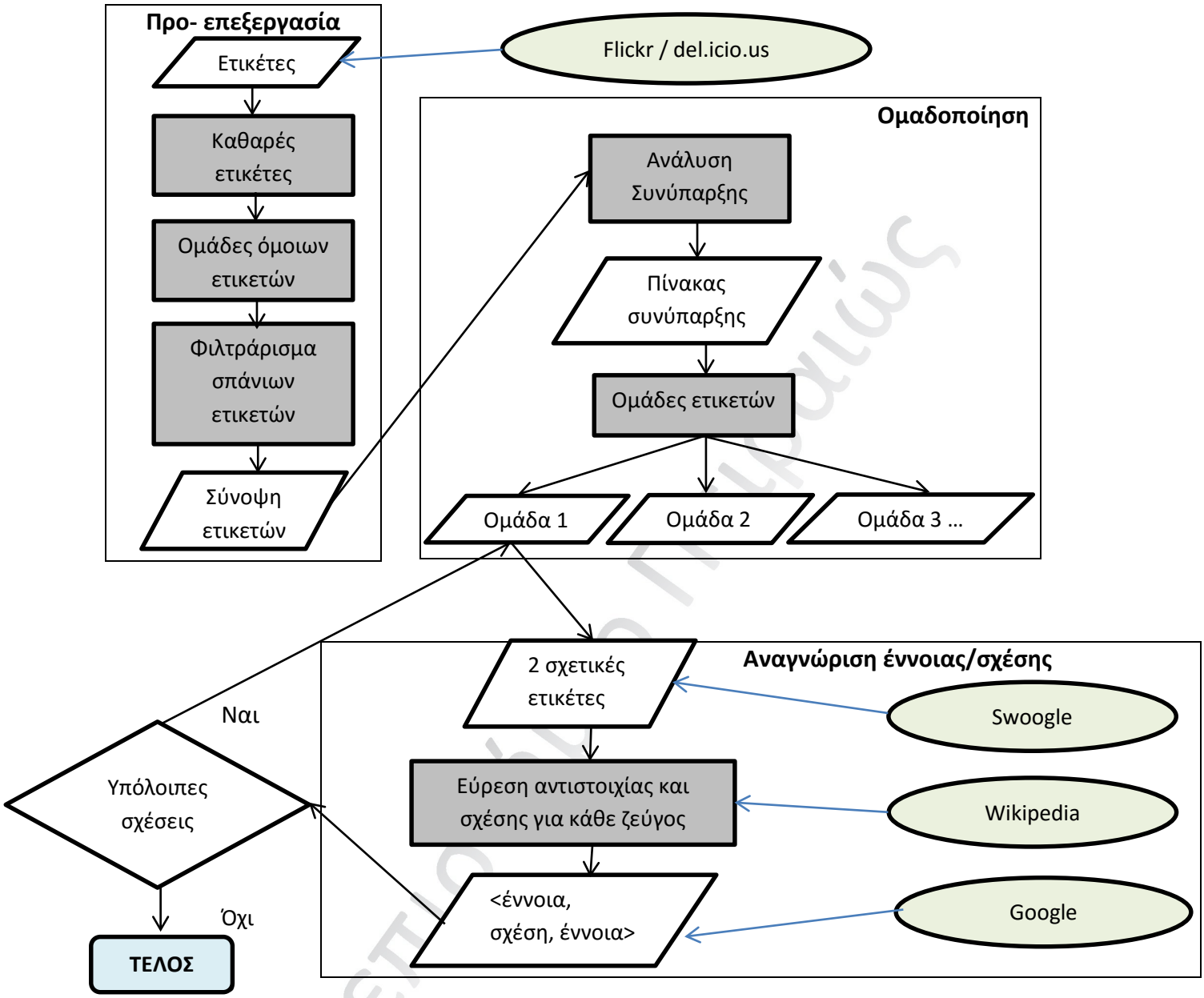
Θεώρησαν πως η κατανόηση της **υποκειμενικής γνώσης** είναι πολύ σημαντική στα σημασιολογικά κοινωνικά δίκτυα. Αυτή αφορά στην αναγνώριση των στοιχείων που αντιστοιχούν στις ετικέτες σήμανσης και στις μεταξύ τους σχέσεις. Υποθέσανε πως μπορεί να προκύψει από τα αποτελέσματα μια στατιστικής ανάλυσης των επισημειώσεων, σε συνδυασμό με ρεαλιστικές πληροφορίες που παρέχονται από τον σημασιολογικό ιστό και

άλλων στοιχείων από εξωτερικές πηγές. Γι αυτό ακριβώς τον λόγο ερεύνησαν τα δεδομένα δύο πολύ γνωστών και ελεύθερων συστημάτων σήμανσης, των *Flickr* και *del.icio.us*. Επιπλέον για τα πειράματα τους χρησιμοποίησαν τις ίδιες ετικέτες για το σύστημα del.icio.us από την έρευνα του Mika P. [21], καθώς και αυτές από το Flickr μεταξύ των περιόδων 01-02-2004 και 01-03-2006. Ο συνολικός αριθμός των ετικετών σήμανσης παρουσιάζεται στον παρακάτω πίνακα:

	Σύνολο		Διακριτές τιμές		
	# εισαγωγές	# ετικέτες	# χρήστες	# πηγές	# ετικέτες
del.icio.us	19605	89978	7164	14211	11960
Flickr	49087	167130	6140	49087	17956

Πίνακας 2: Συνολικός αριθμός ετικετών, με τους αντίστοιχους χρήστες και πηγές, για κάθε σύστημα σήμανσης.

Δύο πολύ σημαντικά χαρακτηριστικά της μεθοδολογίας τους, περιλαμβάνει το γεγονός πως είναι ανεξέλεγκτη, καθώς δεν υποθέτει προηγούμενες αναγνωρισμένες σχέσεις μεταξύ των ετικετών για να λειτουργήσει το σύστημα τους, και δεν απαιτεί κανένα περιεχόμενο παρά μόνο τις ίδιες τις ετικέτες σήμανσης, τις πηγές τους και άλλες ετικέτες που χρησιμοποιούνται σε αυτές (πηγές). Η προσέγγιση τους αποτελείται από τρία βήματα (*Εικόνα 5*): Προ-επεξεργασία (pre-processing), ομαδοποίηση (clustering), αναγνώριση έννοιας/σχέσης (concept/relation identification). Αυτή μπορεί να αποτυπωθεί με το παρακάτω σχήμα:



Εικόνα 5: Αρχιτεκτονική συστήματος των Specia L. και Motta E.

3.5.2.2.1 Προ-επεξεργασία (pre-processing)

Σε αυτό το βήμα πραγματοποιούνται οι παρακάτω διαδικασίες:

- Αρχικά, φιλτράρονται εκτός οι ασυνήθιστες ετικέτες (και οι αντίστοιχες πηγές τους), για να βρεθούν αυτές που έχουν μια γενικότερη εφαρμογή και μπορούν να βρεθούν στις οντολογίες. Γι αυτό ακριβώς τον λόγο ορίζεται ο περιορισμός, πως η ονομασία κάθε

ετικέτας θα πρέπει να ξεκινάει με ένα γράμμα, που θα το ακολουθεί οποιοσδήποτε αριθμός γραμμάτων, αριθμών και συμβόλων (όπως η παύλα, τελεία κτλ.)

- Χρησιμοποιείται η μέτρηση ομοιότητας *Levenshtein* [39] για ομάδες με όμοιες μορφολογικά ετικέτες σήμανσης, για να αντιμετωπιστούν οι μορφολογικές μεταβολές (san Francisco, sanFrancisco) και τα ορθογραφικά λάθη. Κάθε ομάδα ορίζει έναν εκπρόσωπο, με αποτέλεσμα όλες οι ετικέτες να αντικαθίστανται από αυτόν. Τέλος ο κάθε εκπρόσωπος θα πρέπει να είναι μια ετικέτα σήμανσης αποτελούμενη μόνο από γράμματα, που θα την ακολουθούν λέξεις με ίδια σύμβολα, στη συνέχεια συνδυασμός λέξεων και αριθμών και ούτω καθεξής. Για παράδειγμα η ετικέτα typography επιλέγεται ως εκπρόσωπος της ομάδας {tipography typograph typography}.
- Φιλτράρονται οι σπάνιες και μεμονωμένες ετικέτες σήμανσης (και οι αντίστοιχες πηγές). Αυτές αναφέρονται σε ετικέτες που εμφανίζονται λιγότερες φορές σε σχέση με τις αντίστοιχες παραδοσιακές μέσα στα συστήματα σήμανσης ή είναι απομονωμένες (isolated).

3.5.2.2.2 Ομαδοποίηση (clustering)

Στο δεύτερο βήμα εκτελείται μια στατιστική ανάλυση του χώρου ετικετών σήμανσης (tag space), για να προσδιοριστούν οι ομάδες των σχετιζόμενων ετικετών. Η ομαδοποίηση στηρίζεται στην ομοιότητα που προκύπτει από την συνύπαρξη και για να πραγματοποιηθεί, οι ετικέτες σήμανσης οργανώνονται σε ένα **πίνακα συνύπαρξης M** όπου:

- M ένας n x n συμμετρικός πίνακας,
- η μεταβλητή n είναι ο αριθμός των διαφορετικών ετικετών στη βάση δεδομένων,
- η τιμή κάθε στοιχείου m_{ij}, αντιπροσωπεύει την τομή των ετικετών tag1 και tag2 και αντιστοιχεί στον αριθμό των φορών συνύπαρξης του ζευγαριού.

Εάν $tag_{i=}tag_{j}$, τότε η τομή αντιπροσωπεύει την συχνότητα της ετικέτας στη σύνολο δεδομένων.

Κάθε στήλη του πίνακα είναι ένα διάνυσμα που αντιπροσωπεύει καθεμία από τις ετικέτες. Ένα ζεύγος διανυσμάτων μπορεί να υπολογιστεί κατάλληλα με την παρακάτω εξίσωση:

$$angular\_separation_{ij=} \frac{\sum_{k=1}^{n} x_{ik} . x_{jk}}{(\sum_{k=1}^{n} x_{ik}^2 . \sum_{k=1}^{n} x_{jk}^2)^{1/2}}$$

Η **γωνία διαχωρισμού** (angular separation) υπολογίζει πιο συγκεκριμένα το συνημίτονο της γωνίας δύο διανυσμάτων στον πίνακα συνύπαρξης (n x n ζεύγος). Από το αποτέλεσμα της παραπάνω σχέσης, για κάθε ετικέτα αντιστοιχεί μια λίστα από ομοιότητες με όλες τις υπόλοιπες. Επιπλέον χρησιμοποιώντας τον πίνακα συνύπαρξης, λαμβάνονται υπόψη όλες οι άλλες ετικέτες σαν περιεχόμενο. Για να θεωρηθούν όμως όμοιες με την tag_j, η tag_i, θα πρέπει να συνυπάρχει όχι μόνο με την tag_j αλλά και με άλλες που συνυπάρχουν με την tag_j. Απόρροια αυτού είναι ότι και οι δύο ετικέτες έχουν κάτι κοινό που δίνεται από τα διανύσματα συνύπαρξης τους. Αυτές οι ομοιότητες λέγονται μερικές φορές και παραδειγματικοί συσχετισμοί (paradigmatic associations).

Κάθε ετικέτα σήμανσης μπορεί να περιλαμβάνει δύο ή περισσότερους παραδειγματικούς συσχετισμούς (paradigmatic associations), παρουσιάζοντας με την σειρά τους τα διαφορετικά νοήματα ή χρήσεις της ετικέτας. Για παράδειγμα η ετικέτα apple μπορεί να αναφέρεται στην εταιρία υπολογιστών ή στο φρούτο. Σε αυτή την περίπτωση οι ετικέτες σήμανσης που φανερώνουν μεγαλύτερη συσχέτιση με μια συγκεκριμένη ετικέτα, είναι ενδεικτικές των διαφορετικών γνωστικών πεδίων στα οποία αναφέρεται η ίδια ετικέτα (πολυσημία). Γι αυτό ακριβώς τον λόγο, οι πληροφορίες που παρέχονται από τους παραδειγματικούς συσχετισμούς θεωρήθηκαν από τους ερευνητές ελλιπείς στα ζεύγη των ετικετών σήμανσης. Με άλλα λόγια μπορεί να ειπωθεί πως, για μια συγκεκριμένη ετικέτα, υπάρχει ένα σύνολο άλλων ετικετών που σχετίζεται με αυτήν, γεγονός που δεν εγγυάται την ύπαρξη σχέσεων ανάμεσα στην ομάδα. Για να αντιμετωπιστεί αυτό το πρόβλημα χρησιμοποίησαν τον παραδοσιακό **αλγόριθμο ομαδοποίησης K-means**.

Προκειμένου να ομαδοποιηθούν οι ετικέτες που παρουσιάζουν μεγάλο ποσοστό συνύπαρξης, δημιούργησαν ένα **όριο ομοιότητας** που είχε ως στόχο να διαχωρίσει αυτές με μικρό ποσοστό. Ο αλγόριθμος ομαδοποίησης με τη σειρά του, λαμβάνει υπόψη τις αμοιβαίες ομοιότητες ανάμεσα στις ετικέτες και αναγνωρίζει τις ομάδες. Θεωρεί δηλαδή πως κάθε ζεύγος όμοιων ετικετών αποτελούν τη βάση της πρώτης ομάδας και στη συνέχεια αναζητεί άλλες ετικέτες όμοιες και με τις δύο αρχικές, προκειμένου να αυξηθεί το μέγεθος της ομάδας. Η διαδικασία είναι επαναληπτική, και εφόσον δεν υπάρχουν άλλες υποψήφιες ετικέτες για την ομάδα, ένα άλλο ζεύγος θεωρείται βάση και επαναλαμβάνεται η ίδια διαδρομή.

Αποτέλεσμα της παραπάνω διαδικασίας είναι η δημιουργία ενός συνόλου ομάδων με μεγάλη ομοιότητα, που διαφέρουν σε μερικές μόνο ετικέτες. Η ομοιότητα αυτή απορρέει σε πολλές περιπτώσεις από το αποτέλεσμα του φιλτραρίσματος των ανόμοιων ζευγών ετικετών σήμανσης. Στη συνέχεια για να αποφευχθεί ο μεγάλος αριθμός εμφάνισης αυτών των όμοιων ομάδων, χρησιμοποιήθηκαν δύο **ευριστικές στρατηγικές λύσης προβλήματος** (heuristics).

Για κάθε δύο ομάδες:

- Όταν μια ομάδα περιέχει μια άλλη (αυτό συμβαίνει εάν η μεγαλύτερη ομάδα περιέχει όλες τις ετικέτες της μικρότερης), αφαιρεί την μικρότερη.
- Εάν οι ομάδες διαφέρουν σε μικρό ποσοστό, προστίθενται οι ξεχωριστές λέξεις από την μικρότερη -η οποία και αφαιρείται- στην μεγαλύτερη ομάδα.

Οι ευριστικές στρατηγικές λύσης προβλήματος δίνουν τη δυνατότητα συνένωσης δύο ομάδων σήμανσης οι οποίες δεν είναι αρκετά όμοιες (σύμφωνα με το όριο ομοιότητας), που μπορούν όμως να παρουσιάζουν κοινά στοιχεία. Εάν και υπάρχει η δυνατότητα διαγραφής των αποκλίσεων, διατηρούνται οι ομάδες που μοιράζονται ετικέτες και παρουσιάζουν πολλαπλά νοήματα, φανερώνοντας με αυτόν τον τρόπο την ασάφειά τους.

Συνοψίζοντας, ένα πολύ σημαντικό χαρακτηριστικό της ομαδοποίησης του μοντέλου των *Specia L.* και *Motta E.* είναι το ότι δεν απαιτεί τον καθορισμό του αριθμού των ομάδων που παράγονται. Οι μοναδικοί παράμετροι είναι το όριο που ορίζει το μέγεθος της συνύπαρξης των ετικετών σήμανσης και το ποσοστό της διαφοράς που επιτρέπεται για τις όμοιες ομάδες.

3.5.2.2.3 Αναγνώριση Έννοιας/Σχέσης (concept/relation identification)

Εφόσον οι ομάδες προέρχονται από τις πληροφορίες συνύπαρξης, δεν υπάρχει καμία ένδειξη σχέσης μεταξύ των υποσυνόλων τους. Στόχος του τρίτου βήματος είναι η χρησιμοποίηση της γνώσης που προέρχεται από διαφορετικές πηγές (*Wikipedia, Google*), για να ανακαλυφθεί η ύπαρξη και η κατηγοριοποίησή ανάμεσα στις ετικέτες κάθε ομάδας (εάν υπάρχει σχέση). Η διαδικασία περιλαμβάνει αντιστοίχηση των ετικετών στις έννοιες/στιγμιότυπα/ιδιότητες των οντολογιών και εύρεση των πιθανών σχέσεων ανάμεσα σε αυτές. Ως πηγή οντολογιών οι ερευνητές χρησιμοποίησαν μια μηχανή αναζήτησης του σημασιολογικού ιστού, την ***Swoogle*** [40], με την διαδικασία να είναι η ακόλουθη:

1) Συμπεριλαμβάνεται κάθε πιθανό ζεύγος ετικετών στην Swoogle, για να ανακτηθούν οι οντολογίες που περιέχουν και τις δύο. Όλοι οι συνδυασμοί των ζευγών εφαρμόζονται, εφόσον δεν μπορεί κάποιος να γνωρίζει το όριο ομοιότητας.
2) Εάν κάποια ετικέτα δεν μπορεί να βρεθεί από τη μηχανή αναζήτησης, θεωρείται ως ακρωνύμιο, ορθογραφικό λάθος ή παραλλαγή του όρου. Το επόμενο βήμα μιας τέτοιας περιπτώσεις είναι η αναζήτησή της σε εναλλακτικές πηγές:

   2.1) Χρησιμοποιείται ο ιστόχωρος *Wikipedia* για να βρεθεί εάν είναι ακρώνυμο. Για παράδειγμα τα αρχικά NYC στη συγκεκριμένη ιστοσελίδα εμφανίζεται κανονικά η λέξη ως New York City.

   2.2) Στην περίπτωση που η ετικέτα δεν βρεθεί στην *Wikipedia*, θεωρείται ως ορθογραφικό λάθος ή όρος που περιλαμβάνει πολλές λέξεις. Επόμενο βήμα είναι η χρησιμοποίηση της μηχανής αναζήτησης *Google*, για να βρεθεί μία πιθανή σωστή διατύπωση του όρου. Για παράδειγμα στην αναζήτηση για τον ασαφή όρο sanfrancisco, το σύστημα ρωτάει αμέσως : Μήπως εννοείτε san Francisco (που είναι και το σωστό).
3) Εάν δύο ετικέτες (ή οι αντίστοιχοι όροι από την *Wikipedia* ή το *Google*) δεν βρεθούνε μαζί στην *Swoogle*, θεωρούνται πως δεν σχετίζονται και διαγράφονται - εφόσον αναζητηθούν όλοι οι πιθανοί συνδυασμοί- από την ομάδα που δεν σχετίζεται με άλλες ετικέτες.

4) Αντίστροφα, εάν βρεθούν οντολογίες να περιέχουν και τις δύο ετικέτες:

   4.1) Ελέγχεται εάν οι ετικέτες αντιστοιχούν σωστά στα στοιχεία (*concepts, instances, properties*) των οντολογιών.

   4.2) Ανακτούνται πληροφορίες σχετικά με τις ετικέτες σήμανσης για κάθε μια από τις οντολογίες: ο τύπος της (*concept, instance, property*) και ο γονέας, εάν είναι έννοια (*concept*) ή το στιγμιότυπο (*instance*) και το πεδίο εφαρμογής (*domain*), η εμβέλεια των τιμών της, εάν είναι ιδιότητα (property).

5) Για κάθε ζεύγος ετικετών, που η μηχανή αναζήτησης *Swoogle* μπόρεσε να παράγει πληροφορίες, ελέγχονται οι πιθανές μεταξύ τους σχέσεις:

   5.1) Μια ετικέτα είναι "*πρόγονος*" μιας άλλης, όπως για παράδειγμα για την οντολογία Food, το μήλο είναι υποκατηγορία του φρούτου.

   5.2) Μια ετικέτα μπορεί να αποτελεί την εμβέλεια (range) ή την τιμή μιας από τις ιδιότητες για κάθε άλλη ετικέτα. Για παράδειγμα η κατηγορία της οντολογίας Wine, είναι η Zinfandel,η οποία έχει μια ιδιότητα hasColor, που η τιμή της είναι red. Γι αυτό η σχέση hasColor υπάρχει ανάμεσα στις Zinfandel και red.

   5.3) Και οι δύο ετικέτες έχουν τον ίδιο γονέα, όπως για παράδειγμα για την οντολογία FOOD, οι έννοιες apple και orange έχουν τον ίδιο πατέρα (fruit).

   5.4) Και οι δύο έχουν τους ίδιους πρόγονους στο ίδιο επίπεδο.

   5.5) Τέλος και οι δύο έχουν τους κοινούς πρόγονους σε διαφορετικά επίπεδα. Για παράδειγμα στο WordNet η ετικέτα σήμανσης chapterhouse έχει προγόνους της building (1^{ο} επίπεδο) και construction (2^{ο} επίπεδο), ενώ η edifice έχει πρόγονο την construction (1^{ο} επίπεδο).

Αναζητώντας ζεύγη ετικετών σε μία οντολογία (αντί μεμονωμένων ετικετών σήμανσης), επιτυγχάνεται σε μεγάλο βαθμό η εξάλειψη της ασάφειας. Για παράδειγμα στην περίπτωση 5.3, η ετικέτα apple μπορεί να ορίζεται σε άλλες οντολογίες με διαφορετικό νόημα, όπως σαν computer model στην οντολογία Clib-core-office.

Για κάθε οντολογία που περιέχει και τις δύο ετικέτες, οι ερευνητές χρησιμοποίησαν μια απλή ***στρατηγική ανάλυσης***: Δίνεται προτεραιότητα στην οντολογία που περιέχει άλλες

ετικέτες και πιθανόν σχέσεις μέσα στην ομάδα. Εάν υπάρχουν ακόμα πολύπλευρες οντολογίες (σελ 33, faceted ontologies), η πρώτη που εκπληρώνει τον παραπάνω περιορισμό είναι η ζητούμενη. Επιπλέον όσον αφορά τις πολλαπλές σχέσεις μεταξύ των ζευγών, επιλέγεται η πρώτη σύμφωνα με τα βήματα 5.1 έως 5.5. Τέλος, εάν η παραπάνω διαδικασία δεν αποδίδει καμία σχέση συνύπαρξης για ένα δεδομένο ζεύγος ετικετών σε μία τουλάχιστον οντολογία, βγαίνει το συμπέρασμα πως δεν είναι σχετικές, κάτι που και τις αφαιρεί από την ομάδα (εκτός εάν είναι σχετικές με άλλες ετικέτες).

3.5.2.3 Σύνοψη και Μελλοντικά σχέδια

Η προσέγγιση των *Specia L.* και *Motta E.* έχει ως αποτέλεσμα την εύρεση ομάδων αποτελούμενες από ετικέτες με μεγάλο βαθμό συνύπαρξης. Αυτές με την σειρά τους αντιστοιχούν σε στοιχεία οντολογιών, δομημένα σύμφωνα με τις σχέσεις μεταξύ τους και μπορούν να θεωρηθούν ως πολύπλευρες οντολογίες (faceted ontologies). Με άλλα λόγια είναι αυτές που αναπαριστούν μια συγκεκριμένη εννοιολογική γνώση συγκεκριμένου γνωστικού πεδίου. Οι παραγόμενες οντολογίες της προσέγγισής τους διαφέρουν από αυτές που θεωρούνται παραδοσιακές, καθώς προκύπτουν ενώνοντας τμήματα των πολύπλευρων οντολογιών στο σημασιολογικό ιστό. Επιπλέον μπορούν να χρησιμοποιηθούν για την ενίσχυση διαφόρων διαδικασιών στο συστήματα σήμανσης όπως: η επέκταση και η αποσαφήνιση της αναζήτησης για την ομαδοποίηση μέσω των ευριστικών κανόνων που προαναφέραμε, οπτικοποίηση και η υπόδειξη ετικετών σήμανσης.

