Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Τεχνολογιας Πληροφορικης και Υπολογιστων

Εποπτεία γεωγραφικής κάλυψης συζητήσεων σε κοινωνικά δίκτυα

Διπλωματικη Εργασια

του

ΕΜΜΑΝΟΥΗΛ ΛΟΥΚΑΔΑΚΗ

Επιβλέπων: Ιωάννης Βασιλείου
Καθηγητής Ε.Μ.Π.

Εργαστηριο Συστηματων Βασεων Γνωσεων και Δεδομενων
Αθήνα, Απρίλιος 2016

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών
Εργαστήριο Συστημάτων Βάσεων Γνώσεων και Δεδομένων

Εποπτεία γεωγραφικής κάλυψης συζητήσεων σε κοινωνικά δίκτυα

Διπλωματικη Εργασια

του

ΕΜΜΑΝΟΥΗΛ ΛΟΥΚΑΔΑΚΗ

Επιβλέπων: Ιωάννης Βασιλείου
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 8η Απριλίου 2016.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Ιωάννης Βασιλείου	Νεκτάριος Κοζύρης	Ιωάννης Θεοδωρίδης
Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Καθηγητής Παν. Πειραιώς

Αθήνα, Απρίλιος 2016

(Υπογραφή)

.....

Λουκαδακης Εμμανουηλ

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών
Εργαστήριο Συστημάτων Βάσεων Γνώσεων και Δεδομένων

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή κ. Ι. Βασιλείου, καθώς επίσης και τον κ. Ι. Θεοδωρίδη και κ. Ν. Κοζύρη που συμμετείχαν στην επιτροπή εξέτασης.

Ιδιαίτερα, θα ήθελα να ευχαριστήσω τον Κώστα Πατρούμπα για την τεράστια βοήθεια που μου παρείχε προκειμένου να βγεί εις πέρας η παρούσα διπλωματική εργασία.

Επίσης ευχαριστώ ιδιαίτερα τους Χ. Ευσταθιάδη, Α. Μπελεσιώτη και Δ. Σκούτα που μου παρείχαν τα πειραματικά δεδομένα με βάση τα οποία πραγματοποιήθηκε η πειραματική αξιολόγηση.

Τέλος θα ήθελα να ευχαριστήσω την οικογένειά μου και την Αναστασία, για την μεγάλη ψυχολογική υποστήριξη και συμπαράσταση που μου προσέφεραν όλον αυτόν τον καιρό.

Περίληψη

Αντικείμενο της διπλωματικής εργασίας είναι η ανάπτυξη ενός αλγορίθμου για την ανίχνευση δημοφιλών συζήτησεων που εκτυλίσσονται στα κοινωνικά δίκτυα, καθώς και για την εποπτεία της γεωγραφικής τους κάλυψης με την πάροδο του χρόνου. Συγκεκριμένα, η μελέτη επικεντρώθηκε στο κοινωνικό δίκτυο Twitter, καθώς τα σχετικά μηνύματα δημοσιεύονται με πολύ γρήγορο ρυθμό, επομένως θεωρείται ότι συγκροτούν ένα ρεύμα δεδομένων. Όταν διαδραματίζεται ένα γεγονός (λ.χ. ποδοσφαιρικός αγώνας, διαδήλωση, συναυλία), οι χρήστες των κοινωνικών δικτύων έχουν την τάση να δημοσιεύουν μηνύματα και να συζητούν σχετικά με αυτό. Τέτοια γεγονότα μπορούν να χαρακτηριστούν ως δημοφιλείς συζητήσεις και τα σχετικά μηνύματα στο Twitter συνήθως συνοδεύονται από χαρακτηριστικές ετικέτες. Πολλές φορές, μία τέτοια συζήτηση είναι δημοφιλής μόνο κατά τόπους, υποθέτοντας βέβαια ότι είναι γνωστή η γεωγραφική θέση των χρηστών που συμμετέχουν σ΄ αυτήν. Στην παρούσα διπλωματική εργασία, μελετάμε την εξέλιξη τέτοιων συζητήσεων στον χώρο και στον χρόνο, ανανεώνοντας τις περιοχές επιρροής τους σε πραγματικό χρόνο. Παρακολουθώντας τις μεταβολές στην γεωγραφική κάλυψη τέτοιων φαινομένων, μπορούμε να εξάγουμε χρήσιμα συμπεράσματα σχετικά με την ένταση κάθε φαινομένου και τον αντίκτυπο που έχει στην κοινωνία.

Η εργασία επικεντρώνεται κυρίως στην ανάπτυξη τεχνικών για ανίχνευση συζητήσεων, προσδιορισμό των πιο δημοφιλών από αυτές και εποπτεία της γεωγραφικής κάλυψής τους. Ο αλγόριθμος είναι επίτηδες προσεγγιστικός ως προς την εκτίμηση της γεωγραφικής κάλυψης των δημοφιλών συζητήσεων, στοχεύοντας σε βελτιωμένες επιδόσεις ως προς τον χρόνο εκτέλεσης. Πειραματική μελέτη σε πραγματικά δεδομένα από το Twitter για την ευρύτερη περιοχή του Λονδίνου, επιβεβαίωσε ότι η προτεινόμενη μεθοδολογία είναι ικανή να αντιμετωπίσει κλιμακούμενους όγκους μηνυμάτων για διάφορες τιμές παραμέτρων επιστρέφοντας εκτιμήσεις καλής ποιότητας. Ως γενικό συμπέρασμα της εργασίας μπορεί να ειπωθεί ότι ο αλγόριθμος είναι κατάλληλος για ανίχνευση της χωρικής κάλυψης τέτοιων συζητήσεων, θυσιάζοντας την ακρίβεια προς όφελος της έγκαιρης απόκρισης.

Λέξεις Κλειδιά

Κοινωνικά δίκτυα, δημοφιλείς συζητήσεις, χωρική κάλυψη, μηνύματα με γεωγραφικό στίγμα, ρεύματα δεδομένων.

Abstract

The purpose of this diploma thesis is to develop an algorithm for detection of popular topics unfolding in social networks, as well as for monitoring their spatial coverage through time. We particularly focus our study on Twitter, mainly because the posting rate of such messages is extremely high, so they can be considered as a data stream. When an event (e.g., a match, a demonstration, or a concert) is ongoing, many users tend to post messages referring to that event in social networks. Essentially, this creates a discussion over the same topic, which is usually characterized by hashtag(s). Sometimes, such a topic may be popular only locally, assuming that related posts have a geographical reference, which represents the location where the user posted the message. In this diploma thesis, we study how trending topics evolve both in space and time, updating the corresponding local areas in real time. Observing the evolving spatial coverage of such posts may reveal the intensity of the phenomenon and its impact on local communities.

This thesis mainly focuses on developing efficient methods for topic detection, identification of trending topics and monitoring their spatial coverage, aiming at significant reduction in processing costs. The suggested algorithm is deliberately chosen to be approximate with respect to the calculation of spatial coverages. An experimental study was conducted against real-world data from Twitter concerning the greater area of London. These tests confirmed that the proposed methodology is able to perform against scalable volumes of data under various parameter settings and can also yield good-quality results. The overall conclusion of this thesis is that the algorithm is suitable for real time detection of the spatial footprint of evolving topics, where some accuracy may be sacrificed for the benefit of timely response.

Keywords

Social networks, popular topics, spatial coverage, geotagged messages, data streams.

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

Την τελευταία δεκαετία, η χρήση των κοινωνικών δικτύων γνωρίζει ιδιαίτερη άνθηση παγκοσμίως. Ολοένα και περισσότεροι χρήστες καθημερινά δημοσιεύουν μηνύματα στα κοινωνικά δίκτυα εκφράζοντας τα συναισθήματά τους, τις απόψεις τους, σχόλια κτλ. Ενδεικτικά, σήμερα (2016) περίπου 2.000.000.000 χρήστες χρησιμοποιούν τα κοινωνικά δίκτυα. Αυτή η ραγδαία αύξηση της χρήσης τους έχει προσελκύσει και το ενδιαφέρον της επιστημονικής κοινότητας. Η ανάλυση τέτοιων μηνυμάτων μπορεί να δώσει πληροφορίες σχετικά με τα συναισθήματα των χρηστών, την σκιαγράφηση της προσωπικότητάς τους, τις πολιτικές τους πεποιθήσεις, τάσεις στην κοινή γνώμη κ.ά. Κάποια δημοφιλή κοινωνικά στην εποχή μας είναι: Twitter, Facebook, Foursquare κ.ά. Πιο συγκεκριμένα, στο κοινωνικό δίκτυο Twitter, σύμφωνα με επίσημα στατιστικά για το έτος 2015, οι χρήστες δημοσίευαν μηνύματα με ρυθμό 500.000.000 την ημέρα. Λόγω της εφήμερης φύσης τους, τα δεδομένα αυτά θεωρούνται ως *ρεύματα δεδομένων*. Επομένως, για την επεξεργασία τους θα πρέπει να ληφθούν υπόψη τα εξής χαρακτηριστικά:

- Μεγάλος όγκος πληροφορίας (Volume).
- Μεγάλος ρυθμός παραγωγής δεδομένων (Velocity).
- Μεγάλη ποικιλία δεδομένων (Variety).

Η ταχύτητα με την οποία δημιουργείται αυτή η πληροφορία και ο μεγάλος όγκος δεδομένων απαιτεί την ανάπτυξη νέων αποδοτικών μεθόδων επεξεργασίας. Η αποθήκευση των δεδομένων σε ένα κλασσικό σύστημα βάσεων δεδομένων δεν είναι αποδοτική, καθώς τα δεδομένα είναι ανεξάντλητα και ανανεώνονται ανά τακτά χρονικά διαστήματα, επομένως η τήρηση και η ανανέωση των δεδομένων είναι αργή διαδικασία. Έτσι, θα πρέπει να σχεδιαστούν αλγόριθμοι ενός περάσματος (single pass) και η επεξεργασία να πραγματοποιείται στην κύρια μνήμη. Κατά καιρούς έχουν προταθεί μέθοδοι επεξεργασίας δεδομένων σε κοινωνικά δίκτυα. Ενδεικτικά κάποιες από αυτές είναι:

- Top-k χωροκειμενικά ερωτήματα: Ο χρήστης υποβάλλει κάποιο ερώτημα προκειμένου να εντοπιστούν τα k εγγύτερα εστιατόρια τα οποία σερβίρουν πιάτα με την κουζίνα της προτίμησής του.

- Ανίχνευση τοπικών γεγονότων: Με την επεξεργασία των δεδομένων σε κοινωνικά δίκτυα υπάρχει η δυνατότητα ανίχνευσης γεγονότων. Όταν διαδραματίζεται ένα γεγονός, τότε οι χρήστες έχουν την τάση να δημοσιεύουν μηνύματα για αυτό. Η επεξεργασία αυτής της πληροφορίας μας επιτρέπει να εντοπίσουμε και να μάθουμε γεγονότα πολύ πιο γρήγορα από τα μέσα μαζικής ενημέρωσης.

1.1 Αντικείμενο της διπλωματικής εργασίας

Τα μηνύματα που δημοσιεύουν οι χρήστες κοινωνικών δικτύων (social networks) έχουν συνήθως επίκαιρο χαρακτήρα, λ.χ. σχολιάζουν ένα γεγονός που μόλις συνέβη, πυροδοτώντας συχνά μία ακολουθία σχολίων για το ίδιο θέμα *(trending topic)*. Ωστόσο, πολλές φορές μία *συζήτηση* μ' αυτήν την θεματολογία έχει και χωρική πτυχή, όταν τα μηνύματα συνοδεύονται από γεωγραφικό προσδιορισμό, όπως η τρέχουσα θέση του σχολιαστή *(geo-tagged tweets)* ή αναφορά μιας τοποθεσίας στο κείμενο (λ.χ., #Syntagma), η οποία προφανώς συνδέεται με το περιεχόμενο του μηνύματος (π.χ. διαδήλωση στην πλατεία Συντάγματος). Έχει λοιπόν ενδιαφέρον να μελετηθεί πώς εξελίσσονται τέτοιες συζητήσεις από χωροχρονική σκοπιά, παρατηρώντας την μεταβαλλόμενη έκταση και την χωρική διακύμανση των σχετικών μηνυμάτων που καταφθάνουν ως *ρεύμα δεδομένων (data stream)* με την πάροδο του χρόνου. Για παράδειγμα, τα μηνύματα σχετικά με μία πυρκαγιά αρχικά προέρχονται από παρατηρητές κοντά στην αρχική εστία, αλλά σταδιακά μπορεί να καλύψουν ευρύτερη περιοχή (ίσως με διαφορετική συχνότητα κατά τόπους), και τελικά εκλείπουν όταν η πυρκαγιά τεθεί υπό έλεγχο. Σύμφωνα με πρόσφατες επιστημονικές διερευνήσεις [1, 2, 3, 12, 26], η διασπορά και η χωρική έκταση συσχετιζόμενων μηνυμάτων μπορεί ν' αποκαλύψει την ένταση του φαινομένου και την πιθανή επίδρασή του στο κοινωνικό περιβάλλον, προκειμένου να βελτιωθεί η ενημερότητα των χρηστών σε γεγονότα και καταστάσεις με ισχυρό τοπικό αντίκτυπο. Γι' αυτόν τον σκοπό, στην εργασία αυτή επιχειρήθηκε ο σχεδιασμός μιας μεθόδου που να λαμβάνει υπ' όψιν το περιεχόμενο των κειμένων, την χρονική τους διάρκεια, καθώς και την χωρική τους κατανομή, με στόχο την αποτελεσματική *συνάθροιση (aggregation)* και την *ανίχνευση (monitoring)* της εξέλιξής τους.

Από τα διαθέσιμα κοινωνικά δίκτυα επιλέχθηκε να μελετηθεί το κοινωνικό δίκτυο Twitter καθώς η δομή των μηνυμάτων που δημοσιεύονται έχουν μια συγκεκριμένη απλή δομή, η οποία εξυπηρετεί και διευκολύνει την επεξεργασία των μηνυμάτων για την παρούσα διπλωματική εργασία.

1.2 Οργάνωση του τόμου

Η εργασία οργανώνεται σε 8 κεφάλαια. Στα κεφάλαια 2–5 παρέχονται το θεωρητικό υπόβαθρο και η βιβλιογραφική επισκόπηση. Στα κεφάλαια 6–8 πραγματοποιείται η μοντελοποίηση του προβλήματος, η μεθοδολογία που σχεδιάστηκε για την επίλυση και η πειραματική αξιολόγηση της υλοποιημένης μεθόδου. Ειδικότερα:

Στο Κεφάλαιο 2 περιγράφεται το μοντέλο των ρευμάτων δεδομένων, δίνοντας έμφαση

στην αδυναμία των συστημάτων διαχείρισης βάσεων δεδομένων όσον αφορά την αποδοτική επεξεργασία δυναμικά μεταβαλλόμενων δεδομένων. Παρουσιάζονται τα χαρακτηριστικά των ρευμάτων δεδομένων καθώς και οι επεκτάσεις που έγιναν στις συμβατικές βάσεις δεδομένων. Επίσης, γίνεται αναφορά στη διαχείριση των χωροχρονικών βάσεων δεδομένων.

