ΠΑΝΕΠΙΣΤΗΜΙΟ ΜΑΚΕΔΟΝΙΑΣ
ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
ΤΜΗΜΑΤΟΣ ΕΦΑΡΜΟΣΜΕΝΗΣ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΑΝΑΛΥΣΗ ΙΣΤΟΡΙΚΩΝ ΔΕΔΟΜΕΝΩΝ ΣΕ ΚΟΙΝΩΝΙΚΑ ΔΙΚΤΥΑ ΓΙΑ ΤΗΝ ΠΡΟΒΛΕΨΗ ΣΥΝΔΕΣΜΩΝ

Διπλωματική Εργασία

της

Έλενης Μαυροϊδάκη του Μαυρουδή

Θεσσαλονίκη, 10/2018

ΑΝΑΛΥΣΗ ΙΣΤΟΡΙΚΩΝ ΔΕΔΟΜΕΝΩΝ ΣΕ ΚΟΙΝΩΝΙΚΑ ΔΙΚΤΥΑ

Ελένη Μαυροϊδάκη

Πτυχίο Εφαρμοσμένης Πληροφορικής, Πανεπιστήμιο Μακεδονίας, 2010

Διπλωματική Εργασία

υποβαλλόμενη για τη μερική εκπλήρωση των απαιτήσεων του

ΜΕΤΑΠΤΥΧΙΑΚΟΥ ΤΙΤΛΟΥ ΣΠΟΥΔΩΝ ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΠΛΗΡΟΦΟΡΙΚΗ

Επιβλέπουσα Καθηγήτρια
Γεωργία Κολωνιάρη

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 6η Νοεμβρίου 2018

Κολωνιάρη Γεωργία	Ευαγγελίδης Γεώργιος	Γεωργιάδης Χρήστος
.....	.....	.....

Μαυροϊδάκη Ελένη

.....

Περίληψη

Η χρήση των Κοινωνικών Δικτύων, στην εποχή μας, είναι ιδιαίτερα διαδεδομένη τόσο σε επίπεδο προσωπικής χρήσης, αλλά και επαγγελματικής. Τα μέλη των Κοινωνικών Δικτύων απασχολούν, επηρεάζουν και διαμορφώνουν γεγονότα και συμπεριφορές. Η ανάγκη μελέτης τεχνικών εξόρυξης γνώσης μέσα από τα Κοινωνικά Δίκτυα έχει καταστεί σημαντική ως προς το εν δυνάμει αποτέλεσμα που μπορεί να επιφέρει στην κατανόηση, πρόβλεψη αλλά και σχεδιασμό μελλοντικών Στρατηγικών λύσεων. Κύριος στόχος της εργασίας είναι η μελέτη, η καταγραφή, η εφαρμογή αλλά και η σχεδίαση αλγορίθμων εξόρυξης δεδομένων πάνω σε ιστορικά δεδομένα κοινωνικών δικτύων ώστε να αποδειχθεί η συσχετιζόμενη δυναμική του χρόνου και των αποτελεσμάτων που μπορούν να εκμαιευτούν και να χρησιμοποιηθούν σε πολλά σύγχρονα προβλήματα. Συγκεκριμένα, θα πραγματοποιηθεί αρχικά συγκριτική μελέτη σχετικών προσεγγίσεων που εφαρμόζουν τεχνικές εξόρυξης γνώσης και ανάλυσης δεδομένων σε κοινωνικά δίκτυα. Στη συνέχεια, θα διερευνηθεί τόσο η δυνατότητα επέκτασης τέτοιων προσεγγίσεων με στόχο να λαμβάνουν υπόψη και χρονική πληροφορία με στόχο την εκμαίευση καλύτερων ποιοτικά αποτελεσμάτων όσο και ο σχεδιασμός νέων μεθόδων που βασίζονται σε χρονική πληροφορία. Στο πλαίσιο της εργασίας, θα μελετηθούν τεχνικές εξόρυξης δεδομένων από κοινωνικά δίκτυα που αφορούν στο διαδεδομένο πρόβλημα της πρόβλεψης και εύρεσης συνδέσμων (link prediction). Τέλος, θα διεξαχθεί υλοποίηση σχετικών και εναλλακτικών αλγορίθμων σε σχετική πειραματική μελέτη.

Λέξεις Κλειδιά: Ανάλυση ιστορικών δεδομένων σε κοινωνικά δίκτυα, κοινωνικά δίκτυα, αλγόριθμοι εξόρυξης δεδομένων, χρονική πληροφορία στα κοινωνικά δίκτυα, εύρεση συνδέσμων σε κοινωνικά δίκτυα

Abstract

The use of Social Networks in our time is particularly widespread in terms of both personal and business use. Members of the Social Networks employ, influence and shape events and behaviors. The need to study Social Data Mining techniques has become significant in terms of the potential result that can lead to the understanding, anticipation and planning future Strategic Solutions. The proposed approach of this paper is to address, study, record, implement and design data mining algorithms on historical social networking data to demonstrate the associated dynamics of time and results that can be elicited and used in many modern problems. In particular, a comparative study of relevant approaches that implement data mining and analysis techniques on social networks will be carried out. We will then explore the possibility of extending such algorithms in order to take into account both time involved data so as to obtain better quality results and design new methods based on historical data. Our proposed approach is the study of social data mining techniques related to the widespread Link Prediction problem. Finally, relevant and revised algorithms will be implemented and a related experimental study will be carried out.

Keywords: Historical data analysis on social networks, social networks, data mining algorithms, data mining techniques in social network analysis, link prediction problem, time evolving data on social network, link prediction analysis on social networks, social networks

Πρόλογος - Ευχαριστίες

Η παρούσα διπλωματική εργασία εκπονήθηκε στο πλαίσιο του προγράμματος μεταπτυχιακών σπουδών (ΠΜΣ) του τμήματος Εφαρμοσμένης Πληροφορικής, του Πανεπιστημίου Μακεδονίας με την κατεύθυνση ειδίκευσης, Υπολογιστικές Μέθοδοι και Εφαρμογές.

Η ιδέα της εργασίας γεννήθηκε, καλλιεργήθηκε, αναθεωρήθηκε και τελικώς ολοκληρώθηκε σε μία περίοδο προσωπικά δυναμικών αλλά και αμφιλεγόμενων γεγονότων και συγκυριών.

Νιώθω ευγνώμων για την ιδιαίτερα εποικοδομητική συνεργασία που είχα με την επιβλέπουσα της διπλωματικής εργασίας, Επίκουρη Καθηγήτρια κυρία Κολωνιάρη Γεωργία, καθώς τόσο η επικοινωνία μας όσο και η συνεργασία μας απέδωσε καρπούς σε μία δύσκολη χρονική περίοδο.

Δεν θα μπορούσα να παραβλέψω και να ευχαριστήσω ιδιαίτερα την οικογένεια μου τόσο για την ηθική υποστήριξη όσο και για την πρακτική ... καθώς αφιερώνοντας τον πολύτιμο ελεύθερο χρόνο τους για τη φύλαξη του νεογέννητου μωρού μου, στάθηκαν ο λόγος να ολοκληρώσω τη συγγραφή της διπλωματικής εργασίας.

Τέλος, θα ήθελα να **αφιερώσω** όλη αυτή την προσπάθεια των τελευταίων ετών και την ολοκλήρωση των σπουδών μου, στο μονάκριβο μόλις 12 μηνών γιο μου (sir Ηλία Βενέτη), που διευρύνει τους ορίζοντες μου - καθημερινά - με κάθε τρόπο και μου αποδεικνύει ότι τόσο σε Real όσο και σε Social Network Time Evolving Management ... όλα είναι ΔΥΝΑΤΑ !

Και ναι η χρονική πληροφορία μπορεί να διευρύνει τους ορίζοντες μας ...
απλά το καταλαβαίνουμε εκ των υστέρων.

Περιεχόμενα

Κατάλογος Εικόνων

Κατάλογος Πινάκων

1 Εισαγωγή

1.1 Πρόβλημα - Σημαντικότητα του θέματος

Η χρήση των Κοινωνικών Δικτύων (ΚΔ) έχει αλλάξει τον τρόπο που σκεφτόμαστε, που επιλέγουμε, που αποφασίζουμε. Μία νέα γνωριμία ακολουθείται στερεότυπα, πλέον, από μία ερώτηση του τρόπου εύρεσης της ύπαρξης σε κάποιο κοινωνικό δίκτυο (όπως Facebook, Instagram, Twitter, Google + κ.α.). Ακόμη, η εύρεση εργασίας αφορά στην παρουσίαση του εν δυνάμει εργαζομένου σε κάποιο ιστοχώρο επαγγελματικής κοινωνικής δικτύωσης (όπως LinkedIn, Visible Path, Xing κ.α.) καθώς και της επιχείρησης. Μία επιχείρηση "οφείλει" να κατέχει μία δεσπόζουσα θέση σε κάποιο Κοινωνικό Δίκτυο, ιδίως όταν εστιάζει στην ενημέρωση αλλά και εύρεση υποψηφίων πελατών.

Τα τελευταία έτη γιγαντιαία ποσά δαπανούνται από επιχειρήσεις στα Κοινωνικά Δίκτυα προκειμένου να στοχεύσουν τον καταλληλότερο υποψήφιο πελάτη. Ένα πολύ σοβαρό ζήτημα για τους εν δυνάμει Digital Marketing Experts αποτελεί και το "στοίχημα" για **αποδοτικότερες καμπάνιες** με το μεγιστοποιημένο όφελος για τους ίδιους, τις επιχειρήσεις, αλλά και τους υποψήφιους χρήστες / πελάτες.

Συγκεκριμένα :

- Μία **επιχείρηση**, καλείται να αναγνωρίσει τα χαρακτηριστικά του εν δύναμη υποψηφίου πελάτη της, και βάση αυτών να σχεδιάσει και να υλοποιήσει μία σειρά από Καμπάνιες Προώθησης στα Κοινωνικά Δίκτυα με στόχο να "δελεάσουν", να επηρεάσουν και τελικά να πείσουν τον υποψήφιο πελάτη να ολοκληρώσει μία Μετατροπή, δηλαδή την επίτευξη των στόχων που έχουν τεθεί από την επιχείρηση.
- Ένας **Digital Marketing Expert**, καλείται να αναλύσει τους στόχους της επιχείρησης ώστε να σχεδιάσει και υλοποιήσει Καμπάνιες Προώθησης στα Κοινωνικά Δίκτυα, να οριοθετήσει τα χαρακτηριστικά του υποψήφιου group χρηστών που δύναται να αποτελέσουν τους μελλοντικούς πελάτες της επιχείρησης, καθώς και να παρέχει στους τελευταίους ως "δόλωμα", εκείνη την πληροφορία που θα τους οδηγήσει στην Μετατροπή, δηλαδή στην επίτευξη των στόχων που έχουν τεθεί από την επιχείρηση.

- Ένα **υποψήφιος πελάτης** / χρήστης των Κοινωνικών Δικτύων, όντας δραστήριος σε αυτά, ακολουθεί μία σειρά από κινήσεις, οι οποίες "καταγράφονται" στις βάσεις των Κοινωνικών Δικτύων και οριοθετούνται μέσω ετικετών, ως χαρακτηριστικά για τα ενδιαφέροντα και τις προτιμήσεις του χρήστη. Με αυτή την αλληλουχία δραστηριοτήτων κατά το πέρασμα του **χρόνου,** προκύπτει μία σειρά αποθηκευμένων πληροφοριών στη βάση (database) του Κοινωνικού Δικτύου, που αποτελούν την "ταυτότητα" του χρήστη.

Οι τεχνικές Εξόρυξης και χρήσης Ιστορικών Δεδομένων σε Κοινωνικά Δίκτυα δύναται να προσφέρουν όλη εκείνη τη χρήσιμη αλλά και χαμένη πληροφόρηση στους κατάλληλους τύπους και μορφή ώστε να εξαχθούν δυναμικά συμπεράσματα για τον τρόπο βελτιστοποίησης των ενεργειών σε επίπεδο Digital Marketing, Advertising και Branding για τις επιχειρήσεις αλλά και τους Digital Marketing Experts, καθώς και για το πολυπόθητο 5* User Experience των χρηστών.

1.2 Σκοπός - Στόχοι

Ο λόγος εκπόνησης της παρούσας διπλωματικής εργασίας είναι διττός. Αφενός η ανάδειξη του ρόλου της χρονικής πληροφορίας στα Κοινωνικά Δίκτυα σε Ερευνητικό επίπεδο, αφετέρου η πρόταση λύσης που μπορεί να χρησιμοποιηθεί και σε Business Strategic Management επίπεδο.

Ένα επίμαχο και καίριο ζήτημα προς αντιμετώπιση αποτελεί το Πρόβλημα Εύρεσης Συνδέσμων (Link Prediction Problem) στη μορφή εκείνη που θα αποδώσει αποτελεσματικότερα και αποδοτικότητα τη βελτιστοποιημένη πρόταση εύρεσης λύσεων στο πρόβλημα που αντιμετωπίζουν τόσο τα ίδια τα Κοινωνικά Δίκτυα ως μέσα, όσο και οι χρήστες τους από τη σκοπιά του καθενός, αντίστοιχα.

Η διέλευση του χρόνου συντελεί σε ένα σύνολο αλλαγών που απαντούν σε μία πλειάδα γεγονότων, βεβαιώσεων, πράξεων και ζητημάτων προς αντιμετώπιση τόσο στην πραγματική μας ζωή, όσο και στην αντίστοιχη παρουσία του «Εγώ» μας στα Κοινωνικά Δίκτυα.

Θεωρώντας σημαντική την απόδειξη της ορθής και βελτιστοποιημένης μεθόδου εξόρυξης γνώσης για την απόδοση των καμπανιών αλλά και της λειτουργικότητας των Κοινωνικών Δικτύων, έχει τεθεί ως στόχος της παρούσας εργασίας, να χρησιμοποιηθούν εναλλακτικοί αλγόριθμοι εξόρυξης δεδομένων και γνώσης στα Κοινωνικά Δίκτυα με στόχο να απαντηθεί με ποιο τρόπο το Πρόβλημα Εύρεσης Συνδέσμων αποδίδει καλύτερα για τα μέλη των Κοινωνικών Δικτύων αλλά και εάν αποδεικνύεται σημαντικός ο χρονικός παράγοντας κατά το σχεδιασμό Στρατηγικών, λύσεων και αποφάσεων, σε ποιο βαθμό και με ποιο τρόπο.

1.3 Ερωτήματα - Υποθέσεις

Θεωρώντας ότι η διέλευση του χρόνου, επιδρά πάνω στους χρήστες των Κοινωνικών Δικτύων (ΚΝ), θα υλοποιηθεί αλγόριθμος απόδειξης της επιρροής που δύναται να έχει, στο Πρόβλημα Εύρεσης Συνδέσμων (Link Prediction).

Θα γίνει μία προσπάθεια απάντησης των ακόλουθων Business Influence Strategic ερωτημάτων για τη χρήση των εξαγόμενων Ιστορικών Δεδομένων στα ΚΝ :

1. Οι Στρατηγικές Marketing και Advertising μπορούν να εφαρμόσουν την πληροφόρηση που εξάγεται από τη χρήση Ιστορικών Δεδομένων για το πρόβλημα Εύρεσης Συνδέσμων, για την αποδοτικότητα των Καμπανιών και την πρόβλεψη του ενδεχόμενου ρίσκου κατά την τροποποίηση των στρατηγικών ή όχι.
2. Η χρήση των Ιστορικών Δεδομένων σε Κοινωνικά Δίκτυα, αποτελεί εργαλείο ανάδειξης των δυνατών σημείων μίας Στρατηγικής Καμπάνιας έναντι ανταγωνιστικών ή όχι.
3. Η χρήση των Ιστορικών Δεδομένων σε Κοινωνικά Δίκτυα, αποτελεί βασικό στοιχείο αποτελεσματικής ανάδειξης και Εύρεσης Συνδέσμων ή όχι.
4. Η χρήση των Ιστορικών Δεδομένων σε Κοινωνικά Δίκτυα, είναι ικανή να προκαλέσει επιρροές στους προς Εύρεση Συνδέσμους ή όχι.
5. Η αποδοτική χρήση Ιστορικών Δεδομένων σε Κοινωνικά Δίκτυα, μπορεί να επηρεάσει και να αναβαθμίσει/υποβαθμίσει το Brand Reputation ή όχι.