Η παραπάνω μέθοδος μπορεί να υποστηρίξει επιπλέον την εξέλιξη και τον πολλαπλασιασμό των οντολογιών: η νέα και δυναμική γνώση που προέρχεται από τους χρήστες μπορεί να συμπληρώσει αυτή των οντολογιών, με την πρόσθεση εννοιών (ή περιπτώσεις εννοιών) και σχέσεων (ή περιπτώσεις σχέσεων) στην οντολογία. Για αυτό ακριβώς τον λόγο, η προσέγγιση ενσωματώνει folksonomies με τον σημασιολογικό ιστό και θέλει να αποδείξει: (i) ότι οι παραγόμενες οντολογίες μπορούν να χρησιμοποιηθούν για την δομή των folksonomies, (ii) και ότι η δυναμική γνώση που προέρχεται από τα folksonomies μπορεί να χρησιμοποιηθεί ως πηγή για την απόκτηση της κύριας γνώσης, που υποστηρίζει την εξέλιξη

των οντολογιών. Τέλος για την βελτίωση της προσέγγισης τους, οι ερευνητές σχεδιάζουν μελλοντικά να:

- χρησιμοποιήσουν μια νέα τεχνική για τις ομάδες, η οποία θα συνδυάζει την ιεραρχική ομαδοποίηση με ένα όριο για τις ετικέτες που δεν είναι όμοιες με άλλες,
- εφαρμόσουν μια πλήρως αυτοματοποιημένη έκδοση του τελευταίου βήματος της προσέγγισης που αφορά στη διαδικασία αντιστοίχισης των ετικετών με στοιχεία της οντολογίας. Για να πραγματοποιηθεί αυτό θα πρέπει να σχεδιαστούν καλύτερες στρατηγικές για την επιλογή και την αντιστοιχία , καθώς και άλλες τεχνικές με την εξαγωγή πληροφοριών από τις διάφορες πηγές (*Wikipedia, Google*).
- αξιολογήσουν την ποιότητα των νέων αποτελεσμάτων, ενσωματώνοντας τες στο περιεχόμενο των διαδικασιών που αναφέρθηκαν παραπάνω (αποσαφήνιση, οπτικοποίηση κτλ).

3.5.3 Μοντέλο Σήμανσης Οντολογιών (The TagOntology)

3.5.3.1 Γενικά

Ο ***Gruber T.*** [13] παρατηρώντας το λάθος πολλών να έχουν την αντίληψη πως οι οντολογίες και τα Folksonomies είναι διαμετρικά αντίθετα τεχνήματα που δεν μπορούν να συνδυαστούν, επικέντρωσε την έρευνα του σε αυτούς τους δύο όρους. Στο άρθρο του "Ontology of Folksonomy: a mash-up of Apples and Oranges", προσπάθησε να διαχωρίσει τους διαφορετικούς τους ρόλους. Στην συνέχεια εφαρμόζοντας τους μαζί θέλησε να δημιουργήσει οντολογίες για Folksonomies, στοχεύοντας με αυτό τον τρόπο να δείξει πως αυτές οι δύο τεχνικές δεν είναι εντελώς αντίθετες μεταξύ τους, αλλά είναι δυνατόν να αλληλοσυμπληρώνονται.

3.5.3.2 Επιρροές / Διαχωρισμός Ρόλων Οντολογιών - Folksonomies

Ο ερευνητής αναφέρθηκε στις απόψεις του *Shirky K.* [41], ο οποίος όρισε την **οντολογία** "*ως υπερτιμημένη*" και την **σήμανση** "*ως κάτι το διαφορετικό σε σχέση με άλλες στρατηγικές κατηγοριοποίησης*". Εξισώνοντας όμως τον πρώτο όρο με την οργάνωση πληροφοριών, ο *Shirky K.* έδειξε πως τα ιεραρχικά, κεντρικά ελεγχόμενα συστήματα

ταξινόμησης είναι περιορισμένα σε αντίθεση με αυτά που μπορούν εύκολα να τροποποιηθούν. Υιοθετώντας την παραπάνω άποψη, ο *Gruber T.* υποστήριξε πως τα *Folksonomies* παρουσιάζουν μεγαλύτερο ενδιαφέρον σε σχέση με τις ταξινομίες. Αυτές μπορούν να περιορίσουν τις διακρίσεις και τις επιλογές που μπορεί να κάνει κάποιος, με αποτέλεσμα να είναι εκ φύσεως δύσκολο να τηρηθεί η ιεράρχηση των πραγμάτων. Από την άλλη πλευρά στα *Folksonomies* δεν επιβάλλεται καμία συνέπεια στη σήμανση, καθώς κάθε αντικείμενο εύκολα μπορεί να σημανθεί με οποιονδήποτε τρόπο (Κεφάλαιο 2.1.1.2). Το μοναδικό σημείο διαφωνίας ανάμεσα στους δύο ερευνητές, αφορά τη θεώρηση της οντολογίας από τον *Shirky K.* ως συγκεκριμένη καταγραφή εννοιών στη δομή της ταξινομίας. Θεωρεί πως στα σημερινά συστήματα υπάρχει μια νέα πηγή δεδομένων για την εύρεση και την οργάνωση πληροφοριών, η συμμετοχή του χρήστη. Όσον αφορά την εύρεση πληροφοριών, οι ταξινομίες είναι πολύ αυστηρές. Αυτό το πρόβλημα κατάφερε πρώτη να αντιμετωπίσει η Google, με την επαναστατική μέθοδο αναζήτησης της ποιότητας των ιστιοσελίδων μέσω υπερσυνδέσεων (hyperlinks). Οι παλαιότεροι χρήστες του διαδικτύου χρησιμοποιούσαν τους συνδέσμους τους (links) για να δημιουργήσουν έναν κατάλογο με τις ιστοσελίδες που τους ενδιέφεραν. Από την άλλη μεριά οι χρήστες σήμερα χρησιμοποιούν με μεγάλο ενθουσιασμό ετικέτες για την σήμανση φωτογραφιών, σελιδοδεικτών κτλ με ετικέτες ελεύθερου λεξιλογίου, κάτι που ο ερευνητής θεώρησε μια σημαντική πηγή έμπνευσης για εκμετάλλευση.

3.5.3.3 Μεθοδολογία

Σύμφωνα με το όραμα του, μια ομάδα χρηστών – που ανήκουν σε μια κοινότητα - θα πρέπει να αρχίσουν να δουλεύουν πάνω σε μία κοινή οντολογία για σήμανση, την **TagOntology**. Η συγκεκριμένη είναι μια οντολογία που εντοπίζει και φανερώνει την δραστηριότητα σήμανσης στα Folksonomies.

Ο ερευνητής προσπάθησε να αναπαραστήσει την μορφή αυτής, για να αξιοποιήσει τα δεδομένα που δημιουργήθηκαν από τα Folksonomies. Θεώρησε πως η κύρια ιδέα της σήμανσης πρέπει να λαμβάνει υπόψη όλο το περιβάλλον των κοινωνικών δικτύων. Από την πλευρά του χρήστη είναι μία δραστηριότητα κατά την οποία επισημαίνει κάποιο περιεχόμενο

που μπορεί να δημιούργησε, με μια ή περισσότερες ετικέτες. Αυτό υποστήριξε πως μπορεί να αποτυπωθεί με βάση τη σχέση:

Σήμανση (αντικείμενο, ετικέτα)

Η παραπάνω σχέση μπορεί να ισχύει σε μια κλειστή διαδικτυακή κοινότητα. Για να αναπαρασταθεί όμως το συνεργατικό φιλτράρισμα, ο ερευνητής θεώρησε πως χρειάζεται η έννοια του **χρήστη** που κάνει την σήμανση. Γι αυτό ακριβώς τον λόγο η σχέση πήρε την μορφή:

*Σήμανση (αντικείμενο, ετικέτα, **χρήστης**)*

Στη συνέχεια θέλησε να αποδώσει τον τρόπο διάδοσης αυτών των δεδομένων. Υποστήριξε πως εάν συγκρίνει κάποιος δεδομένα από διαφορετικά συστήματα, δεν μπορεί να υποθέσει πως όλα έχουν τα ίδια ακριβώς αντικείμενα, ετικέτες ή χρήστες. Για παράδειγμα εάν δύο εφαρμογές μοντελοποιούσαν τα δεδομένα των ετικετών τους με την παραπάνω σχέση, η μορφή της θα ήταν η ακόλουθη:

Σήμανση (αντικείμενο1, ετικέτα1, χρήστης1)// από το σύστημα 1

Σήμανση (αντικείμενο1, ετικέτα2, χρήστης1)// από το σύστημα 1

Σήμανση (αντικείμενο1, ετικέτα1, χρήστης2)// από το σύστημα 1

Σήμανση (αντικείμενο1, ετικέτα3, χρήστης3)// από το σύστημα 2

Σήμανση (αντικείμενο2, ετικέτα1, χρήστης4)// από το σύστημα 2

Γι αυτό ακριβώς τον λόγο ο ερευνητής κατέληξε στο συμπέρασμα πως θα πρέπει να γίνει σαφής η έννοια του πόρου (source), που μπορεί να θεωρηθεί ως πεδίο εφαρμογής των ονομάτων και της ποσοτικοποίησης για αυτά τα αντικείμενα της παραπάνω σχέσης. Επομένως, προσθέτοντας αυτήν την κατηγορία στη θέση του συστήματος και διατηρώντας την προηγούμενη υπόθεση:

*Σήμανση (αντικείμενο1, ετικέτα1, χρήστης1, **πόρος1**)*

*Σήμανση (αντικείμενο1, ετικέτα2, χρήστης1, **πόρος1**)*

*Σήμανση (αντικείμενο1, ετικέτα1, χρήστης2, **πόρος1**)*

*Σήμανση (αντικείμενο1, ετικέτα3, χρήστης3, **πόρος2**)*

*Σήμανση (αντικείμενο2, ετικέτα1, χρήστης4, **πόρος2**)*

Η σχέση που προκύπτει φανερώνει μια συλλογή από δεδομένα ετικετών σήμανσης, ανεξάρτητα από τις εφαρμογές που προέρχονται.

Για να μη γίνει καμία λανθασμένη υπόθεση από τα συγχωνευμένα ή ανταλλασσόμενα δεδομένα, ο *Gruber T.* θεώρησε πως χρειάζεται μια οντολογική δέσμευση για την σημασιολογία της σήμανσης και των τριών μέρων της. Γι αυτό ακριβώς τον λόγο εφάρμοσε πρώτα το **κριτήριο της εσωτερικής συνοχής** για την ίδια την σχέση, κατά την οποία η κάθε ετικέτα σήμανσης θεωρείται μία ψήφος. Επομένως εάν ισχύει η σχέση *ετικέτα1=ετικέτα2*, μπορεί να επιβεβαιωθεί η συγκεκριμένη ισοδυναμία χωρίς την πρόσθεση οποιασδήποτε πληροφορίας. Αυτό αποτελεί ένα αντιπροσωπευτικό παράδειγμα που φανερώνει τον λόγο για τον οποίο τα συστήματα χρειάζονται να έχουν οντολογικές δεσμεύσεις για το Σημασιολογικό επίπεδο, πέρα από τις συμφωνίες για τις μορφές [42]. Από την άλλη μεριά, εάν κάποιο σύστημα έδινε διαφορετικό νόημα στις ετικέτες του, δεν θα μπορούσε να συνδυάσει τα δεδομένα του. Μια δεύτερη αντίληψη συνυφασμένη με την σήμανση, αφορά το νόημα που διαφέρει από ετικέτα σε ετικέτα ή από χρήστη σε χρήστη. Η σχέση σήμανσης *δεν είναι συμμετρική*: δηλαδή, δεν μπορεί να διατηρηθεί το νόημα εάν αλλάξουν οι χρήστες και οι ρόλοι της διαδικασίας. Γι αυτό ακριβώς τον λόγο, για να διευκρινιστεί δηλαδή η έννοια της σήμανσης θα πρέπει να σχεδιαστεί ένα διαφορετικό είδος σχέσεων ή εξάρτησης για την σήμανση των δεδομένων. Ως εκ τούτου ο ερευνητής θεώρησε πως μπορούν να χρησιμοποιηθούν συστήματα στα οποία μια ετικέτα είναι συνώνυμο μιας άλλης ή δύναται να αντιπροσωπεύσει μια ομάδα άλλων ετικετών. Με άλλα λόγια εκτίμησε πως ο επιτυχημένος διαμοιρασμός της πληροφορίας απαιτεί μόνο την αναγνώριση των πιθανών διαφορών ανάμεσα στις ετικέτες σήμανσης.

Ως επόμενο βήμα του προτύπου του, ο ερευνητής διαπίστωσε πως θα πρέπει με κάποιο τρόπο να αντιμετωπιστεί το συχνό φαινόμενο της αρνητικής σήμανσης, με κάποιο είδος *φιλτραρίσματος.* Αυτό ίσως θα πρέπει να απαιτεί επιβεβαίωση του λάθους ότι μια ετικέτα δεν ισχύει για ένα αντικείμενο. Όμως είναι γενικά δύσκολο να αποδειχθεί πότε δεν εμφανίζεται μια σήμανση. Για αυτό ακριβώς τον λόγο πρόσθεσε μια επιπλέον κατηγορία που σχετιζόταν με την περαιτέρω **ενδυνάμωση (+) ή αποδυνάμωση (-) της δραστηριότητας της σήμανσης** από τον ίδιο τον χρήστη. Απόρροια αυτού, η σχέση τροποποιείται ως:

*Σήμανση (αντικείμενο, ετικέτα, χρήστης, **+ ή -**)*

Υποστήριξε πως εάν η πολικότητα δεν έχει δηλωθεί, μπορεί ανεπίσημα να ισχύει το θετικό πρόσημο ως προεπιλογή για την δραστηριότητα. Αυτό μπορεί να αποτελέσει και μια οντολογική δέσμευση σε σημασιολογικό επίπεδο. Εάν ένα σύστημα επομένως χρησιμοποιεί την σχέση Σήμανση (αντικείμενο, ετικέτα, χρήστης, πόρος) και μια άλλη την Σήμανση (αντικείμενο, ετικέτα, χρήστης, + ή -), η δεύτερη μπορεί να θεωρηθεί ισοδύναμη με την πρώτη και τις παραλλαγές της, με θετική πολικότητα.

Τέλος ο ερευνητής θεώρησε πως η οντολογία του χρειάζεται επίσημους **ορισμούς ταυτότητας**, για τον κάθε πυρήνα των εννοιών της: αντικείμενο, ετικέτα, χρήστης και πόρος. Ο Σημασιολογικός ιστός, προσφέρει ένα εύκολο και βολικό συνάμα πρότυπο καταχώρησης των ονομάτων που χρησιμοποιούνται για την σήμανση με την χρήση URIs. Ισχυρίστηκε πως μπορεί να γίνει η επισημοποίηση των εννοιών, μέσω μιας συνάρτησης αναπαράστασης από τα ονόματα στις ετικέτες. Για παράδειγμα, f("san francisco") = 1^{η} ετικέτα, f("San Francisco") = 2^{η} ετικέτα, f("sanfrancisco") = 3^{η} ετικέτα. Στη συνέχεια μπορούν να αποδοθούν καθαρά όλες οι σχέσεις που ορίζουν, πως ένα συγκεκριμένο σύστημα διαχειρίζεται τα ταυτιζόμενα ονόματα. Για παράδειγμα ένας μπορεί να υποστηρίξει πως ετικέτα 1 = ετικέτα 2, κάποιος άλλος ότι ετικέτα 2 = ετικέτα 3 κτλ. Επίσης θεώρησε πως μπορεί να διαμορφωθεί η παραπάνω λειτουργία με την χρήση **κανονικού ονόματος** (canonical name) για κάθε συγκεκριμένη ετικέτα όπου : cname(tag)= "string". Οι διαφορές ανάμεσα στη μορφή των ετικετών οριοθετούνται στα πλαίσια της εφαρμογής. Με άλλα λόγια δεν είναι τόσο σημαντικό ο τρόπος με τον οποίο αυτές

αναγράφονται ή εμφανίζονται μέσα στην εφαρμογή, αλλά εφόσον εξάγονται και σε άλλα συστήματα θα πρέπει να χρησιμοποιούν μόνο το κανονικό τους όνομα (canonical name).

3.5.3.4 Σύνοψη

Η πλειοψηφία των προσεγγίσεων προσπάθησαν να παράγουν οντολογίες από τα Folksonomies, όμως αρκετές άλλες επηρεάστηκαν και ακολούθησαν την μέθοδο του *Gruber T.* Ουσιαστικά με την TagOntology ο ερευνητής θέλησε να δημιουργήσει ένα πρότυπο σήμανσης, το οποίο επιτρέπει υπηρεσίες ανάλυσης στα δεδομένα των ετικετών από τις αντίστοιχες εφαρμογές.

3.5.4 Εργαλείο Σημασιολογικού Εμπλουτισμού FLOR

3.5.4.1 Γενικά

Η έρευνα της ***Αγγελέτου Σ.*** [6] σχετίζεται με την μελέτη των πιθανών συνδυασμών των ετερογενών τεχνολογιών του Σημασιολογικού ιστού και του Web2.0, με σκοπό να συνεισφέρει στην δημιουργία ενός ανοιχτού και έξυπνου διαδικτυακού περιβάλλοντος.

Παρατηρώντας αρκετές αδυναμίες σε διάφορες άλλες προσεγγίσεις, θέλησε μέσω της δουλειάς της:

- να ορίσει αυτοματοποιημένα τις ετικέτες σήμανσης που ήδη υπάρχουν στα folskonomies,
- και να είναι ανεξάρτητα συνδεδεμένες για να μπορούν να χρησιμοποιηθούν σε συστήματα.

Ο στόχος της είναι να:

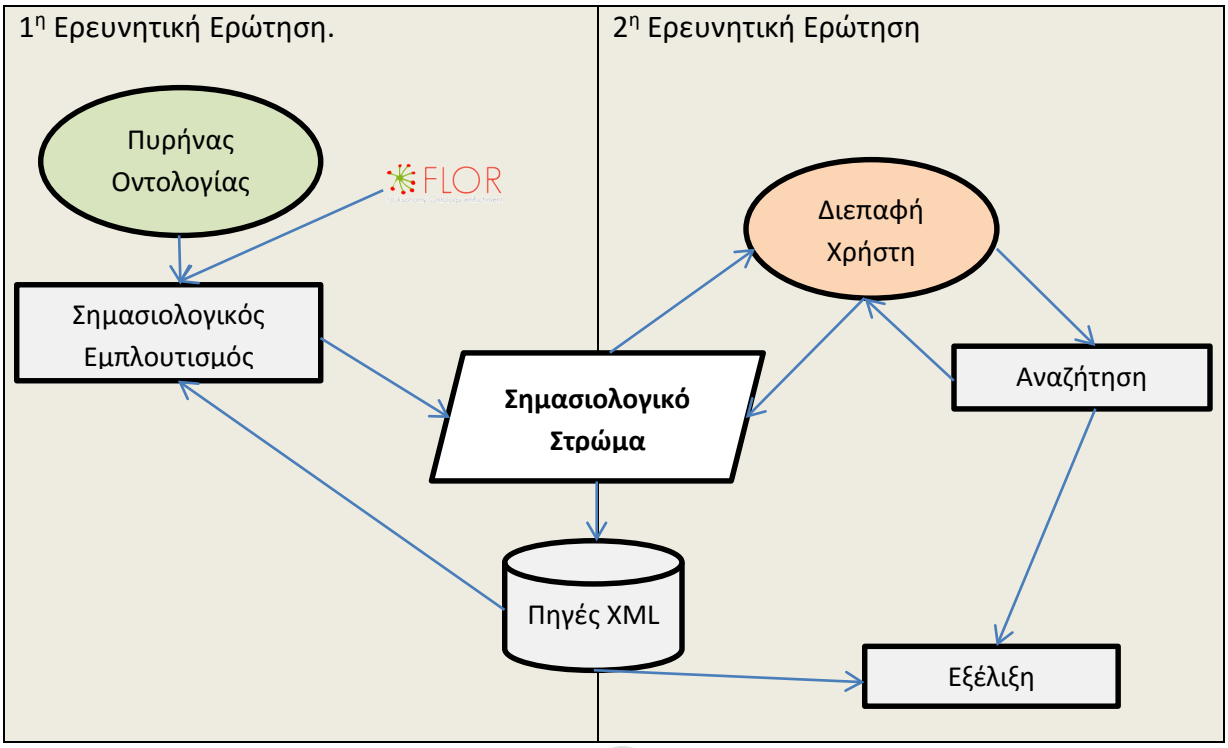
- πετύχει μεγαλύτερη κάλυψη στο εμπλουτισμό των όρων, χρησιμοποιώντας περισσότερες από μια πηγές γνώσης,
- για να αξιολογηθεί στα πραγματικά δεδομένα σήμανσης από τα Folksonomies.

Αρχικά διεξήγαγε ορισμένα πειράματα σχετικά με τα Folksonomies, τα αποτελέσματα των οποίων συνεισέφεραν στη διαμόρφωση της τελικής της μεθοδολογίας. Ο στόχος ήταν διπλός: Από την μία πλευρά θέλησε από το πρώτο στάδιο να διαπιστώσει το βαθμό αυτοματοποίησης του Σημασιολογικού εμπλουτισμού των ετικετών στα Folksonomies και από την άλλη, στην κατανόηση των προβλημάτων που επρόκειτο τυχόν να αντιμετωπίσει στην σχεδίαση του κάθε μέρους της μεθοδολογίας της.

3.5.4.2 Μεθοδολογία

Η προσέγγιση της αφορά στη δημιουργία ενός σημασιολογικού "στρώματος" στο πάνω μέρος των Folksonomies, σαν μια ενδιάμεση διεπαφή ανάμεσα στην αναζήτηση και του χώρου των ετικετών σήμανσης *(Εικόνα 6)*. Με την ενέργεια της αυτή έχει ως στόχο τον σημασιολογικό εμπλουτισμό των ετικετών, ώστε να ενισχυθεί η ανάκτηση περιεχομένου. Όπως φαίνεται στην παρακάτω εικόνα η έρευνα της χωρίζεται σε δύο ερωτήματα :

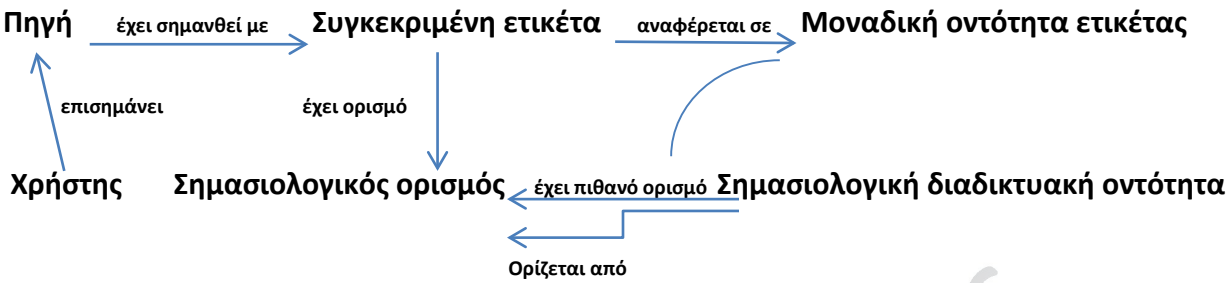
- **1η Ερευνητική Ερώτηση:** *Πως μπορούν οι χώροι ετικετών (tagspace) των Folksonomies να εμπλουτιστούν με σημασιολογικές περιγραφές αυτοματοποιημένα;*
- **2η Ερευνητική Ερώτηση:** *Πως μπορούν οι εμπλουτισμένοι χώροι ετικετών να αξιολογηθούν για την ανάκτηση περιεχομένου σε σύγκριση με αυτούς που δεν έχουν εμπλουτιστεί;*



Εικόνα 6: Σημασιολογικό στρώμα στο χώρο ετικέτας ενός Folksonomy

Κατά την ανάπτυξη της μεθοδολογίας το περιεχόμενο που είναι αποθηκευμένο στα Folksonomies είναι σε μορφή XML, ενώ εκτός από τον τίτλο, την περιγραφή και τις ιδιότητες της πηγής, υπάρχει και ο χρήστης (που έχει σημάνει ή αναρτήσει την πηγή), ο σύνδεσμος (που κάνει προσβάσιμο το περιεχόμενο) και τέλος ένα σύνολο ετικετών σήμανσης για την πηγή πληροφορίας, που έχουν δοθεί από τον ίδιο τον χρήστη.

Ο σημασιολογικός εμπλουτισμός, όπως παρατηρείται και στο παραπάνω σχήμα, δημιουργεί το στρώμα στο πάνω μέρος των πηγών των Folksonomies και απαιτεί εκτός από τα δεδομένα τους, έναν πυρήνα οντολογίας και το σημασιολογικό εργαλείο εμπλουτισμού FLOR, το οποίο θα περιγράψουμε στη συνέχεια. Ο πυρήνας οντολογίας *(Εικόνα 7)* είναι απαραίτητος για την περιγραφή των σχέσεων των οντοτήτων των Folksonomies (χρήστες, πηγές, ετικέτες) που παρέχονται από το σχήμα της XML κατά την είσοδο των δεδομένων.