Στο Κεφάλαιο 3 γίνεται μια επισκόπηση των δομών που έχουν προταθεί κατά καιρούς στην βιβλιογραφία για την τήρηση χωροκειμενικών δεδομένων και στο Κεφάλαιο 4 περιγράφονται διάφοροι μέθοδοι επεξεργασίας που πραγματοποιούνται σε δεδομένα κοινωνικών δικτύων.

Στο Κεφάλαιο 5 μοντελοποιείται και ορίζεται το πρόβλημα που πραγματεύεται η παρούσα διπλωματική εργασία.

Στο Κεφάλαιο 6 παρουσιάζεται ο αλγόριθμος επεξεργασίας που επινοήθηκε για την εποπτεία γεωγραφικής κάλυψης συζητήσεων σε κοινωνικά δίκτυα. Ο αλγόριθμος ανανεώνει τα αποτελέσματα ανά τακτά χρονικά διαστήματα. Τα αποτελέσματα που εξάγει είναι προσεγγιστικά, έχοντας δεχθεί ως είσοδο τα δημοσιευμένα μηνύματα σε πραγματικό χρόνο.

Στο Κεφάλαιο 7 αξιολογείται πειραματικά ο αλγόριθμος πάνω σε πραγματικά δεδομένα από το Twitter και σχολιάζονται διεξοδικά τα αποτελέσματα.

Τέλος, στο Κεφάλαιο 8 εκτίθενται τα γενικά συμπεράσματα καθώς και πιθανές μελλοντικές επεκτάσεις της μεθόδου.

Κεφάλαιο 2

Ρεύματα Δεδομένων

2.1 Εισαγωγή

Οι παραδοσιακές βάσεις δεδομένων αποθηκεύουν στατικά δεδομένα. Ενώ αυτό το μοντέλο δεδομένων είναι επαρκές για εμπορικούς καταλόγους και γενικά για εφαρμογές όπου τα δεδομένα είναι στατικά, πολλές πρόσφατες εφαρμογές απαιτούν την ανάλυση των δεδομένων σε πραγματικό χρόνο πάνω σε δεδομένα που αλλάζουν ραγδαία με την πάροδο του χρόνου. Για αυτό το λόγο έχει προταθεί το μοντέλο ρευμάτων δεδομένων (data stream model), το οποίο μπορεί να διαχειριστεί τέτοιου είδους μεταβαλλόμενες, εφήμερες και ενδεχομένως ελλιπείς πληροφορίες.

Οι πιο χαρακτηριστικές εφαρμογές, στις οποίες εμφανίζεται η ανάγκη της δυναμικής επεξεργασίας δεδομένων, είναι:

- *αισθητήρων* (sensor networks): Οι μετρήσεις που αποστέλλονται έχουν την μόρφή του ρεύματος δεδομένων.
- *Παρακολούθηση κινούμενων αντικειμένων(moving objects)*, π.χ αυτοκίνητα σε οδικά δίκτυα, πλοία στη θάλασσα και αεροπλάνα, άγρια ζώα σε δρυμούς κ.α.
- *Χρηματοπιστωτικές συναλλαγές*, οι οποίες παρακολουθούν σε πραγματικό χρόνο τιμές μετοχών, συσχετίσεις μεταξύ τους κ.α.

Στο κεφάλαιο αυτό θα μελετηθεί η φύση των ρευμάτων δεδομένων και τις απαιτήσεις που πρέπει να ικανοποιεί ένα σύστημα διαχείρισης ρευμάτων δεδομένων(ΣΔΡΔ).Επιπρόσθετα θα μελετηθούν διάφορες τεχνολογίες που έχουν αναπτυχθεί πάνω σε αυτό το μοντέλο δεδομένων και η βελτιστοποίηση των σχετικών ερωτημάτων.

2.2 Ρεύματα δεδομένων

Ρεύμα δεδομένων είναι μια ακολουθία από δεδομένα που παράγεται από μία ή περισσότερες πηγές μέσω ενός καναλιού επικοινωνίας σε κάποιο δέκτη. Τα δεδομένα αυτά καταφτάνουν με συγκεκριμένη σειρά και πολλές φορές μπορούν να εμφανιστούν μόνο για μία φορά.

Ο ακριβής τύπος της πληροφορίας διαφέρει απο εφαρμογή σε εφαρμογή, είναι δυνατό να αποτελεί την παρατήρηση ενός φυσικού φαινομένου ή δευτερογενής πληροφορία όπως πακέτα δεδομένων σε κάποιο δίκτυο. Παρ΄ όλες τις διαφορετικές μορφές, τα ρεύματα δεδομένων υπακούουν σε ένα γενικό μοντέλο. Το μοντέλο αυτό προσδίδει στα ρεύματα δεδομένων την μορφή μια ακολουθίας απο πλειάδες(tuples), με σκοπό να είναι δυνατό να αξιποιηθεί απο τα σύγχρονα ψηφιακά συστήματα.

Για παράδειγμα, το σύστημα παρακολούθησης των πακέτων σε ένα δίκτυο καταγράφει τις κινήσεις των πακέτων σε πραγματικό χρόνο, έπειτα επεξεργάζεται τα δεδομένα που έχει συλλέξει προκειμένου να κάνει κάποιες ενέργειες, όπως λ.χ να εντοπίζει έγκαιρα αν τα αιτήματα που δέχεται είναι μία κυβερνοεπίθεση.

2.2.1 Μοντέλο ρευμάτων δεδομένων

Οπως αναφέραμε παραπάνω,ένα ρεύμα δεδομένων αποτελείται από μια ακολούθια πλειάδων που στις περισσότερες περιπτώσεις περιλαμβάνουν την χρονική στιγμή της καταγραφής τους, το αναγνωριστικό του αποστολέα, που προσδίδει την ταυτότητα του και είναι μοναδικό και την τιμή ή τις τιμές που μετρώνται. Οι πλειάδες ενδέχεται να καταφτάνουν με τυχαία σειρά, ενδεχομένως διαφορετική απο την χρονική στιγμή καταγραφής τους εξαιτίας καθυστερήσεων εντός του δικτύου. Φορμαλιστικά η εκάστοτε πλειάδα σε ένα ρεύμα δεδομένων έχει την μορφή:

$$< t_i, s_i, m >$$

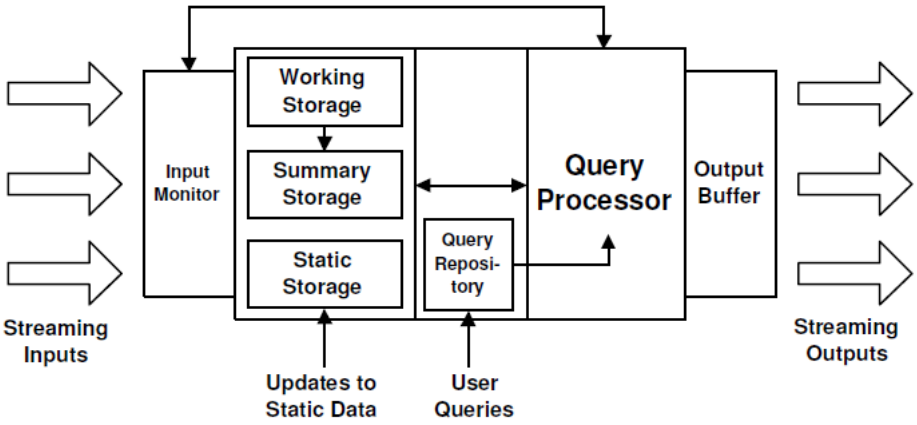
όπου:

- t_i: Η χρονική στιγμή καταγραφής της πληροφορίας,χρονόσημο (timestamp).
- s_i: Το αναγνωριστικό (identifier) του αποστολέα.
- m: Η πληροφορία που καταγράφεται. Μπορεί να είναι και αυτή μία πλειάδα σε περίπτωση που τα μεγέθη ενδιαφέροντος είναι περισσότερα του ενός. Παραδείγματος χάρη, ένας αισθήτηρας που δίνει μετρήσεις συνεχώς όπως θερμοκρασία περιβάλλοντος και υγρασία, την χρονική στιγμή 10:21:32 της ημέρας, αναγράφει θερμοκρασία 17 βαθμούς κελσίου και υγρασία 65%. Σύμφωνα με το μοντέλο ρευμάτων δεδομένων που περιγράψαμε προηγουμένως η μέτρηση έχει την εξής μορφή: $< 10 : 21 : 32, 3314, (17, 65) >$, όπου ο αριθμός 3314 δηλώνει τον αριθμό μετρήσεων που έχουν πραγματοποιηθεί μέχρι την παρούσα χρονική στιγμή.

2.3 Συστήματα διαχείρισης ρευμάτων δεδομένων

2.3.1 Προδιαγραφές συστήματος

Τα ρεύματα δεδομένων διαφέρουν από τα παραδοσιακά ΣΔΒΔ σε αρκετά σημεία. Αυτό οφείλεται στις ειδοποιές διαφορές που έχει το μοντέλο ρευμάτων δεδομένων με τα κλασσικά μοντέλα δεδομένων που διαχειρίζονται οι σχεσιακές βάσεις δεδομένων. Οι κυριότερες δυσκολίες που προκύπτουν έχουν να κάνουν με την μετάδοση ολόκληρης της πληροφορίας στο



Σχήμα 2.1: Σύστημα διαχείρισης ρευμάτων δεδομένων (Πηγή: [15]

σύστημα, τον υπολογισμό απαιτητικών συναρτήσεων της εισόδου με ρυθμό όμοιο με τον ρυθμό άφιξης της πληροφορίας και την αποθήκευση της εισόδου σε προσωρινές δομές ή περιστασιακά στην εξωτερική μνήμη. Οι προδιαγραφές ενός Συστήματος Διαχείρισης Ρευμάτων Δεδομένων (ΣΔΡΔ) είναι οι εξής:

- Τα εισερχόμενα στοιχεία ή πλειάδες καταφθάνουν σε πραγματικό χρόνο (online).
- Τα ρεύματα έχουν δυνητικά απεριόριστο μέγεθος (unbounded size).
- Η επεξεργασία οφείλει να γίνεται στην κύρια μνήμη για λόγους ταχύτητας αποκρίσεων στα ερωτήματα των χρηστών.
- Το σύστημα θα πρέπει να εξετάζει κάθε στοιχείο που εισέρχεται μόνο μία φορά (one pass algorithm).
- Το σύστημα πρέπει να είναι σε θέση να αντιμετωπίσει τυχόν διακυμάνσεις στο ρυθμό άφιξης της πληροφορίας (arrival rate).
- Η χρονική διάταξη των στοιχείων σε ένα ή μεταξύ περισσότερων ρευμάτων δεδομένων, δεν είναι πάντα σωστή, λόγω καθυστερήσεων στη μετάδοση, απωλειών κλπ.
- Το σύστημα οφείλει να ισοσταθμίσει τους πόρους του συστήματος σε χρόνο επεξεργασίας και μνήμη με την ακρίβεια των αποτελεσμάτων που δίνει στα ερωτήματα.
- Η κύρια μνήμη στην οποία γίνεται η επεξεργασία έχει πεπερασμένο μέγεθος.
- Προτιμάται η απάντηση των ερωτημάτων που τίθενται στο σύστημα σε πραγματικό χρόνο, ακόμα και αν αυτό έχει κόστος στην ακρίβεια της.

Κατά καιρούς έχουν αναπτυχθεί πολλά μοντέλα διαχείρισης ρευμάτων δεδομένων, όμως όλα βασίζονται σε ένα γενικό μοντέλο και υπακούουν στις προδιαγραφές που περιγράψαμε προηγούμενως. Το μοντέλο φαίνεται αφηρημένα στο σχήμα 2.1.

2.3.2 Κυριότερα συστήματα διαχείρισης ρευμάτων δεδομένων

Η φύση των ρευμάτων δεδομένων απαιτεί ειδική μεταχείριση που δεν μπορούν να την παρέχουν τα κλασσικά συστήματα διαχείρισης βάσεων δεδομένων. Έτσι αναπτύχθηκαν κατά καιρούς πρωτότυπα ακαδημαικά και εμπορικά συστήματα διαχείρισης ρευμάτων δεδομένων τα οποία ταιριάζουν και δουλεύουν αποδοτικά με το μοντέλο ρευμάτων δεδομένων. Τα κυριότερα ακαδημαικά συστήματα από αυτά είναι [15]:

- AURORA
- STREAM
- TelegraphCQ
- NiagaraCQ
- Gigascope

Ορισμένα από αυτά αποτελούν συστήματα γενικού σκοπού (AURORA, STREAM), ενώ άλλα απευθύνονται κατά κύριο λόγο σε δικτυακές εφαρμογές και δίκτυα αισθητήρων (TelegraphCQ, NiagaraCQ, Gigascope). Αυτά τα συστήματα χρησιμοποιούν ειδικούς τελεστές που είναι ικανοί να διαχειριστούν καλύτερα και προτείνονται επεκτάσεις στο καθιερωμένο πρότυπο της SQL οι οποίες υποστηρίζουν λειτουργίες παραθύρων (ωινδοως). Για παράδειγμα, η CQL (Continuous Query Language), αποτελεί υπερσύνολο της γλώσσας SQL και σχεδιάστηκε στα πλαίσια του STREAM που αναπτύχθηκε στο Stanford, η οποία προσθέτει εξειδικευμένες δομές για υποστήριξη παραθύρων και δειγματοληψίας. Το TelegraphCQ εστιάζει στην ευελιξία και προσαρμοστικότητα εκτέλεσης ερωτημάτων διαρκείας εισάγοντας τον μηχανισμό Εδδψ. Απευθύνεται σε έντονα και ασταθή περιβάλλοντα πληροφορίας. Κατά την ανάπτυξη του STREAM ιδιαίτερη βάση δόθηκε στην αποτελεσματική εκμετάλλευση περιορισμένης μνήμης κατά την επεξεργασία ερωτημάτων. Εξετάζει τρόπους παραγωγής προσεγγιστικών αποτελεσμάτων και επιχειρεί να περιορίσει τις απαιτήσεις σε μνήμη. Τέλος, το AURORA επικεντρώνεται στην προσπάθεια παροχής στους χρήστες της δυνατότητας να επιλέγουν δυναμικά την ποιότητα υπηρεσιών που επιθυμούν (Quality of Service), για την εκτέλεση ερωτημάτων και στη συνέχεια να χρησιμοποιούν αυτά τα μέτρα ποιότητας των απαντήσεων για την κατανομή φόρτου εργασίας. Κατα καιρούς, όπως προαναφέραμε, έχουν αναπτυχθεί και εμπορικά συστήματα. Ενδεικτικά κάποια απο αυτά είναι τα παρακάτω:

- Apache Storm (http://storm.apache.org/)
- Spark Streaming (http://spark.apache.org/streaming/)
- Apache Flink (https://flink.apache.org/)

Όλα τα παραπάνω συστήματα διαχειρίζονται κατανεμημένα τα ρεύματα δεδομένων για υψηλή αποδοτικότητα. Η πλειοψηφία εταιρειών στην σημερινή εποχή χρησιμοποιεί αυτά τα συστήματα.

