ΠΑΝΕΠΙΣΤΗΜΙΟ
ΠΑΤΡΩΝ
UNIVERSITY OF PATRAS

Πολυτεχνικη Σχολη
Τμημα μηχανικων Η/Υ και
πληροφορικης

Αξιοποίηση της ανάλυσης χρονοσειρών στο Twitter για τον υπολογισμό της απόδοσης μιας καμπάνιας

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του Γεράσιμου Λ. Ρόμπολα

Επιβλέπων:
Χρήστος Μακρής
Αναπληρωτής Καθηγητής
Τμήμα Μηχανικών Η/Υ και
Πληροφορικής
Πανεπιστήμιο Πατρών

Συνεπιβλέπουσα:
Ελεάννα Καφέζα
Επίκουρη Καθηγήτρια
Τμήμα Μάρκετινγκ και
Επικοινωνίας
Οικονομικό Πανεπιστήμιο
Αθηνών

Πάτρα, Σεπτέμβριος 2017

Copyright © Γεράσιμος Λ. Ρόμπολας, 2017
Με επιφύλαξη παντός δικαιώματος. All rights reserved.

Απαγορεύεται η αντιγραφή, η αποθήκευση και η διανομή της παρούσας εργασίας, εξ' ολοκλήρου ή τμήματος αυτής, για εμπορικό σκοπό. Επιτρέπεται η ανατύπωση, η αποθήκευση και η διανομή για σκοπό μη-κερδοσκοπικό, εκπαίδευσης ή ερευνητικής φύσης, υπό την προϋπόθεση να αναφέρεται η πηγή προέλευσης και να διατηρείται το παρόν μήνυμα. Ερωτήματα που αφορούν τη χρήση της εργασίας για κερδοσκοπικό σκοπό πρέπει να απευθύνονται προς τον συγγραφέα.

Οι απόψεις και τα συμπεράσματα που περιέχονται σε αυτό το έγγραφο εκφράζουν τον συγγραφέα και δεν πρέπει να ερμηνευθεί ότι αντιπροσωπεύουν τις επίσημες θέσεις του Πανεπιστημίου Πατρών.

Το παρόν πρότυπο LaTeX έχει δημιουργηθεί από τον Γεράσιμο Ρόμπολα.

Περίληψη

Ενω οι επιχειρήσεις διαθέτουν μία πληθώρα μετρικών για τον υπολογισμό της απόδοσης των επιχειρηματικών διαδικασιών τους, πρόσφατα έχει εμφανιστεί η ανάγκη αξιολόγησής τους με βάση το βαθμό επιρροής των καταναλωτών. Στη θεωρία του σύγχρονου μάρκετινγκ τα κοινωνικά δίκτυα αποτελούν ένα από τα σημαντικότερα μέσα προώθησης προϊόντων και υπηρεσιών. Συνεπώς, η εργασία αυτή επικεντρώνεται στην αξιολόγηση της απόδοσης μιας επιχειρηματικής διαδικασίας με βάση την επίδρασή της σε ένα δημοφιλές και δυναμικό κοινωνικό δίκτυο, όπως το Twitter.

Συγκεκριμένα, στην εργασία αυτή προτείνεται ένας νέος τρόπος υπολογισμού των σχέσεων «follow» μεταξύ των χρηστών, στο Twitter. Η νέα αυτή προσέγγιση βασίζεται στις αντιδράσεις των χρηστών ως προς τις δραστηριότητες μίας καμπάνιας μάρκετινγκ, αξιοποιώντας τη συστημική ανάλυση χρονοσειρών και συναισθήματος για τον καθορισμό και τον υπολογισμό της απόδοσής της. Στη συνέχεια, ο κοινωνικός γράφος του δικτύου Twitter ανακατασκευάζεται, αξιοποιώντας τη συμπεριφορά των χρηστών ως προς τις δραστηριότητες μάρκετινγκ της επιχείρησης. Τέλος, για τον προσδιορισμό και την εξαγωγή κοινοτήτων στενά συνδεδεμένων χρηστών χρησιμοποιείται αντίστοιχα ένας αλγόριθμος ανίχνευσης κοινοτήτων και επομένως καθορίζονται οι απαραίτητες μετρικές για τον υπολογισμό της απόδοσης της καμπάνιας μάρκετινγκ.

Για την αξιολόγηση της προτεινόμενης μεθοδολογίας, γίνεται η αξιοποίηση ενός συνόλου δεδομένων βασισμένο σε ένα πολιτικό πρόσωπο, με σκοπό την ανακατασκευή της μάρκετινγκ καμπάνιας του ως ένα σύνολο δραστηριοτήτων, χρησιμοποιώντας τον αλγόριθμος λανθάνουσας κατανομής του Dirichlet (LDA). Τα αποτελέσματα της νέας προσέγγισης δείχνουν ότι η συστημική ανάλυση κοινωνικών δικτύων μπορεί να αποτελέσει μία έγκαιρη και αξιόπιστη μετρική για την αξιολόγηση της απόδοσης μιας επιχειρηματικής διαδικασίας μάρκετινγκ.