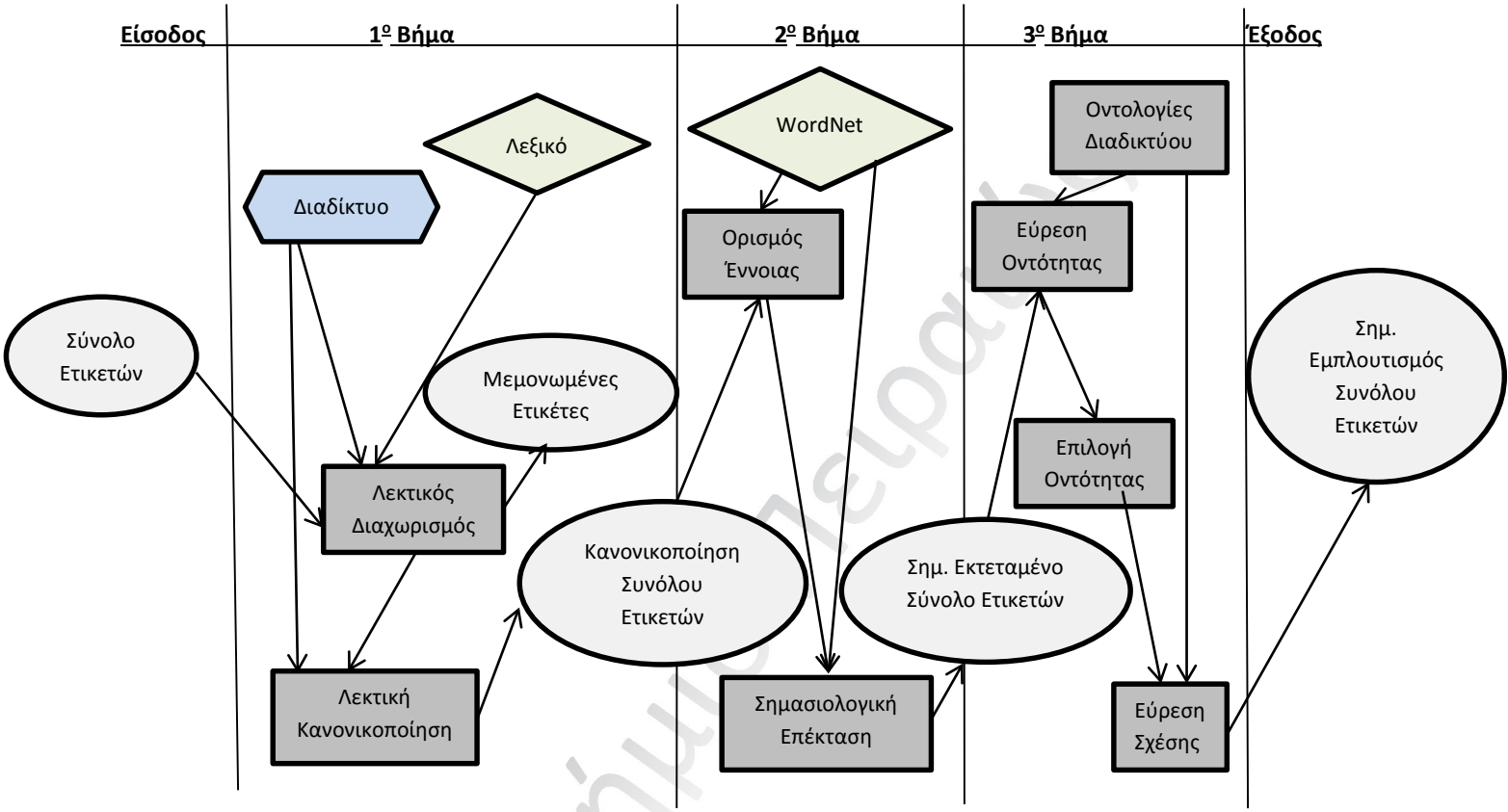
Εικόνα 7: Ο πυρήνας οντολογίας (The Core Ontology)

Οι σχέσεις "*Χρήστης επισημάνει Πηγή"* και "*Πηγή έχει σημανθεί με Συγκεκριμένη ετικέτα"* παρέχονται ήδη από το XML σχήμα. Ορίζει ως "*Συγκεκριμένη ετικέτα"* την κατηγορία για να αναπαραστήσει την ύπαρξη μιας ετικέτας στο περιεχόμενο της πηγής και του χρήστη και "*Μοναδική οντότητα ετικέτας",* την κατηγορία για να αναπαραστήσει την ετικέτα ως μοναδική οντότητα στο σύστημα. Με αυτό τον τρόπο στοχεύεται ο επιπλέον προσδιορισμός συγκεκριμένου ορισμού για τις κατηγορίες "*Συγκεκριμένη ετικέτα"* και "*Μοναδική οντότητα ετικέτας"* για την κατηγορία "*Σημασιολογικός ορισμός".*

3.5.4.3 Εργαλείο FLOR

Το FLOR είναι ένα εργαλείο σημασιολογικού εμπλουτισμού, το οποίο ουσιαστικά δημιουργεί τους αντίστοιχους ορισμούς των ετικετών, συνδέοντας τις αυτόματα με την κατάλληλη σημασιολογική οντότητα. Ο κύριος στόχος του είναι να μετατρέψει το σύνολο των ετικετών ενός Folksonomy σε μια πλούσια σημασιολογική αναπαράσταση, αναθέτοντας σχετικές οντότητες σε κάθε ετικέτα. Η ερευνήτρια θεώρησε πως εκτελεί τρία βασικά βήματα *(Εικόνα 8)*: **την Λεκτική Επεξεργασία** (Lexical Processing), **τον Ορισμό Σημασίας και Σημασιολογικής Επέκτασης** (Sense Definition and Semantic Expansion) και **τον Σημασιολογικό Εμπλουτισμό** (Semantic Enrichment). Κατά το *πρώτο βήμα* ελέγχεται η είσοδος του συνόλου των ετικετών σήμανσης, για να εξαιρεθούν όλες οι ετικέτες που παρουσιάζουν παρόμοιο ή που δεν έχουν κάποιο νόημα (meaningless tags). Απόρροια αυτού, είναι μια λίστα από λεκτικές απεικονίσεις που προέρχονται από την χρήση ενός συνόλου ευριστικών στρατηγικών. Κατά την διάρκεια του *δεύτερου βήματος*, κάθε ετικέτα λαμβάνει μία αντίστοιχη έννοια από το σημασιολογικό λεξικό WordNet [43]. Για να αποδοθεί με μία πλουσιότερη αναπαράσταση

εξάγονται επιπλέον όλα τα συνώνυμα και τα υπώνυμα της. Τέλος στο *τρίτο βήμα*, κάθε ετικέτα συνδέεται με μία σημασιολογική διαδικτυακή οντότητα (Semantic Web Entity).



Εικόνα 8: Οι φάσεις του εργαλείου FLOR

1ο Φάση: Λεκτική Επεξεργασία (Lexical Processing)

Εξαιτίας της ελευθερίας σήμανσης που παρέχουν τα Folksonomies, χρησιμοποιούνται πολλοί διαφορετικοί τύποι ετικετών. Κατά την πρώτη φάση αναγνωρίζονται όλοι αυτοί που χρησιμοποιούνται, για να αποφασιστεί ποιες ετικέτες παρουσιάζουν παρόμοιο ή που δεν έχουν κάποιο νόημα (meaningless tags) για να ληφθούν υπόψη στην βάση της διαδικασίας του Σημασιολογικού εμπλουτισμού. Προηγούμενες έρευνες εντόπισαν διαφορετικές εννοιολογικές κατηγορίες ετικετών (event, location, person), όπως και κατηγορίες που μπορούν να περιγραφούν με συντακτικά χαρακτηριστικά. Για παράδειγμα υπήρχαν πολλές ετικέτες με αριθμούς, πληθυντικό και ενικό αριθμό, συνεχόμενες – χωρίς κενό μεταξύ τους λέξεις, λέξεις με κενά ή γραμμένες σε άλλη γλώσσα. Γι αυτό ακριβώς τον λόγο με την λεκτική επεξεργασία αποκλείονται όλες αυτές που δεν παρουσιάζουν νόημα.

Κατά την διάρκεια του ***λεκτικού διαχωρισμού*** (Lexical Isolation) αναγνωρίζονται είτε οι ετικέτες που πρέπει να αποκλειστούν, ή αυτές που χρειάζονται για την περαιτέρω επεξεργασία. Επιπλέον αποκλείονται όλες οι ετικέτες με νούμερα, με ιδιαίτερους χαρακτήρες (όπως ☺) και αυτές που είναι γραμμένες σε άλλη γλώσσα, καθώς η προτεινόμενη μέθοδος της στηρίζεται σε εξωτερικούς πόρους γνώσης που βασίζονται στην Αγγλική γλώσσα (WordNet, Σημασιολογικές διαδικτυακές οντολογίες).

Χρησιμοποιώντας την ***λεκτική κανονικοποίηση*** (Lexical Normalisation) η ερευνήτρια είχε ως στόχο να λύσει το πρόβλημα ασυμβατότητας μεταξύ των όρων που χρησιμοποιούνται στα Folksonomies, στις οντολογίες και στο WordNet. Κατά την διάρκεια αυτής της φάσης παράγεται μία λίστα με πιθανές λεκτικές αναπαραστάσεις (Lexical Representation) για κάθε ετικέτα. Για παράδειγμα η σύνθετη ετικέτα santabarbara εμφανίζεται στα folksonomies ως Santa-Barbara ή σε πολλές οντολογίες ως Santa + Barbara και Santa Barbara στο WordNet.

2η Φάση: Ορισμός Σημασίας και Σημασιολογική Επέκταση (Sense Definition and Semantic expansion)

Κατά τη διάρκεια της φάσης ***ορισμού της σημασίας και της αποσαφήνισης*** ανακαλύπτεται για κάθε ετικέτα η προοριζόμενη γι αυτή σημασία στο πλαίσιο που εμφανίζεται. Ως πλαίσιο ορίζεται ένα σύνολο από ετικέτες, που για καθεμία - όταν περιγράφεται μια πηγή -συνυπάρχει μια δοσμένη ετικέτα. Για παράδειγμα για το σύνολο ετικετών {apple, orange, banana, pear}, το πλαίσιο της ετικέτας apple είναι {orange, banana,pear}. Στη συνέχεια χρησιμοποιείται το σημασιολογικό λεξικό WordNet, για να υπολογιστούν όλες οι ομοιότητες μεταξύ των ετικετών και να επιτευχθεί η αποσαφήνιση. Σε περίπτωση που μία από αυτές έχει περισσότερες από μία σημασίες, εντοπίζονται οι πιο σχετικές από αυτές με βάση και το πλαίσιο εμφάνισης (τις συν-υπάρχουσες ετικέτες). Γι αυτό ακριβώς τον λόγο κατέληξε στο συμπέρασμα πως μπορεί να μετρηθεί η ομοιότητα μεταξύ των συνδυασμών των ετικετών (προερχόμενα από το λεξικό) με την χρήση της φόρμουλας ομοιότητας των Wu και Palmer [26] στο σχήμα του WordNet. Ο βαθμός ομοιότητας είναι ανάλογος με τον αριθμό των κοινών προγόνων στην ιεραρχία του λεξικού και με το εύρος των συνδεόμενων μονοπατιών (paths) ανάμεσα στις δύο έννοιες. Απόρροια αυτού είναι ένα ζεύγος

εννοιών και ο μεταξύ τους βαθμός. Για την εξέλιξη της διαδικασίας, επιλέγεται ένα τέτοιο ζεύγος που εμφανίζει τη μεγαλύτερη ομοιότητα με την προϋπόθεση να ξεπερνά ένα κατώτερο όριο ομοιότητας. Εάν μια ετικέτα, όταν συγκρίνεται με όλες τις υπόλοιπες από την ομάδα της, έχει ελάχιστα σημεία ομοιότητας, τότε ανατίθεται στην δημοφιλέστερη σημασία στο WordNet.

Η ***σημασιολογική επέκταση (Semantic Expansion)*** περιελάμβανε συνώνυμα και υπερώνυμα για κάθε μία από τις υπάρχουσες ετικέτες στο εργαλείο, δεδομένης της σημασίας που αντιστοιχήθηκε σε κάθε ετικέτα από την προηγούμενη φάση.

3η Φάση: Σημασιολογικός Εμπλουτισμός (Semantic Enrichment)

Κατά την διάρκεια αυτής της φάσης, προσδιορίζονται οι σημασιολογικές διαδικτυακές οντότητες που είναι σχετικές για κάθε ετικέτα, αξιοποιώντας τα αποτελέσματα της λεκτικής κάθαρσης (Lexical Cleaning) και της σημασιολογικής επέκτασης (Semantic Expansion). Για τις ετικέτες που παρουσίαζαν κάποια συσχέτιση, η ερευνήτρια θεώρησε πως θα πρέπει να επιλέγονται από το *Watson Semantic Web gateway* [44]. Στο συγκεκριμένο σύστημα αναζητούνται όλες οι οντότητες των οντολογιών (*Classes, Properties, Individuals*) που περιέχουν μια από τις λεκτικές αναπαραστάσεις ή συνώνυμα της ετικέτας. Σε αυτή τη διαδικασία συχνά παρέχονται ως αποτελέσματα πολλές σχετικές σημασιολογικές διαδικτυακές οντότητες με μεγάλη ομοιότητα μεταξύ τους. Για να μειωθεί επομένως ο μεγάλος αριθμός τους, χρησιμοποιήθηκε η **διαδικασία ολοκλήρωσης της οντότητας** (Entity Integration Process), από την προσέγγιση του *Trillo R.* [45]. Με την συγκεκριμένη εξαλείφονται όλες αυτές που έχουν μεγάλο βαθμό ομοιότητας και για να υπολογιστεί αυτός, συγκρίνονται οι σημασιολογικοί “γείτονες” (superclasses, subclasses, disjoint classes for classes, domain, range, superproperties, subproperties for properties), τα ονόματα και οι ταμπέλες σήμανσης μεταξύ δύο οντοτήτων.

Η **ομοιότητα *simDgr*** για δυο σημασιολογικές διαδικτυακές οντότητες *e1* και *e2* υπολογίζεται ως εξής:

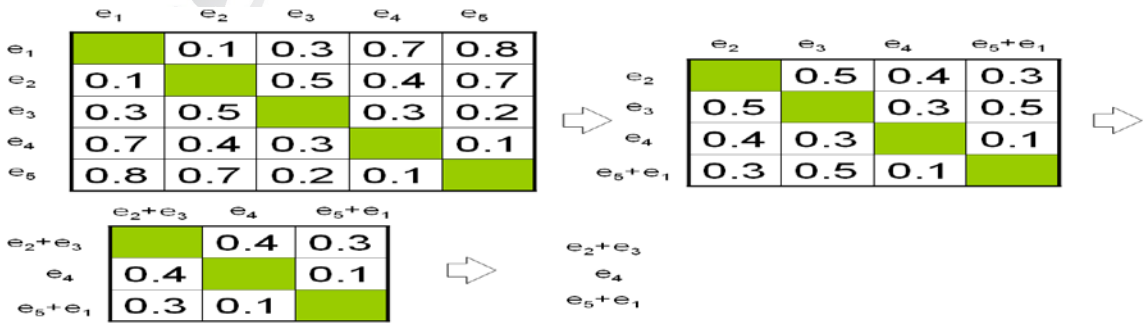
*simDgr = Wi * simLexical(e1, e2) + Wg * simGraph(e1, e2)*

Όπου:

- simLexical(e1, e2) είναι η ομοιότητα μεταξύ της λεκτικής απεικόνισης δύο οντοτήτων, δηλαδή τα ονόματα και οι ταμπέλες σήμανσης τους με την μέθοδο μέτρησης Levenshtein.
- simGraph(e1, e2) είναι η ομοιότητα των "γειτόνων" των οντοτήτων, όπου ο υπολογισμός της αντίστοιχης ομοιότητας για κάθε στοιχείο αυτών, βασίζεται στο αλφαριθμητικό του (string).

Εξαιτίας του γεγονότος πως θεωρείται σημαντικότερη η ομοιότητα των σημασιολογικών γειτόνων από την αντίστοιχη των ταμπελών σήμανσης, οι βαρύτητες (weights) ορίστηκαν ως Wi=0,3 και Wg=0,7. Εάν η ομοιότητα είναι μεγαλύτερη από ένα κατώφλι, αυτές συγχωνεύονται σε μία, ενσωματώνοντας τους "γείτονες" τους σε αυτήν. Στη συνέχεια επαναλαμβάνεται η διαδικασία μέχρι όλες οι οντότητες να διαφέρουν αρκετά μεταξύ τους, δηλαδή ο αριθμός ομοιότητας τους να είναι μικρότερος από το ορισμένο όριο.

Για την καλύτερη κατανόηση της διαδικασίας δίνεται ένα **παράδειγμα** βασισμένο στην παρακάτω εικόνα (*Εικόνα 9*), όπου πέντε σημασιολογικές διαδικτυακές οντότητες συγκρίνονται με ένα κατώφλι 0,5. Αρχικά συγκρίνουμε τα παραγόμενα ζεύγη, παρατηρώντας πως τα ζεύγη (e1,e4), (e1,e5,), (e2,e3) και (e2,e5) έχουν βαθμό ομοιότητας ίση ή κάτω από το όριο. Στη συνέχεια συγχωνεύονται οι δύο πρώτες οντότητες με την μεγαλύτερη ομοιότητα (e1 και e5) σε μία (e1+e5) και υπολογίζεται ο νέος βαθμός για την κάθε μία (νέα οντότητα και εναπομείνουσα). Αυτή η διαδικασία επαναλαμβάνεται ωσότου όλες οι ομοιότητες γίνουν μικρότερες από τα ορισμένα όρια (κάτι που θα δείχνει πως είναι και αρκετά διαφορετικές).



Εικόνα 9: Στρατηγική συγχώνευσης με όριο την τιμή 0,5

Εφόσον δημιουργηθούν όλες οι ομάδες οντοτήτων, κάθε ετικέτα αντιστοιχίζεται με σχετικές οντότητες. Αυτό πραγματοποιείται, συγκρίνοντας τους οντολογικούς "γονείς" (superclasses of the classes, the superproperties of the properties, the classes of individuals) από τις οντότητες, με τα υπερώνυμα (hypernyms) που ανακτώνται από το WordNet. Για παράδειγμα όπως φαίνεται στην *Εικόνα 10*, η ετικέτα moon έχει εμπλουτιστεί με δύο οντότητες. Οι κύριες κατηγορίες (superclasses) και των δύο, αποτελούνται από τα υπερώνυμα της έννοιας moon (από το WordNet). Επιπλέον εκτός από τον ορισμό οντότητα, κάθε ετικέτα εμπλουτίζεται με πληροφορίες από την ίδια την οντότητα (EarthsMoon TypeOf Moon).

moon

Lexical Representations	Synonyms	Hypernyms	Entities
moon		satellite celestial_body heavenly_body natural_object object physical_object entity	**http://www.ida.liu.se/~adrpo/modelica/rdf/inheritance.owl#moon** type (of) http://www.ida.liu.se/~adrpo/modelica/rdf/inheritance.owl#***CelestialBody* **http://www.cyc.com/2003/04/01/cyc#moon** subClassOf http://www.cyc.com/2003/04/01/cyc#***NaturalSatellite*** type http://www.cyc.com/2003/04/01/cyc#***EarthsMoon***

Εικόνα 10: Εμπλουτισμός της ετικέτας moon στο εργαλείο FLOR

3.5.4.4 Συμπεράσματα-Σύνοψη

Η ερευνήτρια με την προσέγγιση της απέδειξε πως ο σημασιολογικός εμπλουτισμός του χώρου ετικετών των Folksonomies -με την βοήθεια του WordNet και των διαδικτυακών οντολογιών- είναι εφικτή, χωρίς την παρέμβαση του χρήστη σε κανένα βήμα της μεθοδολογίας της.

Εφάρμοσε τη μεθοδολογία της σε ένα υποσύνολο από φωτογραφίες του *Flickr*, ενός από τα πιο γνωστά συστήματα διαμοιρασμού δεδομένων. Μετά το τέλος των βημάτων της λεκτικής επεξεργασίας (Lexical Processing) και της σημασιολογικής επέκτασης (Semantic Expansion), συσχετίστηκαν σωστά το 72% (179 από τα 250) από τις ετικέτες των folskonomies με μία τουλάχιστον σημασιολογική διαδικτυακή οντότητα. Εμπλούτισε περίπου το 49% των ετικετών με πάρα πολύ μεγάλη ακρίβεια, ίση με 93%. Αυτό αποδεικνύει μια σημαντική

βελτίωση σε σχέση με προηγούμενες προσπάθειες της, όπου συσχέτιζε τις ετικέτες με σημασιολογικές διαδικτυακές οντότητες χωρίς να τις επεκτείνει με συνώνυμα και υπερώνυμα.

3.5.5 Αλγόριθμος SemTagP: Σημασιολογική Ανίχνευση Κοινότητας στα Folksonomies

3.5.5.1 Γενικά

Οι ***Ereteo G.*** και ***Gandon F.*** [26] πρότειναν τη συγχώνευση *τριών προσεγγίσεων* (*RAK, tag based labeling και folksonomy structure*), έχοντας ως στόχο την ανίχνευση κοινότητας. Μια τέτοια προσπάθεια απέφερε πολλά οφέλη, όχι μόνο ως προς τη δομή των συνδέσμων των κοινωνικών δικτύων, αλλά και ως προς τον σημασιολογικό εμπλουτισμό των Folksonomies. Χρησιμοποίησαν έναν αλγόριθμο ανίχνευσης κοινότητας **SemTagP**, ο οποίος επωφελείται από τα σημασιολογικά δεδομένα κατά την δόμηση των RDF σχημάτων από τα κοινωνικά δίκτυα. Ο αλγόριθμος αυτός, δεν εντοπίζει μόνο αλλά και επισημαίνει κοινότητες, αξιοποιώντας τις ετικέτες και τις σημασιολογικές μεταξύ τους σχέσεις. Μάλιστα, προκειμένου να αξιολογήσουν τα αποτελέσματα της μεθοδολογίας τους εφάρμοσαν τον αλγόριθμο ανίχνευσης κοινότητας SemTagP στο κοινωνικό δίκτυο ADEME Agency.

Η **ανίχνευση της κοινότητας** (community detection) συμβάλει γενικότερα στην κατανόηση της κατανομής των χρηστών (actors) και των δραστηριοτήτων (activities). Ποικίλες εργασίες μπορούν να επωφεληθούν από τη συγκεκριμένη ενέργεια όπως για παράδειγμα η επιχειρηματική ευφυΐα (business intelligence), η παρακολούθηση εξελίξεων στην τεχνολογία (technology monitoring), η παροχή χρήσιμων συμβουλών (consulting). Για να αντιμετωπιστεί αυτό το πρόβλημα χρησιμοποιούνται **είτε ιεραρχικοί ή ευριστικοί αλγόριθμοι** (heuristics). Οι πρώτοι παράγουν ένα δέντρο με τμήματα από την κοινότητα, με διαδοχικό επαναλαμβανόμενο διαχωρισμό του δικτύου σε υποκοινότητες (top-down) ή συγχωνεύοντας τις κοινότητες σε μία μεγαλύτερη (bottom-up). Οι δεύτεροι αξιοποιούν τα χαρακτηριστικά ενός δικτύου για να προσδιορίσουν πυκνές συνδεδεμένες ομάδες κόμβων.

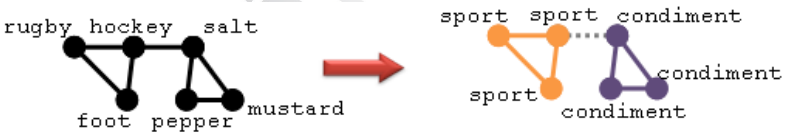
Εκτός από τους παραπάνω υπάρχουν και οι **αλγόριθμοι διάδοσης ετικέτας (label propagation)**, γνωστοί και ως **RAK**, που ανιχνεύουν κοινότητες διαδίδοντας τις ετικέτες τους στα κοινωνικά δίκτυα με τον ακόλουθο τρόπο: (1) Ο αλγόριθμος αποδίδει μια ετικέτα για κάθε κόμβο. (2) Κάθε κόμβος *n* αντικαθιστά την ετικέτα του με αυτήν που χρησιμοποιείται από τους

γειτονικούς του στο γράφημα, εάν η ετικέτα του είναι διαφορετική. (3) Εάν τουλάχιστον ένας κόμβος αλλάξει την ετικέτα του, επαναλαμβάνεται η διαδικασία από το 2ο βήμα. (4) Σε διαφορετική περίπτωση ολοκληρώνεται η διαδικασία.

3.5.5.2 Προσέγγιση

3.5.5.2.1 Σημασιολογική Διάδοση Ετικετών Σήμανσης

Ο *SemTaGP* (*Εικόνα 11*) είναι ένας αλγόριθμος που ανιχνεύει και χαρακτηρίζει κοινότητες από τα γραφήματα των RDF περιγραφών των κοινωνικών δικτύων και των Folksonomies. Είναι μία επέκταση του αλγόριθμου RAK, που ουσιαστικά αλλάζει την απλή διάδοση (label propagation) των ετικετών σε σημασιολογική (semantic propagation). Πιο συγκεκριμένα, αντί να διαδίδονται τυχαίες ετικέτες, παρέχει στους χρήστες αυτές που σχετίζονται οι ίδιοι, όπως παρατηρείται και στο παράδειγμα της παρακάτω εικόνας 11. Η διάδοση αυτή, πραγματοποιείται με τη χρήση γενικευμένων σχέσεων μεταξύ τους (δηλαδή *skos: narrower / skos: broader*), προκειμένου να συγχωνευτούν οι ετικέτες σήμανσης σε εξειδικευμένες κοινότητες και να γενικεύσουν τις ονομασίες τους με κοινά υπερώνυμα (*hyperonyms*).



Εικόνα 11: Σημασιολογική διάδοση ετικετών σήμανσης.