1.4 Συνεισφορά

Η λογική στην οποία στηρίχθηκε η συλλογή και η καταγραφή των επιμέρους τμημάτων της εργασίας, είναι οι παρακάτω : (i) Να αποτελέσει έναν σύντομο οδηγό καταγραφής και ανάλυσης του Προβλήματος Εύρεσης Συνδέσμων, όσον αφορά στα προβλήματα που απαντά στα Κοινωνικά Δίκτυα σε Business Oriented Στρατηγικές Προώθησης, των προεκτάσεων του καθώς και προτάσεων βελτίωσης της απόδοσης ήδη υλοποιημένων μεθόδων εξόρυξης γνώσης. (ii) Η βιβλιογραφική προσέγγιση στηρίχθηκε στο σύνολο της γνώσης που είναι απαραίτητη για την κατανόηση και εμβάθυνση των εναλλακτικών τεχνικών που δύναται να αντιμετωπίσουν με αποδοτικό τρόπο τμήματα του συνόλου του προβλήματος. (iii) Ένας από τους θεμελιώδης στόχους της εργασίας, εκτός από την εν δυνάμει αποτελεσματικότητα των πειραματικών μελετών που θα διενεργηθούν και θα αξιολογηθούν σε επόμενο κεφάλαιο, είναι οι αναγνώστες της εργασίας και οι εν δυνάμει μελλοντικοί ερευνητές αντίστοιχων προβλημάτων, να τη χρησιμοποιήσουν ως ναυαρχίδα για μελλοντικές ενέργειες, μελέτες και προτάσεις.

1.5 Βασική Ορολογία

Κοινωνικά Δίκτυα (Social Network) : Ο όρος Κοινωνικό Δίκτυο χρησιμοποιείται για την περιγραφή web-based υπηρεσιών που επιτρέπουν οντότητες να δημιουργήσουν ένα δημόσιο / ημί-δημόσιο προφίλ σε ένα τομέα (domain), ώστε να είναι δυνατή η επικοινωνία με άλλους χρήστες εντός του δικτύου, για την μεταφορά πληροφοριών.

Εγωκεντρικά δίκτυα (ego-centric networks) : Τα Κοινωνικά Δίκτυα όπως το Facebook, χρησιμοποιούν τον όρο «εγώ» για να υποδηλώσουν ένα πρόσωπο συνδεδεμένο με οποιονδήποτε στο δίκτυο. Ένα δίκτυο του «εγώ» είναι η οργάνωση του κοινωνικού κόσμου από την άποψη του εγώ. Οι υπόλοιποι, που είναι συνδεδεμένοι με ένα δίκτυο του «εγώ» ορίζονται ως οι «άλλοι».

Εξόρυξη Γνώσης (Data Mining) : Η τεχνολογία Εξόρυξης Γνώσης είναι ένα από τα βήματα (το 7ο συγκεκριμένα) της διαδικασίας KDD (Ανακάλυψης Γνώσης από Δεδομένα; Knowledge Discovery from Data), όπου εφαρμόζονται αλγόριθμοι ανάλυσης δεδομένων και εύρεσης που, κάτω από αποδεκτούς περιορισμούς υπολογιστικής απόδοσης, παράγουν μια συγκεκριμένη απαρίθμηση patterns (ή μοντέλων) στα δεδομένα.

1.6 Διάρθρωση της μελέτης

Στο **Κεφάλαιο 2** γίνεται μία εισαγωγική παρουσίαση του Social Network Mining Background καθώς και Data Mining / Τεχνικών εξόρυξης δεδομένων σε Κοινωνικά Δίκτυα. Επίσης εισάγεται η έννοια του Χρόνου και πως μπορεί να επηρεάσει μελλοντικές Στρατηγικές Ανάλυσης και Σχεδιασμού σε Κοινωνικά Δίκτυα, time - evolving analysis.

Στο **Κεφάλαιο 3**, παρουσιάζεται και αναπτύσσεται το Πρόβλημα Εύρεσης Συνδέσμων (Link Prediction Problem) σε Κοινωνικά Δίκτυα. Περιγράφονται εναλλακτικοί αλγόριθμοι Εύρεσης Συνδέσμων, λαμβάνοντας υπόψη και την έννοια του χρόνου.

Στο **Κεφάλαιο 4**, παρουσιάζεται το Experimental μέρος της εργασίας, όπου υλοποιούνται οι ανωτέρω αναφερόμενοι αλγόριθμοί σε δύο διαφορετικά Σύνολα Δεδομένων των Κοινωνικών Δικτύων, του Facebook και του Twitter.

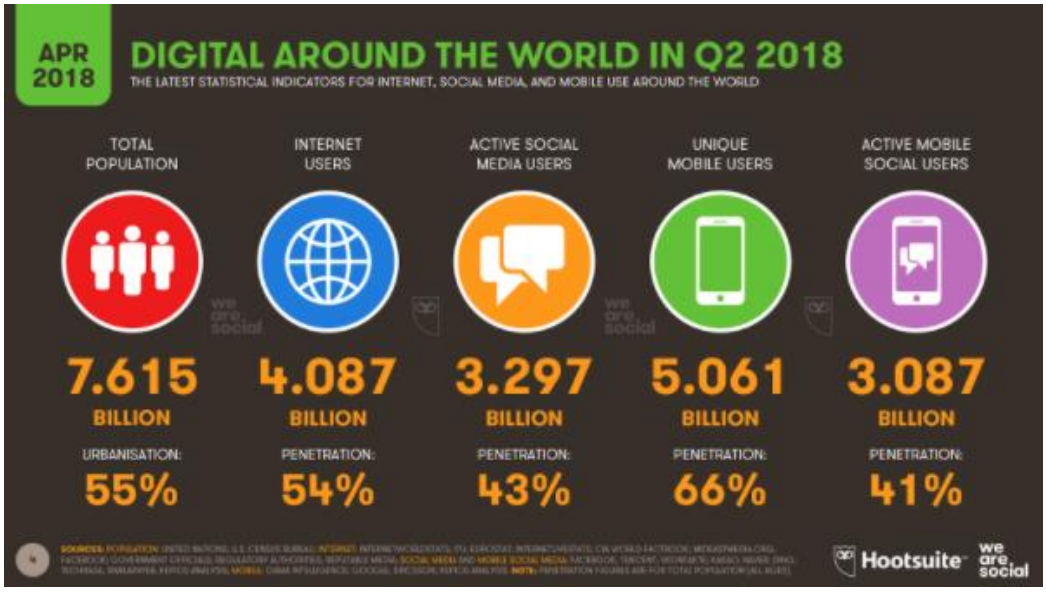
Τέλος, στο **Κεφάλαιο 5**, περιγράφονται τα συμπεράσματα που προέκυψαν από τη διπλωματική εργασία, το Empirical Evaluation καθώς και μελλοντικές κατευθύνσεις για περαιτέρω ανάλυση και υλοποίηση.

2 Βιβλιογραφική Επισκόπηση

Στο Κεφάλαιο αυτό, γίνεται ανάλυση των Κοινωνικών Δικτύων καθώς και των Τεχνικών Εξόρυξης Γνώσης από Δεδομένα. Περιγράφονται Αλγόριθμοι που δύναται να εφαρμοστούν για την εξαγωγή χρήσιμων συμπερασμάτων, για την εργασία. Αναφέρονται βιβλιογραφικά στοιχεία και πληροφορίες που θα βοηθήσουν στην περαιτέρω κατανόηση και εμβάθυνση της εργασίας. Τέλος, γίνεται αναφορά στο Πρόβλημα Εύρεσης Συνδέσμων που θα μας απασχολήσει σε επόμενα κεφάλαια.

2.1 Social Media Networks - Κοινωνικά Δίκτυα

Τα Κοινωνικά Δίκτυα, ως μέσα επικοινωνίας και μεταφοράς πληροφοριών χρησιμοποιούνται ευρέως σε μία παγκόσμια κλίμακα που ξεπερνά τα 3,2 δις χρήστες στο τέλος του 1ου τριμήνου του 2018 (Simon Kemp, 2018), (εικόνα 2-1).

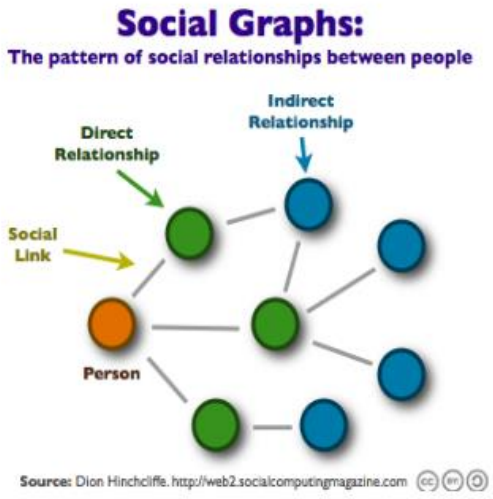


Εικόνα 2-1 Ενεργοί χρήστες στα Κοινωνικά Δίκτυα, στο τέλος του 1ου τριμήνου, 2018.
Πηγή: https://wearesocial.com/blog/2018/04/social-media-use-jumps-in-q1-despite-privacy-fears

Τα Κοινωνικά Δίκτυα αποτελούν μέσα ηλεκτρονικής αλληλεπίδρασης και ανταλλαγής περιεχομένου και συγκεντρώνουν πληροφορίες που επηρεάζουν, αξιολογούν, εκφράζουν απόψεις και συναισθήματα με οποιοδήποτε γραπτό τρόπο

(Adedoyin-Olowe & Gaber & Stahl, 2014). Μεγάλα ποσά δαπανώνται στα Κοινωνικά Δίκτυα για την προβολή διαφημίσεων εταιρειών και επιχειρήσεων. Ένας σημαντικός παράγοντας είναι και το γεγονός της επιρροής που ασκεί αυτού του είδους η αλληλεπίδραση, καθώς έχει παρατηρηθεί ότι οι περισσότεροι χρήστες των Κοινωνικών Δικτύων αναζητούν απαντήσεις στα ερωτήματα τους μέσα από τις πληροφορίες που διοχετεύονται στα Κοινωνικά Δίκτυα, από άγνωστες πολλές φορές οντότητες (Pang & Lee, 2008).

Τα Κοινωνικά Δίκτυα, ως διαδικτυακές εφαρμογές, έχουν βελτιωθεί τόσο σε τεχνολογικό, λειτουργικό αλλά και πολιτισμικό επίπεδο, ιδίως μετά την έλευση του Web 2.0, καθώς επιτρέπουν την επικοινωνία αλλά και μεταφορά πληροφοριών σε κάθε μέσο που δύναται να συνδεθεί στο διαδίκτυο (Kaplan & Haenlein, 2010). Το Κοινωνικό Δίκτυο θα μπορούσε να οριστεί ως ένα γράφημα που αποτελείται από κόμβους (nodes) και συνδέσμους (edges) που αντιπροσωπεύουν τις κοινωνικές σχέσεις σε ιστότοπους κοινωνικών δικτύων. Οι κόμβοι περιλαμβάνουν οντότητες και οι σχέσεις μεταξύ τους αποτελούν τους συνδέσμους (Adedoyin-Olowe & Gaber & Stahl, 2014), (εικόνα 2-2). Η αναπαράσταση του θα ήταν της μορφής G = (V, E), όπου V είναι ένα σύνολο από κόμβους και E ένα σύνολο από ακμές που ενώνουν τους κόμβους.

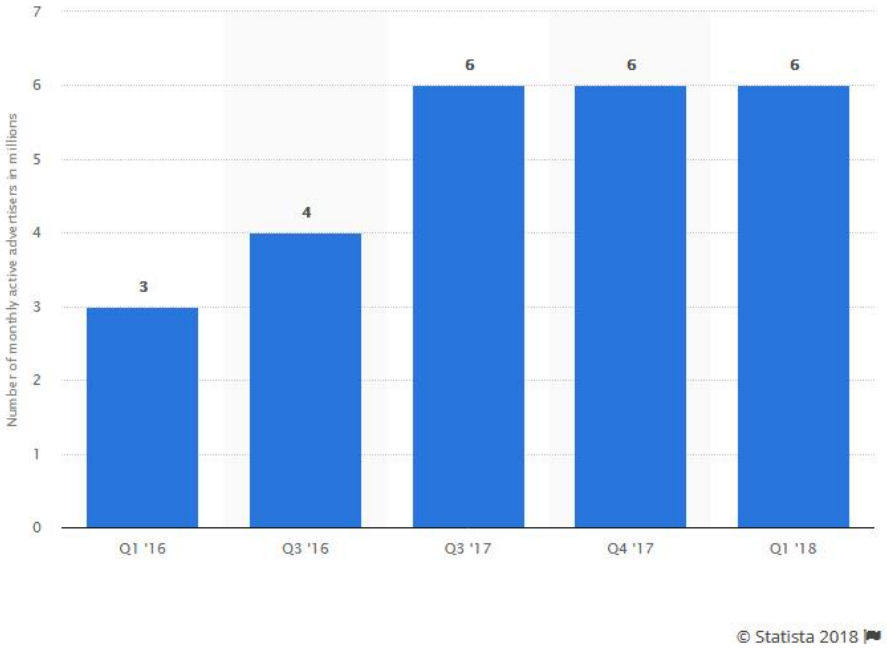


Εικόνα 2-2 Κοινωνικό Γράφημα: Το μοτίβο των κοινωνικών σχέσεων μεταξύ των ανθρώπων
Πηγή : (Kanna Al-Falahi & Yacine Atif & Said Elnaffar, 2010)

Τα Κοινωνικά Δίκτυα, έχουν δημιουργηθεί ώστε οι χρήστες (άνθρωποι, οργανισμοί, επιχειρήσεις κ.α.) που τα χρησιμοποιούν, να ανευρίσκουν άλλους χρήστες

(users) με κοινά ενδιαφέροντα, δραστηριότητες, στόχους, και να μπορούν να αλληλεπιδρούν μεταξύ τους διαμοιράζοντας υλικό και πληροφόρηση υπό διαφορετικές μορφές.

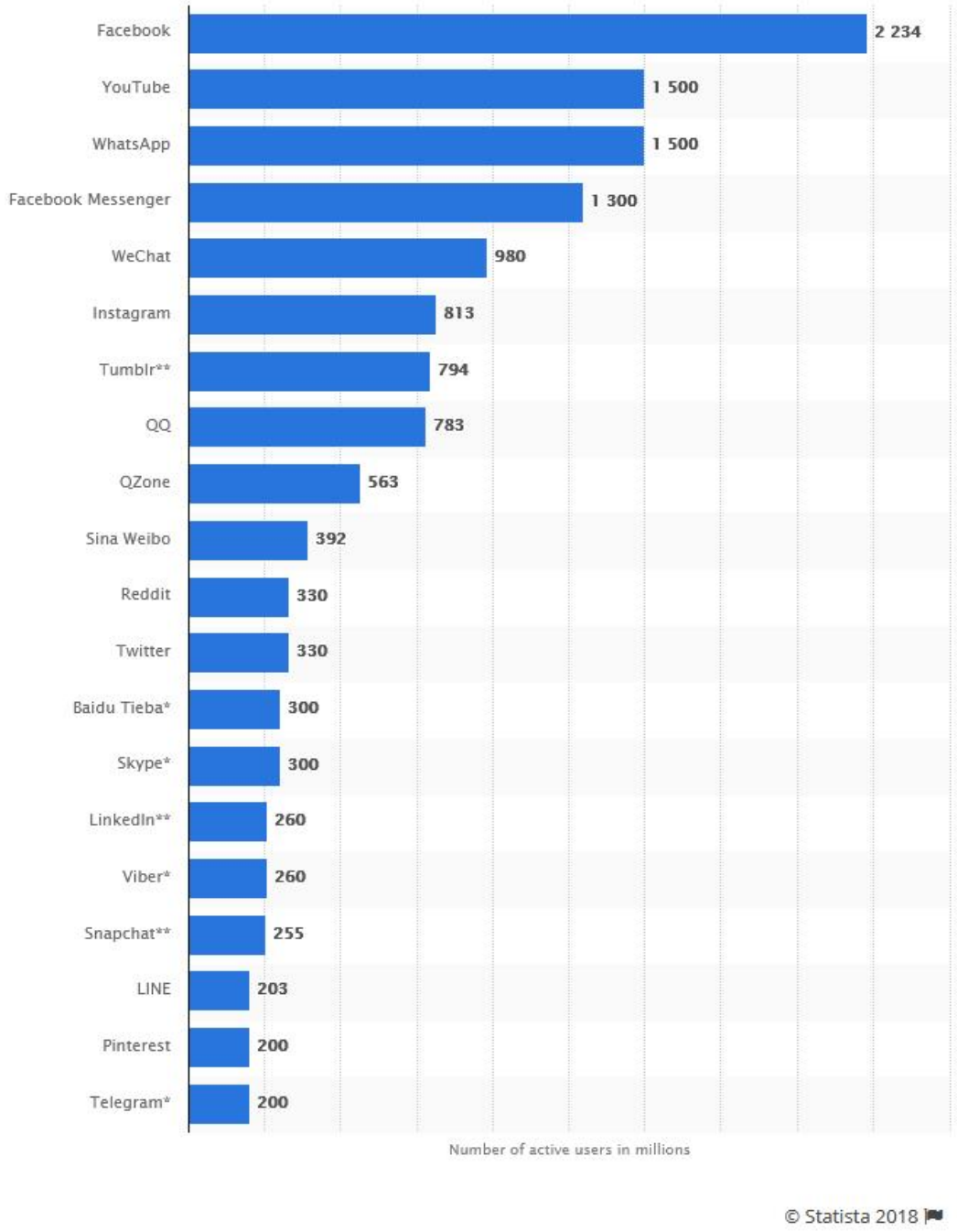
Διαφορετικοί χρήστες, χρησιμοποιούν διαφορετικούς τύπους κοινωνικών δικτύων για διαφορετικούς σκοπούς. Υπό τη σκοπιά αυτή, έντονο Επιχειρηματικό (Business) ενδιαφέρον, έχει αναπτυχθεί προκειμένου οι εταιρείες να καταφέρουν να ανακαλύψουν υποψήφιους πελάτες για την προώθηση των προϊόντων και των υπηρεσιών τους σε ένα δομημένο στρατηγικά (strategic target-marketing) περιβάλλον. Τεράστια ποσά δαπανώνται από επιχειρήσεις προκειμένου να προσεγγίσουν το Στρατηγικά "ιδανικό" κοινό. Χαρακτηριστικό παράδειγμα, αποτελεί η χρήση του Κοινωνικού Δικτύου Facebook από πλήθος επιχειρήσεων, προκειμένου να προσεγγίσει υποψήφιους πελάτες. Το 1ο τρίμηνο του 2018, κατεγράφησαν περί τα 6 εκατομμύρια ενεργοί διαφημιζόμενοι, συγκριτικά με το 2016 όπου την αντίστοιχη περίοδο οι ενεργοί διαφημιζόμενοι ήταν 3 εκατομμύρια (εικόνα 2-3).