Λέξεις Κλειδιά: ανάλυση χρονοσειρών, ανίχνευση κοινοτήτων, επιχειρηματική διαδικασία καμπάνιας μάρκετινγκ, συστημική ανάλυση δεδομένων

Abstract

While businesses have a plethora of metrics to measure the performance of their business processes, recently has emerged the requirement for their evaluation based on the degree of consumer influence. In the theory of modern marketing, social networks are one of the most important means of promoting products and services. Consequently, this work focuses on evaluating the performance of a business process based on its impact on a popular and dynamic social network, such as Twitter.

Specifically, this work proposes a new way of calculating the "follow" relationships between users, on Twitter. This new approach is based on users' reactions to the marketing campaign activities, exploiting time series and sentiment analysis for defining and measuring their performance. Then, Twitter's social graph is being reconstructed, exploiting users' behavior towards business marketing activities. Finally, in order to identify and export communities of densely connected users, a community detection algorithm is used, and therefore the necessary metrics are being defined to measure the performance of the marketing campaign.

In order to evaluate the proposed methodology, a dataset based on a specific politician is being used to rebuild its marketing campaign as a set of activities using the Latent Dirichlet Allocation algorithm (LDA). The results of the new approach show that social network analytics can be a valid and reliable metric for assessing the performance of a marketing campaign business process.

Keywords: community detection, data analytics, marketing campaign business process, time series analysis

"The saddest aspect of life right now is that science gathers knowledge faster than society gathers wisdom."

— Isaac Asimov

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον καθηγητή κ. Χρήστο Μακρή για την επίβλεψη αλλά και για τη συμβολή του στην εκπόνηση αυτής της διπλωματικής εργασίας. Επίσης ευχαριστώ θερμά την καθηγήτρια κα. Ελεάννα Καφέζα για την καθοδήγησή της και για την εξαιρετική συνεργασία που είχαμε.

Τέλος θα ήθελα να ευχαριστήσω τους γονείς μου για τη υποστήριξη και την ηθική συμπαράσταση που μου προσέφεραν όλα αυτά τα χρόνια.

Πάτρα, 1 Σεπτεμβρίου 2017

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

Η ραγδαία ανάπτυξη των τηλεπικοινωνιών και της μεταφοράς δεδομένων στα τέλη του 20ου αιώνα, αποτέλεσε ακρογωνιαίο λίθο για την γέννηση του Διαδικτύου. Η αρχική μορφή του Διαδικτύου περιοριζόταν στην εξυπηρέτηση συστημάτων που βασίζονταν αποκλειστικά σε κείμενο και στις υπηρεσίες ηλεκτρονικού ταχυδρομείου. Ώσπου το 1990, ο Βρετανός Sir Tim Berners-Lee μαζί με τον Βέλγο Robet Cailliau πρότειναν ένα σύστημα υπερκειμένου [2]. Οι δυνατότητες του συστήματος υπερκειμένου κατέστησαν εφικτή τη διασύνδεση ιεραρχημένων στοιχείων, εξασφαλίζοντας τη χρήση εικόνας και ήχου. Το σύστημα αυτό χρησιμοποιείται ως σήμερα και είναι γνωστό ως ο Παγκόσμιος Ιστός.

Το πρώτο στάδιο της εξέλιξης του Παγκόσμιου Ιστού ονομάστηκε ως Web 1.0. Οι χρήστες δρούσαν παθητικά, όντας απλοί καταναλωτές της πληροφορίας, καθώς ο Παγκόσμιος Ιστός ήταν κατακλυσμένος από ένα σύνολο στατικών ιστοσελίδων. Σταδιακά, με την εμφάνιση νέων τεχνολογικών μέσων και εργαλείων εξελίχτηκε στο δεύτερο στάδιο, που ονομάστηκε αντίστοιχα ως Web 2.0. Αυτή η γενιά χαρακτηρίστηκε από την δυναμικότητα που προσέφερε, δίνοντας τη δυνατότητα άμεσης αλληλεπίδρασης μεταξύ των χρηστών. Χαρακτηριστικό στοιχείο της γενιάς αυτής αποτελεί η εφεύρεση των κοινωνικών δικτύων, μέσω των οποίων οι χρήστες μπορούσαν να εκφραστούν και να επικοινωνούν όπως ποτέ άλλοτε. Η γενιά του Web 2.0 είναι εκείνη που επικρατεί ως και σήμερα στον Παγκόσμιο Ιστό. Το επόμενο στάδιο στης εξέλιξής του θεωρείται ότι θα είναι αντίστοιχα το Web 3.0, διαδεδομένο και ως σημασιολογικός ιστός, το οποίο θα προσθέτει σημασιολογικά στοιχεία στη δομή των ιστοσελίδων.