Οι ερευνητές χρησιμοποίησαν τη **διάρθρωση** (*modularity*) σε *RDF* γραφήματα, για να αξιολογήσουν την ποιότητα κάθε διαδρομής διάδοσης (propagation loop) μέσα σε μία κοινότητα. Επομένως όταν ένα κατατμημένο δίκτυο κοινότητας (*partitioned network*) παρουσιάζει υψηλό βαθμό διάρθρωσης, τότε υπάρχουν περισσότερες συνδέσεις ανάμεσα στους κόμβους μέσα σε κάθε κοινότητα, συγκριτικά με τους κόμβους που είναι ανάμεσα σε διαφορετικές κοινότητες. Πιο συγκεκριμένα η ποιότητα κατανομής μετρά τον αριθμό των ακμών μέσα στις κοινότητες του δικτύου, αφαιρώντας την αναμενόμενη τιμή της ίδιας ποσότητας σε ένα άλλο δίκτυο με την ίδια κοινότητα και με τυχαίες συνδέσεις ανάμεσα στους κόμβους.

$$Q = \frac{1}{m}\sum_{i,j\epsilon V}\left[A_{ij} - \frac{d^{out}_{<i>}d^{in}_{<j>}}{m}\right]\delta(c_i, c_j)$$

Η παραπάνω σχέση αποτελεί τον πρώτο ορισμό για τη διάρθρωση όπου: m είναι ο αριθμός των ακμών του δικτύου, A_{ij} ο αριθμός των ακμών μεταξύ των i και j, c_i η κοινότητα των i, δ(c_i και c_j)= 1 εάν c_i = c_j, διαφορετικά ισούται με 0, $d^{out}_{<i>}$ και $d^{in}_{<j>}$ οι βαθμοί της κορυφής i.

Ο αλγόριθμος SemTagP εξαπλώνει επαναληπτικά τις ετικέτες σήμανσης στο δίκτυο, έχοντας ως σκοπό να λάβει μια νέα διάρθρωση: Κατά τη διάρκεια μιας διαδρομής διάδοσης (propagation loop), κάθε χρήστης (actor) επιλέγει την ετικέτα που έχει χρησιμοποιηθεί περισσότερο ανάμεσα στους γείτονές του. Για κάθε ετικέτα t πρέπει να υπάρχει μία εμφάνιση για κάθε γείτονα που την χρησιμοποιεί , και μία εμφάνιση για κάθε γείτονα που χρησιμοποιεί μια ετικέτα skos: narrower ή skos: broader από τις t. Τέλος, το προτελευταίο κατατμημένο δίκτυο αποτελεί και την έξοδο του αλγόριθμου.

Οι ερευνητές, ήδη από προηγούμενες αναλύσεις τους, θεωρούσαν ότι είναι ιδιαίτερα σημαντική για τους χρήστες η ποικιλομορφία και η σημασιολογία των συνδέσμων που χρησιμοποιούν. Ο πολλαπλασιασμός των ετικετών, μέσω των διαφορετικών τύπων των σχέσεων, δύναται να παράγει διαφορετικά τμήματα στην κοινότητα. Κατά συνέπεια, ο αλγόριθμος *SemTagP* ρυθμίζει τις παραμέτρους του από τον τύπο της σχέσης που υπάρχει και έχει την παρακάτω μορφή:

```
Algorithm SemTagP(RDFGraph network, Type rel)
1. DO
2. old_network = network
3. //propagate tags (i.e. compute new
partitions)
4. FOREACH user in network.users
5. user.tag =
mostUsedAdjacentTag(user, rel)
6. END
7. WHILE modularity(network) >
modularity(old_network)
8. RETURN old_network
```

```
Algorithm mostUsedAdjacentTag(User user,
Type rel)
1. resultTag = null; max = 0
2. tagTable = new hashTable()
3. FOREACH agent in user.adjacent[rel]
4. IF tagTable.exists(agent.tag)
5. tagTable[agent.tag] ++
6. ELSE
7. tagTable[agent.tag] = 1
8. IF(max < tagTable[agent.tag]){
9. resultTag = agent.tag;
10. max = tagTable[agent.tag]
11. FOREACH broaderTag in
agent.tag.broaders
12. IF tagTable.exists(broaderTag)
13. tagTable[broaderTag] ++
14. ELSE
15. tagTable[broaderTag] = 1
16. IF max < tagTable[broaderTag]
17. resultTag = broaderTag;
18. max = tagTable[broaderTag]
19. END
20. END
21. RETURN resultTag
```

Στον πρώτο πειραματισμό της εφαρμογή του αλγόριθμου *SemTagP*, οι ερευνητές παρατήρησαν ότι ορισμένες ετικέτες με πολλές *skos: narrower* σχέσεις απορροφούσαν πάρα πολλές ετικέτες κατά τη φάση της διάδοσης (*propagation*), όπως για παράδειγμα η ετικέτα environment που υπάρχει παντού στο σύστημα *ADEME*.

Λαμβάνοντας υπόψη το γεγονός ότι τέτοιες ετικέτες ομαδοποιούν τους χρήστες σε μεγάλες κοινότητες, πρόσθεσαν μια επιλογή για να βελτιώσουν χειροκίνητα τα αποτελέσματα: μετά την πρώτη διαδρομή διάδοσης (*propagation loop*) παρουσίασαν την τρέχουσα κατάτμηση της κοινότητας σε έναν χρήστη, ο οποίος μπορεί να απορρίψει τη χρήση των *skos: narrower* σχέσεων των ετικετών που δημιουργούν τόσο μεγάλες κοινότητες. Έπειτα, επανεκκίνησαν τον αλγόριθμο και επανέλαβαν τη διαδικασία, μέχρις ότου καμία σχέση να μην απορριπτόταν ως την ολοκλήρωσή του. Εν συνεχεία, εφάρμοσαν τον αλγόριθμό τους βασιζόμενοι στη σημασιολογική μηχανή αναζήτησης *KGRAM* [46], η οποία υποστηρίζει την *RDF* γλώσσα αναζήτησης *SPARQL 1.1*.

3.5.5.2.2 Σημασιολογική Ανάθεση Ετικετών

Προκειμένου να μοντελοποιηθούν οι δραστηριότητες των Folksonomies και της κοινωνικής σήμανσης, χρησιμοποιούνται διαφορετικές οντολογίες που παράγουν σχολιασμούς *RDF*. Πιο συγκεκριμένα, οι ερευνητές θεώρησαν ότι η οντολογία *SCOT*, παρέχει ένα συνεκτικό πλαίσιο για να εκφράσει τη κοινωνική σήμανση σε σημασιολογικό επίπεδο, με έναν μηχανικό τρόπο κατανόησης. Οι οντολογίες σήμανσης αναγνωρίζουν τις ετικέτες χάρη στα *URIs* και τις μετατρέπουν σε πραγματικά αντικείμενα (με την έννοια του RDF) που μπορούν να περιγραφούν σημασιολογικά. Εξαιτίας αυτού του γεγονότος μπορεί να αξιοποιηθεί το νόημά τους, χρησιμοποιώντας τες ως θέμα ή αντικείμενο μια τριπλέτας.

Εν προκειμένω, οι ερευνητές ανέδειξαν Σημασιολογικές σχέσεις, ανάμεσα στις ετικέτες, για να δομηθεί το *Folksonomy* με "ελαφριά" σημασιολογία (*light-weight semantic*). Ο σημασιολογικός εμπλουτισμός επιτυγχάνεται, συνδυάζοντας την αυτοματοποιημένη διαδικασία των συνεισφορών από τις ετικέτες και τους χρήστες, μέσω φιλικών διεπαφών για τους τελευταίους. Ο κύκλος αυτός, ξεκινά με μια σύνθετη μέτρηση που συνδυάζει πολλές μετρήσεις μεταξύ συμβολοσειρών (*string)*, προκειμένου να αποκαλυφθούν οι τρείς κύριοι τύποι σχέσης μεταξύ των ετικετών: *skos: related, skos: closeMatch, skos: narrower*. Στη συνέχεια οι χρήστες, μέσω ενός εργαλείου πλοήγησης, μπορούν να επικυρώσουν, να απορρίψουν ή να προτείνουν σημασιολογικές σχέσεις και οι τυχόν αναφυόμενες αδυναμίες/διαφωνίες επιλύονται από έναν χρήστη με συναινετική άποψη. Αυτός ο κύκλος επαναλαμβάνεται για να διατηρήσει μια συναινετικά εμπλουτισμένη folksonomy με σημασιολογικούς ισχυρισμούς.

3.5.5.2.3 Σημασιολογική Διάδοση Ετικετών

Το βήμα διάδοσης (propagation step) συνιστά επαναληπτικά την ανάθεση της συχνότερης ετικέτας, ανάμεσα στο σύνολο των χρηστών, για καθέναν από αυτούς. Για να επιτευχθεί επομένως γενίκευση στις σχέσεις ανάμεσα στις ετικέτες σήμανσης, οι ερευνητές εκτίμησαν πως θα πρέπει να δυναμώσουν τις εμφανίσεις κάθε μιας εξ αυτών με τον αντίστοιχο (πυρήνα) των *skos: narrower* ετικέτες της.

Για παράδειγμα, σχετικά με την ετικέτα "ενέργεια" αξιοποιείται ο όρος με "*skos: narrower* ανανεώσιμες πηγές ενέργειας", μετρώντας μια ακόμα εμφάνιση της πρώτης ετικέτας (ενέργεια) για κάθε εμφάνιση της δεύτερης (ανανεώσιμές πηγές ενέργειας). Αρχικά, οι ερευνητές ξεκίνησαν κάθε διαδρομή (loop) με μια αναζήτηση που αποφέρει σε κάθε χρήστη τις ετικέτες σήμανσης των "γειτόνων" του για μια ορισμένη παραμετροποιημένη σχέση μεταξύ χρηστών και τις ευρύτερες ευρύτερες (broader) αυτών. Στη συνέχεια, όρισαν τα αποτελέσματα για χρήστες και για ετικέτες:

```
1. select ?user ?tag ?y where {
2. ?user param[rel] ?neighbor
3. {{?neighbour scot:hasTag ?tag }
4. UNION
5. {?neighbour scot:hasTag ?tag2
6. ?tag skos:narrower ?tag2
7. filter(exists{?x scot:hasTag
?tag})}}
8. } order by ?user ?tag
```

Διαφορετικά τμήματα για τη λειτουργία του αλγόριθμου mostUsedAdjacentTag() που αναφέρεται και περιγράφεται πιο πάνω (Κεφάλαιο 3.5.5.2.1, σελ.66), έχουν κωδικοποιηθεί σε αυτή την αναζήτηση:

- Η γραμμή 3, κωδικοποιεί την επιλογή κάθε ετικέτας των "γειτόνων" ενός χρήστη,
- οι γραμμές 5 και 7 κωδικοποιούν την επιλογή της ετικέτας που είναι ευρύτερη (broader) συγκριτικά με μια άλλη από αυτές των γειτόνων του χρήστη και
- η γραμμή 8 ορίζει τις προβολές για κάθε χρήστη και ετικέτα, προκειμένου να γίνει πιο εύκολη η διαδικασία της διάδοσης.

Μετά την ολοκλήρωση της παραπάνω διαδικασίας, οι ερευνητές εκτέλεσαν μια επιπλέον διεργασία στο αποτέλεσμα και αντικατέστησαν την ετικέτα κάθε χρήστη με την πιο διαδεδομένη ανάμεσα στους "γείτονες". Για να έχουν την δυνατότητα απόρριψης μιας

γενικευμένης ετικέτας, πρόσθεσαν ένα φίλτρο στις γραμμές 5 έως 7 με σκοπό να αποκλειστεί η χρήση μιας συγκεκριμένης ευρύτερης ετικέτας, δηλαδή filter(?tag!= <http://ademe.fr/energie>).

3.5.5.2.4 Διάρθρωση ενός Γραφήματος RDF

Ένα αρχείο RDF αναλύεται σε μία λίστα από τριπλέτες. Κάθε τριπλέτα αποτελείται από ένα υποκείμενο, ένα κατηγόρημα, και ένα αντικείμενο (subject, predicate, object). Οι τριπλέτες μιας RDF περιγραφής δίνουν μορφή σε ένα κατευθυνόμενο γράφημα σήμανσης, που δύναται να θεωρηθεί και ως «οι ακμές ενός γραφήματος Οντοτήτων-Σχέσεων».

Γράφημα Οντοτήτων-Σχέσεων (ERGraph): Σχετίζεται με ένα σύνολο από ετικέτες L και είναι μια ομάδα, αποτελούμενη από τέσσερις μεταβλητές $G = (E_G, R_G, n_G, 1_G)$ όπου:

- E_G και R_G είναι δύο σύνολα ξένα μεταξύ τους, αποτελούμενα αντίστοιχα από κόμβους και σχέσεις.
- n_G : R_G -> E_G συσχετίζει κάθε σχέση re R_G, ένα ζεύγος από οντότητες $e_{,i}$ e_j ε E_G που λέγονται τα επιχειρήματα της σχέσης. Εάν n_G(r)=(e_1,e_2) σημειώνεται ότι n_{G^i}(r)=$e_{\ i}$ το i^{th} όρισμα του r.
- 1_G : $E^G \cup R^G$ -> L είναι μια λειτουργία σήμανσης των οντοτήτων και των σχέσεων.

Αρθρωτός Σχεδιασμός ενός Γραφήματος Οντοτήτων-Σχέσεων (ERGraph): Ένα γράφημα G=(E_G,R_G,n_G,1_G) σχετικό με ένα σύνολο ετικετών L, για μια δοσμένη ετικέτα σχέσης pεL ισχύει:

$$Q(G,p) = \frac{1}{|R_G^P|} \sum_{i,j \epsilon E_G} [A_{ij}^p - \frac{d_{<p,i>}^{out}(G) d_{<p,j>}^{in}(G)}{|R_G^P|}] \delta(c_i, c_j)$$

Όπου:

- $R_G^P = \{r \epsilon R_G; I_G(r) = p\}$
- $A_{i,j}^P = 1\ \varepsilon\alpha\nu\ \ni r \epsilon R_G^p; n_G^1(r) = i\ \kappa\alpha\iota\ n_G^2(r) = j, \delta\iota\alpha\varphi o\rho\varepsilon\tau\iota\kappa\acute{\alpha}\ \iota\sigma o\acute{\upsilon}\tau\alpha\iota\ \mu\varepsilon\ 0$

- $d^{in}_{<p,i>}(G)$ *και* $d^{out}_{<p,i>}(G)$*είναι αντίστοιχα οι αριθμοί των σχέσεων* $r^{in}, r^{out} \epsilon R^P_G; n^2_G(r^{in}) = i$ *και* $n^1_G(r^{out}) = 1$ ο εσωτερικός (in)και ο εξωτερικός (out)βαθμός των i για την σχέση που σημάνθηκε με p.

Ο παραπάνω ορισμός προκύπτει από την αναζήτηση *RDF* γραφήματος με *SPARQL* ερωτήματα που υπολογίζουν διάφορα μέρη αυτού του τύπου. Σε προηγούμενη εργασία τους οι ερευνητές όρισαν ερωτήματα για τον υπολογισμό μεγεθών του δικτύου, κάτι που επιτρέπει τον υπολογισμό $R^P_G, d^{in}_{<p,i>}(G)$ και $d^{out}_{<p,i>}(G)$. Πρώτα υπολογίζεται το R^P_G με ένα ερώτημα που απλά ανακτά τον αριθμό των ζεύγων των RDF πόρων, τα οποία συνδέονται με την ιδιότητα p. Στη συνέχεια ανακτώνται οι in και out βαθμοί όλων των RDF πόρων που συνδέονται με την ιδιότητα p, με δύο ερωτήματα που υπολογίζουν $d^{in}_{<p,i>}(G)$ *και* $d^{out}_{<p,i>}(G)$ για κάθε πιθανή τιμή των i. Τέλος υπολογίζεται ο τύπος (*formula*) επαναλαμβάνοντας τα δύο αποτελέσματα που ακολουθούν.

Το **παρακάτω ερώτημα (1)** ανακτά όλα τα ζεύγη των πόρων που ανήκουν στην ίδια κοινότητα, για την ιδιότητα που δίνεται ως παράμετρος:

```
1. select ?user1 ?user2 ?tag where{
2.   ?user1 param[property] ?user2
3.   ?user1 scot:hasTag ?tag
4.   ?user2 scot:hasTag ?tag
5. }group by ?user1 ?user2 ?tag
```

Το **δεύτερο ερώτημα (2)** που ακολουθεί ανακτά όλα τα ζεύγη των πόρων που ανήκουν στην ίδια κοινότητα, για την ιδιότητα που δίνεται ως παράμετρος:

```
1. select ?user1 ?user2 ?tag where{
2.   ?user1 scot:hasTag ?tag
3.   ?user2 scot:hasTag ?tag
4.   filter(?user1 != ?user2)
5.   filter(not exists{?user1
param[property] ?user2})
6. } group by ?user1 ?user2 ?tag
```

Στο κεφάλαιο που ακολουθεί παρουσιάζεται η επεξεργασία στα αποτελέσματα των παραπάνω ερωτήσεων **(1),(2)**, για να υπολογιστεί ο αρθρωτός σχεδιασμός των αντίστοιχων τμημάτων της κοινότητας (ADEME Ph.D).

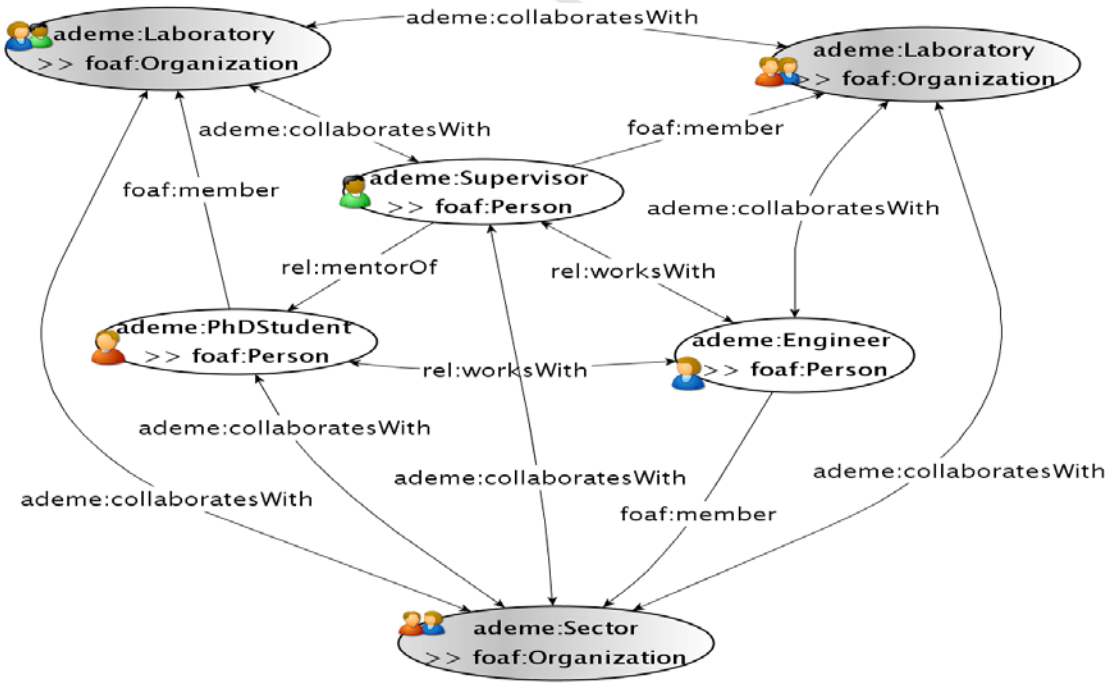
3.5.5.3 Πειραματισμός - Συμπεράσματα

Προκειμένου να αναδείξουν όλα τα πλεονεκτήματα και τα οφέλη της προσέγγισής τους οι ερευνητές εφάρμοσαν τον αλγόριθμό τους στο κοινωνικό δίκτυο *ADEME Agency* (*Εικόνα 12*). Το τελευταίο είναι ένα δίκτυο που ταξινομεί ετικέτες σήμανσης και είναι μορφοποιημένο από υπαλλήλους και ακαδημαϊκούς ερευνητές. Οι ακαδημαϊκοί χρήστες είναι οι φοιτητές και οι επιβλέποντες διδακτορικού καθώς και τα εργαστήρια στα οποία ανήκουν. Από την πλευρά της *ADEME*, κάθε διατριβή εκπονείται από έναν μηχανικό και συνδέεται με μια εσωτερική οργάνωση που ονομάζεται "secteur" (τομέας). Επιπροσθέτως, για κάθε μια από τις διατριβές, χρησιμοποιούνται ελεύθερες ετικέτες σήμανσης που εξυπηρετούν καλύτερα στην ταξινόμησή τους. Γι' αυτόν ακριβώς τον λόγο, οι ερευνητές εξήγαγαν, από το σύνολο δεδομένων, ένα γράφημα *RDF* (το οποίο περιλαμβάνει το Folksonomy και μια περιγραφή του δικτύου) εφαρμόζοντας στη συνέχεια τέσσερις διαφορετικούς αλγόριθμους, με σκοπό να κατανοήσουν ευκολότερα τη δομή της κοινότητας (την κατανομή δηλαδή των χρηστών και των δραστηριοτήτων που επισημαίνονται με ετικέτες από αυτούς).

Πιο συγκεκριμένα οι ερευνητές ανέλυσαν το σύνολο των δεδομένων ADEME, από μια σχεσιακή βάση δεδομένων (relation database), χρησιμοποιώντας μια παλιότερη μέθοδό τους [47] για να δημιουργήσουν τις αντίστοιχες RDF περιγραφές. Αυτή αφορά στην επέκταση της ανάλυσης των χρηστών -operators- των κοινωνικών δικτύων με την χρήση πλαισίων Σημασιολογικού Ιστού, για να φανερωθούν οι σχέσεις και οι αλληλεπιδράσεις μέσα σε αυτά τα δίκτυα.

Στην Εικόνα 13 παρουσιάζεται ένα σύνολο εννοιών που χρησιμοποίησαν οι ερευνητές θέλοντας να αναπαραστήσουν το δίκτυο ADEME Ph.D (στο οποίο διεξάγεται και το τελικό τους πείραμα, όπως αυτό περιγράφεται στη συνέχεια), με τις οντολογίες FOAF (αναπαριστά τα προφίλ των χρηστών και τις μεταξύ τους σχέσεις) [48] και Ademe domain (η οποία σχεδιάστηκε αποκλειστικά για να συνεισφέρει στην ευκολότερη ανάλυση του δικτύου).

Επιπρόσθετα, η παρακάτω εικόνα περιγράφει και με ποιον τρόπο εμπλουτίστηκαν σημασιολογικά οι RDF περιγραφές του ADEME Ph.D προκειμένου να επιτευχθεί η δόμηση του αντίστοιχου σημασιολογικού δικτύου. Πιο συγκεκριμένα, οι ερευνητές σύνδεσαν δύο χρήστες που εργάζονται στο ίδιο Ph.D με την ιδιότητα rel:worksWith. Στη συνέχεια, όρισαν την ιδιότητα ademe:collaboratesWith με στόχο να συνδέσουν δύο χρήστες (foaf:Person ή foaf:Organization) που εμπλέκονται στην ίδια διατριβή. Παράλληλα, δύο μηχανικοί του ίδιου τομέα συνδέθηκαν με την ιδιότητα rel:colleagueOf, με τους ερευνητές να δομούν στη συνέχεια όλες τις παραπάνω κοινωνικές συνδέσεις με τη δήλωση της ιδιότητας rel:worksWith ως υπο-ιδιότητα του ademe:collaboratesWith. Τέλος, δημιούργησαν ένα folksonomy με την επισύναψη των ετικετών του Ph.D από όλους τους χρήστες, χρησιμοποιώντας την ιδιότητα scot:hasTag. Το συγκεκριμένο folksonomy εμπλουτίστηκε σημασιολογικά με skos:narrower σχέσεις, οι οποίες υπολογίστηκαν από τον F.Limpens [47}.



Εικόνα 12: Μορφή κοινωνικού δικτύου ADEME Ph.D.

Στη συνέχεια οι ερευνητές, πειραματίστηκαν στο συγκεκριμένο δίκτυο επιθυμώντας να αξιολογήσουν την προσέγγισή τους, χρησιμοποιώντας τη μέθοδο της σημασιολογικής ανάλυσης κοινωνικού δικτύου [48] για τη μέτρηση των χαρακτηριστικών των πιο «ενεργών»

μελών του συγκεκριμένου δικτύου (ακαδημαϊκοί επιβλέποντες PhD, μηχανικοί ADEME). Θέλοντας να αξιολογήσουν τα οφέλη της σημασιολογίας στο βήμα διάδοσης (3.5.5.2.3 Σημασιολογική Διάδοση Ετικετών, σελ.69), σύγκριναν την κοινότητα που ανίχνευσαν με τέσσερις διαφορετικούς αλγόριθμους στο συγκεκριμένο σύνολο δεδομένων της. Οι αλγόριθμοι 2,3,4 είναι παραλλαγές και χρησιμοποιήθηκαν στον πειραματισμό μόνο για λόγους σύγκρισης:

1. **RAK**: Τυχαία διάδοση σήμανσης.
2. **TagP (Tag Propagation)**: Διάδοση ετικετών σήμανσης χωρίς την αξιοποίηση των σημασιολογικών μεταξύ τους σχέσεων.
3. **SemTagP**: Χωρίς ανθρώπινη παρέμβαση.
4. **Ελεγχόμενος SemTagP**: Εισάγει έναν μη αυτοματοποιημένο έλεγχο, για να αποφευχθεί η χρήση ορισμένων σχέσεων ανάμεσα στις ετικέτες. Χρησιμοποιείται ο συμβολισμός SemTagP (tag1,tag2…) για να καθοριστούν οι ετικέτες, για τις οποίες αγνοούνται οι *skos: narrower* σχέσεις. Για παράδειγμα ο SemTagP(env,energy,model) εξαιρεί τις skos:narrower σχέσεις με τις ετικέτες environment, energetique και modelisation.