2.4 Ερωτήματα σε ρεύματα δεδομένων

Τα ερωτήματα που υποβάλλονται σε ένα σύστημα διαχείρισης ρευμάτων δεδομένων (ΣΔΡΔ) μπορούν να διακριθούν σε δύο βασικές κατηγορίες:

- *Ερωτήματα στιγμιοτύπου* (snapshot or one-time queries), τα οποία εκτελούνται σε τακτά ή μη χρονικά διαστήματα και αφορούν το τρέχον στιγμιότυπο της πληροφορίας. Εξετάζουν δηλαδή την κατάσταση που έχει ένα σύστημα μόνο τη δεδομένη χρονική στιγμή.

- *Ερωτήματα διαρκείας* (continuous queries), τα οποία εκτελούνται συνεχώς από την στιγμή που τίθενται στο σύστημα, και η απάντηση ανανεώνεται με κάθε νέο στοιχείο των εισερχόμενων ρευμάτων.

  Από τις δύο παραπάνω κατηγορίες τα ερωτήματα διαρκείας είναι αυτά που παρουσιάζουν το μεγαλύτερο ενδιαφέρον όταν υποβάλλονται σε ρεύματα δεδομένων. Τα ερωτήματα στιγμιοτύπου δεν διαφέρουν ουσιαστικά από τα ερωτήματα που υποβάλλονται στις σχεσιακές βάσεις δεδομένων. Στα ερωτήματα διαρκείας η απάντηση παράγεται συνεχώς στο χρόνο και αφορά πάντα την πληροφορία που έχει καταφθάσει έως και την τρέχουσα χρονική στιγμή. Μία δεύτερη κατάταξη των ερωτημάτων γίνεται με βάση το χρόνο υποβολής του ερωτήματος και εξετάζεται αν το ερώτημα διατυπώθηκε και υποβλήθηκε πριν ή μετά από την έναρξη της άφιξης των δεδομένων. Οι κατηγορίες αυτές είναι:

- *Προκαθορισμένα ερωτήματα* (predefined queries), τα οποία είναι γνωστά στο σύστημα πριν ξεκινήσει η άφιξη της πληροφορίας προς επεξεργασία.

- *Μη προβλέψιμα ή περιστασιακά ερωτήματα* (ad hoc queries), υποβάλλονται από τον χρήστη ενώ έχουν εισρεύσει ήδη στοιχεία στο σύστημα.

  Η διαφοροποίηση των ερωτημάτων με βάση το πότε τέθηκαν στο σύστημα γίνεται γιατί τα προκαθορισμένα ερωτήματα είναι γνωστά εκ των προτέρων και το σύστημα μπορεί αφενός να επιλέξει κατάλληλες δομές για να ανταποκριθεί αποδοτικά και γρήγορα στα ερωτήματα αυτά και αφετέρου να ισοσταθμίσει την κατανάλωση πόρων (επεξεργαστικός χρόνος, μνήμη) με την ακρίβεια στις απαντήσεις ώστε να ανταποκρίνεται στις απαιτήσεις της εκάστοτε εφαρμογής. Από την άλλη πλευρά, τα ερωτήματα διαρκείας δεν είναι εκ των προτέρων γνωστά στο σύστημα και συνεπώς περιπλέκουν αρκετά την επεξεργασία ερωτημάτων, γιατί δεν μπορούν να συμπεριληφθούν στην κατάστρωση και βελτιστοποίηση του πλάνου εκτέλεσης ερωτημάτων (query optimization). Ένα ακόμα πρόβλημα που υπάρχει στα ερωτήματα διαρκείας είναι πως ενδέχεται να αναφέρονται σε δεδομένα που έχουν ήδη παρέλθει από το σύστημα και η ανάκτησή τους είναι εξαιρετικά δύσκολη. Αυτός ο σκόπελος προσπερνάται είτε με διαβεβαίωση πως το ερώτημα εκτελείται στα δεδομένα που θα καταφθάνουν μετά την υποβολή του, είτε με την χρήση δομών σύνοψης (summaries, synopses) που συγκροτούν επιλεκτικά πληροφορία για παρωχημένα δεδομένα.

2.5 Αλγόριθμοι σε ρεύματα δεδομένων

Όπως προαναφέραμε, το μοντέλο ρευμάτων δεδομένων απαιτεί σχεδιασμό νέων συστημάτων διαχείρισης διαφορετικών από τα παραδοσιακά συστήματα βάσεων δεδομένων. Επομένως δημιουργείται η ανάγκη εκ νέου σχεδιασμού αλγορίθμων για την επεξεργάσια τέτοιων δεδομένων. Οι αλγόριθμοι αυτοί θα πρέπει να έχουν τα εξής χαρακτηριστικά:

- Κατανάλωση της ελάχιστης δυνατής μνήμης, μιας και είναι αδύνατο να κρατηθεί όλη η πληροφορία του ρεύματος δεδομένων στην μνήμη. Έτσι στην μνήμη αποθηκεύεται ένα μικρό κομμάτι του ρεύματος δεδομένων.
- Λόγω του όγκου των δεδομένων κρίνεται επιτακτική ανάγκη να γίνεται μόνο ένα πέρασμα στα δεδομένα.
- Προσαρμοστικότητα του αλγορίθμου για διαφορετικό όγκο δεδομένων. Δηλαδή ο αλγόριθμος πρέπει να είναι σε θέση να εξάγει αποτελέσματα έγκαιρα με το ελάχιστο δυνατό σφάλμα.

Συμπερασματικά καταλήγουμε ότι οι αλγόριθμοι σε ρεύματα δεδομένων οφείλουν να είναι όσο τον δυνατό αποδοτικοί προκειμένου να ακολουθούν την γρήγορη ροή των δεδομένων και να εξάγουν έγκαιρα αποτελέσματα. Αυτό όμως έχει επίπτωση πολλές φορές στην ακρίβεια των αποτελεσμάτων. Οι αλγόριθμοι στην πλειοψηφία τους είναι προσεγγιστικοί και κύριος στόχος τους είναι η ελαχιστοποίηση του σφάλματος με τον πιό αποδοτικό τρόπο. Ετσι υπάρχει ένα trade-off μεταξύ κόστους διαχείρισης και ακρίβειας αποτελέσματος.

2.6 Παράθυρα σε ερωτήματα διαρκείας

Ένα σύστημα διαχείρισης βάσεων δεδομένων εξετάζει ένα συγκεκριμένο τμήμα του συνόλου των πλειάδων που καταφθάνουν ως είσοδο. Είναι σχεδόν αδύνατο να αποθηκεύει όλα τα δεδομένα του ρεύματος εισόδου. Το εξεταζόμενο τμήμα του ρεύματος εισόδου αποτελεί ένα *παράθυρο* (window) επί των πιο πρόσφατων στοιχείων του ρεύματος εισόδου. Συγκεκριμένα, το παράθυρο περιλαμβάνει ένα τμήμα διαδοχικών πλειάδων στις οποίες πραγματοποιείται η επεξεργασία ή το ερώτημα στις οποίες τίθεται. Παραδείγματος χάρην, σε ένα δίκτυο αισθητήρων οι τιμές των θερμοκρασιών που καταγράφονται αφορούν τις τελευταίες μέρες ή το πολύ εβδομάδες, με τα παλαιότερα δεδομένα να διαγράφονται. Έτσι, τα παράθυρα απομονώνουν ένα πεπερασμένο πλήθος δεδομένων απο ένα μεγάλο, πιθανώς άπειρου μήκους ρεύμα δεδομένων. Τα είδη των παραθύρων είναι τα ακόλουθα [14]:

- *Παράθυρα ορόσημου* (Landmark windows): Τα παράθυρα έχουν ως σταθερή αφετηρία κάποιο χρονόσημο, αλλά το πέρας τους παρακολουθεί τη χρονική εξέλιξη των πλειάδων του ρεύματος. Επομένως, το νεότερο άκρο του παραθύρου προχωρεί παράλληλα με το χρόνο, ταυτιζόμενο με την παρούσα χρονική στιγμή, ώστε να καλύπτει συνεχώς την έλευση νέων στοιχείων. Το εύρος του παραθύρου αυξάνεται, λοιπόν, διαρκώς, όπως και ο αριθμός των πλειάδων που περιλαμβάνει.

- *Κυλιόμενα παράθυρα βάσει χρόνου* (Time based sliding windows): Έχουν αφετηρία και πέρας που κινούνται ταυτόχρονα παρακολουθώντας την χρονική εξέλιξη των στοιχείων που συρρέουν στο σύστημα. Έτσι, παλαιότερα δεδομένα απορρίπτονται και καινούρια εισέρχονται με κυμαινόμενο ρυθμό. Το εύρος των παραθύρων παραμένει σταθερό, όμως ούτε το πλήθος των πλειάδων ούτε και τα περιεχόμενά τους διατηρούνται αμετάβλητα. Το συγκεκριμένο είδος παραθύρου χρησιμοποίειται στην παρούσα εργασία. Η ακριβής δομή που χρησιμοποιήθηκε θα περιγραφεί σε επόμενα κεφάλαια.

- *Κυλιόμενα παράθυρα βάσει πλειάδων (Tuple based sliding windows)*: Εχουν αφετηρία και πέρας που κινούνται ταυτόχρονα παρακολουθώντας τις πλειάδες που συρρέουν στο σύστημα.

2.7 Επεξεργασία κειμενικών ρευμάτων δεδομένων

Καθημερινά παράγονται δεδομένα, κυρίως απο το διαδίκτυο, σε μεγάλη κλίμακα και με πολύ γρήγορο ρυθμό. Αυτά τα δεδομένα μπορούν να θεωρηθούν μια ειδική περίπτωση ρευμάτων δεδομένων και η επεξεργασία τέτοιων δεδομένων αποτελεί στις μέρες μας μεγάλη πρόκληση. Με την ανάλυση κειμενικών ρευμάτων δεδομένων μπορούμε να εξάγγουμε σημαντική πληροφορία. Λόγου χάρη, η εξαγωγή περιλήψεων ειδήσεων από διαφορετικές πηγές, η ανάλυση συναισθήματος των χρηστών ενός κοινωνικού δικτύου κ.α. Η επεξεργασία της κειμενικής πληροφορίας είναι πολυδιάστατη σε αντίθεση με το κλασσικό μοντέλο ρεύματος δεδομένων που περιγράψαμε και περιλαμβάνει μονοδιάστατες τιμές. Για τον λόγο αυτό η ευρύτερη επιστημονική κοινότητα έχει στρέψει το ενδιάφερον της στην ανάπτυξη καινοτόμων αλγορίθμων και εφαρμογών. Συνοψίζοντας, τα κειμενικά ρεύματα δεδομένων αποτελούν μια ειδική περίπτωση του μοντέλου ρεύματος δεδομένων που περιγράψαμε σε προηγούμενη ενότητα, με την μόνη διαφορά ότι η πληροφορία που αναγράφεται στην πλειάδα των στοιχείων είναι κειμενική (αποτελείται απο λέξεις). Κυριότερες περιπτώσεις στις οποίες εμφανίζονται κειμενικά ρεύματα δεδομένων είναι:

- *Κοινωνικά δίκτυα (Social networks)*: Οι χρήστες των κοινωνικών δικτύων (Facebook,Twitter κτλ.) παράγουν τέτοια δεδομένα σε μορφή μηνυμάτων που ανταλλάσσουν προκειμένου να επικοινωνήσουν μεταξύ τους.

- *Υπηρεσίες ειδήσεων*: Λαμβάνουν κείμενα και άρθρα ειδήσεων, τα οποία φυσικά είναι κείμενα μεγαλύτερου μεγέθους και πιο δομημένα απο αυτά των κοινωνικών δικτύων.

- *Ανιχνευτές ιστού (Web crawling)*: Συλλέγουν κείμενα μεγάλου όγκου απο το διαδίκτυο και σε υπερβολικά γρήγορο ρυθμό.

Όπως περιγράψαμε προηγουμένως, η επεξεργασία κειμενικών ρευμάτων δεδομένων αποτελεί πολυδιάστατο πρόβλημα. Η επεξεργασία τέτοιων δεδομένων αποτελεί αντικείμενο μελέτης διαφόρων κλάδων όπως: ανάλυση συναισθήματος (sentiment analysis), μηχανική μάθηση (machine learning), εξόρυξη γνώσης (data mining) κ.α. Το κεφάλαιο αυτό περιλαμβάνει μία

επισκόπηση των αλγορίθμων που έχουν αναπτυχθεί τα τελευταία χρόνια για την επεξεργασία τέτοιου τύπου δεδομένων, επικεντρώνεται σε αλγόριθμους εξόρυξης γνώσης και πιο συγκεκριμένα σε αλγορίθμους συσταδοποίησης, οι οποίοι χρησιμοποιήθηκαν και για την δική μας εργασία και κυρίως αφορά επεξεργασία που πραγματοποιείται σε πραγματικό χρόνο.

Ενδεικτικά με την συσταδοποίηση μπορούμε να εντοπίσουμε γεγονότα που διαδραματίζονται και εκφράζονται μέσα απο κείμενα χρηστών στο διαδίκτυο. Όπως θα δούμε και παρακάτω στην εργασία μας, η κάθε συστάδα μπορεί να θεωρηθεί ένα γεγονός. Σημαντική εφαρμογή είναι και η ανάλυση της εξέλιξης των κειμενικών ρευμάτων δεδομένων, δηλαδή η εύρεση προτύπων που εξελίσσονται. Τέτοια πρότυπα μπορεί να είναι χρήσιμα για πολλές εφαρμογές όπως η δημιουργία περιλήψεων των ειδήσεων. Ιδιαίτερα με την ανάπτυξη των κοινωνικών δικτύων πολλές απο τις παραπάνω μεθόδους χρησιμοποιήθηκαν και επεκτάθηκαν αλλά και δημιουργήθηκαν πολλές νέες που θα περιγράψουμε σε επόμενες ενότητες. Τέτοιες μέθοδοι είναι ενδεικτικά οι ακόλουθες [5]:

- *Μέθοδος διασκορπισμού-συγκέντρωσης (Scatter-Gather Method)*: Η τεχνική αυτή συνδυάζει την τεχνική της ιεραρχικής συσταδοποίησης και την τεχνική της διαμέρισης συστάδων. Πιο συγκεκριμένα, αρχικά πραγματοποιείται ιεαραρχική συσταδοποίηση σε ένα δείγμα του συνόλου με σκοπό να δημιουργήσουμε τους σπόρους με τους οποίους μετά θα πραγματοποιθεί η διαμεριστική συσταδοποίηση (παρόμοια με την μέθοδο k-means). Επιπλέον έχουν επινοηθεί μια πληθώρα τεχνικών που καθιστά την μέθοδο αυτή αρκετά αποτελεσματική.