Η δημιουργία των κοινωνικών δικτύων στον Παγκόσμιο Ιστό, οδήγησε τους χρήστες στην άμεση μεταξύ τους επαφή, να μοιράζονται και να ανταλλάσσουν περιεχόμενο σε ένα ελεύθερο περιβάλλον. Η δυνατότητα αυτή των κοινωνικών δικτύων έλκυσε εκατομμύρια ανθρώπους, με αποτέλεσμα ο όγκος του διαμοιραζόμενου περιεχομένου να αυξηθεί αλματωδώς.

Παράλληλα με την καταιγιστική ροή δεδομένων, οι χρήστες απαίτησαν από αυτά πιο εξειδικευμένες πληροφορίες. Χαρακτηριστικό παράδειγμα αποτελεί ο τομέας ανάλυσης και διαχείρισης της αγοράς, όπου η ανάλυση δεδομένων είναι ιδιαίτερα σημαντική για την πορεία των επιχειρήσεων. Συγκεκριμένα, οι επιχειρήσεις ενδιαφέρονται για τη στοχοποίηση των διαφημιστικών τους εκστρατειών σε υψηλά διασυνδεδεμένα σύνολα χρηστών, τον προσδιορισμό των προφίλ και των απαιτήσεων των πελατών, την ανάλυση και τη διαχείριση των κινδύνων, καθώς και την πρόβλεψη των αποθεμάτων τους με βάσει τις πωλήσεις. Περαιτέρω τομείς αφορούν τον εντοπισμό και τη διαχείριση του οικονομικού εγκλήματος, την ανακάλυψη βιολογικής γνώσης αλλά και την πρόβλεψη γεγονότων.

Για την ικανοποίηση αυτών των αναγκών χρησιμοποιούνται ως και σήμερα τεχνικές εξόρυξης δεδομένων και αλγόριθμοι μάθησης, μέσω των οποίων είναι εφικτό να εξαχθεί ή να παραχθεί λειτουργική γνώση που βρίσκεται κρυμμένη μέσα στα δεδομένα.

Κεφάλαιο 2

Θεωρία Γράφων

Τα τελευταία χρόνια υπάρχει ένα διαρκώς αυξανόμενο ενδιαφέρον για τη μελέτη της δομής της σύγχρονης κοινωνίας ως απόρροια της ταχείας ανάπτυξης του Διαδικτύου και του Παγκόσμιου Ιστού. Η ευκολία με την οποία πραγματοποιείται τώρα η παγκόσμια επικοινωνία και η ταχύτητα με την οποία μεταδίδονται οι πληροφορίες και οι ειδήσεις είναι εκπληκτική. Οι άνθρωποι πλέον δίχως να περιορίζονται από γεωγραφικά όρια, έχουν τη δυνατότητα να έρχονται σε άμεση επαφή με άλλους πολιτισμούς, να ανταλλάσσουν και να μοιράζονται ιδέες και απόψεις. Το γεγονός αυτό έχει ως αποτέλεσμα τη δημιουργία ευρύτερων κοινωνικών ομάδων, οι οποίες με βάση τα κίνητρά τους και τη συμπεριφορά τους μπορούν να επέμβουν και να επηρεάσουν σημαντικά τη ζωή άλλων ανθρώπων.

Με γνώμονα αυτές τις παγκόσμιες εξελίξεις, οι επιστημονικοί κλάδοι παρακινήθηκαν και συνεργάστηκαν για να κατανοήσουν τον τρόπο λειτουργίας της σύγχρονης κοινωνίας, αλλά και παρόμοιων υψηλά διασυνδεδεμένων δικτύων. Συγκεκριμένα, από την επιστήμη των υπολογιστών και των μαθηματικών προήλθε ένα συλλογιστικό πλαίσιο ερμηνείας της αύξησης της πολυπλοκότητας των δικτύων. Από τις κοινωνικές επιστήμες εντοπίστηκαν χαρακτηριστικές δομές και αλληλεπιδράσεις που εμφανίζονται μέσα σε κοινωνικές ομάδες. Και από την επιστήμη των οικονομικών ερμηνεύθηκε η επιρροή της συμπεριφοράς των ανθρώπων από τα κίνητρα και τις προσδοκίες τους. Είναι φανερό ότι κάθε επιστημονικός κλάδος συνεισφέρει αποκλειστικά δικές του γνώσεις και τεχνικές. Ωστόσο η σύνθεση όλων των συσχετιζόμενων επιστημονικών κλάδων έχει οδηγήσει στη μελέτη ακόμα πιο σύνθετων δομών δικτύων, αξιοποιώντας ταυτόχρονα

γνώσεις από όλους τους αντίστοιχους κλάδους.