Εικόνα 2-3 Ενεργοί διαφημιζόμενοι στο Facebook
από το 1ο τρίμηνο του 2016 έως το 1ο τρίμηνο του 2018.
Πηγή : https://www.statista.com/statistics/778191/active-facebook-advertisers/

Ένα άλλο δεδομένο που ενισχύει την δυναμική των επιχειρήσεων στη διαμόρφωση ανταγωνιστικού πλεονεκτήματος μέσω της Διαφημιστικής ισχύος στα Κοινωνικά Δίκτυα, είναι και ο τεράστιος αριθμός (ενεργών) χρηστών, με το Facebook να

κερδίζει τις εντυπώσεις, καθώς βρίσκεται στην πρώτη θέση έναντι άλλων Κοινωνικών Δικτύων ξεπερνώντας τα 2,2 δις ενεργούς χρήστες μηνιαίως (εικόνα 2-4).



Εικόνα 2-4 Τα πιο δημοφιλή κοινωνικά δίκτυα παγκοσμίως από τον Απρίλιο του 2018, ταξινομημένα κατά αριθμό ενεργών χρηστών (σε εκατομμύρια)

Πηγή : https://www.statista.com/statistics/693350/social-media-marketers-use-professional-purposes/

Ένας νέος «κόσμος» έχει αναπτυχθεί, αναγκάζοντας τόσο τον επιχειρηματικό κλάδο, όσο και τον ερευνητικό, να αναζητήσουν αποδοτικούς τρόπους χειρισμού όλης

αυτής της νέας δυναμικής που έχει καλλιεργήσει η ακμάζουσα χρήση των Κοινωνικών Δικτύων. Οι τεχνικές Εξόρυξης Γνώσης από Δεδομένα αποτελούν τη ναυαρχίδα των προσπαθειών αυτών.

2.2 Social Media Data Mining - Εξόρυξη Γνώσης από Δεδομένα Κοινωνικών Δικτύων

Όπως προαναφέραμε η εποχή που διανύουμε έχει ένα κύριο χαρακτηριστικό, το οποίο αφορά, στην συγκέντρωση τεράστιων όγκων πληροφορίας, λόγω της υπέρμετρης χρήσης του διαδικτύου και των Κοινωνικών Δικτύων.

Η προσπάθεια ανεύρεσης χρήσιμης πληροφορίας μέσα από έναν τεράστιο όγκο δεδομένων (Sang Jun Lee & Keng Siau, 2001) οδήγησε στην άνθιση της πειραματικής επιστήμης της Τεχνολογίας Εξόρυξης Γνώσης από Δεδομένα ή **Data Mining**. Έως και σήμερα πολλοί ερευνητές έχουν εργαστεί στον κλάδο του Data Mining, και πολλοί ορισμοί έχουν ανακύψει. Ο στόχος του Data Mining, όπως αναφέρει στην διπλωματική του ο (Γακόπουλος Ευθύμιος, 2012) είναι «η εξόρυξη χρήσιμης πληροφορίας από σύνολα ή βάσεις δεδομένων μεγάλου μεγέθους» (Hand et al., 2001) ή «η σύνθετη διαδικασία εξαγωγής συγκεκριμένης αλλά προηγουμένως άγνωστης και δυνητικά ωφέλιμης γνώσης από δεδομένα» (Frawley et al., 1992).

Επιστήμονες από διαφορετικά πεδία επιστημών έχουν έρθει αντιμέτωποι με την πρόκληση ανεύρεσης χρήσιμης πληροφορίας μέσα από έναν τεράστιο όγκο δεδομένων (data-sets). Για αυτό και η επιστήμη της Εξόρυξης Γνώσης από Δεδομένα χρησιμοποιεί μεθόδους και θεωρίες που πηγάζουν από επιστημονικά πεδία όπως των Βάσεων Δεδομένων (Databases), της Αναγνώρισης Προτύπων (Pattern Recognition), Μηχανικής Μάθησης (Machine Learning), της Τεχνητής Νοημοσύνης (Artificial Intelligent), της Στατιστικής (Statistics), των Έμπειρων Συστημάτων (Expert Systems) κ.α.

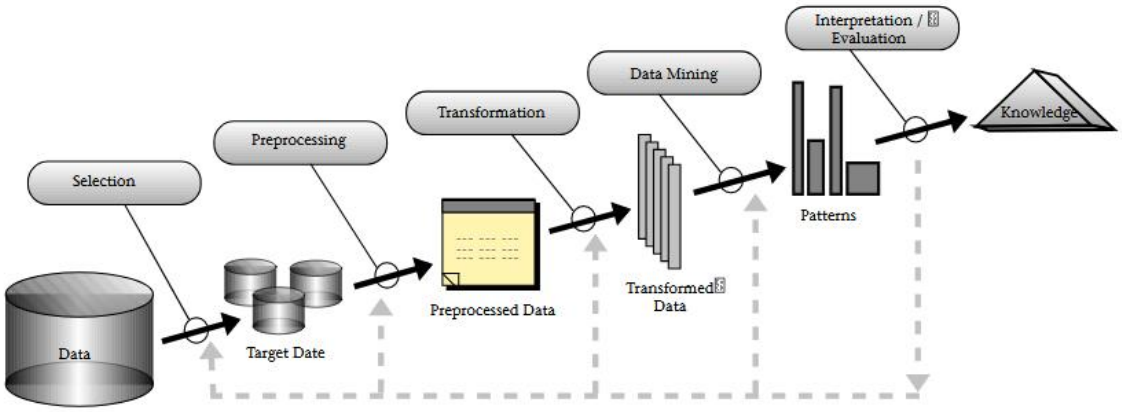
Η διαδικασία Εξόρυξη Γνώσης από Δεδομένα μπορεί να χρησιμοποιεί πολλές περίπλοκες και πολύπλοκες διαδικασίες ανάλυσης (Sang Jun Lee & Keng Siau, 2001), όμως κοινός παρονομαστής των διαδικασιών αυτών είναι τρία βασικά βήματα : η προετοιμασία των δεδομένων, η εκκαθάριση της δειγματοληψίας (αφαίρεση

λανθασμένων δεδομένων, απομάκρυνση θορύβου, εξομάλυνση δεδομένων), και η εξερεύνηση των δεδομένων.

Μία εναλλακτική σκοπιά ορισμένων επιστημόνων αφορά στην αντιμετώπιση της Εξόρυξης Γνώσης από Δεδομένα (Data Mining) ως Ανακάλυψη Γνώσης από Δεδομένα (Knowledge Discovery from Data, KDD) (Jiawei Han & Micheline Kamber, 2006). Η τεχνολογία της KDD περιγράφει ολόκληρη τη διαδικασία εξαγωγής γνώσης από δεδομένα. Στο πλαίσιο αυτό, η γνώση σημαίνει σχέσεις και πρότυπα μεταξύ των στοιχείων δεδομένων.

Η διαδικασία KDD ορίζεται από τα ακόλουθα 9 βήματα (Fayyad & Piatetsky-Shapiro & Smyth, 1996), (εικόνα 2-5). :

1) Αναγνώριση του στόχου της KDD,
2) Δημιουργία Δεδομένων (Data set),
3) Καθαρισμός & προ-επεξεργασία δεδομένων,
4) Μείωση του όγκου των δεδομένων,
5) Αντιστοίχιση των στόχων της KDD (βήμα 1ο) με μία μέθοδο εξόρυξης (summarization, classification, regression, clustering κ.α.),
6) Μοντελοποίηση & διερευνητική ανάλυση με επιλογή υποθέσεων,
7) Διαδικασία Εξόρυξης Δεδομένων ή Data Mining
8) Ερμηνεία / Αξιολόγηση αποτελεσμάτων
9) Ενέργειες στην βάση γνώσης



Εικόνα 2-5 Τα στάδια που απαρτίζουν την διαδικασία Ανακάλυψης Γνώσης από Δεδομένα (KDD)
Πηγή : AI Magazine, 1996, https://www.aaai.org/ojs/index.php/aimagazine/article/viewFile/1230/1131

Η διαδικασία Εξόρυξης Δεδομένων (Data Mining), ως στάδιο της KDD αφορά στην αναζήτηση μοτίβων ενδιαφέροντος σε μία συγκεκριμένη μορφή αναπαράστασης, περιλαμβάνοντας κανόνες ή δένδρα ταξινόμησης (classification rules or trees), ενέργειες παλινδρόμησης (regression) και ομαδοποίησης (clustering) κ.α..

2.3 Τεχνικές Εξόρυξης Γνώσης από Δεδομένα Κοινωνικών Δικτύων

Η επιλογή μίας Τεχνικής Εξόρυξης Γνώσης από Δεδομένα (ΤΕΓ) αφορά τόσο στο είδος του προβλήματος που καλείται να αντιμετωπίσει όσο και στην καταλληλότητα του μοντέλου/μηχανισμού που θα χρησιμοποιηθεί. Οι ΤΕΓ μπορούν να χρησιμοποιηθούν σε Εργασίες που αφορούν στην Πρόβλεψη μελλοντικών τιμών (Predictive Tasks) & στην Περιγραφή ή κατανόηση των δεδομένων (Descriptive Tasks) (Jiawei Han & Micheline Kamber & Jian Pei, 2012). Οι ΤΕΓ που επιλύουν θέματα που αφορούν στην Πρόβλεψη, χρησιμοποιούν ορισμένες μεταβλητές ή πεδία της βάσης δεδομένων (database), προκειμένου να προβλέψουν άγνωστες ή μελλοντικές τιμές άλλων μεταβλητών ενδιαφέροντος. Αντίστοιχα οι ΤΕΓ που επιλύουν θέματα που αφορούν στην Περιγραφή ή κατανόηση των δεδομένων, επικεντρώνονται στην εύρεση προτύπων (Patterns) με ανθρωποκεντρικές ερμηνείες για τα δεδομένα που περιγράφουν.

2.3.1 Εργασίες Εξόρυξης Γνώσης (Data Mining Tasks)

Όπως αναφέρθηκε, οι Εργασίες Εξόρυξης Γνώσης είναι δύο κυρίως ταχυτήτων. Αντιμετωπίζουν προβλήματα που αφορούν στην Πρόβλεψη είτε προβλήματα που αφορούν στην Περιγραφή των δεδομένων. ((Jose Hernandez - Orallo, 2005); (Fayyad & Piatetsky-Shapiro & Smyth, 1996)).

Οι Εργασίες Εξόρυξης Γνώσης με αντικείμενο την Πρόβλεψη αφορούν σε θέματα :

- Ταξινόμησης (Classification Tasks) : Το μοντέλο αυτό χρησιμοποιείται για την πρόβλεψη μίας διακριτής τιμή εξόδου (από μία ή δύο ή

περισσότερες κατηγορίες) με μία ή περισσότερες μεταβλητές εισόδου. Χρησιμοποιείται σε εργασίες κατηγοριοποίησης, κατάταξης, εκμάθησης προτιμήσεων, εκτίμησης πιθανότητας κλάσης κ.α.. Εφαρμόζεται για την κατάτμηση των πελατών, στην επιχειρηματική μοντελοποίηση, στην ανάλυση πιστώσεων και σε πολλές άλλες εφαρμογές.

- Ανάλυση Παλινδρόμησης (Regression Analysis Tasks) : Το μοντέλο αυτό δημιουργεί προγνωστικά μοντέλα. Για παράδειγμα, εκτίμηση της πιθανότητας επιβίωσης ενός ασθενούς δεδομένων των αποτελεσμάτων ενός συνόλου διαγνωστικών εξετάσεων, πρόβλεψη της ζήτησης των καταναλωτών για ένα νέο προϊόν ως συνάρτηση των διαφημιστικών δαπανών, κ.α. Η διαφορά μεταξύ παλινδρόμησης και ταξινόμησης είναι ότι η παλινδρόμηση ασχολείται με αριθμητικά / συνεχή χαρακτηριστικά στόχου, ενώ η ταξινόμηση ασχολείται με διακριτά / κατηγορικά χαρακτηριστικά στόχου.
- Σπουδαιότητα Χαρακτηριστικού (Attribute Importance Tasks) : Το μοντέλο αυτό παρέχει μια αυτοματοποιημένη λύση για τη βελτίωση της ταχύτητας και ενδεχομένως της ακρίβειας των μοντέλων ταξινόμησης που βασίζονται σε πίνακες δεδομένων με μεγάλο αριθμό χαρακτηριστικών. Για παράδειγμα, ανακαλύπτει παράγοντες που σχετίζονται με άτομα που ανταποκρίνονται σε μια προσφορά.
- Εύρεση Συνδέσμων (Link Analysis) : Το μοντέλο αυτό χρησιμοποιείται συνήθως για την ανάλυση συνδέσεων για τη μοντελοποίηση και πρόβλεψη της ατομικής συμπεριφοράς. Χρησιμοποιείται προκειμένου να αναγνωριστεί ένα αίτιο και να προληφθεί το αποτέλεσμα. Για παράδειγμα, χρησιμοποιείται ως Recommender System προκειμένου να προτείνει φίλους ή προϊόντα και υπηρεσίες.

Εργασίες Εξόρυξης Γνώσης με αντικείμενο την Περιγραφή των δεδομένων αφορούν σε θέματα :

- Ομαδοποίησης (Clustering) : Το μοντέλο εντοπίζει "φυσικές" ομάδες στα δεδομένα. Για παράδειγμα, ανακάλυψη ομογενών υποπληθυσμών για τους καταναλωτές σε βάσεις δεδομένων μάρκετινγκ.
- Εύρεσης Σύνοψης (Summarization) : Το μοντέλο περιλαμβάνει μεθόδους για την εύρεση μίας συμπαγούς περιγραφής για ένα υποσύνολο

δεδομένων, όπως η καταγραφή των μέσων τυπικών αποκλίσεων για όλα τα πεδία. Οι πιο εξελιγμένες μέθοδοι περιλαμβάνουν τη συλλογή περιληπτικών κανόνων (summarization rules) .

- Ανάλυσης Συσχέτισης & Παραγόντων (Correlation & factorial analysis) : Το μοντέλο αναγνωρίζει αν υπάρχει σχέση μεταξύ δύο (διμερών) ή περισσότερων (πολυμερών) αριθμητικών μεταβλητών. Για παράδειγμα, το μοντέλο αυτό μπορεί να ανακαλύψει αν η αγορά μιας ομάδας προϊόντων συμβαίνει ταυτόχρονα.
- Ανακάλυψη Συνδέσμων (Association Discovery) : Το μοντέλο προσδιορίζει τη σχέση μεταξύ δύο ή περισσότερων διακριτών μεταβλητών. Ανακαλύπτει κανόνες που συσχετίζονται με συχνά συσχετισμένα αντικείμενα που χρησιμοποιούνται για την ανάλυση του καλαθιού αγοράς, την διασταυρούμενη πώληση και την ανάλυση παραγόντων αιτίας αποτελέσματος. Χρήσιμο για την ομαδοποίηση προϊόντων, την τοποθέτηση νέων προϊόντων και την ανάλυση ελαττωματικών.
- Εύρεση Χαρακτηριστικών (Feature Extraction) : Το μοντέλο προσδιορίζει νέα χαρακτηριστικά ως γραμμικό συνδυασμό των υπαρχόντων. Για παράδειγμα, χρησιμοποιείται σε δεδομένα κειμένου, αποσύνθεση δεδομένων και προβολή δεδομένων καθώς και αναγνώριση προτύπων.
- Κοινωνική Ομοιότητα (Assortativity or Social Similarity) : Το μοντέλο προσδιορίζει το ποσό της επιρροής και της ομοφυλίας μεταξύ συνδεδεμένων οντοτήτων. Μετρά κατά πόσο συνδεδεμένοι είναι μεταξύ τους οι οντότητες (κόμβοι) (Reza Zafarani & Mohammad Ali Abbasi & Huan Liu, 2014).
- Ανάλυση Απόκλισης (Deviation Detection) : Το μοντέλο προσδιορίζει αποκλίσεις οι οποίες εκφράζουν είτε αλλαγές είτε ανωμαλίες στα δεδομένα, με στόχο να εφαρμοστούν διορθωτικές κινήσεις στα νέα δεδομένα. Για παράδειγμα, αυτό το μοντέλο μπορεί να ανακαλύψει αν για παράδειγμα μία πιστωτική κάρτα έχει κλαπεί, καθώς μπορεί να έχει γίνει μία μεγάλη απότομη κίνηση αγορών.

