ΠΑΝΕΠΙΣΤΗΜΙΟ ΜΑΚΕΔΟΝΙΑΣ
ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
ΤΜΗΜΑΤΟΣ ΕΦΑΡΜΟΣΜΕΝΗΣ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΑΝΑΛΥΣΗ ΙΣΤΟΡΙΚΩΝ ΔΕΔΟΜΕΝΩΝ ΣΕ ΚΟΙΝΩΝΙΚΑ ΔΙΚΤΥΑ ΓΙΑ ΤΗΝ ΠΡΟΒΛΕΨΗ ΣΥΝΔΕΣΜΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Μαυροϊδάκη Ελένη του Μαυρουδή

Νοέμβριος 2018

#ΚοινωνικάΔίκτυα

Το πρόβλημα

Εύρεσης τεχνικών εξόρυξης και χρήσης Ιστορικών Δεδομένων σε Κοινωνικά Δίκτυα

Σκοπός - Στόχοι

- Η ανάδειξη του ρόλου της χρονικής πληροφορίας στα Κοινωνικά Δίκτυα σε Ερευνητικό επίπεδο.
- Η πρόταση λύσης που μπορεί να χρησιμοποιηθεί και σε Business Strategic Management επίπεδο.
- Η χρήση εναλλακτικών αλγόριθμων 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.

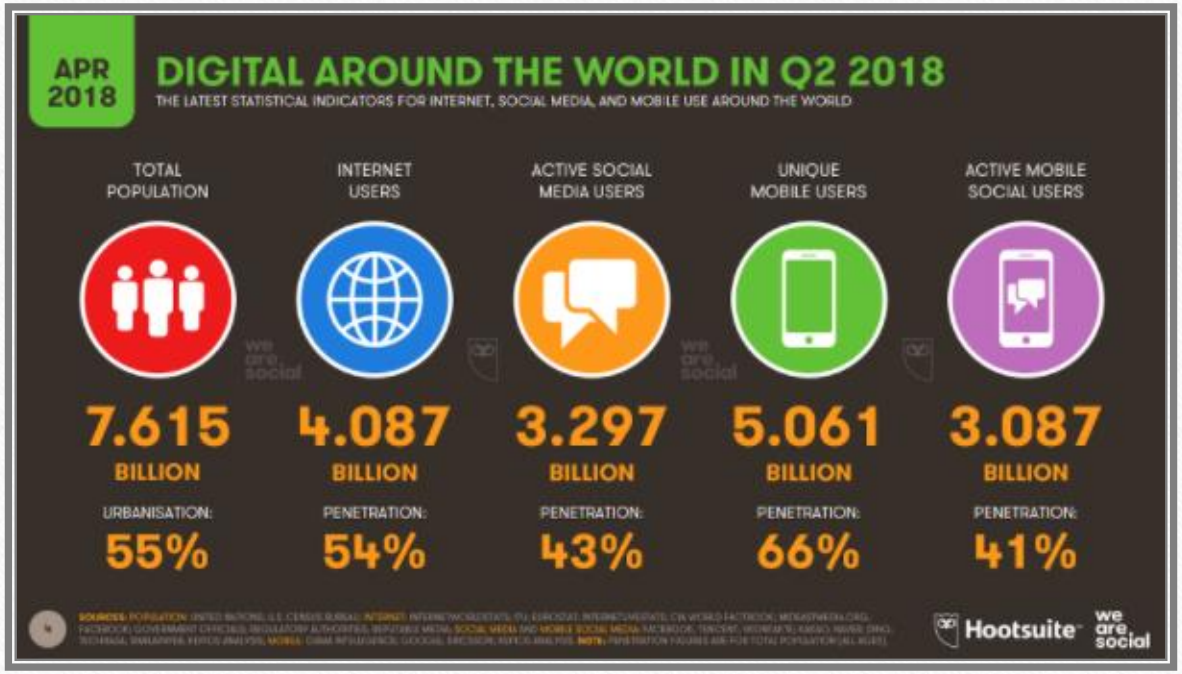
Ερωτήματα :