Η σύνθετη μορφή αυτή του κάθε δικτύου χαρακτηρίζεται από υψηλή πολυπλοκότητα και είναι γενικά δύσκολο να εξαχθεί μια πλήρης εικόνα για τη συνολική μορφή του. Για τη μελέτη της διάρθρωσής των δικτύων είναι απαραίτητη η κατανόηση βασικών εννοιών της θεωρίας γράφων, καθώς αποτελεί το βασικό μοντέλο περιγραφής των δικτύων.

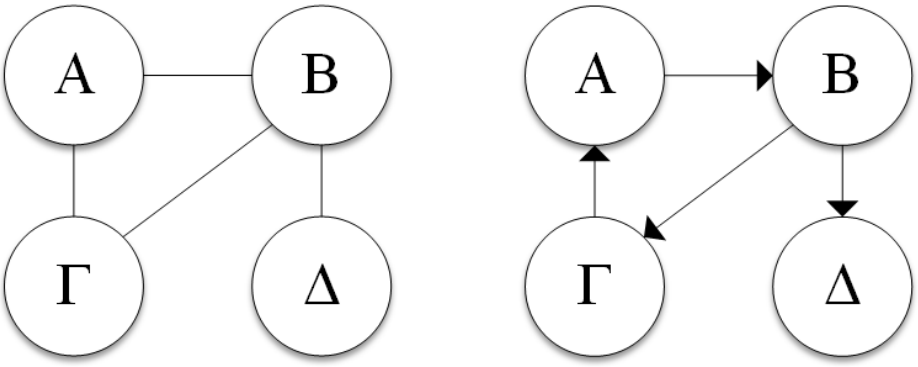
2.1 Βασικές έννοιες

Με σκοπό την περιγραφή και την κατανόηση των δομών των δικτύων, περιγράφονται αρχικά μερικοί από τους βασικούς ορισμούς και συμβολισμούς της θεωρίας γράφων.

2.1.1 (Μη) Κατευθυνόμενοι γράφοι

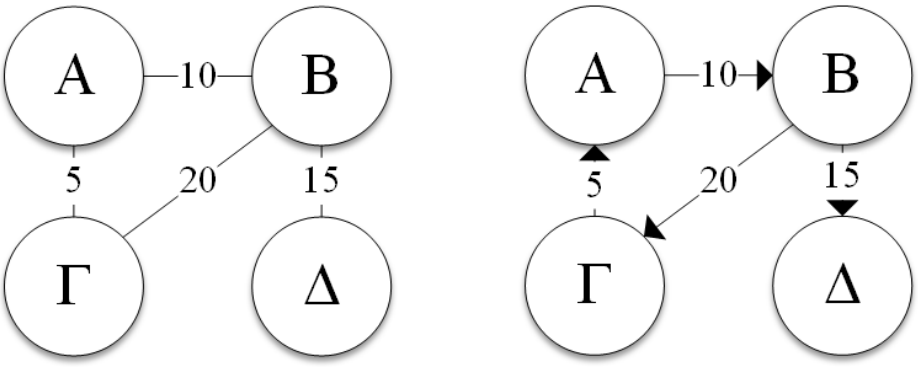
Ένας κοινός ορισμός ενός γράφου είναι η απεικόνιση των σχέσεων μεταξύ ενός συνόλου αντικειμένων και συμβολίζεται συνήθως ως $G = \{V, E\}$. Ο γράφος G αποτελείται από ένα σύνολο σημείων, V, που ονομάζονται κόμβοι, καθώς και από ένα σύνολο συνδέσεων μεταξύ των κόμβων, E, που ονομάζονται ακμές. Δύο κόμβοι ονομάζονται γειτονικοί μόνο εάν συνδέονται μέσω μίας ακμής. Στο Σχήμα 1 αναπαρίσταται ένας τυπικός τρόπος σχεδιασμού ενός γράφου, αναπαριστώντας τους κόμβους με κύκλους και τις ακμές μεταξύ 2 κόμβων με ευθύγραμμα τμήματα. Συγκεκριμένα, ο γράφος G_1 αποτελείται από τέσσερις κόμβους τους Α, Β, Γ και Δ , με τον κόμβο Β να συνδέεται με κάθε έναν από τους υπόλοιπους κόμβους μέσω ακμών, και από μία επιπλέον ακμή μεταξύ των κόμβων Α και Γ.