2.3.2 Αλγόριθμοι Εξόρυξης Γνώσης (Data Mining Algorithms)

Έχοντας κατανοήσει το πρόβλημα προς επίλυση και έχοντας επιλέξει την κατάλληλη Εργασία Εξόρυξης Γνώσης, ένας ή και περισσότεροι Αλγόριθμοι Εξόρυξης Γνώσης (Data Mining Algorithms) θα πρέπει να υλοποιηθούν.

Υπάρχει μεγάλο εύρος αλγορίθμων που έχουν υλοποιηθεί για την επίλυση διαφορετικών προβλημάτων. Ενδεικτικά θα αναφέρουμε κάποιους από τους πιο διαδεδομένους αλγορίθμους, (Reza Zafarani & Mohammad Ali Abbasi & Huan Liu, 2014); (Charlie Berger, 2012), βάση της Εργασίας Εξόρυξης Γνώσης.

- Αλγόριθμοι Ταξινόμησης (Classification Algorithms) : Decision Tree algorithms, Vector Space Model, Naive Bayes Classifier (NBC), Nearest Neighbor classifier, Classification with Network Information , Genetic Algorithms
- Αλγόριθμοι Ανάλυσης Παλινδρόμησης (Regression Analysis Algorithms) : Linear Regression, Multiple Linear Regression, Autoregressive, Logistic Regression.
- Αλγόριθμοι Εύρεσης Συνδέσμων (Link Analysis Algorithms) : Negative Association, Association Discovery, Sequential Pattern Discovery, Matching Time Sequence Discovery
- Αλγόριθμοι Ομαδοποίησης (Clustering Algorithms) : Spectral Clustering, Hierarchical Clustering, Partitioning Clustering, Cohesiveness, Silhouette Index
- Αλγόριθμοι Εύρεσης Χαρακτηριστικών (Feature Extraction Algorithms) : Non negative Matrix Factorization (NMF)

2.4 Ανάλυση Κοινωνικών Δικτύων στο Χρόνο - Time evolving Network Analysis

Αυτή η ενότητα παρέχει μια επισκόπηση αναφορικά με τη χρήση της έννοιας του χρόνου στα Κοινωνικά Δίκτυα. Η βιβλιογραφία των Δικτύων (Networks), και ιδιαίτερα των Κοινωνικών Δικτύων δεν παρέχει πολλές υλοποιημένες μελέτες που εκφράζουν τις δυναμικές πτυχές της ανάπτυξης των σχέσεων ενός Δικτύου, υπό το πρίσμα της επιρροής του χρόνου. Τα περισσότερα μοντέλα δικτύου είναι στατικά, υπό την έννοια ότι ο χρόνος δεν παίζει ρόλο.

Στο σημείο αυτό, θα γίνει αναφορά σε διαφορετικές θεωρίες που μπορούν να εφαρμοστούν στα Κοινωνικά Δίκτυα λαμβάνοντας υπόψη και την έννοια του χρόνου, στο τέλος του κεφαλαίου, των στόχων που υλοποιούν και των μεθόδων που χρησιμοποιούν.

2.4.1 Δεδομένα Χρονοσειρών - Time Series Data

Ως Δεδομένα Χρονοσειρών (Time Series Data) ορίζονται «εγγραφές φαινομένων που διαφοροποιούνται ακανόνιστα στη βάση του χρόνου» σύμφωνα με τον ορισμό που απέδωσε ο (Genshiro Kitagawa, 2010). Η ανάλυση των πειραματικών δεδομένων, η οποία βασίζεται σε παρατηρήσεις σε διαφορετικά χρονικά σημεία, «οδηγεί σε νέα και μοναδικά προβλήματα στη στατιστική μοντελοποίηση και εξαγωγή συμπερασμάτων» (Shumway & Stoffer, 2011). Η ανάλυση αυτών των δεδομένων που περιέχουν χρονική πληροφορία είναι επιπλέον σημαντική για πολλούς διαφορετικούς επιστημονικούς τομείς, με στόχο να προσδιοριστεί το επόμενο βήμα της ανάλυσης και να βρεθούν οι κατάλληλες στρατηγικές για τη στατιστική μοντελοποίηση (Philipp Singer, 2011). Πρωταρχικός στόχος της χρήσης δεδομένων χρονοσειρών είναι να βρεθούν τα κατάλληλα μαθηματικά μοντέλα που θα περιγράψουν τα δεδομένα. Σε επόμενο στάδιο, μπορεί να υλοποιηθεί Πρόβλεψη, βάση των συσχετίσεων που προκύπτουν κατά το πέρασμα του χρόνου μεταξύ των μεταβλητών ενδιαφέροντος.

Στην περίπτωση των Κοινωνικών Δικτύων, υπάρχει ένα σύνολο δεδομένων (dataset) το οποίο διαφοροποιείται στο πέρασμα του χρόνου. Οι πληροφορίες που περιέχονται στο σύνολο των δεδομένων, αφορούν τόσο στο κοινωνικό δίκτυο (social

network) όσο και στους χρήστες οι οποίοι αλληλεπιδρούν με την πάροδο του χρόνου, διαμορφώνοντας διαφορετικές συμπεριφορές και ενέργειες (content network). Η λογική της χρήσης δεδομένων χρονοσειρών (time series data) αφορά στην μοντελοποίηση των διαχρονικών επιρροών ανάμεσα στα κοινωνικά δίκτυα (social networks) και στα δίκτυα περιεχομένου (context networks). Η μοντελοποίηση αυτής της λογικής, υλοποιείται μέσω διαφορετικών μορφών αλγορίθμων που υπακούουν στη θεωρία της Ανάλυσης Παλινδρόμησης (Regression Analysis).

2.4.2 Θεωρία εξέλιξης Γραφημάτων στο χρόνο - Time-evolving graphs

Η θεωρία των γραφημάτων που εξελίσσονται στο χρόνο (time-evolving graphs) αποτελεί μία εναλλακτική πρόταση αντιμετώπισης θεμάτων πρόβλεψης σε δυναμικά εξελισσόμενα δίκτυα, όπως και τα Κοινωνικά Δίκτυα. Το πρόβλημα εντοπισμού προτύπων αλληλεπίδρασης σε ένα Κοινωνικό Δίκτυο, η μελέτη παρακολούθησης προτύπων εξέλιξης των κοινοτήτων (communities) στα Κοινωνικά Δίκτυα, αποτελούν παραδείγματα εφαρμογής της θεωρίας γραφημάτων που εξελίσσονται στο χρόνο.

Η μέθοδος GraphScope (Jimeng Sun & Christos Faloutsos & Spiros Papadimitriou & Philip S. Yu. KDD, 2007), αποτελεί μία πρόταση αυτόματης εξαγωγής των σημείων αλλαγής του χρόνου, καθώς και της διαμέρισης των κόμβων, ώστε να αποκαλυφθεί με ακρίβεια η βασική δομή των κοινοτήτων και των αλλαγών τους κατά την πάροδο του χρόνου. Πρόκειται για μία μέθοδο που δεν απαιτεί τον ορισμό παραμέτρων προκειμένου να εξαχθούν ροές γραφημάτων. Χρησιμοποιεί τις αρχές της MDL (Minimum Description Language) γλώσσας, για την διαμόρφωση των κοινοτήτων και την αναγνώριση του χρόνου τροποποίησης τους. Είναι μία γρήγορη, κλιμακωτή μέθοδος που μπορεί να χρησιμοποιηθεί σε μεγάλου όγκου δεδομένα και δύναται να προσφέρει σημαντικά αποτελέσματα.

2.4.3 Θεωρία Ορθολογικής Επιλογής στο Χρόνο - Rational Choice Theory in Time

Η εξέλιξη των δυναμικών Κοινωνικών Δικτύων στο Χρόνο ως αποτέλεσμα των πράξεων που επενεργούν οι μεμονωμένες οντότητες, περιγράφεται μέσα από τη Θεωρία της Ορθολογικής Επιλογής στο πέρασμα του Χρόνου (Rational Choice Theory). Για

παράδειγμα, τα μέλη μίας ομάδας πρωτοετών φοιτητών ενός Πανεπιστημίου, τη χρονική στιγμή t_0, που αντιστοιχεί στο σημείο έναρξης της σταδιοδρομίας τους, εκτός από κάποιες λίγες φιλίες (friendships) που μπορεί να προϋπάρχουν, δεν γνωρίζονται μεταξύ τους. Αν κατά το πέρασμα 3 εβδομάδων, 2 μηνών, 4 μηνών κ.ο.κ. γίνει μέτρηση των μεταξύ τους σχέσεων, θα έχουν προκύψει πολλές διαφοροποιήσεις.

Για αυτού του είδους τα προβλήματα, η θεωρία της Ορθολογικής Επιλογής στο πέρασμα του Χρόνου υποστηρίζει ότι, «δεδομένου των περιορισμών και των ευκαιριών, η συμπεριφορά μίας οντότητας (actor) μπορεί να διαμορφωθεί, σαν να βασίζονται οι ενέργειες της σε μία ανάλυση κόστους-οφέλους (cost-base analysis)» (Van De Bunt & Van Duijn & Snijders, 1999). Η θεωρία των ανωτέρω επιστημόνων χρησιμοποιεί την διαδικασία Markov σε συνεχή χρόνο σε συνδυασμό με την ευρετική μοντελοποίηση που αφορά στις συμπεριφορές των οντοτήτων (actors) να ενεργήσουν σε ένα συγκεκριμένο χρονικό σημείο t, και διαμορφώνεται ως ένα τυχαίο μοντέλο χρησιμότητας (random utility model), στο οποίο η οντότητα επιλέγει μεταξύ των πιθανών ενεργειών με πιθανότητες που είναι συναρτήσει της αναμενόμενης χρησιμότητας όπως υπολογίζεται από τις μεταβλητές του μοντέλου.

2.4.4 Επιρροή & Συνεργασία στο χρόνο - Influence and Collaboration over time

Η ύπαρξη των Κοινωνικών Δικτύων, διευκολύνει όλο και περισσότερο την ανταλλαγή και προώθηση πληροφοριών, την καταγραφή και αποθήκευση δεδομένων καθώς και την ανάπτυξη συνεργασιών μεταξύ των οντοτήτων. Η μεταβλητή του χρόνου δύναται να επηρεάσει θέματα που αφορούν στην εξέλιξη, διαφοροποίηση και στην επιρροή των συνεργασιών, ενός δυναμικού δικτύου (Kristin R. Eschenfelder & Morgaine Gilchrist Scott & Kalpana Shankar, Greg Downey, 2016). Η θεωρία της Ανάλυσης των Κοινωνικών Δικτύων (Social Network Analysis) εφαρμόζεται σε θέματα επιρροής σχέσεων και συνεργασίας σχέσεων και πως διαφοροποιούνται στο πέρασμα του χρόνου.

2.5 Πρόβλημα Εύρεσης Συνδέσμων - Link Prediction Problem

Τα Κοινωνικά Δίκτυα, όντας δυναμικά εξελισσόμενα δίκτυα στην πάροδο του χρόνου, δύναται να προσφέρουν σημαντική ώθηση σε θέματα Σχεδιασμού, Οργάνωσης, Παρακολούθησης και Εξαγωγής Στρατηγικών και Γνώσης που μπορούν να εφαρμοστούν σε πολλούς ερευνητικούς και επιχειρηματικούς κλάδους. Ένα πολύ συχνό και σοβαρό πρόβλημα που αφορά στην εξέλιξη των δικτύων είναι το Πρόβλημα Εύρεσης Συνδέσμων (Link Prediction Problem).

2.5.1 Περιγραφή Προβλήματος Εύρεσης Συνδέσμων

Το Πρόβλημα Εύρεσης Συνδέσμων, έρχεται να απαντήσει στο ακόλουθο χρονικά εξαρτώμενο (time-evolving) ερώτημα : «Δεδομένου ενός στιγμιότυπου ενός Κοινωνικού Δικτύου τη χρονική στιγμή t, είναι δυνατόν να προβλεφθούν με ακρίβεια οι άκρες (οι εν δυνάμει νέοι σύνδεσμοι) που θα προστεθούν στο δίκτυο κατά τη διάρκεια του χρονικού διαστήματος από το χρόνο t σε ένα δεδομένο μελλοντικό χρόνο t'» (Nowell & Kleinberg, 2007). Εναλλακτικά, το Πρόβλημα Εύρεσης Συνδέσμων σχετίζεται και με το πρόβλημα πρόβλεψης χαμένων συνδέσμων από ένα παρατηρούμενο δίκτυο, το οποίο όμως αφορά σε ένα στατικό στιγμιότυπο του δικτύου και δεν εξετάζει την εξέλιξη του στο πέρασμα του χρόνου.

Το Πρόβλημα Εύρεσης Συνδέσμων έχει εφαρμογές σε πολλούς διαφορετικούς επιστημονικούς και ερευνητικούς κλάδους. Μπορεί να απαντήσει σε ερωτήματα που αντιμετωπίζει το τμήμα Human Resources στην εύρεση του κατάλληλου υποψηφίου εργαζομένου. Στην επιστήμη της Βιολογίας σε θέματα δίκτυα τροφίμων, δίκτυα αλληλεπίδρασης πρωτεϊνών μεταξύ τους, δίκτυα που σχετίζονται με μεταβολικούς και άλλους παράγοντες (Linyuan Lü & Tao Zhou, 2010). Σε θέματα Κοινωνικών Δικτύων όπως πρόταση πιθανών φίλων, πιθανών υπηρεσιών ή και προϊόντων με στόχο τη βελτιστοποίηση της πίστης (loyalty) των χρηστών στους αντίστοιχους Ιστότοπους και της μεγιστοποίησης των κερδών των επιχειρήσεων. Σε θέματα e-commerce ως προτεινόμενα συστήματα (recommendation systems) σε επίπεδο προϊόντων και υπηρεσιών, όπως για παράδειγμα "όσοι αγόρασαν αυτό το προϊόν, αγόρασαν επίσης" ή "τα προϊόντα με τις μεγαλύτερες πωλήσεις". Σε επίπεδο έρευνας της Τεχνητής Νοημοσύνης για θέματα οργάνωσης μεγάλων επιχειρήσεων, σε μία προσπάθεια πρόβλεψης μελλοντικών ωφέλιμων συνεργασιών μεταξύ διαφορετικών τμημάτων μίας

επιχείρησης που δεν είναι ορατές σε παροντικό χρόνο, αλλά εκ των υστέρων θα ήταν ωφέλιμες. Ακόμη, στον κλάδο της Ασφάλειας για την πρόβλεψη μελλοντικών τρομοκρατικών ενεργειών, υπό το πρίσμα της συνεργατικής οργάνωσης οντοτήτων η οποία δεν είναι ευθέως ορατή. Επίσης, στον ακαδημαϊκό κλάδο, πολλές εργασίες έχουν υλοποιηθεί για τα συγγραφικά δίκτυα (co-authorship networks) όπως σε ακαδημαϊκά περιοδικά, όπου οι σύνδεσμοι ενώνουν ζεύγη συγγραφέων που έχουν συγγράψει κάποιο άρθρο από κοινού. Στην περίπτωση αυτή το Πρόβλημα Εύρεσης Συνδέσμων μπορεί να υλοποιηθεί υπό το πρίσμα της πρότασης μελλοντικής συνεργασίας δύο ή περισσοτέρων συγγραφέων (Nowell & Kleinberg, 2007).

Στη δική μας εργασία, το Πρόβλημα Εύρεσης Συνδέσμων, έρχεται να απαντήσει σε μία σειρά Business Oriented ερωτημάτων που αφορούν στην μέτρηση της απόδοσης που μπορεί να έχουν υλοποιημένες Καμπάνιες Προώθησης στα Κοινωνικά Δίκτυα, εφαρμοσμένες σε ένα χρονικά μεταβαλλόμενο δίκτυο, όπως για παράδειγμα στο Facebook. Το Facebook αποτελεί χαρακτηριστικό παράδειγμα ενός χρονικά μεταβαλλόμενου Κοινωνικού Δικτύου, όπου η αναπαράσταση του είναι ένας μη-κατευθυνόμενος γράφος με κόμβους, οι οποίοι εκπροσωπούν τους χρήστες του και συνδέσεις, οι οποίες εκπροσωπούν τις σχέσεις/φιλίες μεταξύ των χρηστών.