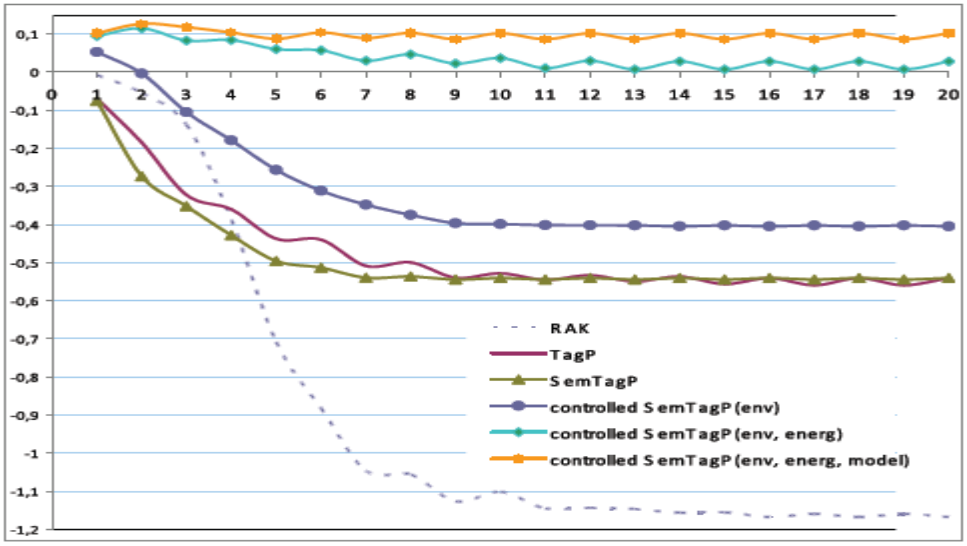
Μέσω της παραπάνω διαδικασίας αναλύθηκαν με τέσσερις αλγόριθμους οι εξελίξεις της διάρθρωσης της κοινότητας κατάτμησης (και τριών ελεγχόμενων SemTagP αλγορίθμων), με τη σύγκριση αυτών των εξελίξεων ώστε να καταδειχθεί η προστιθέμενη αξία (added value) των ετικετών διάδοσης (propagation tags) και η αξιοποίηση της σημασιολογίας τους.

Πιο συγκεκριμένα στην Εικόνα 13, παρουσιάζονται οι μεταβολές της διάρθρωσης (άξονας Y) της κοινότητας που αναφέρθηκε παραπάνω (με τη μορφή καμπυλών), η οποία αποκτήθηκε από κάθε βήμα διάδοσης (propagation loop, άξονας X). Έχουμε τη δυνατότητα να παρατηρήσουμε ότι η διάρθρωση του αλγορίθμου SemTagP(env,energy,model) ξεπερνά τις αντίστοιχες των RAK, TagP και SemTagP ενώ ο αλγόριθμος RAK προσφέρει τη χαμηλότερη ποιότητα (της κατατμημένης κοινότητας) στο σύνολο δεδομένων του δικτύου η οποία περιλαμβάνει χαμηλή πυκνότητα συνδέσμων.

Με άλλα λόγια, οι κοινωνικοί σύνδεσμοι αυτών των δεδομένων, δεν είναι αρκετά επαρκείς για να φανερωθεί η δομή της κοινότητας του συγκεκριμένου κοινωνικού δικτύου με

τη χρήση της τυχαίας διάδοσης ετικετών RAK. Μέσω της παρακάτω εικόνας μπορούμε να παρατηρήσουμε πως οι TagP και SemTagP παράγουν κατατμημένη κοινότητα με αρκετά καλύτερη διάρθρωση, σε σχέση με τον RAK. Παρόλη τη σημασιολογία όμως μεταξύ των ετικετών και του SemTagP, η τιμή της διάρθρωσης παραμένει πολύ κοντά με την αντίστοιχη από τον TagP. Αυτό συμβαίνει εξαιτίας της ευρείας ετικέτας environment, η οποία έχει πολλές skos:narrower σχέσεις, κάτι που αδρανοποιεί τους περισσότερους χρήστες της κοινότητας. Με τον SemTagP(env) εξαιρούνται οι σχέσεις skos:narrower με την ετικέτα environment, γεγονός που βελτιώνει την τιμής της διάρθρωσης.

Συνοψίζοντας ο ελεγχόμενος SemTagP ξεπέρασε τα αποτελέσματα των υπόλοιπων αλγορίθμων, όπως μαρτυρά και η παρακάτω εικόνα (Εικόνα 13), γεγονός που φανερώνει ότι η εισαγωγή των ετικετών και της σημασιολογίας προσφέρουν μια σημαντική βελτίωση του αλγόριθμου RAK.



Εικόνα 13: Διάρθρωση της κατατμημένης (διαρθρωμένης) κοινότητας που παράχθηκε, μετά από κάθε διάδοση διαδρομής με τη χρήση των RAK, TagP, SemTagP και 3 ελεγχόμενων SemTagP αλγορίθμων.

3.5.6 Μέθοδος εξαγωγής έμπειρων προφίλ από τις ετικέτες σήμανσης

3.5.6.1 Γενικά

Η ***Budura A.*** [23] πρότεινε μια νέα προσέγγιση για το πρόβλημα της εξόρυξης της τεχνογνωσίας (expertise mining) σε μια επιχείρηση, αξιοποιώντας τις κοινωνικές εφαρμογές του διαδικτύου που αναπτύσσονται μέσα σε αυτήν.

Οι εφαρμογές αυτές αποτελούνται από κοινωνικό λογισμικό (social software) που αξιοποιείται από Web 2.0 τεχνολογίες, γεγονός που επιτρέπει την ευκολότερη δημιουργία και τον μετέπειτα διαμοιρασμό των δεδομένων. Η ευρεία διάδοση ενός τέτοιου λογισμικού στο διαδίκτυο είχε ως αποτέλεσμα την εισχώρησή του και στο χώρο των επιχειρήσεων, όπου και αναγνωρίζεται ως ένα σημαντικό μέσο ενίσχυσης της «εσωτερικής» συνεργασίας. Ένα μεγάλο πλεονέκτημα ενός ενδο-δικτυακού περιβάλλοντος αφορά στο όνομα χρήστη (user namespace), το οποίο είναι όμοιο σε όλες τις εφαρμογές. Γι' αυτόν ακριβώς τον λόγο, η ερευνήτρια θεώρησε αξιόπιστη τη συσχέτιση της δραστηριότητας του χρήστη σε διαφορετικές εφαρμογές, στηρίζοντας σε αυτή την έρευνά της.

Εμπνευσμένη από τον χώρο των επιχειρήσεων και από τα μοναδικά σύνολα δεδομένων τους, η ερευνήτρια ανέπτυξε μια **πιθανοτική μέθοδο** για την οικοδόμηση προφίλ τεχνογνωσίας βασισμένα στις ετικέτες σήμανσης του εκάστοτε χρήστη. Ο προσδιορισμός της τεχνογνωσίας είναι ένα βασικό ανοιχτό ζήτημα, καθώς οι επιχειρήσεις χρειάζονται συνεχώς δυναμική κατανομή δεξιοτήτων. Γενικότερα οι μέθοδοι έρευνας τεχνογνωσίας (expert finding methods), αποσκοπούν σε ένα προφίλ για κάθε χρήστη που δείχνει την εμπειρία του σε διάφορα θέματα.

Η ερευνήτρια χρησιμοποίησε το πλεονέκτημα των πληροφοριών που προέρχονται από ένα κοινωνικό σύστημα σήμανσης μέσα σε μια επιχείρηση. Η προσέγγισή της βασίζεται στην υπόθεση ότι οι πόροι σήμανσης ενός χρήστη αντιπροσωπεύουν τα ενδιαφέροντά του, τα οποία με τη σειρά τους συνδέονται με την τεχνογνωσία του. Επιθυμώντας να εφαρμόσει το μοντέλο της και να αξιολογήσει τα αποτελέσματα της έρευνάς της, πειραματίστηκε με τα σύνολα δεδομένων από το κοινωνικό δίκτυο σήμανσης *Dogear* και την εφαρμογή *IBMr.*

3.5.6.2 Μεθοδολογία

3.5.6.2.1 Ορισμός προβλήματος

Για την μέθοδο εξαγωγής προφίλ τεχνογνωσίας από τις ετικέτες σήμανσης όρισε ως:

- U, ένα σύνολο χρηστών
- T, ένα λεξιλόγιο ετικετών σήμανσης
- B, ένα σύνολο από σελιδοδείκτες σήμανσης (bookmarks)
- S, ένα λεξιλόγιο δεξιοτήτων

Αρχικά προσδιόρισε το σύνολο των χρηστών U, όπου ο καθένας από αυτούς (u ϵ U) αλληλεπιδρά με το σύστημα σήμανσης αποδίδοντας ετικέτες σε URLs. Κάθε ετικέτα (t ϵ T) αντιπροσωπεύει μία λέξη-κλειδί, από ένα λεξιλόγιο χωρίς περιορισμούς από το σύνολο T. Η σχέση u-> t, δείχνει ότι ο χρήστης u έχει επιλέξει μια ετικέτα t για την σήμανση ενός URL και η συνάρτηση $T_u = \{t|u-> t\}$ αναφέρεται στο σύνολο των ετικετών που χρησιμοποιούνται από αυτόν. Επιπρόσθετα, κάθε όμοια ετικέτα μπορεί να χρησιμοποιηθεί από διαφορετικούς χρήστες ($u_i-> t, u_j-> t$) και ο καθένας από αυτούς με την σειρά του δύναται να τη χρησιμοποιήσει όσες φορές επιθυμεί. Εφόσον η μέθοδος αυτή βασίζεται στις ετικέτες για τον προσδιορισμό της τεχνογνωσίας ενός χρήστη, περιλαμβάνονται μόνο αυτοί όπου $T_u \neq \emptyset$. Μία ετικέτα t χρησιμοποιείται για τη διαδικασία της σήμανσης από έναν χρήστη με τον σελιδοδείκτη b ϵ B και αντιπροσωπεύει ένα συγκεκριμένο χαρακτηριστικό b γι' αυτόν. Κάθε σελιδοδείκτης με την σειρά του έχει τη δυνατότητα να σημανθεί με πολλές ετικέτες από το ίδιο άτομο ενώ το σύνολο των σελιδοδεικτών ενός χρήστη u, ορίζεται ως B_u. Όσον αφορά στο λεξιλόγιο δεξιοτήτων S, ορισμένοι χρήστες επιλέγουν να δημιουργήσουν από μόνοι τους το προφίλ τεχνογνωσίας τους με συγκεκριμένους όρους. Για την μεθοδολογία, υποτίθεται ότι ένα υποσύνολο χρηστών, $U_s \subset U$, καθορίζουν προφίλ τεχνογνωσίας τέτοιας μορφής ενώ χρησιμοποιούνται περαιτέρω και οι σχέσεις u->s, s ϵ S, για να οριστεί πως ο χρήστης u έχει επιλέξει s ως μια από τις περιοχές τεχνογνωσίας. Τέλος, το προφίλ εμπειρίας για έναν χρήστη (u ϵ U_s), είναι ένα σύνολο όρων $S_{u=}\{s|s\epsilon S \Lambda u \rightarrow s\}$.

Συνοψίζοντας, από τις παραπάνω ιδιότητες είναι δυνατόν να οριστεί ο **στόχος της μεθοδολογίας ως** ακολούθως: Η τεχνογνωσία για έναν χρήστη u ϵ {U\U_s}, ο οποίος δεν έχει

ορίσει ένα προφίλ εμπειρίας, βασίζεται στο σύνολο των ετικετών σήμανσής του (T_u) και στην επιπρόσθετη γνώση που λαμβάνει από τα προφίλ τεχνογνωσίας άλλων χρηστών.

3.5.6.2.2 Αρχικοί Πειραματισμοί

Η προαναφερθείσα ερευνήτρια, προτού παρουσιάσει την προσέγγισή της, διεξήγαγε μια σειρά από πειραματισμούς επιθυμώντας να εξετάσει τις σχέσεις ανάμεσα στις ετικέτες και τις δεξιότητες. Θεώρησε λοιπόν ότι μέσω αυτών μπορεί να εξάγει χρήσιμες παρατηρήσεις, που θα μπορούσε στη συνέχεια να τις χρησιμοποιήσει στη λειτουργία βαθμονόμησης (Scoring Function). Για την πραγματοποίησή πειραμάτων και αξιολόγησης της μεθόδου χρησιμοποιήθηκαν τα σύνολα δεδομένων δύο συστημάτων εσωτερικού επιχειρησιακού περιβάλλοντος (*IBM-internal datasets*), του *Dogear* και του *IBMr*.

Το **Dogear** [49] είναι ένα κοινωνικό σύστημα, στο οποίο οι χρήστες έχουν τη δυνατότητα να χρησιμοποιήσουν και να διαμοιράσουν σελιδοδείκτες σήμανσης για ιστοσελίδες. Κατά τη διάρκεια των αρχικών πειραμάτων, το σύνολο των δεδομένων του περιέχει 10.700 χρήστες, 65.000 ετικέτες σήμανσης και 269.000 URLs.

Το σύστημα **IBMr** [50], είναι μια εφαρμογή που παρέχει τη δυνατότητα δημιουργίας προφίλ τεχνογνωσίας, προσφέροντας επιπλέον μια λίστα γνωστικών περιοχών που αφορούν στην τεχνογνωσία του εκάστοτε χρήστη. Κάθε μέλος του συστήματος έχει τη δυνατότητα να επιλέξει ένα προκαθορισμένο λεξιλόγιο 840 όρων καθώς και ένα σύνολο θεμάτων τεχνογνωσίας. Η ερευνήτρια παρατήρησε μάλιστα μια σημαντική διαφορά όσον αφορά στην ονομασία του συνόλου των όρων για καθένα από τα δύο συστήματα και συγκεκριμένα ότι στο *Dogear* χρησιμοποιούνται ως ετικέτες (tags), ενώ στο *IBMr* ως δεξιότητες (skills). Αν και τα δύο σύνολα παρουσιάζουν ομοιότητες θεωρήθηκε ότι υπάρχει σημασιολογική διάκριση ανάμεσά τους. Η εφαρμογή IBMr αποτέλεσε σημαντικό κομμάτι για την ερευνά της *Budura A.*, καθώς συνέβαλε- εκτός από την διάκριση- στην κατανόηση της συσχέτισης ανάμεσα στα σύνολα των όρων.

Για να ορίσει την επίδοση της μεθόδου της η ερευνήτρια αγνόησε τα προφίλ τεχνογνωσίας, προσπαθώντας με τον τρόπο αυτό να ανακτήσει ορισμένα από αυτά βάσει των

ετικετών σήμανσης. Χρησιμοποίησε λοιπόν μια λειτουργία βαθμονόμησης (scoring function) για τη μέτρηση των δεξιοτήτων κάθε χρήστη (σελ 81) την οποία και όρισε ως εξής :

$Precision@k = \frac{\#correct\_skills}{k}$, όπου k είναι ένας βαθμός αποκοπής .

Μια ορισμένη δεξιότητα μπορεί να θεωρηθεί σωστή εάν παρουσιάζει ένα πραγματικό σύνολο (δεξιοτήτων) για κάθε χρήστη.

Για τη διεξαγωγή των πειραμάτων της επέλεξε 1.200 χρήστες, εφαρμόζοντας παράλληλα έναν αλγόριθμο στα σύνολα των όρων, για να υπολογίσει την ομοιότητα ανάμεσα σε λέξεις με ίδια ρίζα. Εξέτασε τα παρακάτω ερωτήματα κάνοντας τις αντίστοιχες παρατηρήσεις:

1. *Οι όροι που ορίζουν τις δεξιότητες είναι και ετικέτες στα κοινωνικά συστήματα σήμανσης (social bookmarking system);*

   Παρατήρηση: Συνολικά, η συμπεριφορά του χρήστη στη σήμανση, αφορά και στα θέματα τεχνογνωσίας.
2. *Σχετίζονται για έναν χρήστη τα σύνολα των ετικετών σήμανσης και δεξιοτήτων που ορίζουν το προφίλ εμπειρίας του;*

   Παρατήρηση: Από όλους τους όρους που χρησιμοποιούνται από αυτόν για την σήμανση, ελάχιστοι περιλαμβάνονται και στο σύνολο των δεξιοτήτων του.
3. *Αναφέρονται, σε γενικές γραμμές, οι δημοφιλέστερες ετικέτες από ένα κοινωνικό σύστημα σήμανσης σε περιοχές τεχνογνωσίας;*

   Παρατήρηση: Εάν μια ετικέτα είναι δημοφιλής, αντιπροσωπεύει μια δεξιότητα.
4. *Αντιστοιχούν για έναν συγκεκριμένο χρήστη οι δημοφιλέστερες ετικέτες σήμανσής του με τις δεξιότητες του;*

   Παρατήρηση: Δεν υπάρχει αντιστοιχία μεταξύ τους.

3.5.6.2.3 Προσέγγιση

Τα στατιστικά γλωσσικά μοντέλα (Statistical language models) έχουν εφαρμοστεί κατά καιρούς σε πολλές έρευνες και ιδιαίτερα στην ανάκτηση πληροφοριών. Η κύρια ιδέα τους αφορά στο ότι κάθε έγγραφο θεωρείται ως ένα γλωσσικό δείγμα το οποίο παράγεται από ένα

συγκεκριμένο γλωσσικό μοντέλο. Επομένως, ένα ερώτημα μπορεί να αναγνωριστεί ως διαδικασία παραγωγής (generation process): η σχετικότητα ενός εγγράφου σε ένα ερώτημα, καθορίζεται από την πιθανότητα το μοντέλο του εγγράφου να παράγει τους όρους του ερωτήματος. Υιοθετώντας την παραπάνω συσχέτιση, η ερευνήτρια εκτίμησε ότι είναι εύκολος ο υπολογισμός του βαθμού κάθε εγγράφου σε σχέση με ένα ερώτημα. Υπέθεσε επομένως ότι υπάρχει ένα σύνολο από αυτά (έγγραφα) $\mathrm{D} = d_1 \dots \mathrm{d_n}$ και ένα ερώτημα που παρουσιάζεται σαν μια σειρά από όρους $\mathrm{Q} = \{t_1, t_2, \dots\dots, \mathrm{t_m}\}$ και υπολόγισε τη σχετικότητα ενός εγγράφου $d \in D$ σε σχέση με το ερώτημα, με την πιθανότητα για d ότι παράγουν όρους για το τελευταίο ως εξής:

$$P(Q|d) = \prod_{t \epsilon Q} P(t|d)$$

Θεωρήθηκε αρχικά ότι κάθε χρήστης $u \, \epsilon \, U$ αντιπροσωπεύεται απευθείας από το σύνολο των ετικετών του T_u, με τις τελευταίες να αποτελούν ξεχωριστά έγγραφα ενώ το λεξιλόγιο των δεξιοτήτων S αντιπροσωπεύει ένα σύνολο ερωτημάτων (queries). Κατά συνέπεια η ερώτηση «Ποιο είναι το προφίλ εμπειρίας για έναν χρήστη u;», μπορεί να μεταφραστεί ως «Ποια είναι τα ερωτήματα που αντιστοιχούν σε ένα συγκεκριμένο έγγραφο;». Αντί επομένως να ορίζονται τα έγγραφα με ένα συγκεκριμένο ερώτημα, εφόσον είναι γνωστά (*queries*), επιτυγχάνεται το αντίστροφο που ισοδυναμεί, σε σχέση με το σύνολο ετικετών του χρήστη T_u, με βαθμονόμηση όλων των δεξιοτήτων στο λεξιλόγιο S.

Ως επόμενο βήμα δημιούργησε ένα **μοντέλο** (Scoring Model) που αντιπροσωπεύει τη συνάφεια ανάμεσα σε μια δεξιότητα και ένα σύνολο ετικετών σήμανσης. Απόρροια αυτής της ενέργειας ήταν η κατασκευή μιας **λειτουργίας βαθμονόμησης** (Scoring Function) που θα λαμβάνει ως όρισμα μια δεξιότητα και ένα σύνολο ετικετών, επιστρέφοντας αμέσως μετά ένα αποτέλεσμα ανάλογο με την πιθανότητα παραγωγής της δεξιότητας από το σύνολο ετικετών. Συνήθως, χρησιμοποιώντας γλωσσικά μοντέλα, η πιθανότητα ενός εγγράφου d να παράγει έναν μεμονωμένο όρο t ορίζεται:

$P(t|d) = \frac{tf(t,d)}{N(d)}$, όπου $tf(t,d)$ είναι ο αριθμός των φορών t που εμφανίζεται σε ένα έγγραφο d και $N(d)$ ο συνολικός αριθμός των όρων στο d. Σε αυτά τα γλωσσικά μοντέλα, εάν ένας όρος δεν εμφανίζεται σε ένα έγγραφο, υπολογίζεται η πιθανότητα εμφάνισης του από το σύνολο των εγγράφων που χρησιμοποιούνται.

Ωστόσο, όπως απέδειξε η ίδια η ερευνήτρια στα αρχικά πειράματά της, η υπόθεση που αναφέρθηκε στην προηγούμενη παράγραφο δεν είναι σωστή. Ως εκ τούτου, θεώρησε ότι θα πρέπει η συγγένεια κάθε όρου να μετράται ανάλογα με το σύνολο των ετικετών του χρήστη και μιας ορισμένης δεξιότητας. Γι' αυτόν ακριβώς τον λόγο η ερευνήτρια εκμεταλλεύτηκε το γεγονός ότι μερικοί χρήστες όρισαν οι ίδιοι τα προφίλ τεχνογνωσίας, χρησιμοποιώντας τη συγκεκριμένη πληροφορία για τον υπολογισμό της βασικής πιθανότητας της παρατήρησης μιας δεξιότητας s ϵ S, για μια συγκεκριμένη ετικέτα t ϵ T_u ως:

$$P(s|t) = \frac{P(t \cap s)}{P(t)} \quad (1), \text{ όπου,}$$

$P(t \cap s)$ θεωρείται ο αριθμός των χρηστών που έχουν μια δεξιότητα s και χρησιμοποιούν μια ετικέτα t, $\{ u \in Us | u -> t \Lambda u -> s \}$, και $P(t)$ αντιπροσωπεύει το σύνολο αυτών που χρησιμοποίησαν μια ετικέτα t, $\{ u \in Us | u -> t \}$. Η προαναφερθείσα πιθανότητα αποτελεί το βασικό δομικό στοιχείο για τη λειτουργία βαθμονόμησης (scoring function) και η ερευνήτρια διαισθητικά εκτίμησε ότι η συγκεκριμένη εξίσωση φανερώνει τον τρόπο με τον οποίο μια ετικέτα t μπορεί να προσφέρει υποστήριξη σε μια ορισμένη δεξιότητα. Εφόσον υπάρχει μια μέθοδος μέτρησης των ατομικών πιθανοτήτων, η ερευνήτρια επέκτεινε την εξίσωση (1) με την **προσθήκη του συνόλου των ετικετών σήμανσης** ως:

$$Score(s, T_u) = 1 - \prod_{t \in T_u} \big(1 - P(s|t)\big) \quad (2)$$

Σύμφωνα με αυτό, κάθε ετικέτα σήμανσης στο T_u προσφέρει υποστήριξη στη δεξιότητα s (δεν μετρώνται ξεχωριστά δηλαδή οι ατομικές ετικέτες). Επειδή αυτό μπορεί να οδηγήσει σε ανεπιθύμητα αποτελέσματα στη λειτουργία βαθμονόμησης, η ερευνήτρια το βελτίωσε

(scoring function) με την ενίσχυση των συνεισφορών από τις ετικέτες που προέρχονται από την ίδια σημασιολογική περιοχή, όπως η δεξιότητα s.

Γενικότερα, προκειμένου να αποδοθεί η σημασιολογική συνάφεια μιας ετικέτας και μιας δεξιότητας στην εξίσωση 2, θεωρείται ότι πολλοί όροι χρησιμοποιούνται και από τις δύο πλευρές (ετικέτα – δεξιότητα). Συνεπώς η ερευνήτρια βασίστηκε στο γεγονός ότι μπορεί να χρησιμοποιηθεί η μέτρηση **συνύπαρξης ετικετών** (tag co-occurrence), η οποία εδράζεται στην εξής παρατήρηση: Εάν δύο ετικέτες σήμανσης συνδέονται με τον ίδιο URL, έχουν σημασιολογική σχέση μεταξύ τους. Αυτή η σημασιολογική συγγένεια αποτελεί και μια επέκταση της 1ης εξίσωσης (σελ.82) και για τον υπολογισμό της είναι αναγκαίος ο αριθμός των URLs που έχουν σημανθεί και από τους δύο όρους (ανεξάρτητα από τους χρήστες), πάνω από το τμήμα των διευθύνσεων URL που έχουν σημανθεί με έναν από τους δύο. Αυτή η δεσμευμένη πιθανότητα (conditional probability) δηλώνεται ως $\mathrm{P_{sem}}(\mathrm{t_1}|t_2)$, και χρησιμοποιείται ο συμβολισμός $(\mathrm{t_1} = \mathrm{s}|\mathrm{t_2} = \mathrm{t})$. Επομένως, η ερευνήτρια βασισμένη στα προαναφερθέντα τροποποίησε την εξίσωση ως εξής :

$$Score(s, T_u) = 1 - \prod_{t \in T_u} 1 - P(s|t))^{Psem(t1 = s|t2 = t)} \quad (3)$$

Στη συγκεκριμένη εξίσωση χρησιμοποιείται ουσιαστικά η σημασιολογική συγγένεια των όρων που αντιπροσωπεύουν μια δεξιότητα και μια ετικέτα, με σκοπό να μετρηθεί η βαρύτητα της υποστήριξης που δίνεται σε μια δεξιότητα από τις ετικέτες σήμανσης που προέρχονται από το ίδιο σημασιολογικό πεδίο γνώσης με τη δεξιότητα αυτή.