Στο γράφο G_1 του Σχήματος 1(α΄), η σχέση μεταξύ δύο διασυνδεδεμένων κόμβων θεωρείται συμμετρική. Ωστόσο, σε πολλές περιπτώσεις υπάρχει η ανάγκη έκφρασης ασυμμετρικών σχέσεων. Για παράδειγμα ο κόμβος Α να δείχνει στον κόμβο Β, αλλά χωρίς να ισχύει το αντίστροφο. Για τον σκοπό αυτό, ορίζεται η έννοια του κατευθυνόμενου γράφου, ο οποίος αποτελείται από ένα σύνολο διασυνδεδεμένων κόμβων μέσω ενός συνόλου κατευθυνόμενων ακμών. Κάθε κατευθυνόμενη ακμή δείχνει από τον έναν κόμβο στον άλλον, δίνοντας σημασία στην κατεύθυνση που δείχνει. Οι κατευθυ-



(α′) Μη κατευθυνόμενος γράφος G_1 (β′) Κατευθυνόμενος γράφος $\hat{G}_1$

Σχήμα 1: Δύο γράφοι αποτελούμενοι από 4 κόμβους και 4 ακμές



(α′) Μη κατευθυνόμενος γράφος G_2 (β′) Κατευθυνόμενος γράφος $\hat{G}_2$

Σχήμα 2: Δύο βεβαρημένοι γράφοι αποτελούμενοι από 4 κόμβους και 4 ακμές

νόμενοι γράφοι σχεδιάζονται συνήθως όπως στο γράφο $\hat{G}_1$ του Σχήματος 1(β′), όπου η κάθε κατευθυνόμενη ακμή αναπαρίσταται με τη χρήση ενός βέλους. Όταν γίνεται αναφορά σε έναν γράφο που δεν είναι κατευθυνόμενος, ονομάζεται ως μη κατευθυνόμενος γράφος.

Η τάξη ενός γράφου ορίζεται ως το συνολικό πλήθος των κόμβων του και συμβολίζεται ως $|V|$. Αντίστοιχα ορίζεται το μέγεθος ενός γράφου, ως το συνολικό πλήθος των ακμών του και συμβολίζεται ως $|E|$. Επίσης, κάθε κόμβος ενός γράφου χαρακτηρίζεται από το βαθμό του, ο οποίος ισούται με το σύνολο των ακμών που συνδέονται σε αυτόν.

Σε πολλές περιπτώσεις, κάθε ακμή ενός γράφου συσχετίζεται με μία αριθμητική τιμή, η οποία ονομάζεται βάρος. Συνήθως, τα βάρη των ακμών είναι μη αρνητικοί ακέραιοι αριθμοί. Ο γράφος που παρουσιάζει αυτή τη δομή ονομάζεται βεβαρημένος και μπορεί να είναι είτε κατευθυνόμενος, είτε μη κατευθυνόμενος, όπως οι γράφοι G_2 και $\hat{G_2}$ αντίστοιχα του Σχήματος 2. Το βάρος μίας ακμής ονομάζεται συχνά και ως κόστος της ακμής. Σε εφαρμογές, το βάρος μιας ακμής μπορεί να αναπαριστά ένα οποιαδήποτε μέτρο ανάμεσα στους δύο αντίστοιχους κόμβους που συνδέει. Για παράδειγμα, μπορεί να μετρά το μήκος ενός δρόμου, τη χωρητικότητα μιας γραμμής, την ενέργεια που απαιτείται για την μετακίνηση μεταξύ των κόμβων κτλ.

2.2 Αναπαράσταση γράφων

Για την πρακτική αξιοποίηση των δομών των γράφων απαιτείται η αναπαράστασή τους σε μια κατάλληλη μορφή, ώστε να είναι επεξεργάσιμη και κατανοητή από τους ανθρώπους και τους υπολογιστές. Για το λόγο αυτό, στην επιστήμη των υπολογιστών αλλά και των μαθηματικών χρησιμοποιούνται αντίστοιχες δομές αναπαράστασης των δεδομένων ανάλογα με τα δομικά χαρακτηριστικά του γράφου. Ως επί το πλείστον, οι δομές αναπαράστασης γράφων που χρησιμοποιούνται είναι τα μητρώα (Adjacency matrices) καθώς και οι λίστες γειτνίασης (Adjacency lists). Κάθε δομή προσφέρει τα δικά της πλεονεκτήματα και μειονεκτήματα ανάλογα με το είδος της απαιτούμενης επεξεργασίας του γράφου. Για την περιγραφή των παραπάνω δομών, ορίζουμε ένα γράφο $G = (V, E)$, όπου V είναι το σύνολο των κόμβων και E είναι το σύνολο των ακμών αντίστοιχα.