Για παράδειγμα, θα ήταν χρήσιμο *να αξιολογηθεί μία προσπάθεια ανεύρεσης μίας πιθανής αυτοματοποιημένης πρότασης εμφάνισης μίας Διαφημιστικής Καμπάνιας σε έναν χρήστη* (του Facebook, για παράδειγμα) *δεδομένου των συνδέσεων που έχει ο χρήστης με άλλους χρήστες και του γεγονότος ότι έχει ανταποκριθεί θετικά σε μία διαφημιστική καμπάνια*, με στόχο την αύξηση της απόδοσης της διαφημιστικής προσπάθειας υπό το πρίσμα της αποτελεσματικότητας της διαφημιστικής προσπάθειας από όλα τα μέρη.

- Υπό το πρίσμα ωφέλειας του χρήστη, μία σωστή, βάση των ενδιαφερόντων του, διαφημιστική πρόταση, αυξάνει το ποσοστό αφοσίωσης (loyalty) του στις υπηρεσίες του Κοινωνικού Δικτύου, καθώς λαμβάνει ωφέλεια κατά τη χρήση των υπηρεσιών του.
- Υπό το πρίσμα ωφέλειας της επιχείρησης, μία διαφημιστική προσπάθεια που αποδίδει σε ένα συγκεκριμένο οικονομικό και χρονικό πλαίσιο, αυξάνει το ποσοστό αφοσίωσης της επιχείρησης στη χρήση των υπηρεσιών του Κοινωνικού Δικτύου, καθώς αυξάνουν τα κέρδη της και μειώνονται οι δαπάνες της.

- Υπό το πρίσμα ωφέλειας του ίδιου του Ιστότοπου ως Κοινωνικό Δίκτυο, όσο αυξάνει το ποσοστό αφοσίωσης των χρηστών του (πρόσωπο, επιχείρηση, οργανισμός), τόσο δύναται να αυξάνουν τα ποσοστά κέρδους του.

2.5.2 Μαθηματική Περιγραφή του Προβλήματος - Mathematical Description

Η αντιμετώπιση του Προβλήματος της Εύρεσης Συνδέσμων μπορεί να γίνει μέσω δύο διαφορετικών προσεγγίσεων, με εποπτεία (Supervised Link Prediction) ή χωρίς εποπτεία (Unsupervised Link Prediction). Στην πρώτη περίπτωση επιθυμούμε να εξάγουμε γνώση από το δίκτυο με στόχο την πρόβλεψη, ενώ στη δεύτερη περίπτωση επιθυμούμε να ορίσουμε ένα σύνολο μέτρων ακρίβειας. Η κάθε προσέγγιση, χρησιμοποιεί διαφορετικούς αλγορίθμους κατάλληλους για την επίτευξη των στόχων του προβλήματος.

Δεδομένου ότι έχουμε ένα Κοινωνικό Δίκτυο $G = <V,E>$ όπου κάθε σύνδεσμος (edge) $e = <x,y> \epsilon E$ (όπου E, ένα σύνολο παρατηρούμενων συνδέσμων) αναπαριστά μία συσχέτιση ανάμεσα στον κόμβο x και y σε μία συγκεκριμένη χρονική στιγμή $t(e)$. Ο στόχος είναι η εύρεση, μέσω πρόβλεψης, της πιθανότητας ενός μη-παρατηρούμενου συνδέσμου e_{xy} να υπάρχει, σε μία μελλοντική στιγμή.

Πιο αναλυτικά : Καταγράφουμε τις πολλαπλές αλληλεπιδράσεις ανάμεσα στους κόμβους x και y σε διαφορετικές τιμές του χρόνου. Για δύο διαφορετικές χρονικές στιγμές $t < t'$, θεωρούμε ότι το $G[t, t']$ υποδηλώνει ένα υπό-γράφημα του G το οποίο αποτελείται από όλους τους συνδέσμους σε κάποια δεδομένη χρονική στιγμή ανάμεσα στο t και t'. Στη συνέχεια, διαχωρίζουμε τα δεδομένα σε K υπό-διαστήματα (subsets). Κάθε χρονική στιγμή ένα από τα K υπό-διαστήματα επιλέγεται ως *διάστημα εκπαίδευσης* (training set) E_{train} και τα υπόλοιπα $K - 1$ ανήκουν στο *διάστημα δοκιμής* (test set) E_{test}. Η διαδικασία αυτή υλοποιείται K φορές και όλα τα υπό-διαστήματα χρησιμοποιούνται μία ακριβώς φορά στο E_{train}. Έστω ότι έχουμε τις χρονικές περιόδους $t_0 < t_0' < t_1 < t_1'$ και εφαρμόζουμε τον επιλεγμένο αλγόριθμο στο δίκτυο $G[t_0,t_0']$. Η έξοδος του αλγορίθμου θα φέρει μία λίστα συνδέσμων, οι οποίοι δεν υπάρχουν στο $G[t_0,t_0']$, αλλά αποτελούν πρόβλεψη των συνδέσμων που θα σχηματιστούν στο $G[t_1,t_1']$. Αναφερόμαστε στο χρονικό διάστημα $E_{train} = [t_0,t_0']$ ως διάστημα εκπαίδευσης και στο $E_{test} = [t_1,t_1']$, ως

διάστημα δοκιμής. Το σύνολο των συνδέσμων E, ορίζεται ως : $E = E_{train}\ U\ E_{test}$ ενώ $E_{train} \cap E_{test} = \emptyset$. Εφαρμόζουμε τον επιλεγμένο αλγόριθμο Εύρεσης Συνδέσμων στο διάστημα εκπαίδευσης E_{train} και στη συνέχεια ελέγχουμε την απόδοση του στο διάστημα δοκιμής E_{test}. Ως έξοδος της υλοποίησης του αλγορίθμου, προκύπτει μία λίστα ανύπαρκτων (προβλεπόμενων προς δημιουργία) συνδέσμων L φθίνουσας κατάταξης, $L : e_L \in U - E_{train}$. Όπου U, ορίζεται το σύνολο των δυνατών συνδέσμων στο γράφημα, $|U| = (|V|\ (|V| - 1)) / 2$. Για να δημιουργηθεί η λίστα L, χρησιμοποιούμε ευρετικούς αλγόριθμους οι οποίοι εκχωρούν έναν πίνακα ομοιότητας S του οποίου η πραγματική είσοδος s_{xy} είναι η Βαθμολογία (score) μεταξύ x και y, $Score\ (x,y)$. Αυτή η βαθμολογία μπορεί να θεωρηθεί ως μέτρο *ομοιότητας* (Proximity or Similarity measure) μεταξύ των κόμβων x και y. Για κάθε ζεύγος κόμβων $x, y \in V$ ισχύει $s_{xy} = s_{yx}$. Στην περίπτωση που η προσέγγιση του προβλήματος γίνεται χωρίς εποπτεία, δεν απαιτείται η διαδικασία μεταφοράς των δεδομένων στο διάστημα εκπαίδευσης.

Για την αξιολόγηση της αποδοτικότητας του αλγορίθμου, χρησιμοποιούνται ευρέως, δύο είδη μετρήσεων (metrics) που εφαρμόζουν στις αρχές της ομοιότητας (similarity) : Το μέτρο AUC (area under the receiver operating characteristic curve) και το μέτρο της Ακρίβειας (Precision) (Linyuan Lü & Tao Zhou, 2010). Πιο αναλυτικά :

i. AUC - Χρησιμοποιείται για την αξιολόγηση των μη-παρατηρούμενων συνδέσμων. Μπορεί να ερμηνευτεί ως η πιθανότητα ένας τυχαία επιλεγμένος χαμένος σύνδεσμος (δηλ. ένας σύνδεσμος που ανήκει στο E_{test}) να βαθμολογηθεί υψηλότερα από έναν τυχαία επιλεγμένο ανύπαρκτο σύνδεσμο (δηλ. ένας σύνδεσμος που ανήκει στο $U - E$). Υπολογίζει μία βαθμολογία για κάθε μη-παρατηρούμενο σύνδεσμο και όχι για ολόκληρη τη λίστα συνδέσμων, L. Η AUC ορίζεται ως ο λόγος $(n' + 0.5n'') / n$, όπου n ο αριθμός των ανεξάρτητων συγκρίσεων, n' ο αριθμός των συγκρίσεων όπου ο χαμένος σύνδεσμος έχει υψηλότερη βαθμολογία, και n'' ο αριθμός συγκρίσεων με την ίδια βαθμολογία.

ii. Ακρίβεια (Precision) - Δεδομένης της λίστας κατάταξης των μη παρατηρούμενων συνδέσεων L, η *Ακρίβεια* ορίζεται ως ο λόγος των σχετικών στοιχείων που έχουν επιλεγεί ως προς τον αριθμό των επιλεγμένων στοιχείων. Συγκεκριμένα, αν επιλέξουμε ως πρόβλεψη, τους κορυφαίους συνδέσμους της λίστας L σε σχέση με τους συνδέσμους που

ανήκουν στη λίστα L_{test} (ως L_{test} θεωρούμε τη λίστα των συνδέσμων του E_{test}), η *Ακρίβεια*, ορίζεται ως ο λόγος L_{test} / L. Όσο μεγαλύτερη είναι η τιμή της *Ακρίβειας*, τόσο ορθότερα χαρακτηρίζονται τα αποτελέσματα.

Για την αξιολόγηση της απόδοσης κάθε αλγορίθμου, στην εργασία αυτή, θα λάβουμε υπόψη την μετρική AUC, για την εξαγωγή των συμπερασμάτων.

Εάν όλες οι βαθμολογίες δημιουργούνται από μια ανεξάρτητη και ταυτόσημη κατανομή, η τιμή της AUC θα πρέπει να είναι περίπου 0,5. Επομένως, ο βαθμός στον οποίο η ακρίβεια υπερβαίνει το 0,5 αποδεικνύει πόσο καλύτερα αποδίδει ο αλγόριθμος από ότι η καθαρή τύχη.

3 Μεθοδολογία

Σε αυτό το σημείο της εργασίας, θα γίνει μία σύντομη ανασκόπηση των εναλλακτικών επιλεγμένων μεθόδων υλοποίησης του Προβλήματος Εύρεσης Συνδέσμων καθώς και των αντίστοιχων πλεονεκτημάτων και μειονεκτημάτων τους. Θα αναπτυχθεί περαιτέρω μία εναλλακτική αναθεωρημένη πρόταση εφαρμογής των εν λόγω αλγορίθμων, υπό τη σκοπιά της εργασίας μας, με την προσθήκη μίας τιμής βάρους στην τιμή του χρόνου, ώστε να προκύψουν συμπεράσματα αναφορικά με την βαρύτητα που μπορεί να προσδώσει στα αποτελέσματα υλοποίησης ο χρονικός παράγοντας υπό τη σκοπιά βελτίωσης των ενεργειών που λαμβάνουν χώρα στα Κοινωνικά Δίκτυα Business Oriented ενέργειες προώθησης και διαφήμισης.

Οι αλγόριθμοι που θα χρησιμοποιηθούν υπακούουν στην αρχή της Ομοιότητας (Similarity) τόσο σε τοπικό όσο και σε καθολικό επίπεδο, λαμβάνοντας υπόψη τη χρονική πληροφορία ως παράγοντα διαμόρφωσης και επιρροής της απόδοσης των αποτελεσμάτων.

3.1 Ομοιότητα - Similarity

Ο όρος Ομοιότητα (Similarity) ορίζεται ως «ένα μέτρο ομοιότητας ή ανομοιότητας ανάμεσα σε δύο ή περισσότερα χαρακτηριστικά στα δεδομένα μίας βάσης» (Sarjon Defit & Mohd Noor Md Sap, 2010). Χρησιμοποιείται σε μοντέλα

Πρόβλεψης και θεωρείται μία σημαντική μέθοδος εφαρμογής στον κλάδο της Εξόρυξης Γνώσης Δεδομένων. Δεδομένου του γεγονότος ότι δεν μπορούν όλα τα χαρακτηριστικά σε μία βάση να χρησιμοποιηθούν για μία πρόβλεψη, είναι σημαντικό να αναγνωριστούν και να εκμαιευτούν εκείνα τα χαρακτηριστικά που έχουν είτε ισχυρή είτε αδύναμη ομοιότητα.

Όπως αναφέρθηκε σε προηγούμενο κεφάλαιο, ένα μέτρο ομοιότητας (similarity metrics) που χρησιμοποιείται για την απόδοση της χρηστικότητας ενός μοντέλου πρόβλεψης, είναι η χρήση βαθμολόγησης του ζεύγους *x,y* κόμβων και ορίζεται ως $S_{xy,}$ ο βαθμός ομοιότητας μεταξύ των κόμβων *x,y*. Όσο υψηλότερη είναι η τιμή του βαθμού, τόσο αποδοτικότερος κρίνεται ο αλγόριθμος υλοποίησης.

3.2 Αλγόριθμοι εφαρμογής στο Πρόβλημα Εύρεσης Συνδέσμων - Link Prediction Algorithms

Οι αλγόριθμοι που θα εφαρμόσουμε για το Πρόβλημα Εύρεσης Συνδέσμων βασίζονται σε δύο διαφορετικές προσεγγίσεις της θεωρίας Ομοιότητας : Ομοιότητα σε τοπικό επίπεδο (Local Similarity) και Ομοιότητα σε καθολικό επίπεδο (Global Similarity). Θα αναφερθούμε στα πλεονεκτήματα της κάθε προσέγγισης και της δυναμικής τους στην επίλυση του προβλήματος μας.

3.2.1 Αλγόριθμοι Ομοιότητας σε Τοπικό Επίπεδο (Local Similarity algorithms)

Η προσέγγιση των αλγορίθμων Ομοιότητας σε Τοπικό Επίπεδο, αφορά στη χρήση ομάδων από γειτονικούς κόμβους (neighbors) και εφαρμόζουν την ιδέα ότι δύο κόμβοι *x,y* είναι πιθανότερο να σχηματίσουν μία σύνδεση (relationship) στο μέλλον, εφόσον οι γειτονικοί τους κόμβοι έχουν μεγάλο αριθμό κοινών κόμβων.

Το πλεονέκτημα των αλγορίθμων αυτών, είναι ότι δύναται να μετρήσουν με ακρίβεια την απόδοση ομοιότητας των χαρακτηριστικών ανάμεσα σε δύο κόμβους. Το μειονέκτημά τους είναι ότι δεν καταφέρνουν να ανταποκριθούν άριστα σε ζητήματα όπου παρουσιάζεται έλλειψη δεδομένων. Παρακάτω, παρουσιάζονται συνοπτικά οι αλγόριθμοι που θα μας απασχολήσουν σε επόμενο κεφάλαιο.