- *Λεκτική και φραστική μέθοδος (Word and Phrase-based method)*: Η τεχνική αυτή βασίζεται στο γεγονός ότι συστάδες απο λέξεις συμβάλλουν στην εύρεση συστάδων απο κείμενα. Έτσι μεταχειρίζεται το πρόβλημα ως δυαδικό. Γνωρίζοντας το σώμα των λέξεων μπορούμε να θεωρήσουμε έναν πίνακα $n \times d$ (n είναι ο αριθμός των λέξεων του σώματος και d ο αριθμός των κειμένων) όπου η εγγραφή (i, j) αναπαριστά τη συχνότητα εμφάνισης του όρου j στο κείμενο i. Ο πίνακας αυτός είναι αρκετά αραιός, αφού ένα κείμενο περιέχει ένα μικρό ποσοστό των λέξεων του σώματος. Στην συνέχεια πραγματοποιείται συσταδοποίηση στις στήλες ή στις γραμμές ή και στις δύο. Αυτή η τεχνική προυποθέτει να γνωρίζουμε το σύνολο του λεξιλογίου εξ' αρχής. Επομένως δεν είναι κατάλληλη για επεξεργασία δεδομένων σε πραγματικό χρόνο.

2.7.1 Συσταδοποίηση κειμενικών ρευμάτων δεδομένων

Το πρόβλημα της συσταδοποίησης (Clustering) κειμενικών ρευμάτων δεδομένων έχει μελετηθεί κατά καιρούς στα πλαίσια αλγορίθμων που επεξεργάζονται αριθμητικά δεδομένα. Στην ορολογία της εξόρυξης γνώσης, με τον όρο συστάδα εννοούμε ένα σύνολο από όμοια αντικείμενα. Μιλώντας για συσταδοποίηση εννοούμε ομαδοποίηση όμοιων αντικειμένων. Άλλες δημοφιλείς μέθοδοι έχουν αναπτυχθέι στα πλαίσια της μηχανικής μάθησης.

Οι περισσότερες μέθοδοι που έχουν αναπτυχθεί βασίζονται και αποτελούν προέκταση του αλγορίθμου k-means. Μια πρόσφατη τεχνική που αναπτύχθηκε είναι ο αλγόριθμος Online Spherical k-Means Algorithm [4], ο οποίος διασπά σε μικρά κομμάτια το ρεύμα δεδομένων και

το καθένα απο αυτά μπορεί να επεξεργαστεί με τον αλγόριθμο του k-means στην κύρια μνήμη. Το πλεονέκτημα αυτής της μεθόδοι έναντι των άλλων είναι οτι απο τη στιγμή που τα μικρά αυτά κομμάτια τοποθετούνται στην κύρια μνήμη τότε μπορούμε να τα προσπελάσουμε παραπάνω απο μία φορά προσφέροντας μας καλύτερα και ακριβέστερα αποτελέσματα. Επιπρόσθετα τα κεντροειδή των συστάδων απο προηγούμενους κύκλους εκτέλεσης επαναχρησιμοποιούνται στις επόμενες εκτελέσεις. Επίσης εισάγεται και ο όρος decay factor που χαρακτηρίζει τα κείμενα (συστάδες) παλιές δίνοντας έτσι μεγαλύτερη έμφαση στις πρόσφατα σχηματιζόμενες συστάδες.

Παρά την αποδοτικότητα του, αυτός ο αλγόριθμος, εξάγει αποτελέσματα ενος επιπέδου. Σε πολλές εφαρμογές καλούμαστε να εντοπίσουμε επιμέρους θέματα σε ένα θέμα. Μεταφράζοντας την προηγούμενη πρόταση στη γλώσσα της εξόρυξης γνώσης, καλούμαστε να σχεδιάσουμε *ιεραρχική συσταδοποίηση* προκειμένου να έχουμε πρόσβαση στα διάφορα επίπεδα της συσταδοποίησης.

2.7.2 ConSTREAM

Ο αλγόριθμος ConSTREAM (CONdensation based STREAM Clustering) [5] που θα περιγράψουμε σε αυτή την ενότητα επεξεργάζεται κατηγορικά δεδομένα και κειμενικά ρεύματα δεδομένων σε πραγματικό χρόνο και προσφέρει καλύτερα αποτελέσματα απο τις προηγούμενες μεθόδους που περιγράψαμε. Πιο συγκεκριμένα δέχεται σαν είσοδο κείμενα (σύνολα απο λέξεις) και τα ομαδοποιεί. Μια παραλλαγή της παρούσας μεθόδου χρησιμοποιηθήκε και στα πλαίσια της διπλωματικής εργασίας, επειδή λαμβάνει υπόψη ότι τα δεδομένα εξελίσσονται με την πάροδο του χρόνου. Η μέθοδος αυτή ουσιαστικά δημιουργεί συνόψεις των κειμενικών δεδομένων, τις οποίες θα περιγράψουμε παρακάτω. Για να προσδώσουμε μεγαλύτερη σημασία στα πιο πρόσφατα δεδομένα, κάθε συστάδα χαρακτηρίζεται απο ένα *βάρος* που είναι συνάρτηση του χρόνου δημιουργίας της $f(t)$.

Στην συνέχεια εισάγουμε κάποιες έννοιες που χρησιμοποιήθηκαν για την ανάπτυξη του αλγορίθμου. Η έννοια της *ημίσειας* (το χρονικό διάστημα που απαιτείται ώστε η τιμή μιας ποσότητας να μειωθεί στο μισό της αρχικής της τιμής), που δεν είναι τίποτα άλλο απο την τιμή $f(t_0) = \frac{1}{2}f(0)$, όπου t_0 είναι η χρονική στιγμή που συμπίπτει με την χρονική στιγμή της *ημίσειας ζωής*. Έπειτα ορίζουμε την έννοια της τιμής απόσβεσης $\lambda = \frac{1}{t_0}$. Η συνάρτηση του βάρους $f(t)$ σχηματίζεται ως εξής: $f(t) = 2^{-\lambda t}$.

Όταν μιά συστάδα σχηματίζεται απο ένα νεοεισερχόμενο στοιχείο τότε αυτομάτως αποκαλείται ακραία. Με τον όρο ακραία εννοούμε ότι ίσως τα δεδομένα της συστάδας δεν αποτελούν κάποιο ουσιαστικό αποτέλεσμα που αξίζει να κρατήσουμε, θα μπορούσε να τα χαρακτηρίσει κανείς 'σκουπίδια'. Κάτα την διάρκεια της περιόδου της ημίσειας ζωής, στην περίπτωση που καταφθάσει ένα τουλάχιστον στοιχείο και ανήκει στην συστάδα που αναφέραμε προηγουμένως τότε η συστάδα χαρακτηρίζεται ως ενεργή. Αντίθετα εφόσον μετά το πέρας της ημίσειας ζωής δεν μεταβληθεί η συστάδα τότε αναγνωρίζεται σαν θόρυβος και διαγράφεται. Στην συνέχεια θα περιγράψουμε την δομή που έχουν οι συστάδες και τον τρόπο που σχηματίζονται. Μια συστάδα $D(t, C)$ του συνόλου των κειμενικών δεδομένων C την χρονική στιγμή t ορίζεται

μια πλειάδα $< DF2, DF1, n, w(t), l >$.Κάθε μέλος της πλειάδας ορίζεται ως εξής:

- $DF2$: Διάνυσμα που περιέχει τα $3wb(wb-1)/2$ στοιχεία του συνόλου C, όπου wb είναι το πλήθος των διακριτών λέξεων του συνόλου C. Κάθε στοιχείο χαρακτηρίζεται από έναν μετρητή που υποδεικνύει το πλήθος εμφανίσεων της εκάστοτε λέξης.
- $DF1$: Διάνυσμα που περιέχει τα $2wb$ στοιχεία που εισήχθησαν.
- n: Ο αριθμός των στοιχείων που ανήκουν στην συστάδα.
- $w(t)$: Περιέχει το άθροισμα των μετρητών (βαρών) της εκάστοτε συστάδας.
- l: Περιέχει τον χρόνο του τελευταίου στοιχείου που προστέθηκε στην συστάδα.

Η παραπάνω δομή της συστάδας δεν υιοθετήθηκε τυχαία. Μας εξυπηρετεί διότι η συγχώνευση και η διάσπαση συστάδων γίνεται εύκολα με μία πρόσθεση και μια αφαίρεση των μελών των πλειάδων.

Στην αρχή του αλγορίθμου διαθέτουμε ένα κενό σύνολο συστάδων. Εφόσον σχηματιστούν k συστάδες τότε περνάμε στο επόμενο στάδιο που έιναι η διατήρηση και η ανανέωση των συστάδων σε πραγματικό χρόνο. Τη στιγμή που ένα καινούργιο στοιχείο καταφθάσει τότε το συγκρίνουμε, με την βοήθεια μιας συνάρτησης ομοιότητας, με τις k συστάδες και βρίσκουμε την συστάδα με την μεγαλύτερη ομοιότητα. Σαν συνάρτηση ομοιότητας μπορούμε να χρησιμοποιήσουμε μια απο τις γνωστές συναρτήσεις που είναι κατάλληλες για κατηγορικά δεδομένα (συνάρτηση συνημιτονου, συνάρτηση Jaccard, κ.α. Στην συνέχεια ελέγχουμε εάν η ομοιότητα είναι επαρκής για να χαρακτηρίσουμε αυτο το δεδομένο ως μέλος της συστάδας. Έφοσον ανήκει τότε εντάσουμε το στοιχείο στην συσταδα και ανανεώνουμε τα στατιστικά που κρατάμε στην πλειάδα. Στην περίπτωση που δεν ανήκει σε καμία συστάδα τότε το στοιχείο σχηματίζει δική του συστάδα, χαρακτηρίζεται ως ανενεργή και διαγράφουμε την συστάδα με το μεγαλύτερο βάρος, δηλαδη την συστάδα που έχει να ανανεωθεί πολύ καιρό, για να δώσουμε χώρο να δημιουργηθεί νέα συστάδα.

Η τεχνική συσταδοποίησης που περιγράψαμε προηγουμένως υπάγεται στην κατηγορία των συσσωρευτικών τεχνικών συσταδοποίησης και ανα πάσα χρονική στιγμή παρέχει τις k-επικρατέστερες και πιο πρόσφατες συστάδες. Αποτελεί ουσιαστικά μια μέθοδος ομαδοποίησης όμοιων κειμένων.

Κεφάλαιο 3

Επεξεργασία χωρικών-κειμενικών δεδομένων

3.1 Εισαγωγή

Η αύξηση των αντικειμένων που περιλαμβάνουν γεωγραφική τοποθεσία προσδιορισμού και κειμενική πληροφορία προσδίδει στο διαδίκτυο χωρική διάσταση. Συγκεκριμένα, χρήστες του διαδικτυου παράγουν πληροφορία που περιλαμβάνει γεωγραφικό προσδιορισμό. Την ίδια στιγμή παράγεται στο διαδίκτυο πληροφορία που δηλώνει σημεία ενδιαφέροντος (point-of-interest) όπως καφετέριες, εστιατόρια, τουριστικά θέρετρα και αξιοθέατα. Τέτοιου είδους πληροφορία καθιστά την ανάγκη για ανάπτυξη τεχνικών και ευρετηρίων που θα διευκολύνουν στην επεξεργασία τους. Στο παρόν κελάφαιο θα περιγράψουμε διάφορα είδη ευρετηρίων που υποστηρίζουν χωρο-κειμενικά δεδομένα (spatio-temporal data) και εξυπηρετούν στον αποδοτικό σχεδιασμό χωρο-κειμενικών ερωτημάτων. Υπάρχουν τρείς τύποι τέτοιων ερωτημάτων [7] που αξίζουν την προσοχή μας:

- *Λογικό k-NN ερώτημα (Boolean k-NN query)* : Ανακτώνται τα k αντικειμένα που βρίσκονται πιο κοντά στην τοποθεσία του χρήστη που θέτει το ερώτημα και περιλαμβάνουν τις λέξεις που επιθυμεί ο χρήστης να περιλαμβάνουν. Παραδείγματος χάρη, ο χρήστης θέλει να εντοπίσει τα πλησιέστερα εστιατόρια που σερβίρουν ιταλικό φαγητό. Επομένως θα θέσει το ερώτημα προκείμενου να ανακτήσει τα k πλησιέστερα, στην τοποθεσία του χρήστη, αντικείμενα (αναπαρίστανται σημειακά) οι περιγραφές των οποίων περιέχουν τις λέξεις κλειδιά *italian*, ρεσταυραντ.

- *Top-kNN ερώτημα* : Εντοπίζει τα k αντικείμενα με το μεγαλύτερο συσχετισμό απόστασης, απο ένα σημείο που τίθεται στο ερώτημα, και κειμενικής συνάφειας μεταξύ των λέξεων κλειδιών του ερωτήματος και της κειμενικής περιγραφής του εκάστοτε αντικειμένου. Ο συσχετισμός εκφράζεται συνήθως ως μια συνάρτηση με μεταβλητές την απόσταση του αντικειμένου απο το σημείο και τη συνάφεια της κειμενικής περιγραφής του αντικειμένου με τις λέξεις κλειδιά που έχουν υποβληθεί στο ερώτημα.

- *Λογικό ερώτημα διαστήματος (Boolean Range Query)* : Ανακτά όλα τα αντικείμενα

που περιλαμβάνουν τις λέξεις κλειδιά που έχουν τεθεί στο ερώτημα και βρίσκονται μέσα στην ακτίνα αποστάσεως που έχει δοθεί απο τπν χρήστη.

Τέτοια χωρο-κειμενικά ερωτήματα (spatial-keyword queries) υποστηρίζονται απο πολλές εφαρμογές στην σημερινή εποχή, όπως είναι τα Google maps όπου σημεία ενδιαφέροντος (POI) μπορούν να ανακτηθούν. Εφαρμογή τέτοιων ερωτημάτων εντοπίζουμε και σε πολλές εφαρμογές κοινωνικής δικτύωσης, όπως Foursquare και Twitter, όπου τα δεδομένα περιλαμβάνουν πολλές φορές χωρική διάσταση (λ.χ συντεταγμένες των χρηστών την στιγμή μιάς δημοσίευσης).

3.2 Κατηγοριοποίηση Ευρετηρίων

Στην ενότητα αυτή θα παρουσιάσουμε τα ευρετήρια που έχουν αναπτυχθεί και χρησιμοποιούνται για την επεξεργασία χωρο-κειμενικών δεδομένων. Η κατηγοριοποίηση των ευρετηρίων έχει πραγματοποιηθεί σύμφωνα με την σειρά που λαμβάνουν υπόψη την κειμενική και την χωρική πληροφορία. με τρία χαρακτηριστικά: τον τρόπο που γίνεται η ευρετηρίαση της χωρικής πληροφορίας, η ευρετηρίαση της κειμενικής πληροφορίας και η υβριδική ευρετηρίαση.

3.2.1 Χωρική μέθοδος δεικτοδότησης

Στην παρούσα κατηγοριοποίηση, διαχωρίζούμε τα ευρετήρια με βάση την δομή ευρετηρίου που χρησιμοποιούμε για να κατατάξουμε τα δεδομένα με βάση την χωρική πληροφορία. Διακρίνουμε τρία είδη ευρετηρίων:

- R-tree based: Στην κατηγορία αυτή χρησιμοποιείται η δομή R-tree. Τα περισσότερα χωρο-κειμενικά ευρετήρια ανήκουν σε αυτή την κατηγορία και για την δεικτοδότηση του κειμένου χρησιμοποιούν την δομή inverted-index. Έτσι τα χωρο-κειμενικά δεδομένα συνδυάζουν την δομή R-tree και την δομή inverted-index.