2.2.1 Μητρώο γειτνίασης

Το μητρώο γειτνίασης χρησιμοποιείται για την αναπαράσταση των κόμβων ενός γράφου, καθώς και για τον τρόπο σύνδεσής τους με τους υπόλοιπους κόμβους. Για έναν πεπερασμένο γράφο $G = (V, E)$, το μητρώο γειτνίασης ορίζεται ως ένας πίνακας διαστάσεων $|V| \times |V|$, του οποίου τα στοιχεία $e_{i,j}$ ορίζονται ως εξής:

$$e_{i,j} = \begin{cases} 1, \ \alpha\upsilon \ \exists \ \alpha\kappa\mu\acute{\eta} \ \mu\epsilon\tau\alpha\xi\acute{\upsilon} \ \tau\omega\upsilon \ \kappa\acute{o}\mu\beta\omega\upsilon \ V_i \ \kappa\alpha\iota \ V_j \\ 0, \ \alpha\lambda\lambda\iota\acute{\omega}s \end{cases}$$

Στην περίπτωση βεβαρημένου γράφου, ο παραπάνω ορισμός των στοιχείων του μητρώου γειτνίασης διαφοροποιείται. Συγκεκριμένα όταν υπάρχει ακμή μεταξύ δύο κόμβων, το στοιχείο του μητρώου $e_{i,j}$ ισούται με το βάρος της αντίστοιχης ακμής μεταξύ των κόμβων V_i και V_j. Στην ειδική περίπτωση που ο γράφος είναι μη κατευθυνόμενος, το μητρώο γειτνίασης που προκύπτει παρουσιάζει συμμετρική μορφή. Για παράδειγμα, στα Σχήματα 3 και 4 παρουσιάζονται τα μητρώα γειτνίασης των γραφών που μελετήθηκαν ως τώρα, των Σχημάτων 1 και 2 αντίστοιχα.

$$A = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \qquad B = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

(α′) Μη κατευθυνόμενου γράφου G_1

(β′) Κατευθυνόμενου γράφου $\hat{G}_1$

Σχήμα 3: Μητρώα γειτνίασης γράφων χωρίς βάρη

2.2.2 Λίστες γειτνίασης

Η λίστα γειτνίασης χρησιμοποιείται για την αναπαράσταση των κόμβων ενός γράφου, καθώς του συνόλου των γειτονικών τους κόμβων. Συγκεκριμένα, η λίστα γειτνίασης ενός πεπερασμένου γράφου $G = (V, E)$ είναι μία λίστα από $|V|$ αταξινόμητες λίστες. Κάθε λίστα αντιστοιχεί και σε έναν κόμβο του γράφου, για τον οποίο η αντίστοιχη λίστα περιλαμβάνει το σύνολο των γειτονικών του κόμβων. Στην περίπτωση βεβαρημένου γράφου, σε κάθε λίστα ενός κόμβου V_i με $i = 0, ..., |V|$, εκτός από την πληροφορία των γειτονικών κόμβων αποθηκεύονται επίσης και τα βάρη των αντίστοιχων ακμών μεταξύ

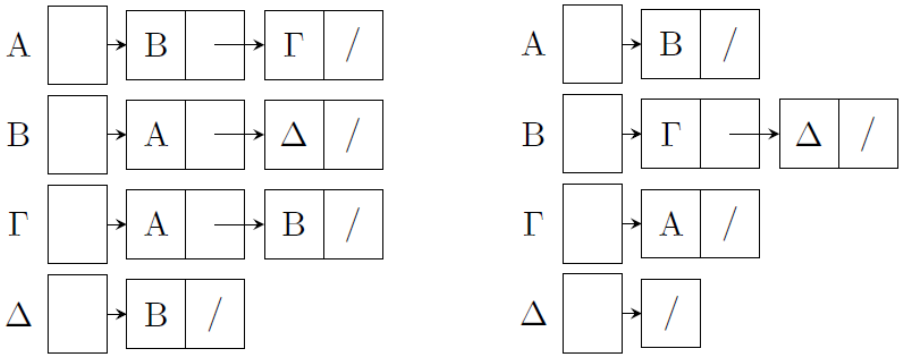
$$\Gamma = \begin{pmatrix} 0 & 10 & 5 & 0 \\ 10 & 0 & 20 & 15 \\ 5 & 20 & 0 & 0 \\ 0 & 15 & 0 & 0 \end{pmatrix} \qquad \Delta = \begin{pmatrix} 0 & 10 & 0 & 0 \\ 0 & 0 & 20 & 15 \\ 5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

(α′) Μη κατευθυνόμενου γράφου G_2

(β′) Κατευθυνόμενου γράφου $\hat{G}_2$

Σχήμα 4: Μητρώα γειτνίασης γράφων με βάρη

του κόμβου V_i και των γειτονικών του κόμβων. Για παράδειγμα, στα Σχήματα 5 και 6 παρουσιάζονται οι λίστες γειτνίασης των γραφών που μελετήθηκαν ως τώρα, των Σχημάτων 1 και 2 αντίστοιχα.