3.2.1.1 Κοινοί Γείτονες - Common Neighbors

Η προσέγγιση του αλγορίθμου Κοινών Γειτόνων (Newman, 2001), βασίζεται στην ιδέα ότι όσο περισσότεροι είναι οι γείτονες ανάμεσα σε δύο κόμβους *x,y*, τόσο μεγαλύτερη είναι και η πιθανότητα μίας μελλοντικής σύνδεσης ανάμεσα στους κόμβους *x,y*. Ο υπολογισμός του μέτρου της ομοιότητας των συνδέσμων ενός μη κατευθυνόμενου γράφου, ορίζεται ως:

$$Similarity(x,y) = |N(x) \cap N(y)|$$

3.2.1.2 Ο Συντελεστής Jaccard - Jaccard's Coefficient

Η προσέγγιση του Παράγοντα του Jaccard (Suphakit Niwattanakul et al, 2013), βασίζεται στην ιδέα του υπολογισμού της πιθανότητας ενός τυχαίου κόμβου *z* να είναι γείτονας και των δύο κόμβων *x,y*, εάν είναι γειτονικός κόμβος τουλάχιστον ενός εκ των δύο *x* ή *y*. Κοινοί σύνδεσμοι / Σύνολο συνδέσμων. Ο υπολογισμός του μέτρου της ομοιότητας των συνδέσμων του γράφου, ορίζεται ως:

$$Similarity(x,y) = |N(x) \cap N(y)| \ / \ |N(x) \ U \ N(y)|$$

3.2.1.3 Δείκτης Adamic/ Adar - Adamic/Adar Index

Η προσέγγιση αυτή (Adamic Lada A. & Adar Eytan, 2003), βασίζεται στην ιδέα ότι ένας κοινός κόμβος των κόμβων *x,y* με χαμηλό βαθμό θα συνεισφέρει περισσότερο σε μία μελλοντική σύνδεση ανάμεσα στους κόμβους *x* και *y*, συγκρινόμενος με έναν άλλο κόμβο με υψηλό βαθμό. Ο υπολογισμός του μέτρου της ομοιότητας των συνδέσμων του γράφου, ορίζεται ως:

$$Similarity(x, y) = \Sigma_{z \in |N(x) \cap N(y)|} \ (1 / \log |\Gamma(z)|)$$

όπου $\Gamma(z)$: δηλώνεται το σύνολο των γειτόνων του κόμβου *z*

3.2.1.4 Προτιμώμενη Προσκόλληση - Preferential Attachment

Η προσέγγιση αυτής της μεθόδου (Barabasi et al, 2008), βασίζεται στην ιδέα ότι η πιθανότητα ύπαρξης ενός συνδέσμου ανάμεσα στους κόμβους *x,y* είναι ανάλογη του βαθμού που κατέχουν οι κόμβοι *x,y*. Η υλοποίηση αυτής της μεθόδου είναι εξαρτώμενη μόνο από τους κόμβους για τους οποίους αναζητείται ο μεταξύ τους σύνδεσμος. Ο υπολογισμός του μέτρου της ομοιότητας των συνδέσμων του γράφου, ορίζεται ως:

$Similarity(x, y) = |\Gamma(x)| * |\Gamma(y)|$

όπου $\Gamma(x), \Gamma(y)$: δηλώνεται το σύνολο των γειτόνων των κόμβων *x,y*

3.2.1.5 Δείκτης Κατανομή Πόρων - Resource Allocation Index

Η προσέγγιση αυτής της μεθόδου (Zhou & Lü & Zhang, 2009), βασίζεται στην ιδέα ότι για τους μη-συνδεδεμένους κόμβους *x,y*, ένας εκ των δύο κόμβων, έστω ο *x* μπορεί να διανέμει κάποιους πόρους στον άλλο κόμβο, στην περίπτωση αυτή στον *y* μέσω των κοινών τους γειτόνων. Θεωρείται ότι κάθε κόμβος έχει έναν πόρο μόνο, τον οποίο εκχωρεί ομοιόμορφα στους γείτονες του. Ο υπολογισμός του μέτρου της ομοιότητας των συνδέσμων του γράφου, ορίζεται ως:

$Similarity(x, y) = \Sigma_{z \in | N(x) \cap N(y) |} \ (1/ |\Gamma(z)|)$

όπου $\Gamma(z)$: δηλώνεται το σύνολο των γειτόνων του κόμβου *z*

3.2.2 Αλγόριθμοι Ομοιότητας σε Καθολικό Επίπεδο (Global Similarity algorithms)

Η προσέγγιση των αλγορίθμων Ομοιότητας σε Καθολικό Επίπεδο, αφορά στην ιδέα της καθολικής συμβολής, δηλαδή της συμμετοχής όλων των κόμβων στην δημιουργία συνδέσμου μεταξύ δύο κόμβων (Nowell & Kleinberg, 2007).

Πλεονεκτούν σε προβλήματα όπου αντιμετωπίζεται το ζήτημα της έλλειψης δεδομένων καθώς το λαμβάνουν υπόψη και υιοθετούν μέτρα διάδοσης της ομοιότητας

(propagating similarity measurement), καθώς αναζητούν περισσότερο όμοιους κόμβους στην περίπτωση αυτή. Παρακάτω, παρουσιάζονται συνοπτικά οι αλγόριθμοι που θα μας απασχολήσουν σε επόμενο κεφάλαιο.

3.2.2.1 Τυχαία περιήγηση με επανεκκίνηση - Random Walk with Restart

Η προσέγγιση αυτής της μεθόδου (Weiping Liu & Linyuan Lu, 2010), βασίζεται στην πιθανότητα ότι ένας κόμβος θα "επισκεφτεί" τον γειτονικό του. Ο υπολογισμός του μέτρου της ομοιότητας των συνδέσμων του γράφου, ορίζεται ως:

$Similarity(x, y) = (\Gamma(x) * \Gamma(y)) / 2 * |E| * |E|$

όπου $|E|$: ο αριθμός των συνδέσμων (edges) στο γράφημα.

3.2.2.2 Εύρεση της Συντομότερης Διαδρομής - Graph Distance, Shortest Path

Η προσέγγιση αυτής της μεθόδου, βασίζεται στην ιδέα ότι οι φίλοι ενός φίλου μπορούν ευκολότερα να γίνουν φίλοι μεταξύ τους. Επομένως, αναζητείται το μήκος της συντομότερης διαδρομής ανάμεσα σε ζεύγη κόμβων *x,y* στο γράφημα G και ορίζεται ως η αρνητική τιμή της συντομότερης απόστασης ανάμεσα στους κόμβους *x,y*. Όσο συντομότερη είναι μία διαδρομή, τόσο υψηλότερη είναι η πιθανότητα να δημιουργηθεί ένας σύνδεσμος. Ο υπολογισμός του μέτρου της ομοιότητας των συνδέσμων του γράφου, ορίζεται ως:

$Similarity(x, y) = |L_{path}(x,y)|$

3.2.2.3 Δείκτης Katz (Exponentially Damped Path Counts)

Η προσέγγιση αυτής της μεθόδου (Katz, 1953), αποτελεί μία εναλλακτική τοποθέτηση της εύρεσης συντομότερης διαδρομής και βασίζεται στην ιδέα ότι όσοι περισσότεροι είναι οι συνδεδεμένοι κόμβοι μεταξύ τους, και όσο μικρότερη είναι η διαδρομή, τόσο ισχυρότερη είναι η σύνδεση μεταξύ τους. Ο υπολογισμός του μέτρου της ομοιότητας των συνδέσμων του γράφου, ορίζεται ως:

$Similarity(x, y) = \Sigma_{i-1 \text{ έως } \infty} \beta^l * |Path^l_{x,y}| = \beta*A + \beta^2*A^2 + \beta^3*A^3 + ..$

Η μεταβλητή l δηλώνει το μήκος του μονοπατιού. Ο συντελεστής β^l έχει μικρή τιμή και δρα επιβραβεύοντας τις διαδρομές με μικρό μήκος, ενώ επιβαρύνει τις διαδρομές με μεγάλο μήκος. $Path^l_{x,y}$ είναι το σύνολο των μονοπατιών μήκους l ανάμεσα στους κόμβους x και y.

Στην πειραματική μας μελέτη θα χρησιμοποιήσουμε την τιμή $\beta^l = 0.1$

3.2.2.4 Δείκτης FriendLink

Η προσέγγιση αυτής της μεθόδου (Papadimitriou & Symeonidis & Manolopoulos, 2012), βασίζεται στην ιδέα της ύπαρξης ομοιοτήτων μεταξύ δύο κόμβων σε ένα μη κατευθυνόμενο γράφημα. Ο δείκτης χρησιμοποιεί ως είσοδο τις συνδέσεις ενός γραφήματος G και εξάγει μια μήτρα ομοιότητας μεταξύ οποιωνδήποτε δύο κόμβων στο G. Ο υπολογισμός του μέτρου της ομοιότητας των συνδέσμων του γράφου, ορίζεται ως:

$$Similarity(x,y) = \Sigma^{l}_{i=1} (1 / i-1)) * |paths^{i}_{x,y}| / \Pi^{i}_{j=2} (n - j)$$

όπου, n : ο αριθμός των κόμβων, l : το μέγιστο προς εξέταση μήκος μονοπατιού και $paths^{i}_{x,y}$: το σύνολο των μονοπατιών μήκους i μεταξύ των κόμβων x,y.

3.3 Προτεινόμενη αναθεωρημένη μέθοδος

Το Πρόβλημα της Εύρεσης συνδέσμων, όπως προαναφέρθηκε, σύμφωνα με τη βιβλιογραφία, έχει αντιμετωπιστεί κυρίως ως ένα στατικό από άποψη χρόνου πρόβλημα καθώς τα περισσότερα μοντέλα δικτύου είναι στατικά και δεν λαμβάνουν υπόψη τους ότι κατά τη διέλευση του χρόνου πιθανόν να δημιουργηθήκαν νέοι σύνδεσμοι μεταξύ των κόμβων που είναι δυνατόν να *αναθεωρήσουν* την προς πρόβλεψη σύνδεση κόμβων.

3.3.1 Αναθεώρηση πρόβλεψης σε αλγορίθμους Ομοιότητας

Η δική μας προσέγγιση, στηρίζεται στη λογική των αλγορίθμων ομοιότητας αλλά λαμβάνει υπόψη την πιθανή αλλαγή της μελλοντικής πρόβλεψης, λόγω χρονικής παλαιότητας ή μη ως προς τη δημιουργία νέου συνδέσμου με κόμβο, ο οποίος δύναται να «επηρεάσει» την εξέλιξη των αποτελεσμάτων πρόβλεψης.

Θα χρησιμοποιήσουμε τη λογική της εφαρμογής ενός penalty, μία πολύ μικρή τιμή, στους συνδέσμους εκείνους με την υψηλότερη κατάταξη στη λίστα *L*, σε διαφορετικές χρονικές στιγμές, $t_0 < t_1 < t_2$, ώστε να επηρεαστούν οι παλαιότερα δημιουργημένες συνδέσεις με στόχο να αποδειχθεί ή όχι αν ο διερχόμενος χρόνος επηρεάζει τη δυναμική επιρροής μίας νέας σύνδεσης ή όχι.

Η λογική της προτεινόμενης μεθόδου, απορρέει από την εμπειρική παρατήρηση[1] της μειωμένης δυναμικής αποτελεσματικότητας σε Καμπάνιες Προώθησης σε Κοινωνικά Δίκτυα, στους χρήστες των οποίων οι φιλίες είναι παλαιότερες στο χρόνο, καθώς τείνουν να επηρεάσουν λιγότερο στη δημιουργία νέων συνδέσεως για τον στοχευμένο χρήστη.

Πιο συγκεκριμένα : Έστω, οι συνδεδεμένοι κόμβοι x,y. Θεωρούμε ότι κατά τη διέλευση του χρόνου, $t_0 < t_1 < t_2$, η δυναμική επιρροής του κόμβου y ως προς τη δημιουργία νέων συνδέσμων για τον κόμβο x φθίνει. Θα εφαρμόσουμε penalties στον πίνακα ομοιοτήτων για τις συνδέσεις που δημιουργήθηκαν παλαιότερα ώστε να δούμε αν επηρεάζονται οι τιμές των μετρικών θετικά, αρνητικά ή και καθόλου.

Η πειραματική μελέτη που έχει διεξαχθεί, θα εφαρμοστεί εκ νέου για τους ίδιους αλγόριθμους ομοιότητας σε τοπικό και καθολικό επίπεδο υπό την επίδραση της τιμής Penalty = 0.0002 που θα εφαρμοστεί στις τιμές της λίστας *L* με την υψηλότερη κατάταξη.

Αναμένουμε να δούμε αν οι τιμές του δείκτη ομοιότητας AUC θα είναι βελτιωμένες ή όχι.

[1] Στατιστικά αποτελέσματα καμπανιών προώθησης σε Κοινωνικά Δίκτυα

4 Πειραματική Μελέτη

Για τον έλεγχο της ορθότητας των προσπαθειών ανάδειξης μίας μεθοδολογίας που μπορεί να απαντήσει στα ερώτημα που απασχολούν την εργασίας μας αναφορικά με το Πρόβλημα Εύρεσης Συνδέσμων, υλοποιήσαμε μία σειρά από πειραματικές μελέτες χρησιμοποιώντας τους αλγόριθμους εξόρυξης δεδομένων που αναφέρονται στο προηγούμενο κεφάλαιο.

Στόχος μας παραμένει να απαντηθεί με ποιο τρόπο το Πρόβλημα Εύρεσης Συνδέσμων αποδίδει καλύτερα για τα μέλη των Κοινωνικών Δικτύων αλλά και εάν αποδεικνύεται σημαντικός ο χρονικός παράγοντας κατά το σχεδιασμό Στρατηγικών, λύσεων και αποφάσεων, σε ποιο βαθμό και με ποιο τρόπο.

4.1 Σύνολα Δεδομένων Κοινωνικών Δικτύων

Στο κεφάλαιο αυτό παρουσιάζουμε τα αποτελέσματα των πειραματικών ενεργειών που βασίστηκαν σε δύο μεγάλα Σύνολα Δεδομένων (Datasets) των Κοινωνικών Δικτύων Facebook & Twitter. Η πηγή των δεδομένων ανήκει στην έρευνα που διενέργησε εκ μέρους του Stanford University[2] ο (Jure Leskovec, 2012) και τα δεδομένα της αφορούν σε εγωκεντρικά-δίκτυα (ego-networks). Οι σύνδεσμοι των κόμβων στο εγωκεντρικό δίκτυο του Facebook είναι μη- κατευθυνόμενοι (undirected) ενώ στο Twitter είναι κατευθυνόμενοι (directed) (καθώς ισχύει ότι, ο κόμβος *x* ακολουθεί το *y*).

4.1.1 Σύνολο Δεδομένων του Facebook

Το Σύνολο Δεδομένων του Κοινωνικού Δικτύου - Facebook αποτελείται από 'κύκλους' ή 'λίστες συνδέσμων' (circles or friendship lists). Τα δεδομένα προήλθαν από έρευνα που διενεργήθηκε σε χρήστες της εφαρμογής, Facebook[3]. Το Σύνολο των Δεδομένων περιλαμβάνει χαρακτηριστικά κόμβων (που αντιστοιχούν σε προφίλ χρηστών), κύκλους και εγωκεντρικά δίκτυα. Τα δεδομένα, στη συνέχεια, μετατράπηκαν

[2] Διαθέσιμο Online στη διεύθυνση : snap.stanford.edu

[3] Διαθέσιμη Online στη διεύθυνση : https://www.facebook.com/home.php

σε ανώνυμες τιμές πληροφορίας για λόγους δεοντολογίας. Παρότι κατεγράφησαν χαρακτηριστικά ιδιοτήτων των χρηστών, η ερμηνεία τους έχει αλλάξει, αλλά οι χρήστες που έχουν μία κοινή ιδιότητα μπορούν να καθοριστούν.

Από τη Βάση Δεδομένων του Facebook εξήλθαν δεδομένα από 10 εγωκεντρικά-δίκτυα, αποτελούμενα από 193 κύκλους (circles) και 4.039 χρήστες (users). Το σύνολο των συνδέσμων μεταξύ των χρηστών είναι 88.234 (πίνακας 4.1.1).

Πίνακας 4.1.1.1 Σύνολο Δεδομένων Facebook

Πληροφορίες Συνόλου Δεδομένων του Facebook	
Κόμβοι (Nodes)	4.039
Σύνδεσμοι (Edges)	88.234
Μέσος Συντελεστής Ομαδοποίησης (Average clustering coefficient)	0,6055
Αριθμός Τριγώνων (Number of triangles)	1.612.010
Κλειστά Τρίγωνα (Fraction of closed triangles)	0,2647
Μήκος Μεγαλύτερης Διαδρομής (Diameter)	8
90-εκατοστιαία αποτελεσματικότητα διαμέτρου (90-percentile effective diameter)	4,7

Για λόγους χρονικών και τεχνολογικών περιορισμών, επιλέχθηκε ένα μικρό και ένα μεγαλύτερο διάστημα του Συνόλου Δεδομένων του Facebook για τις πειραματικές μελέτες. Το μικρό διάστημα αποτελείται από 547 (χρήστες) κόμβους που αντιστοιχούν σε 9.626 συνδέσμους. Ενώ για το μεγαλύτερο διάστημα επιλέχθηκε ένα σύνολο από 1.856 κόμβους που αντιστοιχούν σε 46.024 συνδέσμους.

4.1.2 Σύνολο Δεδομένων Twitter

Το Σύνολο Δεδομένων του Κοινωνικού Δικτύου - Twitter αποτελείται από 'κύκλους' ή 'λίστες' (circles or lists). Τα δεδομένα ανιχνεύτηκαν από δημόσιες πηγές. Το Σύνολο των Δεδομένων περιλαμβάνει χαρακτηριστικά κόμβων (προφίλ), κύκλους, και εγωκεντρικά δίκτυα.

Από τη Βάση Δεδομένων του Twitter εξήλθαν δεδομένα από 1.000 εγωκεντρικά-δίκτυα, αποτελούμενα από 4.869 κύκλους (λίστες [10,19,27,31]) και 81.362 χρήστες (users). Το σύνολο των συνδέσμων μεταξύ των χρηστών είναι 1.768.149 (πίνακας 4.2).

Πίνακας 4.1.2.1 Σύνολο Δεδομένων Twitter

Πληροφορίες Συνόλου Δεδομένων του Twitter	
Κόμβοι (Nodes)	81.306
Σύνδεσμοι (Edges)	1.768.149
Μέσος Συντελεστής Ομαδοποίησης (Average clustering coefficient)	0,5653
Αριθμός Τριγώνων (Number of triangles)	13.082.506
Κλειστά Τρίγωνα (Fraction of closed triangles)	0,06415
Μήκος Μεγαλύτερης Διαδρομής (Diameter)	7
90-εκατοστιαία αποτελεσματικότητα διαμέτρου (90-percentile effective diameter)	4,5

Για λόγους χρονικών και τεχνολογικών περιορισμών, επιλέχθηκε ένα μικρό και ένα μεγαλύτερο διάστημα του Συνόλου Δεδομένων του Twitter για τις πειραματικές μελέτες. Το μικρό διάστημα αποτελείται από 246 κόμβους που αντιστοιχούν σε 9.630 συνδέσμους, ενώ το μεγαλύτερο διάστημα αποτελείται από 3.478 κόμβους και 51.931 συνδέσμους.