3.5.6.2.4 Τελικά Πειράματα

Η ερευνήτρια χρησιμοποίησε για τα τελικά πειράματά της ένα υποσύνολο από 1.200 χρήστες, οι οποίοι παρουσίαζαν τα παρακάτω χαρακτηριστικά: (i) Είχαν κάνει σήμανση σε τρείς τουλάχιστον URLs ($|\mathrm{B_u}| \geq 3$) και είχαν ένα σύνολο από τουλάχιστον 20 ετικέτες ($|T_u| \geq 20$) και (ii) δεν είχαν άδεια προφίλ εμπειρίας ($|S_u| \neq \emptyset$). Αρχικά τους χώρισε σε δύο ομάδες για τις οποίες οι 1.100 χρήστες αποτελούσαν το σύνολο εκμάθησης, πάνω στο οποίο

εκπαίδευσε το μοντέλο της, και το υπόλοιπο σύνολο (100) αφορούσε στο σύνολο ελέγχου των πειραμάτων της.

Κατά τη διάρκεια της πρώτης φάσης (training phase) υπολόγισε -για την ομαλή λειτουργία βαθμονόμησης (scoring function)- τις τιμές συνύπαρξης ανάμεσα στα ζεύγη δεξιότητα- ετικέτα και τη σημασιολογική συγγένεια ανάμεσά τους (semantic relatedness). Εφόσον υπολογίστηκαν οι τιμές αυτές, λήφθηκαν υπόψη μόνο οι όροι που εμφανίστηκαν πάνω από 20 φορές στο σύνολο δεδομένων. Αυτό πραγματοποιήθηκε προκειμένου να διασφαλιστεί ότι οι τιμές που υπολογίστηκαν για τη δεσμευμένη πιθανότητα (conditional probability) της σημασιολογικής συγγένειας μεταξύ των ετικετών, παρουσιάζουν κάποιο νόημα.

Στην συνέχεια, εφαρμόστηκε ο **αλγόριθμος 1** *(Εικόνα 14)* με σκοπό να υπολογιστεί για καθέναν από τους εκατό χρήστες του συνόλου ελέγχου μια λίστα κατάταξης δεξιοτήτων αλλά και για να συγκριθεί στη συνέχεια η συγκεκριμένη λίστα με το σύνολο αυτών (δεξιότητες) που ορίστηκαν με την παρέμβαση του χρήστη (manually). Επομένως για κάθε χρήστη u, επιλέχθηκαν όλες οι ετικέτες που χρησιμοποιήθηκαν από αυτόν T_u. Στη συνέχεια, επαναλήφθηκε η ίδια διαδικασία για το γενικό λεξιλόγιο δεξιοτήτων S, υπολογίζοντας το βαθμό κάθε μιας δεξιότητας και προσθέτοντάς την παράλληλα στην ενδεχόμενη λίστα L_{skills}, για τον χρήστη u. Κατά το επόμενο βήμα της μεθόδου μειώθηκε η λίστα μέσω του αλγορίθμου επιστρέφοντας τις top - k καταχωρήσεις της.

Algorithm 1 getTopKSkills(User u ∈ U, {U \ US}, k)

List<skill, score> *Lskills* = ∅;
Tu = {*t*/*u* → *t*};
S ← all skills;
for *s* ∈ *S* {
scores = *Score*(*s*, *Tu*);
Lskills.add(*s*, *scores*);
}
Lskills.sort();
return *topk*(*Lskills*)

Εικόνα 14: Κύρια βήματα αλγορίθμου 1

Ο πίνακας 3, που παρουσιάζεται παρακάτω, αναφέρεται στον μέσο όρο των τιμών ακρίβειας για τους 100 χρήστες (από το σύνολο του πειραματισμού της ερευνήτριας) για διαφορετικές τιμές των k. Για τον υπολογισμό τους (Precision @ k – k), χρησιμοποιήθηκαν τρία διαφορετικά μοντέλα βαθμονόμησης:

- **Baseline**, αποτελεί το αρχικό μοντέλο, με τη βαθμολόγησή του να βασίζεται στο ποσοστό δημοφιλίας των χρηστών.
- **Intermediate Scoring Model**, το ενδιάμεσο μοντέλο που λαμβάνει υπόψη μόνο την δεσμευμένη πιθανότητα παρατήρησης μιας δεξιότητας μέσω μιας δεδομένης ετικέτας t (Εξίσωση 2).
- **Final Scoring Model**, το οποίο λαμβάνει υπόψη τη σημασιολογική συγγένεια μεταξύ των ετικετών σήμανσης και των δεξιοτήτων (Εξίσωση 3).

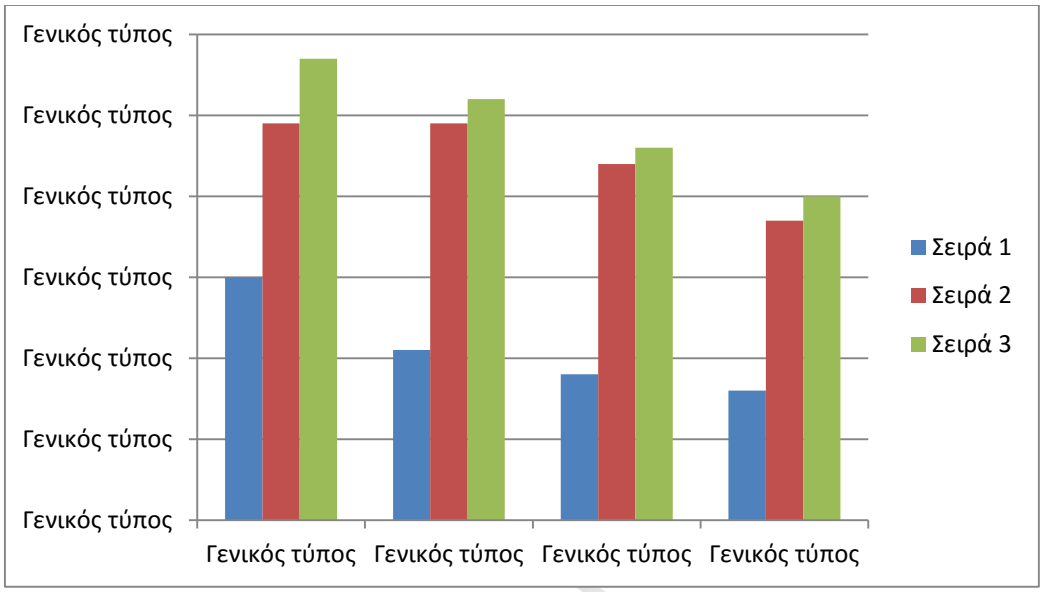
k	Precision@k		
	Baseline	**Intermediate**	**Final**
1	0,30	0,49	0,57
3	0,21	0,49	0,52
5	0,18	0,44	0,46
7	0,17	0,39	0,42
10	0,16	0,37	0,40

Πίνακας 3: Ποσοστά ακρίβειας από τρία διαφορετικά μοντέλα βαθμολόγησης.

Τα ποσοστά που προέρχονται από το δεύτερο μοντέλο (Intermediate Scoring Model), εμφανίζουν βελτίωση στην ακρίβεια σε σχέση με το πρώτο. Η παραπάνω διαπίστωση επιβεβαιώνει τη διαίσθηση της ερευνήτριας που αφορούσε στη χρησιμοποίηση συσχετίσεων μεταξύ των ετικετών και των δεξιοτήτων. Το αρχική ποσοστό ακρίβειας έχει την τιμή 0,49 σε k = 1 και όπως αναμενόταν τα υπόλοιπα ποσοστά μειώνονται με την αύξηση του k. Επίσης, στα αποτελέσματα της τελευταίας λειτουργίας βαθμολόγησης (*Final Scoring Model*) το ποσοστό ακρίβειας είναι μεγαλύτερο σε σχέση με τα δύο προηγούμενα, καθώς παρατηρείται η ύπαρξη της σημασιολογικής συγγένειας. Επιπλέον, το αρχικό ποσοστό ακρίβειας έχει την τιμή 0,57 σε k = 1 το οποίο μειώνεται με την αύξηση του k.

Όλες οι παραπάνω τιμές υπολογίζουν τα προφίλ των χρηστών που οι ίδιοι έχουν ορίσει στο σύστημα *IBMr*. Επειδή όμως ορισμένα από αυτά τα προφίλ δεν είναι ολοκληρωμένα, τα αποτελέσματα μπορούν να ερμηνευθούν ως σχετικά ακριβή. Η ερευνήτρια άλλωστε έκρινε ότι οι απόλυτες τιμές *Precision@k* μπορούν να υπολογιστούν με ανθρώπινη παρέμβαση στην αξιολόγηση με την προϋπόθεση ότι τα ποσοστά αυτά θα είναι μεγαλύτερα ή ίσα με αυτά της προτεινόμενης μεθόδου. Εν τέλει, στην *Εικόνα 15* που ακολουθεί, παρατίθενται όλα τα ποσοστά ακρίβειας και από τις τρείς λειτουργίες βαθμολόγησης. Πιο συγκεκριμένα, με μπλε

χρώμα αναπαριστώνται τα ποσοστά του αρχικού μοντέλου (*Baseline*), με κόκκινο αυτά του ενδιάμεσου (*Intermediate*) και με πράσινο χρώμα τα αντίστοιχα της τελικής μεθόδου (*Final*).



Εικόνα 15: Ποσοστό ακρίβειας για συνολικά εκατό χρήστες και από τις τρείς λειτουργίας βαθμολόγησης.

3.5.6.3 Σύνοψη και παρατηρήσεις

Η προσέγγιση της *Budura A.* αποτελεί μια γενική μέθοδο για τη δημιουργία προφίλ τεχνογνωσίας που βασίζονται στις ετικέτες σήμανσής. Τα αποτελέσματα που εξήγαγε από τις έρευνες που αναφέραμε είναι ιδιαίτερα ελπιδοφόρα καθώς, με την εξέλιξη της μεθόδου το ποσοστό ακρίβειας για την επίτευξη του στόχου της αυξήθηκε. Παρόλα αυτά, η ερευνήτρια ανέφερε ότι θα πρέπει να βελτιωθεί το μικρό μέγεθος των συνόλων των ετικετών που χρησιμοποίησε με τον συνδυασμό της μεθόδου της με αυτοματοποιημένες ετικέτες για τα έγγραφα. Κατά συνέπεια ο εμπλουτισμός των ετικετών σήμανσης, από τις οποίες αντλούνται οι δεξιότητες, δύναται να παράγει μεγαλύτερη ακρίβεια στα αποτελέσματα της προσέγγισης.

3.5.7 Το μοντέλο Χρήστης – Έννοια – Στιγμιότυπο των Οντολογιών

3.5.7.1 Γενικά

Η πλειοψηφία των προσεγγίσεων σύνδεσης υποστηρίζει ότι οι προσωπικές ενέργειες των χρηστών στα Σημασιολογικά Κοινωνικά Δίκτυα μπορούν να οδηγήσουν με επιτυχία στον εμπλουτισμό των οντολογιών. Ο ***Mika P.*** [21] επισήμανε ότι εντοπίζεται μεγάλη έλλειψη

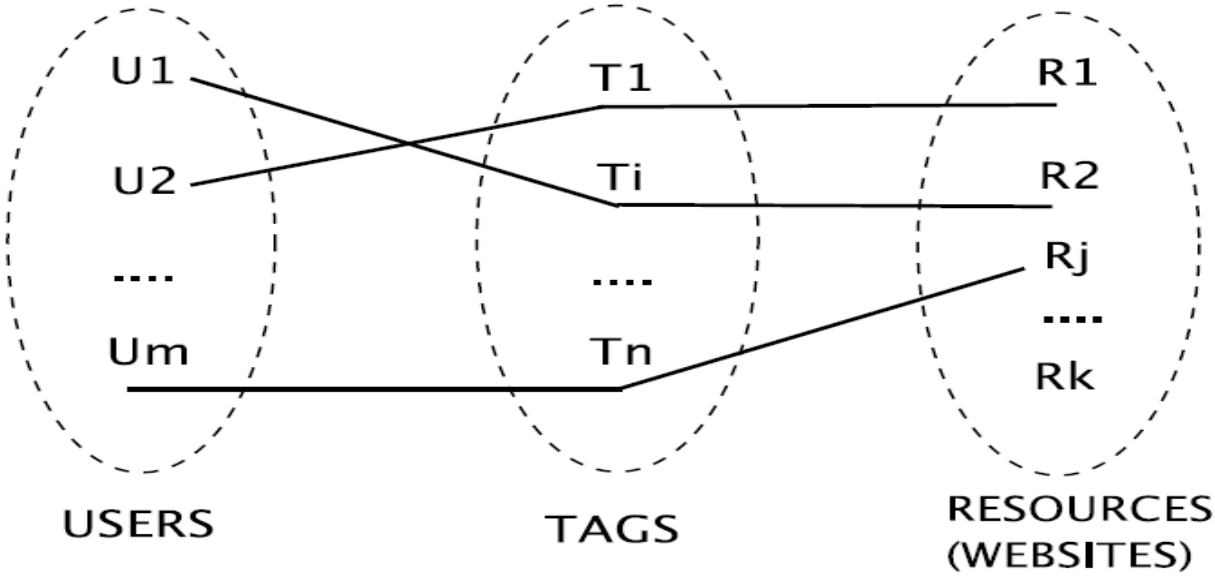
τέτοιων μοντέλων που θα μπορούσαν να εξηγήσουν λεπτομερώς τη διαδικασία αυτή. Ως εκ τούτου, πρότεινε ένα νέο, με την ονομασία *Actor - Concept - Instance* μοντέλο οντολογιών. Εμπνευσμένος από τους σημασιολογικούς μηχανισμούς σήμανσης παρουσίασε ορισμένα Κοινωνικά Δίκτυα (όπως το del.icio.us) με τη μορφή τριμερούς γραφήματος (χρήστης, έννοια, στιγμιότυπο). Με τον τρόπο αυτό πίστεψε ότι είναι δυνατόν να επιτευχθεί η επέκταση της παραδοσιακής έννοιας των οντολογιών (έννοιες, στιγμιότυπο) με κοινωνική διάσταση. Ουσιαστικά ο ερευνητής, με μια πρωτοποριακή δουλειά και έχοντας ως στόχο να λάβει ευρύτερες (broader) και στενότερες (narrower) σχέσεις μεταξύ των ετικετών σήμανσης, ερεύνησε τα Folksonomies ως «ελαφριές οντολογίες» (lightweight ontologies) που αναδύονται μέσω των αλληλεπιδράσεων στις κοινότητες.

3.5.7.2 Μεθοδολογία

3.5.7.2.1 Τριμερές Μοντέλο Οντολογιών

Αρχικά ο ερευνητής, για να μοντελοποιήσει τα δίκτυα των Folksonomies, παρουσίασε ένα κοινωνικό σύστημα σήμανσης ως τριμερές γράφημα (*Εικόνα 16*) με υπερακμές (hyperedges) επεκτείνοντας ουσιαστικά το παραδοσιακό διμερές μοντέλο πόρος - ετικέτα με την πρόσθεση των χρηστών. Πέραν από αυτούς, περιέλαβε τις ετικέτες (tags) και έναν πόρο που αντιπροσωπεύει ένα στιγμιότυπο σήμανσης (resource-instance). Το γεγονός αυτό είχε ως αποτέλεσμα τον διαχωρισμό του συνόλου των κορυφών σε τρία ασύνδετα σύνολα $A = \{a_1, \dots, a_k\}, C = \{c_1, \dots, c_l\}, I = \{i_1, \dots, i_m\}$, που αντιστοιχούν στο σύνολο των χρηστών (users), στο σύνολο των εννοιών (ετικέτες, λέξεις-κλειδιά) και στο αντίστοιχο των αντικειμένων (σελιδοδείκτες, φωτογραφίες) που σημαίνονται.

Στα κοινωνικά δίκτυα σήμανσης, οι χρήστες (users) έχουν τη δυνατότητα να επισημαίνουν αντικείμενα (objects) με έννοιες (concepts) με αποτέλεσμα την ύπαρξη κάποιων σχέσεων ανάμεσά τους. Στηριζόμενος στη συγκεκριμένη διαπίστωση, ο ερευνητής εκτίμησε ότι ένα Folksonomy (F) μπορεί να οριστεί ως ένα σύνολο σχολιασμών (annotations) όπου: $F \subseteq A \times C \times I$. Επιπλέον, η αντίστοιχη υπερακμή του (H) παριστάνεται ως εξής: $H(F) = < V, E >$, $\text{όπου } V = A \cup C \cup I \text{ και } E = \{\{a, c, i\} | (a, c, i) \in F\}$.



Εικόνα 16: Η δομή του τριμερούς γραφήματος (Tripartite graph)

Θεωρώντας όμως πως τα τριμερή γραφήματα και οι υπερακμές θα ήταν δύσκολο να εφαρμοστούν στην προσέγγισή του, ο ερευνητής χώρισε το δεύτερο προαναφερόμενο όρο (υπερακμή) σε τρία διμερή γραφήματα (bipartite graphs) αποτελούμενα από έννοιες και στιγμιότυπα (concepts – instances, γράφημα CI), χρήστες και έννοιες (actors – concepts, γράφημα AC), χρήστες και στιγμιότυπα (actors – instances, AI). Για παράδειγμα, το γράφημα AC ορίστηκε ως: $\boldsymbol{AC} = (A \times C, E_{ac})$, $\boldsymbol{E_{ac}} = \{(a,c)|\exists i \in I: (a,c,i) \in E\}, \boldsymbol{w}: E \to N, \forall e = (a,c) \in E_{ac}$, $\boldsymbol{w(e)} \coloneqq |\{i: (a,c,i) \in E\}|$. Με άλλα λόγια το AC συνδέει τους χρήστες με τις έννοιες, που έχουν χρησιμοποιήσει για τη σήμανση ενός τουλάχιστον αντικείμενου. Η βαρύτητα κάθε συνδέσμου (link) προκύπτει από τον αριθμό των φορών που έχει χρησιμοποιήσει ο χρήστης την έννοια σαν ετικέτα.

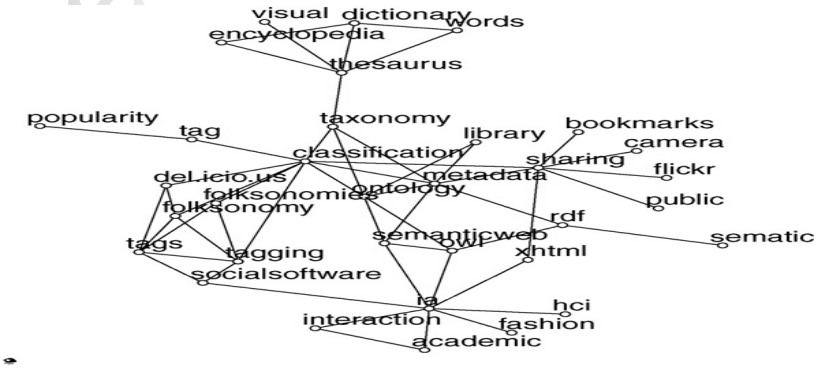
Πιο συγκεκριμένα, ο ερευνητής στηρίχθηκε στην άποψη ότι τα διμερή αυτά δίκτυα αντιστοιχούν σε δύο διαφορετικούς τρόπους προβολής της τριμερούς δομής των folksonomies:

- Ο πρώτος καθορίζει τις σχέσεις μεταξύ των ετικετών σήμανσης (tags) μέσω του προτύπου συνύπαρξης (co-occurrence) στους πόρους (resources) με τους οποίους αυτές συνδέονται. Η πρόβλεψη αυτή αναπαριστάνεται με έναν πίνακα που αποτελείται

από πλήθος R (resources) γραμμών και T (tags) στηλών. Κατά την συμπλήρωση των γραμμών του πίνακα, για κάθε πόρο r_i, υπολογίζεται ο αριθμός των φορών (weight) που κάθε ετικέτα t_j συνδέθηκε με r_i.

- Ο δεύτερος επέτρεπε την ομαδοποίηση κοινοτήτων κοινού ενδιαφέροντος, δηλαδή υποσυνόλων χρηστών που χρησιμοποιούσαν την ίδια ετικέτα για σήμανση. Αυτός ο τρόπος πρόβλεψης αναπαριστάνεται με έναν άλλο πίνακα που αποτελείται από πλήθος U (users) γραμμών και T (tags) στήλες. Κατά τη συμπλήρωση των γραμμών του πίνακα, για κάθε χρήστη u_i υπολογίζεται ο αριθμός των φορών που αυτός χρησιμοποίησε κάθε ετικέτα t_j.

Στη συνέχεια ο ερευνητής, εξήγαγε από τον πρώτο τρόπο προβολής ένα βεβαρυμένο γράφημα (weighted one-mode graph) συνδέοντας ετικέτες που βασίζονται σε συνδέσεις πόρων και από τον δεύτερο ένα αντίστοιχο γράφημα συνδέοντας όμως ετικέτες που βασίζονται στις συνδέσεις των χρηστών. Στην περίπτωση της ανθρωποκεντρικής (user-based) σύνδεσης των ετικετών σήμανσης αναδύθηκε η άποψη ότι για ένα ορισμένο ζεύγος από αυτές, η βαρύτητα (weight) -για το γράφημα- προκύπτει από τον αριθμό των χρηστών που χρησιμοποίησαν και τις δύο ετικέτες τουλάχιστον μία φορά. Η εικόνα που ακολουθεί (*Εικόνα 17*) αποτελεί ένα αντιπροσωπευτικό παράδειγμα τέτοιων γραφημάτων ετικετών (tag graphs) τα οποία προκύπτουν από τις αντίστοιχες ετικέτες στο σύστημα del.icio.us. Στο συγκεκριμένο γράφημα, ένας σύνδεσμος εξετάζεται από τον ερευνητή ανάμεσα σε ένα ζεύγος (ετικετών) εφόσον η βαρύτητα (weight) υπερβαίνει ένα όριο.



Εικόνα 17: Ετικέτες σήμανσης από το σύστημα del.icio.us βάσει των συνδέσεων χρηστών.

Συνοψίζοντας, το γράφημα AC το υπαγωγικό δίκτυο δηλαδή των χρηστών και των εννοιών δύναται να αναπαρασταθεί σε δύο γραφήματα: ένα κοινωνικό δίκτυο βασισμένο στα κοινά στοιχεία των συνόλων των πόρων και μια «ελαφριά» οντολογία (lightweight ontology) των εννοιών που βασίζεται στα κοινά στοιχεία των συνόλων των κοινοτήτων. Σύμφωνα με όλα τα παραπάνω ο ερευνητής κατάφερε να επιτύχει τη σύνδεση των κοινωνικών δικτύων με τη σημασιολογία στο μοντέλο του καθώς το παραδοσιακό διμερές γράφημα πόρος - ετικέτα περιέχει όλες τις πληροφορίες για τη δημιουργία τέτοιων δικτύων.

Τα υπόλοιπα διμερή γραφήματα αναπαραστάθηκαν με παρόμοιο τρόπο. Το γράφημα CI (concept-instances) οδήγησε ειδικότερα σε ένα άλλο σημασιολογικό δίκτυο στο οποίο οι σύνδεσμοι (links) ανάμεσα στους όρους υπολογίζονται από τον αριθμό των στιγμιότυπων (instances) που είχαν σημανθεί και με τους δύο (όρους). Ο συγκεκριμένος τύπος σημασιολογικού δικτύου θα λέγαμε ότι μιμείται τη βασική μέθοδο που εφαρμόζεται γενικότερα στον τομέα της εξόρυξης πληροφορίας από κείμενο, κατά τον οποίο οι όροι συνδέονται μεταξύ τους συνήθως με τη συνύπαρξή τους σε έγγραφα.

Το γράφημα AI (actors-instances) τέλος, αποτέλεσε ένα ακόμη κοινωνικό δίκτυο χρηστών στο οποίο ο αριθμός βαρύτητας (weight) ενός ζεύγους υπολογίζεται από τον αριθμό των κοινών αντικειμένων που έχουν σημανθεί. Απόρροια αυτού είναι ένα επιπλέον δίκτυο στιγμιότυπων (instances) με τις συνδέσεις τους να παριστούν τον αριθμό των χρηστών που έχουν επισημάνει ένα συγκεκριμένο ζεύγος στιγμιότυπων.