Η χρήση των ιστορικών Δεδομένων σε Κοινωνικά Δίκτυα

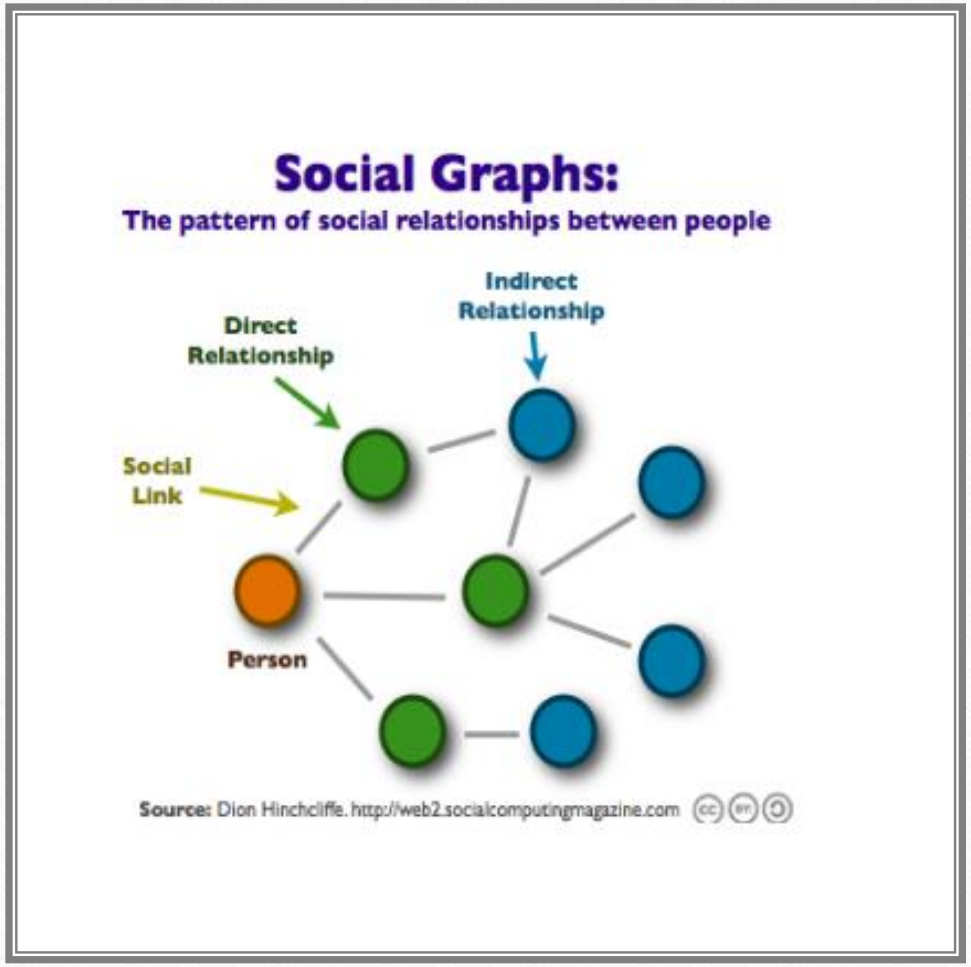
- μπορεί να εφαρμοστεί στις Στρατηγικές Marketing και Advertising για το πρόβλημα Εύρεσης Συνδέσμων, για την αποδοτικότητα των Καμπανιών και την πρόβλεψη του ενδεχόμενου ρίσκου κατά την τροποποίηση των στρατηγικών ή όχι.
- αποτελεί εργαλείο ανάδειξης των δυνατών σημείων μίας Στρατηγικής Καμπάνιας έναντι ανταγωνιστικών ή όχι.
- αποτελεί βασικό στοιχείο αποτελεσματικής ανάδειξης και Εύρεσης Συνδέσμων ή όχι.
- είναι ικανή να προκαλέσει επιρροές στους προς Εύρεση Συνδέσμους ή όχι.
- μπορεί να επηρεάσει και να αναβαθμίσει/υποβαθμίσει το Brand Reputation ή όχι.