4.2 Βήματα Πειραματικών Μελετών

Για τις πειραματικές μελέτες χρησιμοποιήθηκαν τα υποσύνολα των Συνόλων Δεδομένων των Κοινωνικών Δικτύων Facebook & Twitter. Το Σύνολο Δεδομένων του Facebook είναι ένας μη κατευθυνόμενα (undirected), ενώ του Twitter είναι κατευθυνόμενα (directed). Σημείο έναρξης της διαδικασίας, ορίζεται η χρονική στιγμή *t=0* που αφορά στα δεδομένα του γραφήματος *G* των Κοινωνικών Δικτύων. Κατά τη διέλευση του χρόνου, θεωρούμε, ότι έχουν δημιουργηθεί νέοι σύνδεσμοι μεταξύ των κόμβων του γραφήματος *G*.

Για την υλοποίηση των πειραματικών μελετών, ακολουθήσαμε την ακόλουθη διαδικασία Εξόρυξης Γνώσης:

1. Προεπεξεργαζόμαστε το Σύνολο Δεδομένων του κάθε Κοινωνικού Δικτύου και διαχωρίζουμε τους συνδέσμους (edges) που προϋπήρχαν τη χρονική στιγμή *t=0*. Τα δεδομένα που αφορούν στους υπόλοιπους συνδέσμους θεωρούμε ότι δημιουργήθηκαν κατά τη διέλευση του χρόνου στη χρονική στιγμή *t'*. Το εν λόγω Σύνολο Δεδομένων είναι μία λίστα συνδέσμων (edge list - graph).
2. Μοιράζουμε τη λίστα συνδέσμων σε *διάστημα εκπαίδευσης* E_{train} (Training Set) και *διάστημα δοκιμής* E_{test} (Test set), επιλέγοντας 50% του Συνόλου Δεδομένων σε κάθε περίπτωση (και για τα δύο Σύνολα Δεδομένων G_s, G_b)
3. Εφαρμόζουμε τους (ανωτέρω αναφερόμενους) αλγόριθμους στο *διάστημα εκπαίδευσης* E_{train} και στη συνέχεια ελέγχουμε την απόδοση του στο *διάστημα δοκιμής* E_{test}. Υπολογίζουμε τον πίνακα ομοιότητας για κάθε ζεύγος κόμβων x,y (graph).
4. Χρησιμοποιούμε τον πίνακα ομοιότητας για να προτείνουμε τους k (k=5) top κόμβους ως "άτομα που ίσως γνωρίζει" ο χρήστης που βρίσκεται στο *διάστημα δοκιμής* E_{test} (Test set).
5. Αξιολογούμε την αποτελεσματικότητα του αλγορίθμου με τη χρήση του δείκτη ομοιότητας *AUC*.

Τελικό βήμα της διαδικασίας, αποτελεί η συγκριτική μελέτη μεταξύ των διαφορετικών επιλεγμένων αλγορίθμων για την ανάδειξη των αποτελεσμάτων.

4.3 Αποτελέσματα Πειραματικών Μελετών σε Facebook & Twitter

Στο σημείο αυτό θα παρουσιάσουμε τα αποτελέσματα των πειραματικών διεργασιών για τα δύο διαφορετικά Σύνολα Δεδομένων των Κοινωνικών Δικτύων Facebook & Twitter. Το Σύνολο Δεδομένων του Facebook είναι ένας μη κατευθυνόμενος γράφος, ενώ του Twitter είναι ένας κατευθυνόμενος γράφος.

Οι συγκριτικές μελέτες υλοποιηθήκαν υπό το πρίσμα δύο διαφορετικών κατευθύνσεων με στόχο να παρουσιαστεί εάν η δυναμική του χρόνου μπορεί να επηρεάσει τα αποτελέσματα και ιδιαίτερα αν είναι ικανή να προσφέρει βελτιώσεις στις Μεθόδους Εξόρυξης Γνώσης για το Πρόβλημα Εύρεσης Συνδέσμων.

Λόγω χρονικών και τεχνολογικών περιορισμών χρησιμοποιήσαμε ένα μέρος των δεδομένων των Κοινωνικών Δικτύων σε ένα πολύ μικρό και σε ένα μεγαλύτερο διάστημα. Δεν επιτεύχθηκε η υλοποίηση της μελέτης στο σύνολο των δεδομένων.

4.3.1 Α' κύκλος Πειραματικών Μελετών

Στον Α' κύκλο πειραματικών μελετών εφαρμόστηκαν οι προαναφερόμενοι (Κεφάλαιο 3.2) αλγόριθμοί Ομοιότητας σε Τοπικό (Local Similarity) και Καθολικό επίπεδο (Global Similarity) στα δύο υποσύνολα (G_s: το μικρό σύνολο δεδομένων, G_b: το μεγαλύτερο σύνολο δεδομένων) των δύο Συνόλων Δεδομένων των Κοινωνικών Δικτύων Facebook & Twitter.

Ως δείκτες απόδοσης των τεχνικών που εφαρμόστηκαν, επιλέχθηκε η τιμή του δείκτη *AUC*. Όπως προαναφέρθηκε, επιλέχθηκε να εφαρμοστούν οι μελέτες σε ένα μικρό G_s (αρχικά) και ένα μεγαλύτερο G_b (στη συνέχεια) υποδιάστημα των Συνόλων Δεδομένων των δύο Κοινωνικών Δικτύων.

4.3.1.1 Δείκτης AUC - Αποτελέσματα πειραματικής μελέτης

Στους ακόλουθους πίνακες 4.3.1.1 & 4.3.1.2, παρουσιάζονται τα αποτελέσματα της συγκριτικής μελέτης για το Δείκτη αξιολόγησης *AUC*.

Στον πίνακα 4.3.1.1, παρατίθενται τα αποτελέσματα που αφορούν στα Σύνολα Δεδομένων του Κοινωνικού Δικτύου Facebook.

Πίνακας 4.3.1.1 Αποτελέσματα Συγκριτικής Μελέτης του Δείκτη AUC για τα Σύνολα Δεδομένων του Κοινωνικού Δικτύου, Facebook.

Δείκτης AUC	Facebook G_s	Facebook G_b
Common Neighbors	0.524	0.552
Jaccard's Coefficiency	0.537	0.643
Adamir-Adar	0.529	0.619
Preferential Attachment	0.526	0.621
Resource Allocation	0.528	0.623
Random Walk with Restart	0.538	**0.648**
Graph Distance	0.519	0.533
Katz	**0.539**	0.647
FriendLink	0.531	0.645

Τα αποτελέσματα της πειραματικής μελέτης υποδεικνύουν ότι το μεγαλύτερο G_b Σύνολο Δεδομένων για το Κοινωνικό Δίκτυο Facebook απέδωσε καλύτερα στη χρήση του αλγορίθμου Random Walk with Restart (Τυχαία περιήγηση με επανεκκίνηση) με τιμή (**0.648**) έναντι των υπολοίπων μεθόδων. Στην δεύτερη θέση με ελαφρώς μικρότερη τιμή είναι ο αλγόριθμος Katz (0.647). Στην περίπτωση όμως του μικρότερου Συνόλου Δεδομένων G_s ο αλγόριθμος που απέδωσε καλύτερα αποτελέσματα είναι ο Katz με μία τιμή (**0.539**) ελαφρώς καλύτερη από του αλγόριθμου Random Walk with Restart (0.538).

Γενικότερα μπορούμε να πούμε ότι οι αλγόριθμοι Random Walk with Restart και Katz αποδίδουν καλύτερα στο Πρόβλημα Εύρεσης Συνδέσμων των Συνόλων Δεδομένων για το Κοινωνικό Δίκτυο Facebook.

Το αποτέλεσμα αυτό δεν μας εκπλήσσει καθώς οι δύο αυτοί μέθοδοι εφαρμόζουν μία κοινή παραδοχή ότι οι διπλανοί κόμβοι ή οι κόμβοι των οποίων η διαδρομή είναι η μικρότερη έχουν μεγαλύτερη δυναμική επιρροής άρα και μεγαλύτερο βαθμό ομοιότητας.

Τα αποτελέσματα της παραπάνω πειραματικής μελάτης υποδεικνύουν ότι ενδεχομένως απαιτείται μεγαλύτερο μέγεθος Συνόλων Δεδομένων για τη λήψη των σωστότερων αποτελεσμάτων, καθώς οι τιμές που λάβαμε δεν είναι ιδιαίτερα υψηλών τιμών.

Στον πίνακα 4.3.1.2, παρατίθενται τα αποτελέσματα που αφορούν στα Σύνολα Δεδομένων του Κοινωνικού Δικτύου Twitter.

Πίνακας 4.3.1.2 Αποτελέσματα Συγκριτικής Μελέτης του Δείκτη AUC για τα Σύνολα Δεδομένων του Κοινωνικού Δικτύου, Twitter.

Δείκτης AUC	Twitter G_s	Twitter G_b
Common Neighbors	0.530	0.535
Jaccard's Coefficiency	0.531	0.577
Adamir-Adar	0.532	0.579
Preferential Attachment	0.531	0.574
Resource Allocation	0.532	0.577
Random Walk with Restart	**0.562**	**0.612**
Graph Distance	0.525	0.538
Katz	0.551	0.563
FriendLink	0.561	0.609

Τα αποτελέσματα της πειραματικής μελέτης υποδεικνύουν ότι τα καλύτερα αποτελέσματα προήλθαν από τον αλγόριθμο Random Walk with Restart (Τυχαία περιήγηση με επανεκκίνηση) με τιμή (**0.562**) για το μικρότερο G_s, και τιμή (**0.612**) για μεγαλύτερο G_b Σύνολο Δεδομένων για το Κοινωνικό Δίκτυο Twitter.

Στην περίπτωση του Κοινωνικού Δικτύου Twitter ο αλγόριθμος Katz δεν ακολούθησε με εξίσου καλές τιμές. Ο αλγόριθμος FriendLink όμως ακολουθεί με τις αμέσως καλύτερες τιμές και υπερτερεί έναντι του Katz.

Γενικότερα μπορούμε να πούμε ότι ο αλγόριθμος Random Walk with Restart αποδίδει καλύτερα στο Πρόβλημα Εύρεσης Συνδέσμων των Συνόλων Δεδομένων για το Κοινωνικό Δίκτυο Twitter.

4.3.2 Β' κύκλος Πειραματικών Μελετών

Στον Β' κύκλο πειραματικών μελετών εφαρμόστηκε η πρόταση αναθεώρησης (Κεφάλαιο 3.3) των αλγορίθμων ομοιότητας σε τοπικό και καθολικό επίπεδο (Κεφάλαιο 3.2) στα δύο Σύνολα Δεδομένων. Η επιλογή αυτού του κύκλου πειραματικής μελέτης έγκειται στην προσπάθεια επιβεβαίωσης ή όχι μέσω συγκριτικής μελέτης της βελτιωμένης απόδοσης του Προβλήματος Εύρεσης Συνδέσμων μέσα από τεχνικές που υιοθετούν την επιβολή penalty, με στόχο να απαντηθούν οι υποθέσεις της εργασίας μας (Κεφάλαιο 1.3).

Η τιμή της επιβολής Penalty που θα χρησιμοποιήσουμε στην πειραματική μελέτη είναι 0.0002. Σε περίπτωση που λάβουμε βελτιωμένα αποτελέσματα, έναντι των παραδοσιακών μεθόδων, θα υλοποιήσουμε εναλλακτικές τιμές Penalty έως ότου καταλήξουμε στα βέλτιστα δυνατά αποτελέσματα. Στην περίπτωση αυτή θα θεωρήσουμε ότι η προτεινόμενη αναθεωρημένη εκδοχή μας αποδίδει.

Στον παρακάτω πίνακα 4.3.2.1, παρατίθενται τα αποτελέσματα που αφορούν στα Σύνολα Δεδομένων του Κοινωνικού Δικτύου Facebook για την αναθεωρημένη μέθοδο.

Πίνακας 4.3.2.1 Αποτελέσματα Αναθεωρημένης Συγκριτικής Μελέτης του Δείκτη AUC για τα Σύνολα Δεδομένων του Κοινωνικού Δικτύου, Facebook.

Δείκτης AUC+	Facebook G_s	Facebook G_b
Common Neighbors +	0.520	0.536
Jaccard's Coefficiency +	0.536	0.608
Adamir-Adar +	0.517	0.613
Preferential Attachment +	0.524	0.589
Resource Allocation +	0.523	0.599
Random Walk with Restart +	0.531	**0.621**
Graph Distance +	0.506	0.520
Katz +	**0.539**	0.611
FriendLink +	0.529	0.597

Τα αποτελέσματα της αναθεωρημένης πειραματικής μελέτης υποδεικνύουν ότι το μεγαλύτερο G_b Σύνολο Δεδομένων για το Κοινωνικό Δίκτυο Facebook απέδωσε καλύτερα στη χρήση του αλγορίθμου Random Walk with Restart (Τυχαία περιήγηση με επανεκκίνηση) με τιμή (**0.621**) έναντι των υπολοίπων μεθόδων. Στην περίπτωση όμως του μικρότερου Συνόλου Δεδομένων G_s ο αλγόριθμος που απέδωσε καλύτερα αποτελέσματα είναι ο Katz με μία τιμή (**0.539**).

Γενικότερα μπορούμε να πούμε ότι οι αλγόριθμοι Random Walk with Restart+ και Katz+ αποδίδουν καλύτερα και στο αναθεωρημένο Πρόβλημα Εύρεσης Συνδέσμων των Συνόλων Δεδομένων για το Κοινωνικό Δίκτυο Facebook.

Τα αποτελέσματα των πειραματικών μελετών σε αυτό το σημείο δεν ήταν καλύτερα έναντι των κλασσικών μεθόδων παρόλο που αναμέναμε να είναι κοντά ή και ελαφρώς καλύτερα.

Στον πίνακα 4.3.1.2, παρατίθενται τα αποτελέσματα που αφορούν στα Σύνολα Δεδομένων του Κοινωνικού Δικτύου Twitter.

Πίνακας 4.3.2.2 Αποτελέσματα Αναθεωρημένης Συγκριτικής Μελέτης του Δείκτη AUC για τα Σύνολα Δεδομένων του Κοινωνικού Δικτύου, Twitter.

Δείκτης AUC+	Twitter G_s	Twitter G_b
Common Neighbors +	0.523	0.523
Jaccard's Coefficiency +	0.529	0.558
Adamir-Adar +	0.524	0.556
Preferential Attachment +	0.520	0.565
Resource Allocation +	0.529	0.564
Random Walk with Restart +	**0.545**	**0.598**
Graph Distance +	0.511	0.526
Katz +	0.533	0.542
FriendLink +	0.552	0.548

Τα αποτελέσματα της αναθεωρημένης πειραματικής μελέτης υποδεικνύουν ότι τα Σύνολα Δεδομένων για το Κοινωνικό Δίκτυο Twitter απέδωσαν καλύτερα στη χρήση του

αλγορίθμου Random Walk with Restart (Τυχαία περιήγηση με επανεκκίνηση) με τιμή (**0.545**) για το μικρότερο *Gs* και (**0.598**) για μεγαλύτερο G_b Σύνολο Δεδομένων αντίστοιχα.

Και σε αυτό το σημείο παρατηρούμε ότι τα αποτελέσματα της πειραματικής έρευνας για την τιμή Penalty 0.0002 δεν είναι ορθότερα των κλασσικών μεθόδων. Σύμφωνα με την υπόθεση μας αναμέναμε αποτελέσματα καλύτερα ή πολύ κοντά σε αυτά που λάβαμε με τις κλασσικές μεθόδους, για να θεωρήσουμε ότι η τοποθέτηση μας έχει βάση για να συνεχιστούν τα πειράματα με άλλους συντελεστές penalty.

Μία υπόθεση αυτής της αστοχίας μπορεί να θεωρηθεί η ενδεχόμενη ασυμβατότητα της βάσης της θεωρίας μας που βασίζεται κυρίως σε εμπειρική παρατήρηση μέσα από τη χρήση σε Business Oriented επίπεδο του Κοινωνικού Δικτύου Facebook σε πραγματικό παροντικό χρόνο και της παλαιότητας του Συνόλου Δεδομένων καθώς τα δεδομένα αυτά έχουν καταγραφεί το 2012 όπου οι χρήστες των Κοινωνικών Δικτύων ήταν ενδεχομένως λιγότερο influenced στη χρήση τους. Μία σύγχρονη Βάση Δεδομένων μπορεί να οδηγήσει σε καλύτερα αποτελέσματα.