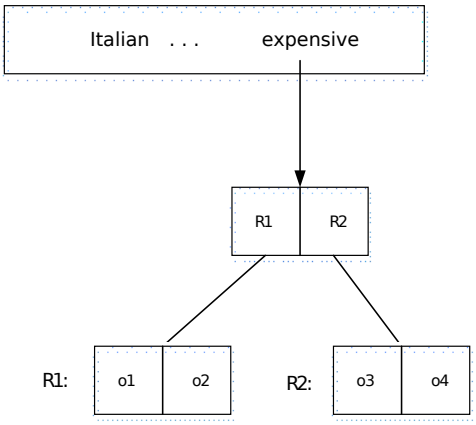
- Grid based: Τέτοιου είδους ευρετήρια συνδυάζουν την δομή του ομοιόμορφου καννάβου με κειμενικές δομές δεικτοδότησης (inverted index κ.α.). Ο κάνναβος διαχωρίζει τον χώρο σε ομοιόμορφα τμήματα.

- Space filling curve based: Στην περίπτωση αυτή αντί για την χρησιμοποίηση του καννάβου χρησιμοποιούμε καμπύλη χώρου. Τέτοιες καμπύλες είναι η καμπύλη του Hilbert, Z-curved κ.α.

3.2.2 Κειμενική μέθοδος δεικτοδότησης

Στην περίπτωση αυτή τα ευρετήρια διακρίνονται με βάση τη δομή ευρετηρίου που χρησιμοποιούμε για την δεικτοδότηση της κειμενικής πληροφορίας.

- Inverted file: Κάθε inverted file περιλαμβάνει ένα λεξικό και κάθε όρος του λεξικού σχετίζεται με ένα αντικείμενο με τη χρήση ενός inverted index.



Σχήμα 3.1: Ευρετήριο τύπου IF-R* [7]

- Bitmaps: Κάποια ευρετήρια που χρησιμοποιούν δομές R-tree, χρησιμοποιούν τα bitmaps προκειμένου να δεικτοδοτήσουν την κειμενική πληροφορία σε υποδέντρα. Κάθε bit αναπαριστά την παρουσία (1) ή την απουσία του όρου (0).

3.3 Υβριδικά ευρετήρια

Έπειτα απο την κατηγοριοποίηση των ευρετηρίων, στην συνέχεια θα αναφέρουμε ορισμένα απο τα πιο σημαντικά χωρο-κειμενικά ευρετήρια που έχουν προταθεί στην βιβλιογραφία.

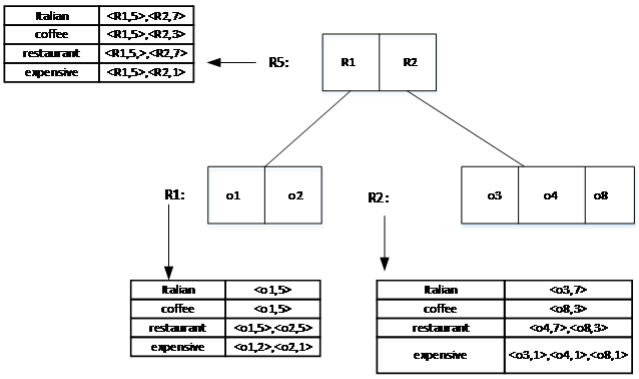
3.3.1 Ευρετήρια βασισμένα σε R-tree

Η δομή R-tree είναι η πιο κύριαρχη δομή για δεικτοδότηση χωρικών δεδομένων και η δομή inverted file αντίστοιχα για κειμενικά δεδομένα. Έτσι ο συνδυασμός των δύο δομών θα μπορούσε να φέρει ικανοποιητικά αποτελέσματα για την επεξεργασία χωρο-κειμενικών δεδομένων. Τέτοια υβριδικά ευρετήρια είναι:

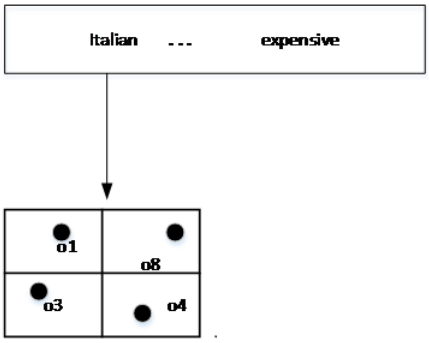
- IF-R* και R*-IF: Το IF-R* όπως φαίνεται και στο σχήμα 3.1 είναι ένα ευρετήριο που χρησιμοποιεί το R-tree για την δεικτοδότηση των αντικειμένων χωρίς να λαμβάνει την κειμενική διάσταση και έπειτα σε κάθε φύλλο του δένδρου κατασκευάζεται ένα inverted file για το κειμενικό μέρος των δεδομένων. Αντίστοιχα, η δομή IF-R κατασκευάζει αρχικά το inverted file αγνοώντας την χωρική πληροφορία και στην συνέχεια δεικτοδοτεί την δενδρική δομή για την εκάστοτε λέξη που ανήκει στο inverted file.

- IR^2-tree: Αυτή η δομή ευρετηριού (βλ. σχήμα 3.2) ενσωματώνει την δομή bitmap που περιγράψαμε προηγουμένως στους κόμβους του R-tree. Κάθε bitmap που περιλαμβάνει ένας κόμβος της δενδρικής δομής είναι αποτέλεσμα της ένωσης των bitmaps που περιλαμβάνουν οι κόμβοι παιδιά. Η δομή αυτή χρησιμοποείται για την απάντηση λογικών k-NN ερωτημάτων και Top-kNN ερωτημάτων.

- Υβριδική χωρο-κειμενική δεικτοδότηση (Hybrid spatial keyword indexing): Όπως και προηγουμένως χρησιμοποιούμε συνδιασμό των δομών bitmap και R-tree και πρόκειται



Σχήμα 3.2: Ευρετήριο τύπου IR^2-tree [7]



Σχήμα 3.3: Ευρετήριο τύπου Text Primary [7]

ουσιαστικά για μια επέκταση της δενδρικής δομής R. Κάθε κόμβος του δένδρου που είναι πατέρας ένος φύλλου τότε καλείται υπερκόμβος (supernode). Κάθε υπερκόμβος (supernode) σχετίζεται με μια μορφή bitmap ενός inverted file. Συγκεκριμένα κάθε αντικείμενο που βρίσκεται στα φύλλα του δένδρου σχετίζεται με ένα bitmap. Στην περίπτωση που το αντικείμενο περιέχει την λέξη κλειδί τότε τίθεται η λογική τιμή 1, διαφορετικά η λογική τιμή 0.

3.3.2 Grid based ευρετήρια

Τα ευρετήρια που χρησιμοποιούν πλέγμα για την δεικτοδότηση αντικειμένων με βάση την χωρική διάσταση. Ενδεικτικά κάποια απο αυτά είναι:

- Spatial Primary Index και Text Primary Index: Χρησιμοποιούν πλέγμα για την δεικτοδότηση των αντικειμένων με βάση την χωρική διάσταση και inverted index για την δεικτοδότηση με βάση την κειμενική διάσταση. Ένα TPS ευρετήριο φαίνεται στο σχημα 3.3.
- Spatial Keyword inverted file (SKIF): Αποθηκεύει τα δεδομένα σε μια δομή παρόμοια με την inverted file προκειμένου να διαχειρίζεται το κειμενικό και το χωρικό μέρος των δεδομένων ταυτόχρονα.

Κεφάλαιο 4

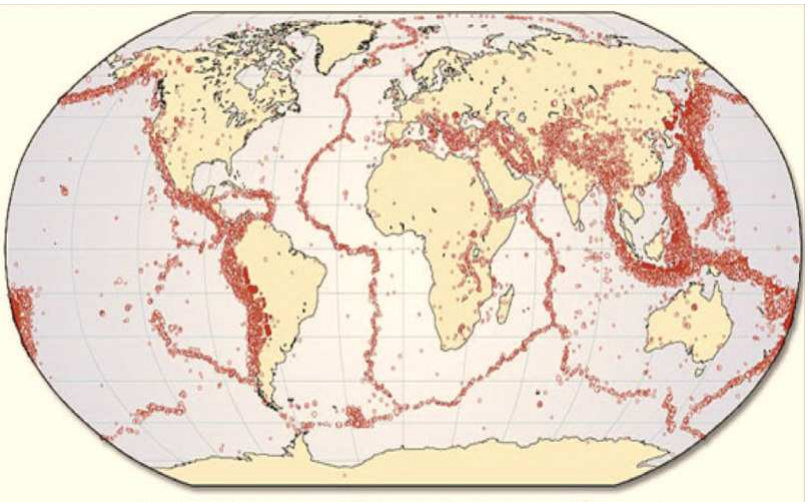
Επεξεργασία δεδομένων στα κοινωνικά δίκτυα

4.1 Εισαγωγή

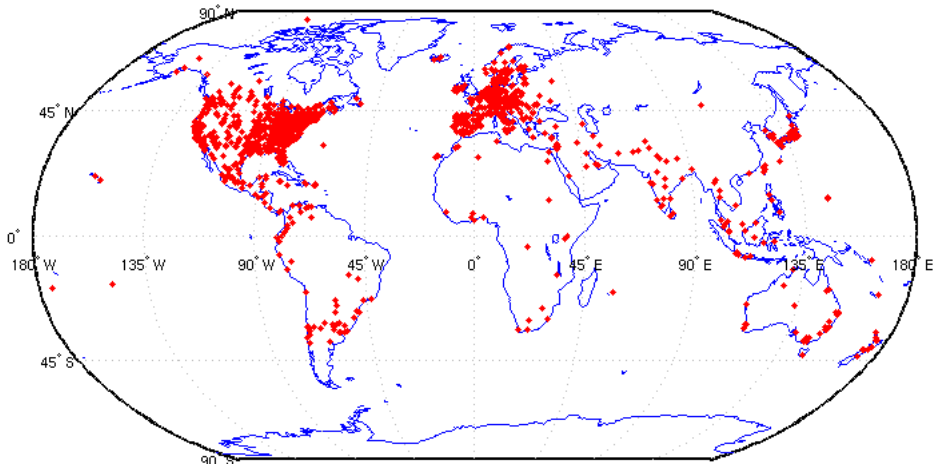
Στην σημερινή εποχή, με την ραγδαία ανάπτυξη των τεχνολογιών διαδικτύου, ολοένα και περισσότεροι χρήστες του διαδικτύου χρησιμοποιούν τα κοινωνικά δίκτυα ως μέσο επικοινωνίας. Η ραγδαία αυτή αύξηση της χρήσης των κοινωνικών δικτύων έχει ως αποτέλεσμα να δημιούργειται μεγάλος όγκος πληροφορίας καθημερινά. Η επεξεργασία αυτής της πληροφορίας και η εφαρμογή ερωτημάτων σε τέτοια δεδομένα μπορεί να μας δώσει κάποια πολύ σημαντικά αποτελέσματα. Χαρακτηριστικά ερωτήματα και αναλύσεις που πραγματοποιούνται σε τέτοια δεδομένων και έχουν προταθεί κατά καιρούς στην βιβλιογραφία είναι οι ακόλουθες:

- *Ανίχνευση τοπικών γεγονότων*: Ένα απο αυτά τα αποτελέσματα είναι η ανίχνευση πραγματικών γεγονότων που εκφράζονται μέσα απο την πληροφορία που δημιουργούν οι ίδιοι οι χρήστες. Όταν ένα γεγονός εξελίσσεται τότε οι χρήστες που συμμετέχουν ή το παρατηρούν, έχουν την τάση να το δημοσιοποιούν και να παράγουν μηνύματα σχετικά με αυτό στα κοινωνικά δίκτυα μέσω του κινητού τους. Η παραπάνω παρατήρηση είναι πολύ όμοια με την λειτουργία ενός αισθητήρα. Έπομένως συμπεραίνουμε πως πολλές φορές εκμεταλλευόμενοι την πληροφορία των κοινωνικών δικτύων μπορούμε να εντοπίσουμε και να μάθουμε για γεγονότα πολύ πιο γρήγορα από τα μέσα μαζικής ενημέρωσης. Ένα χαρακτηριστικό παράδειγμα απεικονίζεται στην εικόνα 4.1 που δείχνει τον χάρτη με τις σεισμογενείς περιοχές σε αντιστοιχία με τον χάρτη της εικόνας 4.2 των δημοσιεύσεων που αναφερουν την λέξη κλειδί earthquake [23]. Απο την σύγκριση είναι προφανές πως οι δύο χάρτες δεν έχουν μεγάλες διαφορές.

- *k-NN λογικά ερωτήματα χωροχρονικών όρων*: Τα ερωτήματα αυτά δεν είναι τίποτε άλλο από λογικά k-NN ερωτήματα που είχαμε συναντήσει σε προηγούμενο κεφάλαιο. Στην προκειμένη περίπτωση αναφερομάστε σε δημοσιεύσεις χρηστών απο blogs. Έτσι με τα ερωτήματα αυτά εντοπίζουμε τις k κορυφαίες δημοσιεύσεις που είναι πιο κοντά σε μια περιοχή ενδιαφέροντος που τίθεται στο ερώτημα.



Σχήμα 4.1: Χάρτης σεισμογενών περιοχών ανά τον κόσμο [23]



Σχήμα 4.2: Χάρτης απο tweets αναφερόμενα στην λέξη earthquake [23]

- *Top-k ερωτήματα χωροχρονικών λεκτικών όρων* [26]: Τα ερωτήματα αυτά, απευθύνονται στο χώρο του διαδικτύου και των κοινωνικών δικτύων, εντοπίζουν τους k κορυφαίους λεκτικούς όρους που εμφανίζονται πιο συχνά σε μια περιοχή ενδιάφεροντος. Αποτελούν ερωτήματα διαρκείας, επομένως είναι αναγκαίο η επεξεργασία να γίνεται σε πραγματικό χρόνο. Με αυτό τον τρόπο μπορούμε να εντοπίσουμε κάποιο τοπικό επίκαιρο θέμα συζήτησης που εκτυλλίσεται την συγκεκριμένη χρονική στιγμή που τίθεται το ερώτημα. Στην συνέχεια του κεφαλαίου γίνεται αναλυτικότερη περιγραφή της μεθοδολογίας που αναπτύχθηκε για την προσέγγιση του προβλήματος αυτού.