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 2ου τριμήνου του 2018



Πηγή: https://wearesocial.com



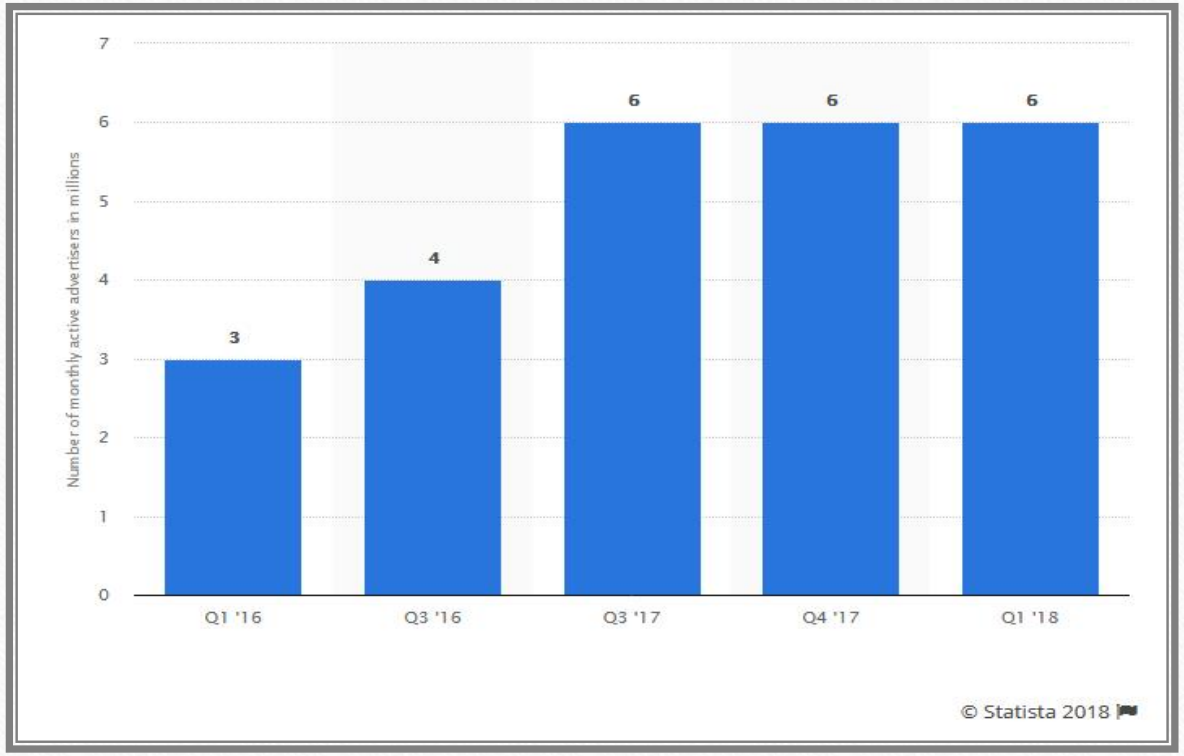
Πηγή : (Kanna Al-Falahi & Yacine Atif & Said Elnaffar, 2010)

Κοινωνικό Δίκτυο

- 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)

Επιχειρηματικό ενδιαφέρον Κοινωνικών Δικτύων

- Οι χρήστες (άνθρωποι, οργανισμοί, επιχειρήσεις κ.α.) των ΚΔ ανευρίσκουν άλλους χρήστες (users) με κοινά ενδιαφέροντα, δραστηριότητες, στόχους, και να μπορούν να **αλληλεπιδρούν μεταξύ τους διαμοιράζοντας υλικό** και πληροφόρηση υπό διαφορετικές μορφές.
- Οι εταιρείες στοχεύουν στην ανεύρεση υποψηφίων πελατών με στόχο την προώθηση των προϊόντων & υπηρεσιών τους σε ένα δομημένο στρατηγικά (strategic target-marketing) περιβάλλον.



Facebook. 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

Πηγή : https://www.statista.com/

Social Media Data Mining
Εξόρυξη Γνώσης από Δεδομένα Κοινωνικών Δικτύων

- Ο στόχος του Data Mining, είναι «**η εξόρυξη χρήσιμης πληροφορίας από σύνολα ή βάσεις δεδομένων μεγάλου μεγέθους**» (Γακόπουλος Ευθύμιος, 2012) ή «η σύνθετη διαδικασία εξαγωγής συγκεκριμένης αλλά προηγουμένως άγνωστης και δυνητικά ωφέλιμης γνώσης από δεδομένα» (Frawley et al., 1992).
- Η επιστήμη του Data Mining 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.

Τεχνικές Εξόρυξης Γνώσης από Δεδομένα ΚΔ

- 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).