3.5.7.2.2 Εμπλουτισμός Οντολογιών

Προτού ο ερευνητής εφαρμόσει το μοντέλο του σε συστήματα σήμανσης εξέτασε δύο ελαφριές οντολογίες για τη διαδικασία του σημασιολογικού εμπλουτισμού τους, βασιζόμενες σε κοινά στοιχεία κοινοτήτων (O_{ac}) και στιγμιότυπων (O_{ci}).

Αρχικώς, προέβαλλε την ιδέα ότι η ελαφριά οντολογία O_{ac} (community based), παρουσίαζε ορισμένες ιδιομορφίες στην αναπαράσταση της γνώσης. Γι’ αυτόν ακριβώς τον λόγο έκρινε ότι η συνειρμική του οντολογία που θα δημιουργηθεί, μπορεί να συγκριθεί με την **EAT** (Edinburgh Associative Thesaurus) [51], η οποία έχει ως βάση παραγόμενα εμπειρικά δεδομένα και όχι ένα σημασιολογικό δίκτυο, όπως το WordNet. Πιο συγκεκριμένα είναι ένας

θησαυρός (Thesaurus) στον οποίο η αναζήτηση επιτυγχάνεται με δύο τρόπους: Κάθε χρήστης έχει την δυνατότητα να εισάγει τη λέξη ερέθισμα για να βρει τις παραγόμενες λέξεις (από ερέθισμα σε απάντηση), ή μπορεί να κάνει το αντίστροφο (από απάντηση σε ερέθισμα).

Ο ερευνητής κατέληξε στα παρακάτω συμπεράσματα:

- Η συνειρμική του οντολογία ήταν παρόμοια με την EAT, καθώς οι αριθμοί βαρύτητας των συνδέσμων ανάμεσα στους όρους εκφράζονται από τον αντίστοιχο αριθμό των χρηστών που χρησιμοποιούσαν και τους δύο όρους για την επισήμανση.
- Ενώ η συλλογή EAT ζητά με σαφήνεια από τους χρήστες να δημιουργήσουν συνδέσμους μεταξύ των εννοιών στην οντολογία συμπεραίνονται σύνδεσμοι απλά παρατηρώντας τη συμπεριφορά της σήμανσης.
- Τα αποτέλεσμα και των δύο μεθόδων (EAT, συνειρμική οντολογία ερευνητή) έχουν μια πάρα πολλή σημαντική ιδιότητα και συγκεκριμένα ότι το αποτέλεσμά τους εξαρτάται από την κοινότητα των ανθρώπων που παίρνουν μέρος στην πειραματική διαδικασία. Η συλλογική νοοτροπία επομένως, διαμορφώνεται από τον γνωστό νόμο σχηματισμού κοινότητας, ότι δηλαδή η αλληλεπίδραση δημιουργεί ομοιότητα, όπως και το αντίστροφο.
- Η EAT είναι μια αναδυόμενη οντολογία με βάση εμπειρικά δεδομένα όπως ακριβώς όλες οι παραγόμενες γραφικές παραστάσεις που σχετίζονται με την εξέλιξη του μοντέλου του.
- Η δομή και των δύο είναι ιδιαίτερα απλή, παρουσιάζοντας ένα σημαντικό μειονέκτημα στην ετερογένεια των όρων. Η ετερογένεια αφορά σε συγκεκριμένους ή γενικούς όρους, οι οποίοι δεν καθορίζουν με σαφήνεια το περιεχόμενο (στιγμιότυπα όπως Christos).

Επηρεαζόμενος από την προαναφερθείσα παρατήρηση ο ερευνητής ανέφερε ότι ο εμπλουτισμός των οντολογιών δύναται να επιτευχθεί με τις παρακάτω διαδικασίες:

- Θα πρέπει αρχικά να διακριθούν οι σαφείς και ασαφείς όροι. Αυτό πραγματοποιείται με τον υπολογισμό του συντελεστή ομαδοποίησης (clustering

coefficient) ή της μεταξύ τους ομοιότητας (betweenness centrality). Οι παραπάνω ενέργειες προσφέρονται από πακέτα ανάλυσης όπως το Pajek [52] (το οποίο χρησιμοποιεί ο ερευνητής στην εφαρμογή της μεθόδου του, Κεφάλαιο 3.5.7.2.3) και το UCINET [53].

- Μέσω των ευρύτερων ή στενότερων (broader/narrower) σχέσεων των όρων εξάγεται η ιεραρχία τους που βασίζεται στις σχέσεις υπο-κοινότητας. Ο ερευνητής προέβαλλε την άποψη ότι οι σχέσεις διαφέρουν κατά την ανάλυση είτε της οντολογίας O_{ci} είτε της O_{ac}. Στην πρώτη περίπτωση (O_{ci}) όλα τα (ή τα περισσότερα) αντικείμενα που υπάγονται στον στενότερο όρο (narrower) εμφανίζονται υπό τον ευρύτερο όρο (broader). Στη δεύτερη περίπτωση O_{ac} όλα τα πρόσωπα που συνδέονται με στενότερο όρο (narrower) συνδέονται παράλληλα και με τον ευρύτερο όρο (broader).

3.5.7.2.3 Αξιολόγηση Αποτελεσμάτων

Γενικότερα η αξιολόγηση των αποτελεσμάτων της εκμάθησης ή της αντιστοίχησης οντολογιών αποτελεί μια αρκετά επίπονη διαδικασία, καθότι απαιτεί τη συμβολή της κοινότητας ή των κοινοτήτων, της οποίας ή των οποίων οι έννοιες είναι ήδη γνωστές. Για τον λόγο αυτό, η αξιολόγηση των αποτελεσμάτων του μοντέλου του *Mika P.* διενεργήθηκε και ολοκληρώθηκε με τη συμβολή 61 ερευνητών. Από αυτούς, η πλειοψηφία ήταν μέλη του διεθνούς συνέδριου σημασιολογικού ιστού (International Semantic Web Conference-ISWE), ενώ οι υπόλοιποι, αποτελούσαν τον πυρήνα της κοινότητας. Η αξιολόγηση αυτή, σχετιζόταν ουσιαστικά με την απάντηση μια συγκεκριμένης ερώτησης:

Όσον αφορά στις συσχετίσεις μεταξύ των εννοιών ποια οντολογία του Σημασιολογικού Ιστού, που σχετίζεται με αυτές, θεωρείται ακριβέστερη;

Τα αποτελέσματα της αξιολόγησης παρατίθενται στον πίνακα που ακολουθεί:

	N	O_{ac}	O_{ci}	Ποσοστό (%)	Ακρίβεια
Όλοι	30	22	8	73.3	0.0055
ISWC	23	18	5	78.3	0.0040
ISWC πυρήνας	15	13	2	86.7	0.0032

Πίνακας 4: Αποτελέσματα έρευνας αξιολόγησης

Βασιζόμενος στα παραπάνω δεδομένα, ο ερευνητής κατέληξε στα εξής συμπεράσματα:

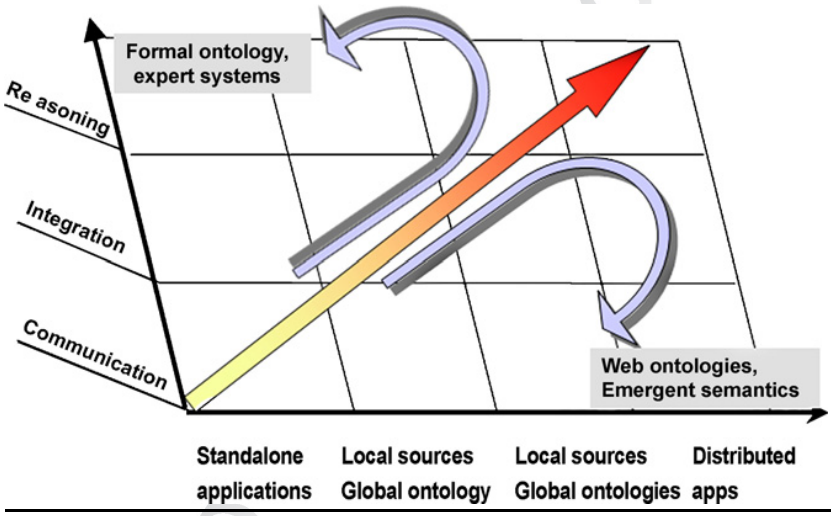
- Οι ερευνητές θεώρησαν ότι το δίκτυο O_{ac} είναι πιο ακριβές στην αναπαράσταση των συσχετίσεων ανάμεσα στις έννοιες συγκριτικά με το O_{ci}.
- Συγκεκριμένα, εκτιμάται ως πιο ακριβές το O_{ac} από αυτούς, των οποίων τα ονόματα χρησιμοποιήθηκαν κατά την διαδικασία εξαγωγής (extraction process).
- Τέλος, το δίκτυο O_{ac} αντικατοπτρίζει καλύτερα την εννοιολογική αναπαράσταση από τους χρήστες που βρίσκονται πιο κοντά στον πυρήνα της κοινότητας.

Μέσω των συγκεκριμένων ευρημάτων που προέκυψαν από την αξιολόγηση του μοντέλου Actor - Concept - Instance, επιβεβαιώθηκε η άποψη του Mika P. ότι το δίκτυο O_{ac} υποδηλώνει ουσιαστικά την εννοιολογική αναπαράσταση των χρηστών που εμπλέκονται ενεργά στην έρευνα του σημασιολογικού ιστού.

3.5.7.3 Τελικά Συμπεράσματα

Οι οντολογίες και τα Folksonomies έχουν απασχολήσει κατά καιρούς πολλούς ερευνητές. Οι διαφορετικοί υποστηρικτές των δύο όρων αποζητούν ουσιαστικά μια συγκεκριμένη επιλογή για την σύνδεση της Σημασιολογίας στο διαδίκτυο. Σε αντίθεση με την πλειοψηφία των ερευνητών ο Mika P. θεωρεί ότι τα Folksonomies είναι από μόνα τους οντολογίες. Για να υποστηρίξει μάλιστα αυτή την άποψη βασίστηκε στον ορισμό της οντολογίας, όπως αυτός διατυπώθηκε από από τους Elst L. και Abecker A. [55], οι οποίοι ανέφεραν ότι η απόκτηση της γνώσης παρουσιάζει μεγάλη δυσκολία εάν είναι τυπική, δυναμική και ευρέως διαδεδομένη. Σύμφωνα με τον Mika P. η κατάταξη αυτή μπορεί να

απλοποιηθεί σε δύο διαστάσεις (*Εικόνα 18*), όπου οι εφαρμογές που βασίζονται σε οντολογίες δύναται να ταξινομηθούν ανάλογα με το επίπεδο εκφραστικότητας (Semantic dimension) και κατανομής (Web dimension). Εκτελώντας διεργασίες που εδράζονται στα αποτελέσματα υπαρκτών εφαρμογών, ο ερευνητής διαπίστωσε ότι είναι δύσκολο να προχωρήσουν και οι δύο διαστάσεις ταυτόχρονα. Συμπέρανε μάλιστα ότι η επίσημη (formal) πληροφορία απαιτεί κάποια δέσμευση εν αντιθέσει με τις οντολογίες που είναι ευρέως διαμοιραζόμενες, όπως η «ελαφριά» οντολογία FOAF. Στη διαπίστωση αυτή οδηγήθηκε λαμβάνοντας υπόψη κυρίως τη σταθερότητα (stability), καθώς βασίζονται σε αυτή μεγάλα δυναμικά (dynamic) δίκτυα (Παγκόσμιος Ιστός), αποτελούμενα από επίσημες οντολογίες.



Εικόνα 18: Η ταξινόμηση σύμφωνα με τους ερευνητές Mika P. και Akkerman H.

Με βάση τη θεωρία του υποστήριξε πως τα Folksonomies:

- Είναι «**ελαφριές»** (lightweight) και όχι επίσημες (οντολογίες) αναπαραστάσεις πληροφοριών, γεγονός που οδηγεί πολλές φορές στην σε μεγάλο βαθμό παρεξήγησή τους από ορισμένους χρήστες και ερευνητές. Οι μέθοδοί για την εξόρυξη της σημασιολογίας προέρχονται από την ανάλυση του διαδικτύου. Με άλλα λόγια χαρακτηρίζονται ως «**ελαφριές οντολογίες**» και συνδέουν το νόημα των όρων.
- Είναι **δυναμικές** (dynamic), καθώς η πλειοψηφία των όρων περιλαμβάνει ένα καλά ορισμένο νόημα που παραμένει σταθερό με την πάροδο του χρόνου. Επομένως,

εκτιμάται ότι περιέχουν μια σταθερή «ραχοκοκαλιά», εάν εξεταστούν για μεγάλα χρονικά διαστήματα.

- Όσον αφορά στον διαμοιρασμό, **έχουν περιορισμένη εμβέλεια** (limited scope) σε σχέση με τις παραδοσιακές οντολογίες, με τη δόμησή τους να αποτελεί ένα ιδιαίτερα δύσκολο εγχείρημα.

Εν τέλει, κατέληξε στο συμπέρασμα ότι ένα folksonomy, με όλα τα προαναφερθέντα χαρακτηριστικά, είναι ένα διαφορετικό είδος οντολογίας συγκριτικά με τα πιο επίσημα και σταθερά λεξιλόγια που υποστηρίζονται από την κοινότητα του Σημασιολογικού Ιστού στο διαδίκτυο. Αν και δεν βασίζονται σε λογικές προσεγγίσεις, ο ερευνητής υποστήριξε την άποψη ότι δύναται να προκαλέσουν σημαντικό αντίκτυπο στην υλοποίηση του αρχικού οράματος του Σημασιολογικού Ιστού, ως επέκταση του Παγκοσμίου Ιστού.

Κεφάλαιο 4

4.1 Συνοπτική Παρουσίαση

Η σύνδεση της Σημασιολογικής πληροφορίας με τα Κοινωνικά Δίκτυα έχει απασχολήσει κατά καιρούς πάρα πολλούς ερευνητές. Η πλειοψηφία των προσπαθειών για την ανάπτυξη των μεθόδων και των εφαρμογών αναφέρθηκε στο Κεφάλαιο 3.4, με τις σημαντικότερες από αυτές να αναλύονται λεπτομερώς στο αμέσως επόμενο Κεφάλαιο (3.5). Οι σημαντικότερες λοιπόν, που επιλέχθηκαν κυρίως με γνώμονα την αναγνωρισιμότητα, την πρωτοπορία και την αποτελεσματικότητά τους, μπορούν να συγκριθούν με βάση ορισμένα βασικά χαρακτηριστικά που αφορούν:

1. στο είδος της έρευνάς τους, στο εάν δηλαδή είναι ποσοτική ή ποιοτική,
2. στα διαφορετικά συστήματα σήμανσης στα οποία εξετάζουν οι ερευνητές τη μέθοδό τους και τους εξωτερικούς πόρους που πιθανόν χρησιμοποιούν, οι οποίοι σχετίζονται κυρίως με εργαλεία ανάλυσης (WordNet).
3. στην παρέμβαση ή μη (αυτοματοποιημένες) του χρήστη στη διαδικασία του Σημασιολογικού εμπλουτισμού των folksonomies και
4. στις αναδυόμενες σημασιολογικές σχέσεις των ετικετών σήμανσης για κάθε μεθοδολογία (υπαγωγικές, ομαδοποίηση).

4.1.1 Είδος Έρευνας

Το πρώτο χαρακτηριστικό στο οποίο εδράζεται η σύγκρισή μας, αφορά στο είδος έρευνας κάθε προσέγγισης. Όπως γνωρίζουμε από την βιβλιογραφία [56], οι έρευνες διακρίνονται σε ποσοτικές και ποιοτικές, οι οποίες διαφέρουν ανάλογα με το είδος των πληροφοριών και με τον τρόπο που τις συλλέγουν οι ερευνητές. Πιο συγκεκριμένα, οι ποσοτικές έρευνες επικεντρώνονται σε αριθμητικά δεδομένα και σε στατιστικές συγκρίσεις και στη μέτρηση των θεωρητικών εννοιών και απόψεων, κυρίως μέσω εργαλείων, προκειμένου να εξαχθούν τα επιθυμητά αποτελέσματα.

Από την άλλη μεριά, οι ποιοτικές έρευνες βασίζονται εν μέρει στις ποσοτικές, δίνοντας έμφαση ωστόσο στην εξέλιξη των διαδικασιών μέσα από τις οποίες διαμορφώνονται συγκεκριμένες αντιλήψεις. Όσον αφορά στις σημαντικότερες προσεγγίσεις, η έρευνα του Gruber T. [13] δύναται να χαρακτηριστεί ως ποιοτική. Ο συγκεκριμένος ερευνητής ουσιαστικά επιθυμούσε να δημιουργήσει ένα πρότυπο σήμανσης, στηριζόμενο, εν μέρει, σε χαρακτηρίστηκα, που προκύπτον από την ανάλυση των ετικετών σήμανσης. Στον αντίποδα αυτής της προσπάθειας όλοι οι υπόλοιποι ερευνητές, στήριξαν τις μεθόδους τους είτε σε στατιστικές συγκρίσεις είτε σε αριθμητικά δεδομένα που προέκυπταν από τις εφαρμογές αυτών.

4.1.2 Χρήση Συστημάτων Σήμανσης και Εξωτερικών Πηγών

Τα Κοινωνικά Συστήματα Σήμανσης διαδραμάτισαν ιδιαίτερα σημαντικό ρόλο στην πλειοψηφία των προσεγγίσεων. Χρησιμοποιήθηκαν είτε πριν από την εκάστοτε μεθοδολογία για πρότερη έρευνα, παρατήρηση και πιθανή διόρθωση ενός ήδη υπάρχοντος μοντέλου (Schmitz P. [24]), είτε στο τελευταίο στάδιό της για την αξιολόγηση των αποτελεσμάτων. Εκτός του Gruber T. [13], όλοι οι υπόλοιποι ερευνητές ασχολήθηκαν είτε με δημοφιλή ελεύθερα Κοινωνικά Συστήματα Σήμανσης είτε με λιγότερο γνωστά και διαδεδομένα (Ereteo G. / Gandon F. [27], Budura A. [23]).

Πιο συγκεκριμένα, οι Schmitz P. [24], Αγγελέτου Σ. [6] και Specia L. / Motta E. [22] επικέντρωσαν τις εργασίες τους στο Flickr για την εφαρμογή του εκάστοτε μοντέλου τους στο τέλος της μεθοδολογίας τους. Ο πρώτος, εξέτασε το υπαγωγικό του μοντέλο για να αξιοποιήσει τις στατιστικές γλωσσικές τεχνικές επεξεργασίας προκειμένου να δημιουργήσει οντολογίες μέσα από τη βάση δεδομένων του Flickr. Η Αγγελέτου Σ. [6] χρησιμοποίησε το λογισμικό ομαδοποίησης του συγκεκριμένου συστήματος στα αρχικά της πειράματα, για να οδηγηθεί σε χρήσιμα συμπεράσματα στον εμπλουτισμό των folksonomies που θα την βοηθούσαν στην ολοκλήρωση του εργαλείου της FLOR. Οι Specia L. / Motta E. [22] θεώρησαν ότι η κατανόηση της υποκειμενικής γνώσης είναι πολλή σημαντική για την μέθοδό τους, που αφορούσε στην ενσωμάτωση των folksonomies με τη βοήθεια του σημασιολογικού ιστού. Γι' αυτόν ακριβώς τον λόγο ερεύνησαν και αυτοί με την σειρά τους τα δεδομένα του

συγκεκριμένου συστήματος, όπως και αυτά του κοινωνικού δικτύου σελιδοσήμανσης del.icio.us. Με το τελευταίο ασχολήθηκε εν μέρει η Αγγελέτου Σ. [6] στα πρότερα πειράματά της – καταλήγοντας όμως στο Flickr-, σε αντίθεση με τον Mika P. [21] που χρησιμοποίησε το υποσύνολο ετικετών σήμανσής του για να δημιουργήσει γραφήματα σχετικότητας. Ο συγκεκριμένος ερευνητής εξέτασε επιπλέον και το κοινωνικό δίκτυο σημασιολογικού ιστού Flink για την εξαγωγή και τη δημιουργία των οντολογιών O_{ac} και O_{ci}. Τέλος, η Budura A. [23] πειραματίστηκε πάνω στην ολοκλήρωση της μεθοδολογίας της με τα σύνολα ετικετών σήμανσης των Dogear και IBMr, ενώ οι Ereteo G. / Gandon F. [27] εφάρμοσαν και αξιολόγησαν τον αλγόριθμό τους SemTagP στο κοινωνικό δίκτυο ADEME.

Αρκετοί ερευνητές επιπροσθέτως χρησιμοποίησαν πολλούς διαδικτυακούς εξωτερικούς πόρους όπως για παράδειγμα το WordNet, το Google και την Wikipedia. Οι ερευνητές Specia L./Motta E. [22] αξιοποίησαν πόρους όπως το διαδικτυακό λεξικό WordNet (εύρεση εννοιών), την Wikipedia (ορθογραφικά λάθη), τις μηχανές αναζήτησης Google (σωστή διατύπωση όρου) και Swoogle (πηγή οντολογιών) για να συνάξουν σημασιολογικές σχέσεις μεταξύ των ετικετών σήμανσης.

Το σύστημα FLOR της Αγγελέτου Σ. [6] διαφέρει με αυτό της προσέγγισης των Specia L. / Motta E. [22], καθώς ξεπερνά τη φάση ομαδοποίησης όμοιων ετικετών και ενσωματώνει μια φάση αποσαφήνισης αυτών με τη στήριξη του WordNet. Στην συγκεκριμένη ($2^{η}$ φάση FLOR), ορίζεται κάθε ετικέτα με βάση την ομάδα στην οποία ανήκει και εξάγονται όλα τα σχετικά συνώνυμα και υπερώνυμα για τον εμπλουτισμό της. Τέλος ο Mika P. [21] θεώρησε πως η διάκριση των σαφών και ασαφών εννοιών από τις αναδυόμενες οντολογίες του δύναται να επιτευχθεί με τον υπολογισμό του συντελεστή ομαδοποίησης, της μεταξύ τους (έννοιες) ομοιότητας ή των περιορισμών των όρων στο δίκτυο με τα πακέτα ανάλυσης Pajek και UCINET.

4.1.3 Ανθρωποκεντρικές και Αυτοματοποιημένες Διαδικασίες

Η κεντρική ιδέα όλων των προσεγγίσεων αφορούν στη σύνδεση του Σημασιολογικού Ιστού με τα Κοινωνικά Δίκτυα, με άλλα λόγια δηλαδή στον σημασιολογικό εμπλουτισμό των folksonomies. Πολλές από αυτές στηρίζονται στην παρέμβαση του χρήστη (ανθρωποκεντρικές), ενώ ορισμένες προτείνουν αυτοματοποιημένες (είτε σε ορισμένο βαθμό)

διαδικασίες για τον εμπλουτισμό. Η παραπάνω διαπίστωση αποτελεί και το βασικό κριτήριο για την σύγκριση που ακολουθεί.

Ο Schmitz P. [24] υιοθέτησε την ανθρωποκεντρική διαδικασία, καθώς τα δέντρα ομαδοποίησης των ετικετών σήμανσης που προκύπτουν από την εφαρμογή του υπαγωγικού του μοντέλου στο σύστημα Flickr, αξιολογούνται με την παρέμβαση του χρήστη. Με αυτό ακριβώς το κριτήριο ο ερευνητής θεώρησε πως το μοντέλο του μπορεί να αναπτυχθεί καλύτερα όχι σαν μια απλή αυτοματοποιημένη διαδικασία αλλά σαν ένα παραγωγικό εργαλείο με ενθουσιώδεις συντονιστές.

Από την άλλη πλευρά, η Αγγελέτου Σ. [6] προσπάθησε με τη μέθοδό της να ορίσει αυτόματα τις ετικέτες σήμανσης που ήδη υπάρχουν στα folksonomies. Μέσω κάποιων αρχικών πειραμάτων κατέληξε στο συμπέρασμα πως υπάρχει η δυνατότητα μιας τέτοιας διαδικασίας. Με τον τρόπο αυτό, κατάφερε να εμπλουτίσει τα σύνολα ετικετών ενός folksonomy, αυτοματοποιημένα, χωρίς την παραμικρή παρέμβαση του χρήστη. Κάτι τέτοιο επιτεύχθηκε με τη χρήστη του WordNet, των διαδικτυακών οντολογιών και των μεθόδων του εργαλείου της FLOR για τον λεκτικό διαχωρισμό, την αποσαφήνιση, τη σημασιολογική επέκταση και φυσικά για το σημασιολογικό εμπλουτισμό. Οι Specia L./Motta E. [22], προτείνουν επίσης, για τη σύνδεση των ετικετών σήμανσης με διαδικτυακές οντολογίες, μια αυτοματοποιημένη μέθοδο αποτελούμενη από τρείς φάσεις: α) την προ-επεξεργασία, β) την ομαδοποίηση και γ) την αναγνώριση έννοιας – σχέσης. Εφόσον πραγματοποιηθεί η ανάλυση συνύπαρξης, το σύστημα αναζητά αυτόματα κοινά στοιχεία για τις ετικέτες από τις οντολογίες. Σε περίπτωση επιτυχίας αναγνωρίζει τις έννοιες και τις ιδιότητές τους, για να τις εμπλουτίσει στη συνέχεια σε ομάδες με σημασιολογία.