Μία επόμενη υπόθεση που μπορεί να θεωρηθεί ότι έχει επηρεάσει τα αποτελέσματα είναι η μη χρήση ενός μεγάλου Συνόλου Δεδομένων όπου η εναλλαγή του χρόνου έχει σαφέστερα και βαθύτερα σημάδια επιρροής. Για παράδειγμα η καταγραφή σύγχρονων Συνόλων Δεδομένων από τα αντίστοιχα Κοινωνικά Δίκτυα διάρκειας έως 6 μηνών θα μπορούσε να αποτελέσει ιδανικότερο καμβά των πειραματικών μας μελετών.

Συνοψίζοντας, θα καταλήγαμε ότι η θεώρηση της ανωτέρω λογικής ίσως να μην είναι και η ορθότερη δεδομένων των συνθηκών και του υλικού δοκιμών.

Επαναλαμβανόμενες πειραματικές μελέτες διαφορετικών τιμών και ενδεχομένως μεγαλύτερων και πιο σύγχρονων Συνόλων Δεδομένων μπορεί να επιφέρουν καλύτερα αποτελέσματα. Στην εργασία αυτή, υλοποιήθηκε ένας κύκλος πειραμάτων, υπό την έννοια της τυχαία επιλεγμένης τιμής Penalty.

5 Επίλογος

Το αναθεωρημένο, κατά τα τελευταία έτη, Πρόβλημα Εύρεσης Συνδέσμων με χρήση της Χρονικής Πληροφορίας, καλείται να αντιμετωπιστεί ως ένα state-of-the-art πρόβλημα που αντιμετωπίζουν πολλές διαφορετικές ομάδες χρηστών σε παγκόσμια εμβέλεια. Η ανάγκη, αποδοτικών και άμεσων λύσεων είναι επιτακτική καθώς δύναται να επιλύσει προβλήματα που σχετίζονται με οικονομικά, ακαδημαϊκά, ερευνητικά, επιχειρηματικά, κοινωνιολογικά, και σχεδόν κάθε φύσης και κλάδου προβλήματος που διακινείται στο διαδίκτυο μεταξύ Κοινωνικών Δικτύων.

5.1 Σύνοψη και συμπεράσματα

Το Πρόβλημα Εύρεσης Συνδέσμων, σύμφωνα με τα αποτελέσματα των πειραματικών μελετών μας, υπέδειξε ότι η στατική χρήση του χρόνου, στη μοντελοποίηση των μεθόδων δεν επαρκεί για τη απόρροια των βέλτιστων δυνατών προβλέψεων, καθώς κατά τη διέλευση του χρόνου δραματικές αλλαγές μπορεί να προκύψουν σε ένα δίκτυο. Συγκεκριμένα, οι ανωτέρω αναφερόμενοι μέθοδοι χρησιμοποιούν την έννοια του χρόνου, για ένα συγκεκριμένο χρονικό προορισμό που στη συνέχεια φθίνει ως παράγοντας επιρροής με αποτέλεσμα οι προβλέψεις να μην είναι οι βέλτιστες δυνατές για ένα αποδοτικό Σχεδιασμό Στρατηγικών Προωθητικών και άλλων κινήσεων.

5.2 Όρια και περιορισμοί της έρευνας

Η έρευνα που διεξήχθη για την εργασία αυτή, τόσο σε πειραματικό επίπεδο, όσο και σε βιβλιογραφικό ήρθε αντιμέτωπη με πολλές νέες και ανερχόμενες πηγές ερεθισμάτων για περαιτέρω ανάπτυξη και έρευνα. Το ερευνητικό υπόβαθρο ενός τόσο σοβαρού προβλήματος θα πρέπει να αναπτυχθεί περαιτέρω με στόχο να ανακύψουν τα σημεία όπου μπορούν να εφαρμοστούν διαφοροποιημένες πειραματικές μελέτες.

Το πειραματικό μας ενδιαφέρον δεν κατάφερε να ολοκληρωθεί μέσω της υλοποίησης διαφορετικών μεθόδων στο σύνολο των Συνόλων Δεδομένων που

συγκεντρώθηκαν, λόγω χρονικών και τεχνολογικών περιορισμών. Η χρήση μεγαλύτερων και πιο σύγχρονων Συνόλων Δεδομένων μπορεί να επιφέρουν καλύτερα αποτελέσματα.

5.3 Μελλοντικές Επεκτάσεις

Μία σημαντική πρόκληση που συνεχίζει να καλείται να αντιμετωπίσει το Πρόβλημα Εύρεσης Συνδέσμων, είναι η εύρεση μεθόδων που μπορούν να ανταποκριθούν με τον καλύτερο δυνατό τρόπο στα τεράστια σε μέγεθος και πλήρως δυναμικά στο χρόνο Κοινωνικά Δίκτυα. Οι περισσότερες θεωρίες και τεχνικές που βασίζονται στην ομοιότητα (similarity -based methods) για την Πρόβλεψη μελλοντικών Συνδέσμων σε ένα εξελίξιμο χρονικά δίκτυο λαμβάνουν υπόψη ένα χρονικό στιγμιότυπο (snapshot). Παρότι η πλειοψηφία αυτών των μεθόδων αποδίδει όσον αφορά στην πρόβλεψη της δημιουργίας ενός συνδέσμου, δεν είναι τόσο προφανές, ότι μέσα από τη στατική μοντελοποίηση τους, μπορούν να προβλέψουν μελλοντικές επαναλαμβανόμενες συνδέσεις. Για παράδειγμα, η θετική ανταπόκριση ενός χρήστη σε μία προβλεπόμενη διαφημιστική καμπάνια μίας επιχείρησης τον Ιούνιο του 2018, δεν συνεπάγεται την ίδια θετική ανταπόκριση του σε κάθε διαφημιστική καμπάνιας της επιχείρησης κάθε μήνα ή κάθε επόμενο Ιούνιο, ούτε και των σχετιζόμενων φίλων του. Ο βαθμός διαφοροποίησης των ενδιαφερόντων ενός χρήστη και ο βαθμός σημαντικότητας της χρονικής στιγμής t_x δεν αποδεικνύεται ότι σχετίζονται με τη χρονική στιγμή t_j. Τέτοιου είδους ζητήματα δεν μπορούν να απαντηθούν στην τρέχουσα χρονική περίοδο καθώς υπάρχει ανάγκη περαιτέρω εμβάθυνσης και κατανόησης όλων εκείνων των < Προσωρινών [4]> χαρακτηριστικών που επιδρούν στα ενδιαφέροντα και στις συμπεριφορές των χρηστών και εξελίσσονται με την πάροδο του χρόνου. Αυτό το πεδίο ενδιαφέροντος θα μπορούσε να αποτελέσει μία νέα μελλοντική προσπάθεια εξερεύνησης και ανεύρεσης προτεινόμενων λύσεων.

[4] Θεωρούμε ότι πρόκειται για δεδομένα που επιδρούν στο χρήστη, αλλά η διάρκεια τους είναι σύντομη.

6 Βιβλιογραφία

A. L. Barabasi, H. Jeong, Z. Neda, E. Ravasz, A. Schubert, and T. Vicsek, 2008. *Evolution of the social network of scientific collaborations* [pdf] Available at : <https://arxiv.org/pdf/cond-mat/0104162.pdf> [Accessed 15 May 2018]

Adam Perer, Ben Shneiderman, 2006. *Balancing Systematic and Flexible Exploration of Social Networks*, IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS, VOL. 12, NO. 5, SEPTEMBER/OCTOBER 2006

Adamic Lada A., Adar Eytan, 2003. Friends and Neighbors on the Web. [Online] Available at : <http://www.hpl.hp.com/research/idl/papers/web10/fnn2.pdf> [Accessed 15 May 2018]

Alexis Papadimitriou, Panagiotis Symeonidis, Yannis Manolopoulos, 2012. *Fast and accurate link prediction in social networking systems*. [pdf] Available at : <http://newiranians.ir/57a45bb21932d-Haron%20Abadi_10.1016-j.jss.2012.04.019-1.pdf > [Accessed 10 June 2018]

David Hand, Heikki Mannila, Padhraic Smyth, 2001. *Principles of Data Mining.* Cambridge. [pdf] Available at : <https://doc.lagout.org/Others/Data%20Mining/Principles%20of%20Data%20Mining%20%5BHand%2C%20Mannila%20%26%20Smyth%202001-08-01%5D.pdf> > [Accessed 10 May 2018]

Fayyad M.Usama, Piatesky-Shapiro Gregory, Smyth Padhraic, 1996. *'From Data Mining to Knowledge Discovery: An Overview*', in Fayyad, Piatetsky-Shapiro, Smyth, Uthurusamy, Advances in Knowledge Discovery and Data Mining, AAAI Press / The MIT Press, Menlo Park, CA, 1996, pp.1-34

Frawley, W - Piatetsky-Shapiro, G. - Matheus, C., 1992. *Knowledge Discovery in* Databases : An Overview, AI Magazie 13.3, 57-70, 1992. [pdf] Available at : <https://pdfs.semanticscholar.org/7a7b/51b86e22d0077215287980c7ba793b09e4cd.pdf> [Accessed 22 March 2018]

Genshiro Kitagawa, 2010. *Introduction to Time Series Modeling*, (Chapman & Hall/CRC Monographs on Statistics & Applied Probability), Chapman and Hall/CRC, 2010

Itai Himelboim, 2017 : *Social Network Analysis (Social Media)*, University of Georgia, USA, [Online] Available at : <https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118901731.iecrm0236> [Accessed 15 January 2018]

J. Leskovec and J. McAuley, 2012. *Learning to Discover Social Circles in Ego Networks*. NIPS, 2012. [pdf] Available at : <http://i.stanford.edu/~julian/pdfs/nips2012.pdf> [Accessed 15 January 2018]

Jiawei Han & Micheline Kamber & Jian Pei, 2012. *Data Mining Concepts and Techniques*, 3rd Ed., 2012, ISBN 978-0-12-381479-1. [pdf] Available at : <http://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf> [Accessed 22 March 2018]

Jimeng Sun, Christos Faloutsos, Spiros Papadimitriou, Philip S. Yu. KDD, 2007. *GraphScope: Parameter-Free Mining of Large Time-Evolving Graphs*. [pdf] Available at : <http://www.cs.cmu.edu/~spapadim/pdf/gscope_kdd07.pdf> [Accessed 15 January 2018]

Jose Hernandez - Orallo, 2005. '*Knowledge Discovery from Databases*', in Laura C. Rivero, Jorge Horacio Doorn, and Viviana E. Ferraggine, Encyclopedia of Database Technologies and Applications, IGI Global, June 2005, Chapter 54, p 313-318, ISBN13: 9781591405603

Jure Leskovec, Mary McGlohon, Christos Faloutsos, Natalie Glance, Matthew Hurs, 2006. *Patterns of Cascading Behavior in Large Blog Graphs*. [pdf] Available at : <https://cs.stanford.edu/~jure/pubs/blogs-sdm07.pdf> [Accessed 15 January 2018]

Kanna Al-Falahi, Yacine Atif, Said Elnaffar, 2010. *Social Networks Challenges and New Opportunities*. 2010 [Online] Available at : <https://ieeexplore.ieee.org/document/5724921/> [Accessed 8 May 2018]

Kaplan, A.M. and Haenlein, M., 2010. *Users of the world unite! The challenges and opportunities of social media*. Science direct, 53, 59 - 68, 2010. [pdf] Available at : <http://michaelhaenlein.eu/Publications/Kaplan,%20Andreas%20-%20Users%20of%20the%20world,%20unite.pdf> [Accessed 15 January 2018]

Kristin R. Eschenfelder, Morgaine Gilchrist Scott, Kalpana Shankar, Greg Downey, 2017. *Social Science Data Archives: A Historical Social Network Analysis*.[Online] Available at : <http://irserver.ucd.ie/bitstream/handle/10197/8695/vol_40_1_eschenfelder.pdf?sequence=1> [Accessed 15 January 2018]

Linyuan Lü, Tao Zhou, 2010. *Link Prediction in Complex Networks: A Survey*. [pdf] Available at : <https://arxiv.org/pdf/1010.0725.pdf> [Accessed 15 May 2018]

Mariam Adedoyin-Olowe, Mohamed Medhat Gaber, Frederic Stahl, 2014. *A Survey of Data Mining Techniques for Social Network Analysis*. Journal of Data Mining & Digital Humanities. [pdf] Available at : < https://arxiv.org/vc/arxiv/papers/1312/1312.4617v1.pdf > [Accessed 20 April 2018]

Newman, M.E.J., 2001. *Clustering and preferential attachment in growing networks*. Physical Review E 64 [pdf] Available at : <https://arxiv.org/pdf/cond-mat/0104209.pdf> [Accessed 15 January 2018]

Nowell, D.L. , Kleinberg, J., 2007. *The Link-Prediction Problem for Social Networks*. Journal of the American society for information science and technology,58(7): 1019-1031 [pdf] Available at : <http://www.cs.carleton.edu/faculty/dlibenno/papers/link-prediction/link.pdf> [Accessed 15 January 2018]

Pang, B and Lee L., 2008. *Opinion mining and sentiment analysis.* Foundations and trends in information Retrieval. Vol. 2, Nos. 1-2, 1-135, 2008 [pdf] Available at : <http://www.cs.cornell.edu/home/llee/omsa/omsa.pdf> [Accessed 15 May 2018]

Philipp Singer, 2011. *Time Series Analysis of Online Social Network Data and Content*, Knowledge Management Institute Graz University of Technology. 2011. [pdf] Available at : <https://www.philippsinger.info/files/masterarbeit_psinger.pdf > [Accessed 15 January 2018]

Sang Jun Lee, Keng Siau, 2001. *A review of data mining techniques* [pdf] Availble at : <https://pdfs.semanticscholar.org/7f3f/408d3d2c89f0d36e807622efeee880cab95f.pdf> [Accessed 15 May 2018]

Sarjon Defit, Mohd Noor Md Sap, 2010. *Predictive Data Mining Based On Similarity and Clustering Methods.* [pdf] Available at : <http://eprints.utm.my/id/eprint/8711/1/SarjonDefit2000_PredictiveDataMiningBasedOnSimilarity.pdf> [Accessed 15 June 2018]

Reza Zafarani, Mohammad Ali Abbasi, Huan Liu., 2014. *Social media mining An introduction*, Cambridge University Press, April 2014 [Online] Available at : <http://dmml.asu.edu/smm/SMM.pdf> [Accessed 15 March 2018]

Shumway and Stoffer, 2011 : *'Time Series Analysis and Its Applications With R Examples*', Springer Texts in Statistics, 3rd Edition, 2011, pp 1 - 44, DOI 10.1007/978-1-4419-7865-3. [pdf] Available at : <http://db.ucsd.edu/static/TimeSeries.pdf> [Accessed 15 May 2018]

Simon Kemp, We are social, 2018. *Social Media use Jumps in Q1 despite privacy fears*. April 2018 [Online] Available at : <https://wearesocial.com/blog/2018/04/social-media-use-jumps-in-q1-despite-privacy-fears> [Accessed 24 May 2018]

Suphakit Niwattanakul, Jatsada Singthongchai, Ekkachai Naenudorn and Supachanun Wanapu, 2013. Using *of Jaccard Coefficient for Keywords Similarity*, [pdf] Available at <https://www.researchgate.net/profile/Ekkachai_Naenudorn/publication/317248581_Using_of_Jaccard_Coefficient_for_Keywords_Similarity/links/592e560ba6fdcc89e759c6d0/Using-of-Jaccard-Coefficient-for-Keywords-Similarity.pdf> [Accessed 15 May 2018]

T. Zhou, L. Lü and Y.-C. Zhang, 2009. *Predicting missing links via local information*, in European Physical Journal B, vol. 71, pp. 623-630, October 2009, [pdf] Available at : <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.876.2255&rep=rep1&type=pdf> [Accessed 25 May 2018]

Van De Bunt, Van Duijn, Snijders, 2009. *Friendship Networks Through Time: An Actor-Oriented Dynamic Statistical Network Model*, Volume 5, Number 2, July 1999, pp. 167-192. [pdf] Available at : <https://pdfs.semanticscholar.org/9262/4e3a989db16421af5b5c78fe84911dba7981.pdf> [Accessed 15 January 2018]

Weiping Liu, Linyuan Lu, 2010. *Link Prediction Based on Local Random Walk*. [pdf] Available at : <https://arxiv.org/pdf/1001.2467.pdf> [Accessed 1 June 2018]

Γακόπουλος Ευθύμιος, 2012. *Εφαρμογή τεχνικών Data Mining σε δεδομένα κυκλοφορίας οδικού δικτύου*. Μεταπτυχιακή Εργασία στην Επιχειρηματική Πληροφορική, Σχολή Εφαρμοσμένης Πληροφορικής, Πανεπιστήμιο Μακεδονίας. [pdf] Available at : <https://dspace.lib.uom.gr/bitstream/2159/14897/3/GakopoulosEuthymiosMsc2012.pdf> [Accessed 1 June 2018]