Αλγόριθμοι Εξόρυξης Γνώσης (Data Mining Algorithms)

- Έχοντας κατανοήσει το πρόβλημα προς επίλυση 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
- Αλγόριθμοι Ταξινόμησης, Αλγόριθμοι Ανάλυσης Παλινδρόμησης, Αλγόριθμοι Εύρεσης Συνδέσμων, Αλγόριθμοι Ομαδοποίησης, Αλγόριθμοι Εύρεσης Χαρακτηριστικών …

Πρόβλημα Εύρεσης Συνδέσμων
Link Prediction Problem (1/3)

- Δεδομένου ενός στιγμιότυπου ενός ΚΔ 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).

Πρόβλημα Εύρεσης Συνδέσμων
Link Prediction Problem (2/3)

- Human Resources : εύρεση του κατάλληλου υποψηφίου εργαζομένου.
- Βιολογία : θέματα δίκτυα τροφίμων, δίκτυα αλληλεπίδρασης πρωτεϊνών μεταξύ τους, δίκτυα που σχετίζονται με μεταβολικούς και άλλους παράγοντες (Linyuan Lu & Tao Zhou, 2010).
- Κοινωνικά Δίκτυα : 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 & της ωφέλειας του τελικού χρήστη.
- e-commerce : 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".

Πρόβλημα Εύρεσης Συνδέσμων
Link Prediction Problem (3/3)

- Τεχνητή Νοημοσύνη : οργάνωση 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
- Ασφάλεια : πρόβλεψη μελλοντικών τρομοκρατικών ενεργειών, υπό το πρίσμα της συνεργατικής οργάνωσης οντοτήτων η οποία δεν είναι ευθέως ορατή.
- Ακαδημαϊκός κλάδος : 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).

Πρόβλημα Εύρεσης Συνδέσμων

Business Oriented ερωτήματα που αφορούν
στην μέτρηση της απόδοσης που μπορεί να έχουν
υλοποιημένες Καμπάνιες Προώθησης στα Κοινωνικά Δίκτυα

Ζητούμενο

- Αξιολόγηση μίας προσπάθειας 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
- Στόχος : η αύξηση της απόδοσης της διαφημιστικής προσπάθειας υπό το πρίσμα της αποτελεσματικότητας της από όλα τα μέρη
- Βελτιστοποίηση της ωφέλειας (χρήστης – ΚΔ – επιχείρηση).

Μαθηματική Περιγραφή του Προβλήματος (1/3)

- Έχουμε ένα Κοινωνικό Δίκτυο $G = \langle V, E \rangle$ όπου κάθε σύνδεσμος (edge) $e = \langle x, y \rangle \boldsymbol{\epsilon} E$ (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.
- Καταγράφουμε τις πολλαπλές αλληλεπιδράσεις ανάμεσα στους κόμβους x και y σε διαφορετικές τιμές του χρόνου.
- 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}.

Μαθηματική Περιγραφή του Προβλήματος (2/3)

- 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*.

Μαθηματική Περιγραφή του Προβλήματος (3/3)

- 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 $s_{xy} = s_{yx}$.

Αξιολόγηση αλγορίθμου - AUC

- AUC - area under the receiver operating characteristic curve
- Χρησιμοποιείται για την αξιολόγηση των μη-παρατηρούμενων συνδέσμων.
- Ορίζεται ως ο λόγος $(n' + 0.5n'') / n$,
  n : ο αριθμός των ανεξάρτητων συγκρίσεων,
  n' : ο αριθμός των συγκρίσεων όπου ο χαμένος σύνδεσμος έχει υψηλότερη βαθμολογία,
  n'' : ο αριθμός συγκρίσεων με την ίδια βαθμολογία.

Μεθοδολογία

Σύντομη ανασκόπηση των εναλλακτικών επιλεγμένων
μεθόδων υλοποίησης του Προβλήματος Εύρεσης Συνδέσμων
καθώς και των αντίστοιχων πλεονεκτημάτων και μειονεκτημάτων τους

Αλγόριθμοι εφαρμογής στο Πρόβλημα Εύρεσης Συνδέσμων
Link Prediction Algorithms

- **Αλγόριθμοι Ομοιότητας σε Τοπικό Επίπεδο (Local Similarity algorithms)**
- Οι κόμβοι x,y είναι πιθανότερο να σχηματίσουν μία σύνδεση στο μέλλον, εφόσον οι γειτονικοί τους κόμβοι έχουν μεγάλο αριθμό κοινών κόμβων.
- Το πλεονέκτημα τους 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.