4.2 Top-k ερωτήματα χωροχρονικών λεξικών όρων

Όπως αναφέραμε και στην εισαγωγή του κεφαλαίου η απάντηση τέτοιου είδους ερωτημάτων μας δίνουν πληροφορίες για τα θέματα που συζητιούνται πιο πολύ στο διαδίκτυο, ποια ειναι τα δημοφιλή κ.α. Τα τελευταία χρόνια έχουν αναπτυχθεί συστήματα που επεξεργάζονται

δεδομένα των κοινωνικών δικτύων με παρόμοιο τρόπο που περιγράψαμε προηγουμένως. Το πιο σύγχρονο και αποδότικότερο σύστημα που αναφέρεται στην βιβλιογραφία ειναι το σύστημα υλοποίησης AFIA. Στην συνέχεια θα αναλύσουμε βήμα βήμα την μεθοδολογία που ακολουθήθηκε για την αποδοτική υλοποίηση τέτοιων ερωτημάτων και συγκεκριμένα την υλοποίηση του AFIA. [26]

4.2.1 Μέτρηση συχνότητας αντικειμένων

Το παρόν πρόβλημα ουσιαστικά αποτελεί την εύρεση των k πιο συχνών αντικειμένων σε μια ορισμένη περιοχή. Όπως είναι φανερό, το πρόβλημα έγκειται στο γεγονός πως τα δεδομένα που επεξεργαζόμαστε καταφθάνουν σε πραγματικό χρόνο με αποτέλεσμα να μην γνωρίζουμε εξάρχής το πλήθος τους. Σε συνδυασμό με το γεγονός πως διαθέτουμε πεπερασμένη μνήμη στον υπολογιστή μας είναι εμφανές πως ακριβής μέτρηση της συχνότητας των λέξεων είναι αδύνατη. Αυτό έχει ως αποτέλεσμα να καταφεύγουμε σε προσεγγιστικούς αλγόριθμους για την εύρεση των k πιο συχνών αντικειμένων. Οι αλγόριθμοι για μέτρηση συχνότητας αντικειμένων μπορούν να διαχωριστούν σε δύο κατηγορίες:

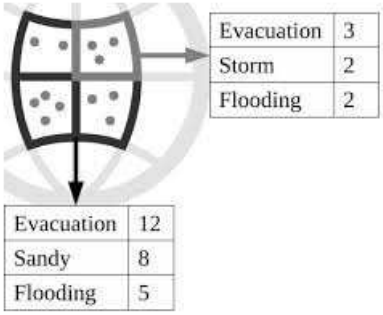
- Απαρίθμησης (Counting): Ο αριθμός των μετρητών είναι προκαθορισμένος και σταθερός. Στην περίπτωση που ένα αντικείμενο είναι συχνό και δεν υπάρχει στους μετρητές τότε αφαιρείται ένας μετρητής και προστίθεται ο μετρητής του καινούργιου αντικειμένου. Τέτοιοι αλγόριθμοι είναι οι LoosyCounting, Frequent, SpaceSaving.

- Σκιαγράφισης: Σε αυτή τη μέθοδο τηρούνται όλες οι προσεγγιστικές συχνότητες των αντικειμένων και κάθε κατακερματισμένος μετρητής ανανεώνεται εφόσον εντοπιστεί το αντίστοιχο αντικείμενο.

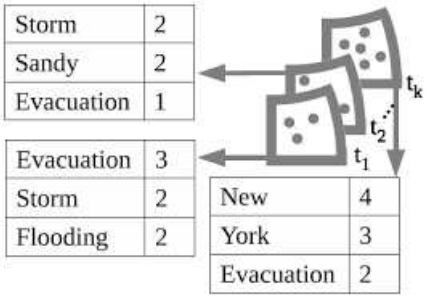
4.2.2 Συνάθροιση στον χώρο και στον χρόνο

Το δεύτερο στάδιο που απαιτείται για την υλοποίηση της παραπάνω μεθοδολογίας είναι να εντοπίζουμε τις πιο συχνές λέξεις σε μια ορισμένη περιοχή. Μια γνωστή πρακτική που εφαρμόζεται είναι να διαμερίσουμε τον χώρο και τον χρόνο σε ομοιόμορφα τμήματα με πολλαπλά επίπεδα. Για την διάσταση του χώρου, τον χωρίζουμε σε πολλαπλά επίπεδα απο ομοιόμορφους καννάβους με κλικακούμενα στο μέγεθος κελιά. Με αυτό τον τρόπο μπορούμε να θέτουμε ερωτήματα σε περιοχές με διαφορετικό μέγεθος συναθροίζοντας τα αποτελέσματα που βρήκαμε σε μικρότερους υποχώρους. Για την διάσταση του χρόνου, αφού έχουμε ορίσει την μικρότερη διάσταση στον χρόνο (μικρότερο χρονικό διάστημα), κατασκευάζουμε ομοιόμορφα πλέγματα για κάθε τέτοια καινούργια χρονικά διάστηματα που προκύπτουν με την πάροδο του χρόνου. Επομένως, για να εξάγουμε ένα αποτέλεσμα σε ένα ερώτημα αρκεί να συναθροίσουμε τα ελάχιστα χρονικά διαστήματα που εμπίπτουν στο χρονικό διάστημα του ερωτήματος.

Το σύστημα υλοποίησης AFIA [26] επεκτείνει τον γνωστό αλγόριθμο Space Saving και παρέχει τα πιο ακριβή αποτελέσματα απο όλα τα προαναφερθέντα συστήματα.Ο χρήστης υποβάλει ένα ερώτημα top-k ερώτημα και το σύστημα επιστρέφει του k κορυφαίους λεξικούς όρους



Σχήμα 4.3: Πολλαπλά χωρικά επίπεδα



Σχήμα 4.4: Πολλαπλά χρονικά επίπεδα [26]

σε συχνότητα εμφάνισης. Η καινοτομία που διαχωρίζει τον AFIA απο τα άλλα συστήματα είναι πως οι πρώτοι k_g όροι είναι ακριβείς ενώ οι υπόλοιποι $k - k_g$ είναι προσεγγιστικοί.

4.3 Εντοπισμός τοπικών γεγονότων στο Twitter

Η τεχνική αυτή αποσκοπεί στον εντοπισμό γεγονότων που εξελίσσονται στο κοινωνικό δίκτυο Twitter και έχουν τοπικό χαρακτήρα. Δηλαδή εντοπίζονται σε μία περιορισμένη γεωγραφική περιοχή.Η τεχνική αυτή εντοπίζει τοπικά γεγονότα σε πραγματικό χρόνο εξετάζοντας τις πιο πρόσφατες δημοσιεύσεις χρησιμοποιώντας ένα χρονικά κυλιόμενο παράθυρο. Τα τοπικά γεγονότα αποτελούνται από ένα σύνολο απο λέξεις-κλειδια, την χρονική στιγμή που ξεκίνησε το γεγονός και τέλος την γεωγραφική τους τοποθεσία. Στην τεχνική αυτή χρησιμοποείται κάνναβος για την αποτελεσματικότερη επεξεργασία των δεδομένων [3, 30].

4.3.1 Εύρεση λέξεων-κλειδιά

Σε κάθε κύκλο εκτέλεσης εξάγονται λέξεις-κλειδιά που είναι κατάλληλα να περιγράψουν ένα γεγονός. Λαμβάνοντας και την ιστορική πληροφορία εξάγονται λέξεις σύμφωνα με έναν δείκτη εκρηκτικότητας που υποδηλώνει ότι η συχνότητα εμφάνισης της λέξης είναι ασυνήθιστα υψηλή. Πιο συγκεκριμένα αξιοποιώντας τα ιστορικά δεδομένα, συγκεκριμένου μεγέθους, που έχουν συλλεγεί απο προηγούμενους κύκλους εκτέλεσης και θεωρώντας ότι ακολουθούν ένα είδος κατανομής εξάγεται η προβλεπόμενη συχνότητα που θα είχε η λέξη στο τρέχοντα κύκλο εκτέλεσης. Στην συνέχεια για την παρούσα εκτέλεση υπολογίζεται η πραγματική συχνότητα. Στην περίπτωση που αποκλίνει πάνω από ένα κατώφλι τότε η λέξη αυτή θεωρείται λέξη-κλειδί.

Στην συνέχεια φιλτράρονται οι υποψήφιες λέξεις-κλείδια για να εντοπίστουν οι λέξεις-κλειδιά που θα μπορούσαν να περιγράψουν ένα τοπικό γεγονός. Για τον σκοπό αυτό υπολογίζεται μια χωρική σφραγίδα (timestamp) για κάθε λέξη κλειδί. Η χωρική σφραγίδα αποτελεί την χωρική κατανομή πυκνότητας της λέξης κλειδιού σε ένα συγκεκριμένο χώρο, που στην παρούσα περίπτωση είναι το κελί του κάνναβου. Στόχος είναι να μείνουν οι λέξεις κλειδιά που βρέθηκαν και να διατηρηθούν οι λέξεις που έχουν μια χωρική τοπικότητα. Έτσι υπολογίζοντας την εντροπία της χωρικής σφραγίδας, που αναφέραμε προηγουμένως, ουσιαστικά υπολογίζεται η διασπορά της χωρικής πυκνότητας της λέξης στο χώρο. Επομένως αν υπάρχει μεγάλη εντροπία σημαίνει ότι υπάρχει και μεγάλη διασπορά και αυτές οι λέξεις αποκόπτονται.

4.3.2 Εύρεση τοπικών γεγονότων

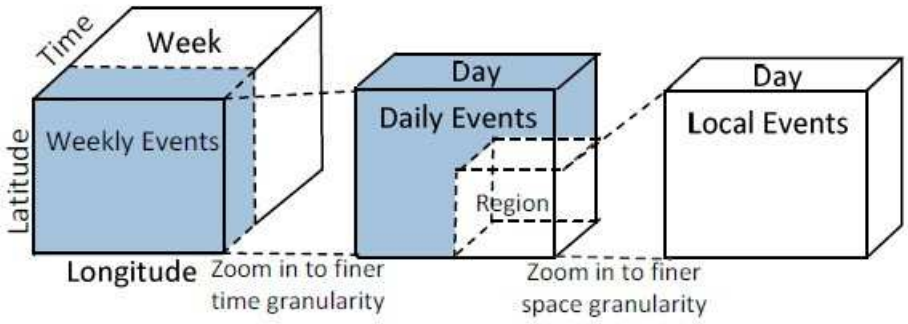
Για να εντοπιστούν τοπικά γεγονότα γίνεται η εξής παραδοχή: Λέξεις-κλειδιά που ανήκουν στο ίδιο τοπικό γεγονός τείνουν να έχουν παρόμοια χωρική σφραγίδα. Αυτό διαισθητικά είναι αποδεκτό, αφού ουσιαστικά σημαίνει να έχουν χωρική εγγύτητα μεταξύ τους. Στην συνέχεια, πραγματοποίεται συσταδοποίηση όμοια με αυτή του BIRCH [31]. Έτσι ομαδοποιούνται λέξεις-κλειδία με συνάρτηση ομοιότητας του συνημιτόνου. Κάθε φορά συγκρίνονται το κεντροειδές της συστάδας με την χωρική σφραγίδα της εκάστοτε λέξης. Αν η λέξη-κλειδί ανήκει σε μία συστάδα ενσωματώνεται μεταβάλλοντας το κεντροειδές της, διαφορετικά αποτελεί μια συστάδα η ίδια η λέξη.

4.4 Εξερεύνηση γεγονότων με ιεραρχική συσταδοποίηση

Η μέθοδος που προτάθηκε στο [12] επιτρέπει να έξερευνούμε΄ στο κοινωνικό δίκτυο Twitter και να εντοπίζουμε γεγονότα για διαφορετικές διαστάσεις του χώρου και χρόνου. Πιο συγκεκριμένα δίνεται η δυνατότητα να εστιάσμού στον χάρτη προκειμένου να εντοπίστούν τοπικά γεγονότα είτε να σμικρύνσης του χάρτη προκειμένου να εντοπίστούν ευρέως διαδεδομένα γεγονότα όπως φαίνεται και στην εικόνα 4.5. Αντίστοιχα δίνεται η δυνατότητα μεταβολής της χρονικής διάστασης προκειμένου να εντοπίστούν γεγονότα με διαφορετική επιρροή στον χρόνο. Η επεξεργασία των δεδομένων γίνεται όπως και προηγουμένως σε πραγματικό χρόνο. Η παρούσα μεθοδολογία μεταχειρίζεται τα γεγονότα σαν συστάδες από ένα σύνολο *ετικετών* (hashtags). Για την ευελιξία στον χωρό και στον χρόνο υλοποιείται μια παραλλαγή της δομής του κύβου (OLAP) που χρησιμοποιείται στις αποθήκες δεδομένων με όνομα STREAMCUBE.

4.4.1 Χωρική και χρονική ιεραρχικοποίηση

Σε πολλές εφαρμογές που διαχειρίζονται χωροχρονικά δεδομένα είναι αποτελεσματικό να χρησιμοποιούμε δομές με ιεράρχιση του χώρου και του χρόνου. Μια γνώστη τεχνική, που χρησιμοποιείται και στις αποθήκες δεδομένων είναι ο κύβος δεδομένων (OLAP). Ο κύβος δεδομένων είναι κατάλληλος για οργάνωση και εξερεύνηση πολυδιάστατων δεδομένων. Στην



Σχήμα 4.5: Κύβος χωροχρονικών διαστάσεων [12]

προκειμένη περίπτωση οι διαστάσεις είναι δύο, ο χώρος και ο χρόνος. Για την επεξεργασία δεδομένων που είναι αποθηκευμένα στον κύβο, υπάρχουν μια σειρά απο λειτουργίες που είναι οι ακόλουθες:

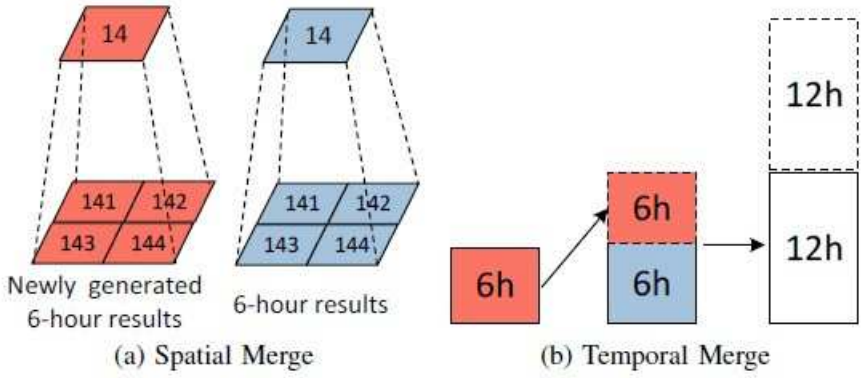
- Τεμαχισμός: Βλέπουμε έναν υποκύβο για να εκλάβουμε πιο ειδικές πληροφορίες. Αυτό γίνεται επιλέγοντας μία διάσταση.
- Συναθροιστική άνοδος: Επιτρέπει στον χρήστη να θέτει ερωτήματα κινούμενος στην ιεραρχία προς τα πάνω.
- Συναθροιστική κάθοδος: Με αυτή την λειτουργία ο χρήστης λαμβάνει περισσότερες λεπτομέρειες κατευθυνόμενος προς τα κάτω στην ιεραρχία.