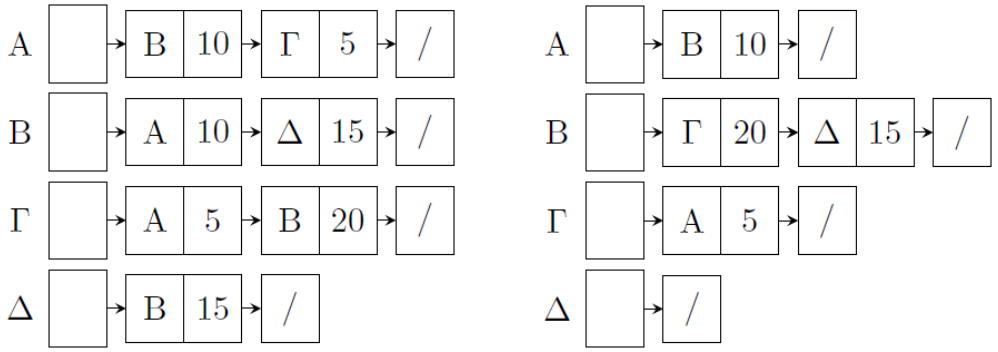


(α′) Μη κατευθυνόμενος γράφος

(β′) Κατευθυνόμενος γράφος

Σχήμα 5: Λίστες γειτνίασης γράφων χωρίς βάρη

Για την περαιτέρω κατανόηση των διαφορών μεταξύ των μητρώων και των λιστών γειτνίασης παρατίθεται ο Πίνακας 1. Ο Πίνακας 1 παρουσιάζει συγκεντρωτικά τα κόστη της χρονικής πολυπλοκότητας εκτέλεσης διαφόρων βασικών ενεργειών, που εφαρμόζονται συχνά πάνω στους γράφους. Συγκεκριμένα, στη γενική περίπτωση οι λίστες γειτνίασης προτιμώνται για την αναπαράσταση αραιών γράφων. Ωστόσο στη χειρότερη περίπτωση όπου ο γράφος είναι πυκνός, τότε η πολυπλοκότητά τους ισούται με αυτή των μητρώων γειτνίασης. Το σημαντικότερο πλεονέκτημα των λιστών γειτνίασης είναι



(α′) Μη κατευθυνόμενος γράφος (β′) Κατευθυνόμενος γράφος

Σχήμα 6: Λίστες γειτνίασης γράφων με βάρη

η άμεση προσπέλαση των γειτονικών κόμβων ενός δοθέντος κόμβου, σε σχέση με ένα μητρώο γειτνίασης. Η ιδιότητα αυτή των λιστών γειτνίασης είναι ιδιαίτερα σημαντική σε πρακτικές υλοποιήσεις πολλών αλγορίθμων.

Πίνακας 1: Χρονική πολυπλοκότητα ενεργειών πάνω στους γράφους

	Λίστα Γειτνίασης	**Μητρώο Γειτνίασης**
Αποθήκευση γράφου	$O(\|V\| + \|E\|)$	$O(\|V\|^2)$
Προσθήκη κόμβου	$O(1)$	$O(\|V\|^2)$
Προσθήκη ακμής	$O(1)$	$O(1)$
Αφαίρεση κόμβου	$O(\|E\|)$	$O(\|V\|^2)$
Αφαίρεση ακμής	$O(\|V\|)$	$O(1)$
Έλεγχος γειτνίασης 2 κόμβων	$O(\|V\|)$	$O(1)$

2.2.3 Απλό διανυσματικό μοντέλο (VSM)

Το διανυσματικό μοντέλο (Vector Space Model) προέρχεται από το χώρο της ανάκτησης πληροφορίας, όπου είναι ιδιαίτερα διαδεδομένο [1]. Η φιλοσοφία του μοντέλου βασίζεται στη δημιουργία ενός λεξικού, στο οποίο καταγράφεται για κάθε κείμενο το πλήθος εμφάνισης της κάθε λέξης.

Η λειτουργία του χαρακτηρίζεται από δύο στάδια. Αρχικά, προηγείται ένα στάδιο ευρετηρίασης όλων των κειμένων, όπου γίνεται καταγραφή όλων των λέξεων w_i, $i \in [1, M]$ που απαρτίζουν κάθε κείμενο, ενώ στη συνέχεια σε κάθε λέξη w_i ενός κειμένου d_j, $j \in [1, N]$, ανατίθεται ένα βάρος TF-IDF (Term Frequency – Inverse Document Frequency) σύμφωνα με τη Σχέση 1.

$$v_{i,j} = \begin{cases} (1 + log f_{i,j}) log\left(\frac{N}{n_i}\right), \ \alpha v \ f_{i,j} > 0 \\ 0, \quad \alpha\lambda\lambda\iota\acute{\omega}s \end{cases} \tag{1}$$