- **Αλγόριθμοι Ομοιότητας σε Καθολικό Επίπεδο (Global Similarity algorithms)**
- Καθολική συμμετοχή 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 καθώς αναζητούν περισσότερο όμοιους κόμβους στην περίπτωση αυτή.

Αλγόριθμοι Ομοιότητας σε Τοπικό Επίπεδο

(Local Similarity algorithms)

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:

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

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:

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

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:

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

όπου *Γ(z)*: δηλώνεται το σύνολο των γειτόνων του κόμβου *z*

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:

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

όπου *Γ(x), Γ(y)* : δηλώνεται το σύνολο των γειτόνων των κόμβων *x,y*

Δείκτης Κατανομής Πόρων – Resource Allocation Index

- Η προσέγγιση αυτής της μεθόδου (Zhou & Lu & 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:

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

όπου *Γ(z)*: δηλώνεται το σύνολο των γειτόνων του κόμβου *z*

Αλγόριθμοι Ομοιότητας σε Καθολικό Επίπεδο

(Global Similarity algorithms)

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:

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

όπου */E/*: ο αριθμός των συνδέσμων (edges) στο γράφημα.

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:

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

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:

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

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 $\beta^l = 0.1$

Δείκτης FriendLink

- Η προσέγγιση αυτής της μεθόδου (Papadimitriou & Symeonidis & Manolopoulos, 2012), βασίζεται στην ιδέα της ύπαρξης ομοιοτήτων μεταξύ δύο κόμβων.
- 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:

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

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*.

Προτεινόμενη αναθεωρημένη μέθοδος

Αναθεώρηση πρόβλεψης σε αλγορίθμους Ομοιότητας (1/3)

- 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, σε συνδέσμους, 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.

Αναθεώρηση πρόβλεψης σε αλγορίθμους Ομοιότητας (2/3)

- Η λογική της προτεινόμενης μεθόδου, απορρέει από την εμπειρική παρατήρηση 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.

Αναθεώρηση πρόβλεψης σε αλγορίθμους Ομοιότητας (3/3)

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

Πειραματική Μελέτη

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

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

Πηγή των δεδομένων Stanford University,2012, snap.stanford.edu

Σύνολα Δεδομένων Κοινωνικών Δικτύων

- Σύνολο Δεδομένων του Facebook
- 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.

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

Πηγή των δεδομένων Stanford University, 2012 snap.stanford.edu

Σύνολα Δεδομένων Κοινωνικών Δικτύων

- Σύνολο Δεδομένων του Twitter
- 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
- Το μεγαλύτερο διάστημα αποτελείται από 3.478 κόμβους και 51.931 συνδέσμους.

Υλοποίηση Πειραματικών Μελετών (1/2)

- 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).
- Μοιράζουμε τη λίστα συνδέσμων σε *διάστημα εκπαίδευσης* E_{train} (Training Set) και *διάστημα δοκιμής* E_{test} (Test set)

Υλοποίηση Πειραματικών Μελετών (2/2)

- Εφαρμόζουμε τους αλγόριθμους ομοιότητας 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
- 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).
- 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.

Αποτελέσματα Πειραματικών Μελετών σε Facebook & Twitter

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

Α' κύκλος Πειραματικών Μελετών

- Υλοποίηση αλγόριθμων Ομοιότητας σε Τοπικό και Καθολικό επίπεδο στα δύο υποσύνολα (G_s: το μικρό σύνολο δεδομένων, G_b: το μεγαλύτερο σύνολο δεδομένων) των δύο Συνόλων Δεδομένων των Κοινωνικών Δικτύων Facebook & Twitter.

Δείκτης AUC
Αποτελέσματα πειραματικής μελέτης ΣΔ του ΚΔ : Facebook (1/2)

- Το μικρότερο ΣΔ *Gs* απέδωσε καλύτερα στη χρήση του αλγόριθμου Katz ενώ ακολουθεί ο αλγόριθμος Random Walk with Restart με μία μικρή απόκλιση τιμής.
- Το μεγαλύτερο ΣΔ *Gb* απέδωσε καλύτερα στη χρήση του αλγορίθμου Random Walk with Restart. Στην δεύτερη θέση με ελαφρώς μικρότερη τιμή είναι ο αλγόριθμος Katz
- Δεδομένου ότι αναμέναμε τιμή μεγαλύτερη του 0.5 που αντιστοιχεί στην καθαρή τύχη, δεν λάβαμε αποτελέσματα υψηλών τιμών.