Η μεθοδολογία αυτή χρησιμοποιεί μια προέκταση της δομής του κύβου δεδομένων, με το όνομα STREAMCUBE. Ο STREAMCUBE κατασκευάστηκε λαμβάνοντας υπόψην την χρονική και χωρική διάσταση. Για την διακριτοποίηση του χώρου και την ιεράρχιση του χρησιμοποιήθηκε μια δομή παρόμοια με αυτή του τετραδικού δέντρου (quad-tree) όπως απεικονίζεται στο σχήμα 4.6. Το πρώτο επίπεδο αναπαριστά τον ενιαίο χώρο. Καθώς πλοηγούμαστε προς τα κάτω στην ιεραρχία ο χώρος διασπάται σε τέσσερις ίδιου μεγέθους περιοχές. Αντίστοιχα για την ιεράρχιση του χρόνου, στο πρώτο επίπεδο τοποθετούμε την μεγαλύτερη διάρκεια του χρόνου που έχει οριστεί. Στην παρούσα μεθοδολογία αυτή είναι μία ημέρα. Το δεύτερο επίπεδο χωρίζεται σε δύο τμήματα που το καθένα αναπαριστά ένα χρονικό διάστημα 12 ωρών κ.ο.κ.

Τέλος, θα πρέπει να υπάρχει μία σύνδεση μεταξύ της χωρικής και της χρονικής ιεράρχησης. Αυτό επιτυγχάνεται με τον εμφωλιασμό της χωρικής ιεράρχησης στην χρονική. Δηλαδή κάθε τμήμα της χρονικής διάστασης δείχνει σε μια δόμη ιεραρχησης του χώρου. Οπως είναι φανερό οι δομές αυτές που προαναφέραμε μας επιτρέπουν να μεταβάλλουμε τις διάστασεις του χώρου και του χρόνου για την ανίχνευση γεγονότων.

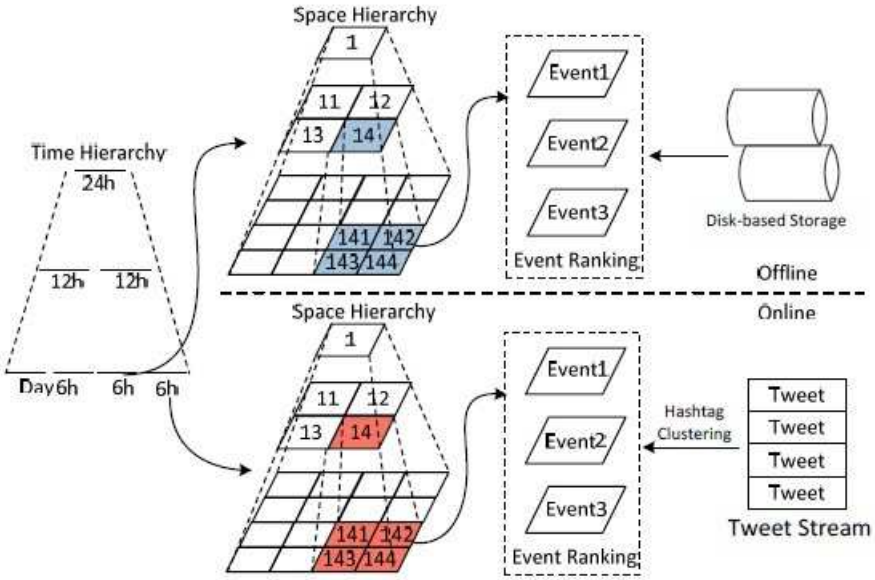
4.4.2 Συσταδοποίηση από Hastags

Για αυτή την μεθοδολογία θεωρήσαμε πως ένα γεγονός είναι ένα σύνολο απο hashtags που επικεντρώνονται στο ίδιο θέμα. Πιο συγκεκριμένα ένα γεγονός αναπαρίσταται απο βεβαρυμένα



Σχήμα 4.6: Χρονική και χωρική ιεαρχικοποίηση [12]

hashtags. Επίσης η προαναφερθείσα μεθοδολογία λαμβάνει υπόψη και την εξέλιξη του περιεχομένου των hashtags για ένα θέμα. Παραδείγματος χάρη, τα hashtags Obama,Elections είναι άρρηκτα συνδεδεμένα κατά την διάρκεια των εκλογών όμως για άλλη χρονική περίοδο ίσως δεν έχουν σχέση. Κατά την επεξεργασία, όταν ένα ηασηταγ καταφθάνει και δεν έχει ξαναεμφανιστεί τότε σχηματίζουμε μια καινούργια συστάδα και ονομάζουμε την συστάδα ανενεργή. Αν το ηασηταγς ανήκει σε μία απο τις υπάρχουσες συστάδες τότε το ενσωματώνουμε στην συστάδα και ελέγχουμε την κατάσταση της. Αν το γεγονός δεν περιέχει πολλά tweets τότε το κρατάμε ανενεργό. Στην περίπτωση μας ανενεργή συστάδα είναι η συστάδα που περιέχει λιγότερα απο 30 tweets. Αν η συστάδα απο ανενεργή έγινε ενεργή τότε ελέγχουμε αν πρέπει να ενταχθεί στη λίστα των ενεργών γεγονότων. Αν παρέμεινε ενεργή τότε ελέγχουμε αν πρέπει να διασπαστεί ή να συγχωνευτεί. Τέλος η λίστα των ενεργών γεγονότων ταξινομείται με βάση κάποιες στατιστικές μετρικές προκειμένου να δίνοντα στον χρήστης τα αποτελεσμάτα ταξινομημένα. Συνοπτικά η μεθοδολογία υποδεικνύεται σχηματικά στο σχήμα 4.7.



Σχήμα 4.7: Μεθοδολογία [12]

Κεφάλαιο 5

Μοντελοποίηση του προβλήματος

5.1 Εισαγωγή

Αντικείμενο της παρούσας διπλωματικής αποτελεί ο σχεδιασμός και η υλοποίηση ενός αλγορίθμου που αποσκοπεί στην ανίχνευση δημοφιλών συζητήσεων που εκτυλίσσονται στα κοινωνικά δίκτυα, αλλά και στην διαρκή εποπτεία τέτοιων συζητήσεων με σκοπό την μελέτη της εξέλιξής τους στον χώρο και στον χρόνο.

Σήμερα, γνωρίζουν άνθηση πολλά κοινωνικά δίκτυα όπως το Twitter [28], το Facebook [11], το Foursquare [13] κ.ά. Από αυτά επιλέξαμε να μελετήσουμε το φαινόμενο στο Twitter διότι ο τρόπος λειτουργίας και η δομή του εξυπηρετούν καλύτερα την παρούσα μελέτη. Πιο συγκεκριμένα, το Twitter αποτελεί ένα ρεύμα δεδομένων (δεν είναι τυχαίο που το αποκαλούν Twitter Stream), καθώς οι δημοσιεύσεις χρηστών ανά τον κόσμο γίνονται σε πολύ γρήγορο ρυθμό. Ενδεικτικά, κατά το έτος 2015, οι χρήστες δημοσίευσαν μηνύματα με ρυθμό 500.000.000 ανά ημέρα σε παγκόσμια κλίμακα. Επιπλέον, οι δημοσιεύσεις (tweets) των χρηστών αποτελούνται μέχρι 140 χαρακτήρες, κάνοντας έτσι την επεξεργασία της πληροφορίας πιο εύκολη σε σχέση με τα άλλα κοινωνικά δικτύα στα οποία αυτός ο περιορισμός δεν υπάρχει.

Δίνεται η δυνατότητα μελέτης του φαινομένου αυτού θέτοντας διαφορετικές παραμέτρους ως είσοδο στον αλγόριθμο. Τα δεδομένα που επεξεργαζόμαστε είναι οι δημοσιεύσεις των χρηστών (tweets). Για τον σκοπό της μελέτης λάβαμε υπ΄ όψιν την επεξεργασία των δημοσιεύσεων που συμπεριλαμβάνουν την χωρική πληροφορία προκειμένου ο αλγορίθμος να έχει τη δυνατότητα να τα εξετάσει και ως προς την εξελιξή τους στο χώρο. Τα μηνύματα θεωρείται ότι δημοσιεύονται σε πραγματικό χρόνο, οπότε είναι ανάγκη η επεξεργασία των δεδομένων να είναι αποδοτική αποδεχόμενοι ότι τα αποτελέσματα θα είναι κατ΄ ανάγκη προσεγγιστικά. Επομένως για την σχεδίαση της μεθοδολογίας επιδιώκουμε ελαχιστοποίηση της πολυπλοκότητας του αλγορίθμου. Η μεθοδολογία εξάγει ενημερωμένα αποτελέσματα ανά τακτά χρονικά διαστήματα και ανταποκρίνονται στην συνεχή ροή μηνυμάτων.

Στο κεφάλαιο αυτό γίνεται μια σαφής διατύπωση του μοντέλου του προβλήματος που μελετάται στην παρούσα διπλωματική εργασία. Για το λόγο αυτό περιγράφονται τα μοντέλα των δεδομένων που διαχειριζόμαστε, όπως επίσης και οι υποθέσεις και θεωρήσεις που έχουμε

κάνει για την επίλυση του προβλήματος.

5.2 Μοντέλο συστήματος

Στην συνέχεια περιγράφονται οι υποθέσεις και οι θεωρήσεις που έχουμε κάνει όσον αφορά τη φύση και το περιεχόμενο των δεδομένων, αλλά και τις παραμέτρους που εμπλέκονται στον προτεινόμενο αλγόριθμο.

5.2.1 Ρεύμα δημοσιευμένων μηνυμάτων (Twitter Stream)

Οι δημοσιεύσεις των χρηστών σχηματίζουν ένα ρεύμα δεδομένων της μορφής S=(..., s_i, ..., s_j, ...). Κάθε μήνυμα (tweet) $s = < \tau, uid, loc, W >$ δηλώνει την:

- την χρονική στιγμή τ που ο χρήστης δημοσίευσε το tweet,
- τον κωδικό χρήστη uid, που είναι μοναδικός και προσδίδει ταυτότητα στον χρήστη,
- προαιρετικά το γεωγραφικό στίγμα loc της δημοσίευσης, (την τοποθεσία που βρισκόταν ο χρήστης τη στιγμή της δημοσίευσης),
- ένα σύνολο *λέξεων* $W = \{w_1, w_2, ..., w_k\}$ που αποτελεί το κειμενικό περιέχομενο της δημοσίευσης.

Παράδειγμα μιας δημοσίευσης στο Twitter είναι η ακόλουθη:

$s =< 31/12/2012\ 23:32:42, 143134834343, (53.432, 43.732),$
$\#NewYorkNewYearsEveHappyNewYear >$

Στην πραγματικότητα μόνο το 2% των δημοσιεύσεων στο Twitter περιλαμβάνουν την γεωγραφική τοποθεσία του χρήστη για λόγους εμπιστευτικότητας, ασφάλειας ή απουσία μηχανισμού εύρεσης τοποθεσίας. Δημοσιεύσεις που περιλαμβάνουν γεωγραφική τοποθεσία μπορεί να είναι κάποιες δημοσιεύσεις που αναφέρονται σε έναν ποδοσφαρικό αγώνα και οι φίλαθλοι δημοσιεύουν απο το χώρο διεξαγωγής του αγώνα, είτε μια πορεία διαμαρτυρίας, όπου το γεωγραφικό στίγμα των δημοσιεύσεων σχετίζεται με την τοποθεσία που διαδραματίζεται. Για το σκοπό της διπλωματικής εργασίας λάβαμε υπόψη μόνο τις δημοσιεύσεις που δηλώνουν ρητώς γεωγραφικό στίγμα. Δεν γίνεται σημασιολογική ανάλυση του μηνύματος για την εξαγωγή της γεωγραφικής πληροφορίας. Εφόσον λ.χ. στο μήνυμα εμφανίζεται η λέξη Syntagma μπορούμε με τεχνικές ανάλυσης κειμένου (text mining) να εξάγουμε την ακριβή γεωγραφική τοποθεσία με συντεταγμένες $< 23.734, 37.975 >$. Τέλος υποθέτουμε ότι κάθε δημοσίευση προέρχεται απο έναν συγκεκριμένο γεωγραφικό χώρο G. Ένας τέτοιος χώρος μπορεί να είναι μια γεωγραφική έκταση που περιλαμβάνει μία πόλη, μία χώρα και μία ολόκληρη ήπειρο. Ουσιαστικά ο χώρος αυτός υποδηλώνει την γεωγραφική περιοχή που παρουσιάζει ενδιάφερον να μελετήσουμε.

Η επεξεργασία όλης της κειμενικής πληροφορίας ξεφεύγει από την μελέτη του παρόντος προβλήματος καθώς εμπίπτει σε κλάδους επιστημών όπως εξόρυξη γνώσης απο κειμενικά δεδομένα (text mining), ανάλυση συναισθήματος (sentimemt analysis) κ.α. Επομένως δεν

λάβαμε υπόψη όλη την κειμενική πληροφορία αλλά απο το σύνολο W εξάγαμε το σύνολο των hashtags $H = \{h_1, h_2, \dots h_j\}$ όπου ισχύει $H \subseteq W$. Αυτή η παραδοχή δεν απέχει πολύ απο την πραγματικότητα, αφού οι συζητήσεις χαρακτηρίζονται με βάση τις *ετικέτες* (hashtags) των δημοσιεύσεων. Για παράδειγμα, την περίοδο εκλογών στην Αμερική όλες οι συζητήσεις που αφορούσαν στο θέμα αυτό, περιελάμβαναν ετικέτες όπως #Obama, #RomneyElections, #ObamaVsRomney. Δηλαδή συζητήσεις που εξελίσσονται στο Twitter περιλαμβάνουν τα ίδια ή παρόμοια σύνολα ετικετών. Επομένως η δομή καθενός tweet για τις ανάγκες της επεξεργασίας διαμορφώνεται ως εξής: $s =< \tau, uid, loc, H >$.