όπου:

$f_{i,j}$ είναι η συχνότητα της λέξης w_i στο κείμενο d_j

n_i είναι το πλήθος των κειμένων στο οποίο εμφανίζεται η λέξη w_i

N είναι το συνολικό πλήθος των κειμένων

Τα βάρη TF-IDF των λέξεων ενός κειμένου d_j σχηματίζουν μια διανυσματική περιγραφή του κειμένου, ως $\vec{d_j} = (v_{1,j}, v_{2,j}, ..., v_{M,j})$. Ενώ αντίστοιχα, τα βάρη TF-IDF των λέξεων του κάθε κειμένου d_j, για $\forall j \in [1, N]$, σχηματίζουν ένα μητρώο βαρών $U_{i,j}$ διαστάσεων $M \times N$, μεταξύ των λέξεων w_i και των κειμένων d_j ως εξής:

$$U_{i,j} = \begin{pmatrix} v_{1,1} & v_{1,2} & \dots & v_{1,M} \\ v_{2,1} & v_{2,2} & \dots & v_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ v_{N,1} & v_{N,2} & \dots & v_{N,M} \end{pmatrix}$$

Κεφάλαιο 3

Ανάλυση Χρονοσειρών

Μια χρονοσειρά είναι μία ακολουθία από καλά ορισμένα σημεία δεδομένων σε σταθερά χρονικά διαστήματα μίας χρονικής περιόδου. Η ανάλυση χρονοσειρών αναφέρεται στη χρήση στατιστικών μεθόδων για την μελέτη αυτών των χρονικών σημείων, με σκοπό την εξαγωγή χρήσιμων στατιστικών και χαρακτηριστικών των δεδομένων. Το μεγαλύτερο πλεονέκτημα της ανάλυσης των χρονοσειρών αποτελεί η δυνατότητα αξιοποίησής τους για την κατανόηση των ήδη υπαρχουσών δεδομένων, αλλά και για τη μελέτη των μελλοντικών τιμών.

3.1 Βασικές έννοιες

Επίσημος ορισμός μίας χρονοσειράς δεν υπάρχει, καθώς υπάρχει μία πληθώρα εναλλακτικών ορισμών. Πολλοί ορισμοί περιγράφουν μία χρονοσειρά ως μία ακολουθία αποκλειστικά αριθμητικών τιμών, ενώ άλλοι θεωρούν ότι οι τιμές αποτελούν σαφή ορισμένα και ισαπέχοντα χρονικά σημεία. Συνεπώς, για την περιγραφή ενός γενικού μοντέλου χρονοσειράς δίνονται οι Ορισμοί 1 και 2 [3].

Ορισμός 1: *Δοθείσης μιας διάστασης δεδομένων, D μια* ***χρονοσειρά*** *είναι ένα σύνολο από n τιμές:* $\{< t_1, d_1 >, < t_2, d_2 >, ..., < t_n, d_n >\}$. *Σε κάθε χρονική στιγμή* t_i, *όπου* $i \in [1, n]$, *αντιστοιχεί μία τιμή δεδομένων* d_i. *Στην περίπτωση όπου οι τιμές ορίζονται σε συγκεκριμένες και καλά ορισμένες χρονικές στιγμές, τότε οι τιμές αυτές μπορούν να αναπαρασταθούν σαν ένα διάνυσμα* $< d_1, d_2, ..., d_n >$.

Παράδειγμα παρουσιάσης αλγορίθμου:

Αλγόριθμος 1: Κατασκευή του Twitter keepup γράφου

Είσοδος: Ένα σύνολο κόμβων N, ένα κατώφλι th, μία συνάρτηση για τον υπολογισμό της $απόστασης$

Έξοδος: Ο keepup γράφος $G_{th}(N,E)$

1: $E = \text{κενό}$
2: **για κάθε** $(n_i, n_j) \in N$ **κάνε**
3: $keepup(n_i, n_j) = 1 - (|distance(N_c, N_i) - distance(N_c, N_j)|)$
4: $e_{ij} = keepup(n_i, n_j)$
5: **αν** $e_{ij} <= th$ **τότε**
6: $E = E \cup e_{ij}$
7: **τέλος αν**
8: **τέλος για**

Βιβλιογραφία

[1] Baeza-Yates, R., Ribeiro-Neto, B., et al. (1999). *Modern information retrieval*, volume 463. ACM press New York.

[2] Connolly, R. and Hoar, R. (2014). *Fundamentals of Web Development*. Always learning. Pearson.

[3] Dunham, M. (2003). *Data Mining Introductory and Advanced Topics*. An Alan R. Apt book. Prentice Hall/Pearson Education.