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

Δείκτης AUC
Αποτελέσματα πειραματικής μελέτης ΣΔ του ΚΔ : Twitter

- Και τα δύο ΣΔ *Gs, Gb* έδωσαν την καλύτερη τιμή για τον αλγόριθμο ομοιότητας Random Walk with Restart.
- Ο αλγόριθμος FriendLink όμως ακολουθεί με τις αμέσως καλύτερες τιμές
- Δεδομένου ότι αναμέναμε τιμή μεγαλύτερη του 0.5 που αντιστοιχεί στην καθαρή τύχη, δεν λάβαμε αποτελέσματα υψηλών τιμών.

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

Β' κύκλος Πειραματικών Μελετών

- Υλοποίηση αλγόριθμων Ομοιότητας σε Τοπικό και Καθολικό επίπεδο στα δύο υποσύνολα (G_S, G_D) των δύο ΣΔ των Κοινωνικών Δικτύων Facebook & Twitter
- Η υλοποίηση έλαβε χώρα υπό το πρίσμα της επιβολής Penalty λόγω χρονικής παλαιότητας 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

Δείκτης AUC
Αποτελέσματα αναθεωρημένης πειραματικής μελέτης ΣΔ του ΚΔ : Facebook (1/2)

- Το μικρότερο ΣΔ *Gs* απέδωσε καλύτερα στη χρήση του αλγορίθμου Katz
- Το μεγαλύτερο ΣΔ *Gb* απέδωσε καλύτερα στη χρήση του αλγορίθμου Random Walk with Restart.
- Τα αποτελέσματα των πειραματικών μελετών υστερούν έναντι των κλασσικών μεθόδων παρόλο που αναμέναμε να είναι κοντά ή και ελαφρώς καλύτερα.

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

Δείκτης AUC
Αποτελέσματα αναθεωρημένης πειραματικής μελέτης ΣΔ του ΚΔ : Twitter

- Και τα δύο ΣΔ *Gs, Gb* απέδωσαν καλύτερα στη χρήση του αλγόριθμου ομοιότητας Random Walk with Restart.
- Τα αποτελέσματα των πειραματικών μελετών υστερούν έναντι των κλασσικών μεθόδων παρόλο που αναμέναμε να είναι κοντά ή και ελαφρώς καλύτερα.

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

Αξιολόγηση αποτελεσμάτων (1/2)

- Τα αποτελέσματα της αναθεωρημένης πειραματικής έρευνας για την τιμή Penalty 0.0002 δεν είναι ορθότερα των κλασσικών μεθόδων.
- Λόγος αστοχίας : Ασυμβατότητα της βάσης της θεωρίας VS παλαιότητα του Συνόλου Δεδομένων (2012), όπου οι χρήστες των Κοινωνικών Δικτύων ήταν λιγότερο influenced στη χρήση τους.
- Μία σύγχρονη Βάση Δεδομένων μπορεί να οδηγήσει σε καλύτερα αποτελέσματα.
- 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.

Αξιολόγηση αποτελεσμάτων (2/2)

- 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.

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.
- 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).
- Οι αλλαγές στα ενδιαφέροντα των χρηστών κατά τη διέλευση του χρόνου αποτελεί ένα μεγάλο πεδίο ενδιαφέροντος τόσο για ερευνητική όσο και για Business oriented κατεύθυνση και περαιτέρω μελέτη.
- Παρότι τα αποτελέσματα των μελετών μας δεν αποδείχθηκαν σπουδαίας σημασίας, παραμένει η αίσθηση ότι περισσότερες μελέτες θα πρέπει να επικεντρωθούν στην αναζήτηση της σημασίας του χρόνου και της επίδρασης τους στη δομή των Κοινωνικών Δικτύων.

Βιβλιογραφία (1/3)

- 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]
- 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]
- 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/51b86e22d007721528798 0c7ba793b09e4cd.pdf> [Accessed 22 March 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]

Βιβλιογραφία (2/3)

- Linyuan Lu, 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]

Βιβλιογραφία (3/3)

- 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. Lu 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]
- 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]

Σας ευχαριστώ