Η Budura A. [23] πρότεινε μια νέα προσέγγιση για το πρόβλημα της εξόρυξης της τεχνογνωσίας. Ανέπτυξε μια αυτοματοποιημένη μέθοδο πιθανοτήτων για την οικοδόμηση προφίλ τεχνογνωσίας για κάθε χρήστη βασιζόμενο στις ετικέτες σήμανσής του (scoring model). Αντίθετα στις παραδοσιακές προσεγγίσεις, κάθε προφίλ παράγεται με αντίστοιχο τρόπο ως μια συλλογή όρων από τα έγγραφα του καθενός. Επιπλέον σε αυτά, το σύστημα εύρεσης τεχνογνωσίας (expert finding system) είναι αυτό που ερμηνεύει και αναγνωρίζει την

τεχνογνωσία κάθε χρήστη. Παρόλα αυτά η ερευνήτρια χρησιμοποίησε την προσέγγισή της με τέτοιο τρόπο ώστε να είναι άμεσα εφαρμόσιμη σε σύνολα ετικετών άλλων εφαρμογών πληροφορίας, όπως σε ορισμένα Κοινωνικά επιχειρησιακά συστήματα σήμανσης (Dogear και IBMr). Χρησιμοποίησε στα πειράματά της με αυτά τα συστήματα έναν αλγόριθμο για να υπολογίσει αυτόματα για καθέναν από τους χρήστες ενός ορισμένου συνόλου μια λίστα κατάταξης δεξιοτήτων.

Η προσέγγιση των ερευνητών Ereteo G. / Gandon F. [27] διαφέρει από τις υπόλοιπες, καθώς εμπλουτίζει σημασιολογικά τις ετικέτες σήμανσης με τον συνδυασμό αυτόματης διαδικασίας και των συνεισφορών των χρηστών (περιγραφές RDF με την βοήθεια δικτύου ADEME). Αυτή η μέθοδος ξεκινά με μια σύνθετη μέτρηση (string-based metrics) για να αποκαλύψει τρείς βασικούς τύπους σχέσεων ανάμεσα στις ετικέτες: tags: skos:related, skos:closeMatch και skos:narrower. Στη συνέχεια οι χρήστες έχουν τη δυνατότητα να αξιολογήσουν, να απορρίψουν ή να προτείνουν σημασιολογικές σχέσεις μέσω ενός εργαλείου πλοήγησης. Οι αναδυόμενες αδυναμίες λύνονται με τη συναίνεση ενός ορισμένου χρήστη. Η μέθοδος είναι επαναληπτική ωσότου επιτευχθεί ο στόχος.

Σε αντίθεση με τις παραπάνω προσεγγίσεις, οι ερευνητές Gruber T. [13] και Mika P. [21] κινούνται σε διαφορετική κατεύθυνση. Πιο συγκεκριμένα, ο Mika P. [21] συνέταξε μια οντολογία βασισμένη στην κοινότητα χρησιμοποιώντας το κοινωνικό εργαλείο del.icio.us και δημιουργώντας επιπλέον δύο ελαφριές οντολογίες (χρήστης – έννοια και έννοια - στιγμιότυπο). Ο στόχος του πειράματός του αφορούσε στο να δείξει πως οι οντολογίες μπορούν να σχηματισθούν με τη χρήση του περιεχομένου της κοινότητας στην οποία και δημιουργούνται. Με το ίδιο σκεπτικό ο Gruber T. [13] δημιούργησε την TagOntology, μια οντολογία για folksonomy.

4.1.4 Σχέσεις μεταξύ των Ετικετών Σήμανσης

Ένα ακόμα σημαντικό κριτήριο στο οποίο μπορεί να βασιστεί η σύγκριση των κυριότερων προσεγγίσεων, σχετίζεται με τις αναδυόμενες σημασιολογικές σχέσεις ανάμεσα στις ετικέτες σήμανσης. Οι συγκεκριμένες, μπορούν να χωριστούν σε τρείς υποκατηγορίες που αφορούν:

- στη μέτρηση της ομοιότητας μεταξύ των ετικετών, με άλλα λόγια δηλαδή στον βαθμό συνύπαρξής τους (co-occurrence),
- στην ύπαρξη πιθανόν υπαγωγικών σχέσεων (subsumption relations),
- στους διαφορετικούς τρόπους ομαδοποίησης που μπορεί να υπάρχουν ανάμεσα σε ισοδύναμες (παραλλαγές του ίδιου όρου) ή όμοιες ετικέτες.

Ο Schmitz P. [24] επηρεάστηκε από το στατιστικό μοντέλο υπαγωγής των Sanderson M. και Croft, B [34], όπου η ετικέτα X υπάγεται στην Y εάν P(x|y >= 0.8) and P(y|x < 1). Για να προκαλέσει μια ιεραρχία από Flickr - ετικέτες ο Schmitz P. [24] προσάρμοσε τη μέθοδο του, ενσωματώνοντας νέα στατιστικά όρια για να μετρήσει την ιδιομορφία των folksonomies (Κεφάλαιο 3.5.1.3).

Οι Specia L. / Motta E. [22] χρησιμοποίησαν στη μεθοδολογία τους εξωτερικές πηγές για την εύρεση του κατάλληλου εκπροσώπου σε μια ομάδα με ισοδύναμες ετικέτες. Εφάρμοσαν επιπλέον μια τεχνική ομαδοποίησης (2η φάση μεθόδου), όπου η κάθε ομάδα περιέχει μια ετικέτα ως βάση (seed tag) και στη συνέχεια προστίθεται μια άλλη, εφόσον ο βαθμός ομοιότητας είναι πάνω από το ορισμένο όριο με όλες τις άλλες ετικέτες της ομάδας. Χρησιμοποίησαν επίσης ευριστικές στρατηγικές λύσης προβλήματος (heuristics), για να συγχωνεύσουν ομάδες που παρουσίαζαν μεγάλη ομοιότητα (μεγάλο ποσοστό ισοδύναμων ετικετών).

Ο Gruber T. [13] θεώρησε πως υπάρχει δυνατότητα δημιουργίας μια κοινής αντίληψης για την ομαδοποίηση των ετικετών σήμανσης, που μπορεί να καλύψει την πλειοψηφία των αδυναμιών που εμφανίζονται σε πολλά συστήματα. Γι’ αυτόν ακριβώς τον λόγο πρότεινε δύο τρόπους αντιμετώπισης τους συγκεκριμένου προβλήματος που αφορούν :

- στην αναπαράσταση μιας λειτουργίας από τα ονόματα στις ετικέτες, δηλαδή σε μια αναπαράσταση που θα ισοδυναμεί με άλλα λόγια με όμοιες σημάνσεις. Για παράδειγμα "town of Argos" = 1η ετικέτα, "Town Of Argos" = 2η ετικέτα και "TownofArgos" = 3η ετικέτα, τότε μπορεί να ισχυριστεί κάποιος ότι 1η ετικέτα=2η ετικέτα, 2η ετικέτα=3η ετικέτα κτλ.

- στην χρήση κανονικού ονόματος για κάθε ετικέτα της μορφής cname(tag)="string".

Η Αγγελέτου Σ. [6] θεώρησε, μέσω της μεθοδολογίας της, πως η ανάκτηση περιεχομένου θα μπορούσε περεταίρω να βελτιωθεί, κάνοντας σαφείς τις σχέσεις ανάμεσα στις ετικέτες σήμανσης. Με αυτή την ενέργεια έκρινε πως οι σχετιζόμενες με αυτές, θα μπορούσαν να παρέχουν ένα βασικό τρόπο αναζήτησης ή ανάκτησης. Όπως αναφέρθηκε και παραπάνω (4.1.2), το σύστημά της δεν επιφέρει κάποια ομαδοποίηση ανάμεσα στις ετικέτες αλλά ενσωματώνει μια φάση ορισμού και αποσαφήνισής τους, μέσω ορισμένων εξωτερικών πηγών. Πιο συγκεκριμένα, εφόσον εξαιρεθούν όλες οι ενδεχόμενες νοηματικές (1η φάση) ετικέτες, εμπλουτίζεται κάθε μια από τις εναπομένουσες με το λεξικό WordNet (2η φάση). Ο εμπλουτισμός αυτός περιλαμβάνει συνώνυμα και υπερώνυμα για κάθε ετικέτα, σχηματίζοντας κατά κάποιο τρόπο υπαγωγικές σχέσεις. Κατά την τελευταία φάση υπολογίζεται ο βαθμός ομοιότητας των οντολογιών τους με τη μέτρηση Levenshtein, για να εντοπιστούν και να εμπλουτιστούν σημασιολογικά αυτές που συνδέονται με τις ετικέτες.

Οι Ereteo G. / Gandon F. [27] θεώρησαν πως υπάρχει η δυνατότητα να συμπεράνουμε τις σημασιολογικές σχέσεις μεταξύ των ετικετών προκειμένου να εμπλουτιστεί ένα folksonomy με ελαφριά σημασιολογία. Για τον λόγο αυτό, χρησιμοποιώντας αρκετές μετρήσεις (string-based), ομαδοποίησαν τις ετικέτες με βάση τρείς σχέσεις: α) skos:related, β) skos:closeMatch και γ) skos:narrower.

Ο Mika P. [21] εφάρμοσε την ανάλυση των κοινωνικών δικτύων σε διαφορετικές προβολές της τριμερούς δομής των folksonomies. Συγκέντρωσε παρόμοιες κοινότητες ενδιαφέροντος (χρήστες με κοινές ετικέτες σήμανσης) έτσι ώστε να αντλήσει υπαγωγικές ιδιότητες ανάμεσα στις ετικέτες με την ενσωμάτωση αυτών.

Τέλος η Budura A. [23], χρησιμοποίησε μια εξίσωση (εξίσωση 3 σελ. 83) για να μετρηθεί η βαρύτητα της υποστήριξης που δίνεται σε μια δεξιότητα από τις ετικέτες σήμανσης που προέρχονται από το ίδιο σημασιολογικό πεδίο γνώσης με αυτήν (δεξιότητα).

4.2 Χαρακτηριστικά Σημαντικότερων Μεθόδων και Εφαρμογών Σύνδεσης

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Schmitz P. [24]		X	**Υπό Συνθήκη Πιθανότητα**	X		X			X
Specia L./Motta E. [22]		X	**Κατανομή (tag context)**		X		X	X	X
Gruber T. [13]	X				X				
Angeletou S. [6]		X	**Λεξιλογικές Αναπαραστάσεις**	X			X	X	X
Ereteo G. / Gandon F. [27]		X	**Μετρήσεις string-based**		X	X	X		X
Budura A. [23]		X	**Υπό Συνθήκη Πιθανότητα**				X		X
Mika P. [21]		X	**Βασιζόμενη στο Διαδίκτυο**	X				X	X

(1) αναφέρεται σε ποιοτικές έρευνες, **(2)** αναφέρεται σε ποσοτικές έρευνες, **(3)** αφορά την μέτρηση ομοιότητας ανάμεσα στις ετικέτες σήμανσης, **(4)** εμφάνιση υπαγωγικών σχέσεων, **(5)** ομαδοποίηση ετικετών, **(6)** παρέμβαση χρήστη, **(7)** αυτοματοποιημένες διαδικασίες, **(8)** χρήση εξωτερικών πόρων, **(9)** χρήση κοινωνικών συστημάτων σήμανσης.

4.3 Χαρακτηριστικά Υπόλοιπων Μεθόδων και Εφαρμογών Σύνδεσης

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Heymann P. / Molina H. [25]		X	**Κατανομή Πλαισίου Πόρων**	X			X		X
Wu X. [26]		X	**Σημασιολογικός Δείκτης Latent**		X	X			X
Jaschke R.[17]		X	**Αλγόριθμος Trias**	X	X		X		X
Begelman G. [28]		X	**Συνύπαρξη**		X		X		X
Xu Z./Fu Y. [29]		X	**"Tag browsing via filtering"**		X	X	X		X
Mathes A. [18]	X								X
Cattuso C. [31]		X	**Κατανομή 3 πλαισίων**	X		X		X	X
Damme C. [32]		X	**Αλγόριθμοι για την Αντιστοίχηση Οντολογιών**	X	X	X	X	X	X
Marihno L. / Buza K. [33]		X	**Συντελεστής Jaccard**	X	X		X	X	X
Knerr T. [34]		X	**Tag Υποκατηγορία skos:Concept**	X	X	X	X		X
Φωντόπουλος Γ. [5]		X	**Οντολογία SKOS**	X	X	X			X
Markines B. [35]		X	**Κοινές Πληροφορίες**	X		X		X	X

Alves H. / Santanche A. [36]		**X**	**Αλγόριθμος Ομοιότητας**	**X**	**X**	**X**		**X**	**X**
Hotho A. [37]		**X**	**Αλγόριθμος FolkRank**				**X**	**X**	**X**
Golder S. / Huberman B. [30]	**X**								**X**

(1) αναφέρεται σε ποιοτικές έρευνες, **(2)** αναφέρεται σε ποσοτικές έρευνες, **(3)** αφορά την μέτρηση ομοιότητας ανάμεσα στις ετικέτες σήμανσης, **(4)** εμφάνιση υπαγωγικών σχέσεων, **(5)** ομαδοποίηση ετικετών, **(6)** παρέμβαση χρήστη, **(7)** αυτοματοποιημένες διαδικασίες, **(8)** χρήση εξωτερικών πόρων, **(9)** χρήση κοινωνικών συστημάτων σήμανσης.

Κεφάλαιο 5

5.1 Συμπεράσματα - Επίλογος

Εξαιτίας του μεγάλου όγκου των δεδομένων στον Παγκόσμιο Ιστό, πολλοί χρήστες θεώρησαν πως θα πρέπει να αλλάξουν οι παλιές μέθοδοι κατηγοριοποίησής τους. Την άποψη τους αυτή τη στήριξαν στο γεγονός ότι ένα μεγάλο μέρος αυτού παρέμενε ανοργάνωτο, με τις περισσότερες παραδοσιακές μηχανές αναζήτησης να μην έχουν τη δυνατότητα να ικανοποιήσουν τις ανάγκες των χρηστών.

Τα τελευταία χρόνια παρατηρείται μια ραγδαία αύξηση του ενδιαφέροντος για τα Κοινωνικά Δίκτυα, τα οποία παρέχουν πρόσφορο έδαφος για τη συνεργασία ανάμεσα στα μέλη μιας κοινότητας. Με άλλα λόγια μέσω αυτών αναπτύχθηκε η ιδέα ότι το νόημα μπορεί να προέλθει καλύτερα από τους απλούς χρήστες, παρά από τους ειδικούς του περιεχομένου.

Αν και όλοι θεωρούνε πως οι ειδικοί θα πρέπει σίγουρα να έχουν μια σημαντική θέση στις βασικές τεχνολογίες του διαδικτύου, διατυπώθηκε η άποψη ότι θα πρέπει να διερευνηθούν περισσότερο νέοι τρόποι σήμανσης, όπως τα Folksonomies, παράλληλα με άλλους τομείς όπως ο Σημασιολογικός Ιστός. Αυτή την ιδέα υποστήριξε και η πλειοψηφία των ερευνητών, η οποία έκρινε οτι ο συνδυασμός των τεχνολογιών του Σημασιολογικού Ιστού, των Κοινωνικών Δικτύων και του περιεχομένου δύναται να οδηγήσει στην ολοκληρωμένη λειτουργία του Παγκόσμιου Ιστού.

Στην παρούσα διπλωματική εργασία επιχειρήσαμε να διερευνήσουμε το σύνολο των προσπαθειών για την ανάπτυξη μεθόδων και εφαρμογών σύνδεσης και επηρεαζόμενοι από την δημοφιλία και την αποτελεσματικότητά τους, προβήκαμε την ανάλυση των σημαντικότερων από αυτές. Κλείνοντας αξίζει να επισημάνουμε την προσπάθεια μας να προσεγγίσουμε και να συγκρίνουμε τις σημαντικότερες μεθόδους και εφαρμογές και να αναδείξουμε μέσω αυτής τα κοινά και τα διαφορετικά σημεία που τις διακρίνουν.

5.2 Μελλοντικά Σχέδια

Η παρούσα διπλωματική εργασία αποτελεί μια καλή βάση για περαιτέρω ενασχόληση με τον Σημασιολογικό Κοινωνικό Ιστό. Επηρεαζόμενοι από τις μεθόδους και εφαρμογές που ερευνήσαμε, στοχεύουμε μελλοντικά στο να προτείνουμε μια νέα προσέγγιση σύνδεσης της Σημασιολογικής πληροφορίας με τα Κοινωνικά Δίκτυα.

Κεφάλαιο 6

6.1 Βιβλιογραφία

[1] Del.icio.us: Social bookmarking. http://delicious.com.

[2] Flickr: Photo sharing. http://Flickr.com.

[3] Bibsonomy: Social bookmarking. http://www.bibsonomy.org/

[4] CiteULike: Social bookmarking. http://www.citeulike.org/

[5] Fountopoulos G.I. (2007), *"RichTags: A Social Semantic Tagging System",* A dissertation submitted in partial fulfilment of the degree of MSc Web Technology,

[6] Angeletou S., Sabou M. & Motta E. (2008). "*Semantically enriching folksonomies with FLOR*". In CISWeb Workshop at Europ. Semantic Web Conf, 2008.

[7] Berners-Lee T., Hendler J., and Lassila O. (2001),*"The semantic web".*

[8] Neches, R.; Fikes, R.; Finin, T.; Gruber, T.; Patil, R.; Senator, T.; Swartout, W.R. "*Enabling Technology for Knowledge Sharing"*. AI Magazine. Winter 1991. Σελ. 36-56.

[9] Swartout B.; Patil R.; Knight k.; Russ T. (1997) "Toward Distributed Use of Large-Scale Ontologies Ontological Engineering", AAAI-97 Spring Symposium Series, Σελ. 138-148.

[10] Bernaras A.; Laresgoiti I.; Correra J. *"Building and Reusing Ontologies for Electrical Network Applications"* ECAI96. 12th European conference on Artificial Intelligence. Ed. John Wiley & Sons, Ltd. Σελ. 298-302.

[11] Smith B., Welty C., *"Ontology: towards a new synthesis, in: Formal Ontology in Information Systems"*, ACM Press, Ongunquit, Maine, 2001, παράγραφοι iii–x.

[12] Gruber T. (1993) "*A translation Approach to portable ontology specifications",* Knowledge Acquisition. Vol. 5. Σελ. 199-220.

[13] Gruber T. (2005)*"Ontology of folksonomy: A mash-up of apples and oranges".*

[14] O'Reilly T. (2005) *"What is web2.0? ".*

[15] Berners-Lee T., Hendler J., and Lassila O.(2001), *"The semantic web".*

[16] Vanderwal T. (2007), *"Folksonomy coinage and definition".*

[17] Jaschke R., Hotho A., Schmitz C., Ganter B., Stumme G., *"Trias—an algorithm for mining iceberg tri-lattices"*, in: Proceedings of the Sixth IEEE International Conference on Data Mining (ICDM06), IEEE Computer Society, Hong Kong, 2006.

[18] Mathes A. (2005), "*Folksonomies - Cooperative Classification and Communication Through Shared Metadata*". Rapport interne, GSLIS, Univ. Illinois Urbana-Champaign.

[19] Shachter, J. Del.cious.us. http://del.icio.us/joshua. See also IDG article on del.icio.us: Del.icio.us: Social bookmarking phenomenon. InfoWorld, November 15, 2005, διαθέσιμο στο http://www.infoworld.com/article/05/11/15/HNdel.icio.us_1.html.

[20] Breslin J. (2011), "What Is The Social Semantic Web, And Why Do We Need It?", διαθέσιμο στην ιστοσελίδα http://newtechpost.com/2011/09/09/what-is-the-social-semantic-web-and-why-do-we-need-it.

[21] Mika P. (2005), "*Ontologies are Us: a Unified Model of Social Networks and Semantics*".

[22] Specia L. and Motta E. (2007), *"Integrating folksonomies with the semantic web"*.

[23] Budura A., Bourges D., Riordan J. (2009), *"Deriving Expertise Profiles From Tags".* International Conference on Computational Science and Engineering.

[24] Schmitz P. (2006), *"Inducing ontology from flickr tags".* In Collaborative Web Tagging Workshop (WWW '06).

[25] Heymann P. and Garcia-Molina H.(2006), *"Collaborative creation of communal hierarchical taxonomies in social tagging systems".* Technical report, Stanford University.

[26] Wu Z. and Palmer M. (1994), *"Verb semantics and lexical selection".*

[27] Ereteo G, Gandon F, Buffa M. (2004), *"SemTag: Semantic Community Detection in Folksonomies".*

[28] Begelman G., Keller P. & Smadja F. (2006), "Automated tag clustering: Improving search and exploration in the tag space".

[29] Xu, Z., Fu, Y., Mao, J., Su, D (2006.), *"Towards the Semantic Web: Collaborative Tag Suggestions"*. Proc. Of WWW2006, Collaborative Web Tagging Workshop.

[30] Golder, S., and Huberman (2005), *" The Structure of Collaborative Tagging Systems".* HP Labs technical report.

[31] Cattuto C., Benz D., Hotho A. & Stumme G. (2008), *"Semantic grounding of tag relatedness in social bookmarkingsystems".* 7th International Semantic Web Conference, Karlsruhe, Germany.

[32] Van Damme C., Hepp M. & Siorpaes K. (2007), *"Folksontology: An integrated approach for turning folksonomies into ontologies"*. In Bridging the Gep between Semantic Web and Web 2.0 (SemNet 2007), Σελ. 57–70.

[33] Marihno L.B., Buza K., Schmidt – Thieme L., "Folksonomy-based Collabulary Learning".

[34] Knerr T., "Tagging Ontology – Towards a Common Ontology for Folksonomies".

[35] Markines B., Cattuto C., Menczer F., Benz D., Hotho A. & Stumme G. (2009). "Evaluating similarity measures for emergent semantics of social tagging". In 18th International World Wide Web Conference, Σελ. 641–650.

[36] Alves H., Santanche A., "Folksonomized Ontologies – from social to formal".

[37] Hotho A., Jäschke R., Schmitz C. & Stumme G. (2006). Information Retrieval in Folksonomies: Search and Ranking.

[38] Sanderson M., Croft, B. (1999), *"Deriving concept hierarchies from text"*, In: Proceedings of the 22nd ACM Conference of the Special Interest Group in Information Retrieval, Σελ. 206-213.

[39] Levenshtein distance: string metric, http://www.miislita.com/searchito/levenshtein-edit-distance.html

[40] Ding, L., Finin, T., Joshi, A., Pan, R., Scott Cost, R., Peng, Y., Reddivari, P., Doshi, V.C., and Sachs, (2004), *"Swoogle: A Search and Metadata Engine for the Semantic Web".* 13th ACM Conference on Information and Knowledge Management, Washington D.C.

[41] Shirky, C. (2005), *"Ontology is Overrated: Categories, Links, and Tags".* Από shirkey.com blog. http://shirky.com/writings/ontology_overrated.html

[42] Gruber, T. (1993), *"A Translation Approach to Portable Ontology Specifications".* Knowledge Acquisition, Σελ. 199-220.

[43] Wordnet: Lexical database, http://wordnet.princeton.edu/

[44] Watson, *"The Semantic Web Gateway: search engine",* http://kmi-web05.open.ac.uk/WatsonWUI/

[45] Trillo R., Gracia J., Espinoza M., Mena E. (2007), *"Discovering the Semantics of User Keywords".*

[46] KGRAM, http://www-sop.inria.fr/edelweiss/software/corese/kgram/index.php.

[47] Limpens, F., Gandon, F., Buffa, M. (2010): "Helping online communities to semantically enrich folksonomies". In Proc. Of WebSci10, Raleigh

[48] Erétéo, G., Buffa, M., Gandon,, F., Corby, O. (2009): "Analysis of a Real Online Social Network Using Semantic Web Frameworks", In Proc. Of ISWC'2009.

[49] Dogear: Social Bookmarking, http://www.eigology.com/dogear/ .

[50] IBMr: Institute of Business Management & Research, http://www.ibmr.org/.

[51] EAT: Interactive Associative Thesaurus, http://www.eat.rl.ac.uk/.

[52] Batagelj V., Mrvar A. (1998), "*Pajek—program for large network analysis*". Σελ. 47–57.

[53] Borgatti S., Everett M., Freeman L., "*Ucinet for Windows: Software for Social Network Analysis*", Analytic Technologies, Harvard.

[54] Mika P. (2005), "Flink: Semantic Web technology for the extraction and analysis of social networks".

[55] van Elst L., Abecker A. (2002), "*Ontologies for information management: balancing formality, stability, and sharing scope*", Expert Syst. Appl. 23 (4) Σελ. 357–366.

[56] Είδη έρευνας: el.wikipedia.org/wiki/Έρευνα_αγοράς.

[57] Davies J., Grobelnik M., Mladevic D.(2009), "Semantic Knowledge Management", Εκδόσεις Springer – Velrag Berlin Heidelberg, Σελ. 3-21 και 129-140

[58] Antoniou A., van Harmelen F. (2009), "Εισαγωγή στο Σημασιολογικό Ιστό", Εκδόσεις Κλειδάριθμος, Σελ. 19-37.