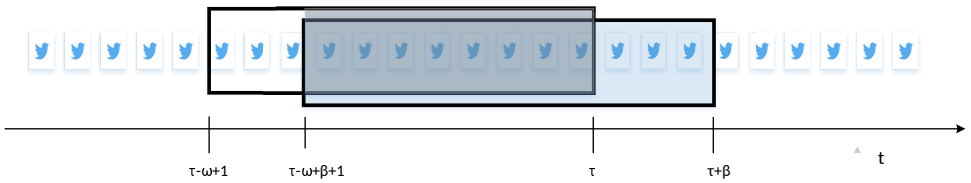
5.2.2 Χρονικά κυλιόμενο παράθυρο

Για την καλύτερη και αποδοτικότερη αντιμετώπιση του μεγάλου όγκου δεδομένων απο νεοεισερχόμενα μηνυμάτα (tweets) χρησιμοποιήθηκε χρονικά *κυλιόμενο παράθυρο (sliding window)*. Αυτό το παράθυρο συμβαδίζει με την εξέλιξη των δεδομένων στο χρόνο και έχει διακριτοποιηθεί σε χρονικά μη επικαλυπτόμενα *πλαίσια (panes)* [20] προκαθορισμένου ορίζοντα $(\tau_{now} - \omega, \tau_{now}] \rightarrow (\tau_{now} - \omega, \tau_{now} - \omega + \beta], (\tau_{now} - \omega + \beta, \tau_{now} - \omega + 2\beta], \dots, (\tau_{now} - \beta, \tau_{now}]$ όπου τ_{now} είναι τρέχουσα χρονική στιγμή, ω το *χρονικό εύρος (range)* του παραθύρου και β το *χρονικό βήμα (sliding step)*. Από τη δομή του κυλιόμενου παραθύρου είναι φανερό πως ο αριθμός των μη επικαλυπτόμενων χρονικών πλαισίων που απαρτίζουν το εύρος του κυλίομενου παραθύρου πρέπει να είναι ακέραιο πολλαπλάσιο, αλλά και επίσης το χρονικό εύρος και το χρονικό βήμα θα πρέπει να είναι ακέραιοι αριθμοί. Η παραπάνω πρόταση ουσιαστικά διατυπώνει τον εξής περιορισμό:

$$\frac{\omega}{\beta} = \lambda, \text{όπου } \lambda, \omega, \beta \in \mathbb{Z}.$$

Η δομή του κυλιόμενου παραθύρου φαίνεται στο Σχήμα 5.1. Το παράθυρο του σχήματος 5.1 εφαρμόζεται για χρονικό ορίζοντα ω, δηλαδή περιλαμβάνει δημοσιεύσεις που ανιχνεύτηκαν την συγκεκριμένη χρονική περίοδο (λ.χ μία ώρα). Στην συνέχεια το παράθυρο μετακυλίεται με βάση το χρονικό βήμα προκειμένου να συμπεριλάβει τις πιο πρόσφατες δημοσιεύσεις και οι πιό παλιές να διωχθούν. Επειδή το εύρος ω παρακολουθεί την εξέλιξη του χρόνου και το άνω άκρο συμπίπτει με το παρόν σε κάθε μετακύλιση, αναπόφευκτα τα παλαιότερα μηνύματα εκπίπτουν όταν το παράθυρο προχωρά (λ.χ όσα ελήφθησαν έως μία ώρα πρίν). Ο αριθμός των δημοσιεύσεων σε κάθε μετακύλιση του παραθύρου κυμαίνεται ανάλογα με το πλήθος των δημοσιεύσεων που παράγονται στην χρονική περίοδο που καλύπτει το εκάστοτε παράθυρο. Λ.χ την ημέρα θα υπάρχουν λογικά περισσότερα μηνύματα, ενώ την νύχτα λιγότερα κτλ.

Συμπερασματικά θα λέγαμε πως το κυλιομένο παράθυρο, συμβάλλει στην αποτελεσματική επεξεργασία του μεγάλου όγκου δεδομένων και μας προσφέρει την δυνατότητα της μελέτης των συζητήσεων ως προς την χρονική τους διάσταση, διότι καθώς μετακυλίεται το παράθυρο ο αριθμός των δημοσιεύσεων μεταβάλλεται, αφού νέες δημοσιεύσεις καταφθάνουν και άλλες εκλείπουν με την πάροδο του χρόνου.



Σχήμα 5.1: Κυλιόμενο παράθυρο εύρους ω και βήματος β τον χρόνο τ

5.2.3 Δημοφιλείς Συζητήσεις

Αν συλλογιστούμε πιο προσεκτικά τη σημασία του όρου *συζήτηση* στην καθημερινή ζωή, θα λέγαμε πως είναι ένα σύνολο προτάσεων που αναφέρονται στο ίδιο θέμα και προέρχονται απο δύο ή περισσότερους ανθρώπους. Στην προκειμένη περίπτωση, ανάγοντας την παραπάνω διατύπωση με όρους που αφορούν τα κοινωνικά δίκτυα θα λέγαμε πως μια *συζήτηση* είναι ένα σύνολο από δημοσιεύσεις που προέρχονται απο πολλαπλούς χρήστες, αναφέρονται στο ίδιο θέμα και για την εργασία μας θεωρήσαμε πως περιορίζονται σε μιά συγκεκριμένη περιοχή ενδιαφέροντος. Στην προηγούμενη ενότητα είχαμε αναφέρει πως οι συζητήσεις που εκτυλίσσονται στο Twitter παρουσιάζουν μεγάλη *ομοιότητα* ως προς το σύνολο των ετικετών που περιέχουν. Επομένως θα μπορούσαμε να θεωρήσουμε ότι μία συζήτηση στο Twitter είναι ένα σύνολο από ετικέτες που παρουσιάζουν *χωρική εγγύτητα* και αυξημένο *βαθμό ομοιότητας* ως προς το σύνολο των ετικετών. Παραδείγματος χάρη, δύο δημοσιεύσεις με σύνολα ετικετών $s_1.H = \{\#tubestrike, \#london\}$ και $s_2.H = \{\#notrain, \#london, \#tubestrike\}$ αποτελούν πιθανότατα μέρος μιας συζήτησης που αφορά την απεργία του μετρό στο Λονδίνο. Το σύνολο των ετικετών της συζήτησης είναι η ένωση των δύο συνόλων των ετικέτών: $\{\#tubestrike, \#london\} \cup \{\#notrain, \#london, \#tubestrike\}$, η οποία δίνει τελικά $\{\#tubestrike, \#london, \#notrain\}$.

Στο σημείο αυτό κρίνεται επιτακτική ανάγκη να δώσουμε έναν πιο αυστηρό ορισμό της έννοιας της *συζήτησης*, κάτι που θα μας βοηθήσει στην σχεδίαση της μεθοδολογίας. Έστω το ρεύμα δημοσιευμένων μηνυμάτων στο Twitter $S = \{\ldots, < \ \tau, uid, loc, H \ > \ldots\}$ και $\mathcal{T} = \{T_1, T_2, \ldots T_k, \ldots T_n\}$ το σύνολο των συζητήσεων που ανιχνεύθηκαν απο προηγούμενη εκτέλεση του αλγορίθμου. Για κάθε μήνυμα $s \in S$ του ρεύματος δημοσιεύσεων S, καλούμαστε να εντοπίσουμε αν το s ανήκει σε κάποια συζήτηση T_k αλλιώς εφόσον δεν ανήκει σε κάποια συζήτηση, δημιουργείται μια νέα συζήτηση T_{n+1}. Έστω μια συνάρτηση ομοιότητας $sim(s.H, T_k)$ η οποία δέχεται ως ορίσματα το μέρος του μηνύματος που περιέχει τις *ετικέτες* $s.H$ και μία συζήτηση T_k απο το σύνολο των τρεχουσών δημοσιεύσεων. Η συνάρτηση ομοιότητας επιστρέφει έναν πραγματικό αριθμό που εκφράζει τον βαθμό ομοιότητας του μηνύματος με την εξεταζόμενη συζήτηση. Ορίζουμε ένα κατώφλι ομοιότητας θ, το οποίο είναι ένας πραγματικός αριθμός και ισχύει $0 \leqslant \theta \leqslant 1$. Ένα νεοεισερχόμενο μήνυμα s αφορά την συζήτηση T_k εφόσον η συνάρτηση ομοιότητας sim επιστρέψει τιμή μεγαλύτερη του κατωφλίου θ. Πιθανόν το ίδιο μήνυμα s να ταιριάζει με διάφορες προϋπάρχουσες συζητήσεις στο $\mathcal{T}$, όμως τελικά θα ενταχθεί σε μία μόνο συζήτηση T_k, αυτή με την οποία έχει τον υψηλότερο βαθμό ομοιότητας. Όλα αυτά συμπυκνώνονται στον παρακάτω ορισμό που προσδιορίζει την έννοια της συζήτησης:

$$T_k = \{s.H, s \in S : (\forall s_k.H \in T_k, sim(s_k.H, s.H) > \theta) \bigwedge$$
$$(\forall s_i \in C \setminus T_k, sim(s_i.H, s_k.H) < sim(s_k.H, s.H))\}$$

Στην συνέχεια, θα ορίσουμε τον όρο *δημοφιλής συζήτηση*. Προφανώς μία συζήτηση είναι δημοφιλής εφόσον υπάρχει απήχηση στον κόσμο, δηλαδή συμμετέχουν σε αυτήν πολλά άτομα. Έτσι και στα κοινωνικά δίκτυα, δημοφιλής μπορούμε να πούμε πως είναι η συζητήση στην οποία ο αριθμός των μηνυμάτων που συμμετέχουν ξεπερνά μια *ελάχιστη τιμή δημοφιλίας* $\phi \in \mathbb{R}^+$. Στην εργασία μας, η ελάχιστη τιμή δημοφιλίας ϕ εκφράζει το ελάχιστο πλήθος μηνυμάτων που θα πρέπει να περιλαμβάνει μια συζήτηση προκειμένου να χαρακτηριστεί δημοφιλής επομένως οι τιμές που λαμβάνει είναι πραγματικοί αριθμοί. Θεωρούμε ότι κάθε μήνυμα που επεξεργαζόμαστε προέρχεται απο διαφορετικό χρήστη, κάθε συζήτηση εξελίσσεται για συγκεκριμένο χρονικό διάστημα και είναι δημοφιλής σε μια συγκεκριμένη περιοχή ενδιαφέροντος. Αναφερόμενοι στο προηγούμενο παράδειγμα, μια δημοφιλής συζήτηση στο Twitter θα μπορούσε να αφορά την απεργία των υπαλλήλων του μετρό και χαρακτηρίζεται από την ετικέτα #tubestrike ή συγγενικές της όπως {#tubestriketoday, #tubestrikeagain}, ενώ ο αριθμός των μηνυμάτων που αναφέρονται στις παραπάνω ετικέτες είναι μεγαλύτερος απο δέκα και η περιοχή ενδιαφέροντος είναι το Λονδίνο.

Για την επίλυση του προβλήματος θεωρήσαμε δημοφιλή μια συζήτηση για την οποία το σύνολο των ετικετών (hashtags) των μηνυμάτων που την απαρτίζουν παρουσιάζουν έναν αυξημένο *βαθμό ομοιότητας* μεταξύ τους (κατώφλι θ), *χωρική εγγύτητα*, ενώ επίσης ο αριθμός των μηνυμάτων ξεπερνά την *ελάχιστη τιμή δημοφιλίας* ϕ στον προκαθορισμένο χρονικό ορίζοντα που θέλουμε να μελετήσουμε. Έστω μια περιοχή ενδιαφέροντος A και ω το χρονικό εύρος του κυλιόμενου παραθύρου. Δοθείσης μιας ελάχιστης τιμής δημοφιλίας ϕ, ένα ήδη υπάρχον θέμα συζήτησης T_k θεωρείται δημοφιλές εφόσον ικανοποείται η παρακάτω σχέση:

$$popularity(A, \omega) = \{s_k \in T_k, ((s_k.loc \in A) \bigwedge (s_k.\tau \in (\tau_{now} - \omega, \tau_{now}]))\}$$

και

$$|popularity(A, \omega)| > \phi.$$

Η συνάρτηση *popularity* εκφράζει πόσο δημοφιλές είναι ένα θέμα συζήτησης κατά την διάρκεια του τρέχοντος παραθύρου ω στην δεδομένη περιοχή A. Αυτό συμβαίνει όταν το σύνολο των μηνυμάτων που συγκροτούν αυτό το θέμα και εντοπίζονται εντός της περιοχής A υπερβαίνουν την προκαθορισμένη τιμή ϕ.

5.2.4 Χωρική κάλυψη και εποπτεία εξέλιξης τους

Το πρόβλημα της εύρεσης δημοφιλών συζητήσεων έγκειται στο γεγονός ότι δεν γνωρίζουμε εκ των προτέρων την περιοχή κάλυψης απο την οποία θα συμπεράνουμε την δημοτικότητα μιας συζήτησης T_k. Απεναντίας, η *χωρική κάλυψη (spatial coverage)* πρέπει να ανιχνευθεί βάσει των γεωγραφικών στιγμάτων των μηνυμάτων της συζήτησης που εξετάζεται. Ας μην

ξεχνάμε ότι για την εργασία αυτή θεωρήσαμε πως όλα τα μηνύματα (tweets) φέρουν γεωγραφικές συντεταγμένες. Παραδείγματος χάρη, μια τέτοια κάλυψη που αφορά συζήτηση και συγκεκριμένα πλήθος μηνυμάτων για τον ιό έμπολα απεικονίζεται στην εικόνα του Σχήματος 5.2. Οι περιοχές με κίτρινο χρώμα αναπαριστούν τις περιοχές ενδιαφέροντος στις οποίες η συζήτηση για τον ιό έμπολα είναι δημοφιλής.

Η εξαγωγή αποτελεσμάτων όπως του Σχήματος 5.2 δεν είναι τίποτα άλλο από μία διαδικασία *χωρικής συσταδοποίησης (spatial clustering)*. Κατά καιρούς έχουν παρουσιαστεί διάφορες μέθοδοι χωρικής συσταδοποίησης με πιο διαδεδομένη την μέθοδο DBSCAN [24]. Η εφαρμογή της μεθόδου αυτής θα παρήγαγε καλύτερα αποτελέσματα απο άποψη ακρίβειας. Παρ΄ όλα αυτά, δεν ενδείκνυται για την εργασία μας καθώς η πολυπλοκότητα του DBSCAN είναι απαγορευτική για ανάλυση δεδομένων σε πραγματικό χρόνο, αφού επεξεργάζεται μεγάλο πλήθος σημείων με πολυπλοκότητα $O(nlogn)$, όπου το n το πλήθος των σημείων που πρέπει να εξεταστεί σε κάθε κύκλο εκτέλεσης. Στην προκειμένη περίπτωση, το πλήθος n μπορεί να δηλώνει εκατοντάδες χιλιάδες ή και εκατομμύρια μηνύματα (tweets) που έχουν ληφθεί εντός παραθύρου (λ.χ ανά ώρα), γεγονός που καθιστά την χωρική συσταδοποίηση μη αποδοτική. Απεναντίας, θα είναι προτιμότερο ο χώρος να διακριτοποιηθεί σε ένα μικρό πλήθος απο μη επικαλυπτόμενες περιοχές σταθερού μεγέθους προκειμένου να γίνεται σε αυτές όλη η επεξεργασία των μηνυμάτων, και όχι η ακριβή χωρική συσταδοποίηση κάθε νεοσεισερχόμενου μηνύματος. Ένας ακόμη λόγος μή εφαρμογής μιας μεθόδου χωρικής συσταδοποίησης είναι πως δεν ενδιαφερόμαστε μόνο να εντοπίσουμε την χωρική κάλυψη του φαινομένου αλλά και πιθανές μεταβολές στην χωρική έκταση (συρρίκνωση, εξάπλωση) με την πάροδο του χρόνου.

Είμαστε πλέον σε θέση να διατυπώσουμε την έννοια της *χωρικής κάλυψης* μιας δημοφιλούς συζήτησης. Έστω G ένας ομοιόμορφος κάνναβος που τοποθετείται στην περιοχή παρακολούθησης (area of monitoring) και την διαμερίζει σε ένα σύνολο απο κελιά $g \times g$. Ο κάνναβος επιτρέπει να θεωρήσουμε την περιοχή κάλυψης A μιας συζήτησης T_k ως ένα σύνολο Q από τέτοια κελιά g:

$$Q = \{g \in G : |popularity(g, \omega)| > \phi\}$$

Έπειτα, εντοπίζονται συμπαγείς περιοχές κάλυψης, καθεμία από τις οποίες (έστω A μία από αυτές) αποτελείται από πολλαπλά γειτονικά κελιά του καννάβου όπου η συγκεκριμένη συζήτηση T_k είναι δημοφιλής:

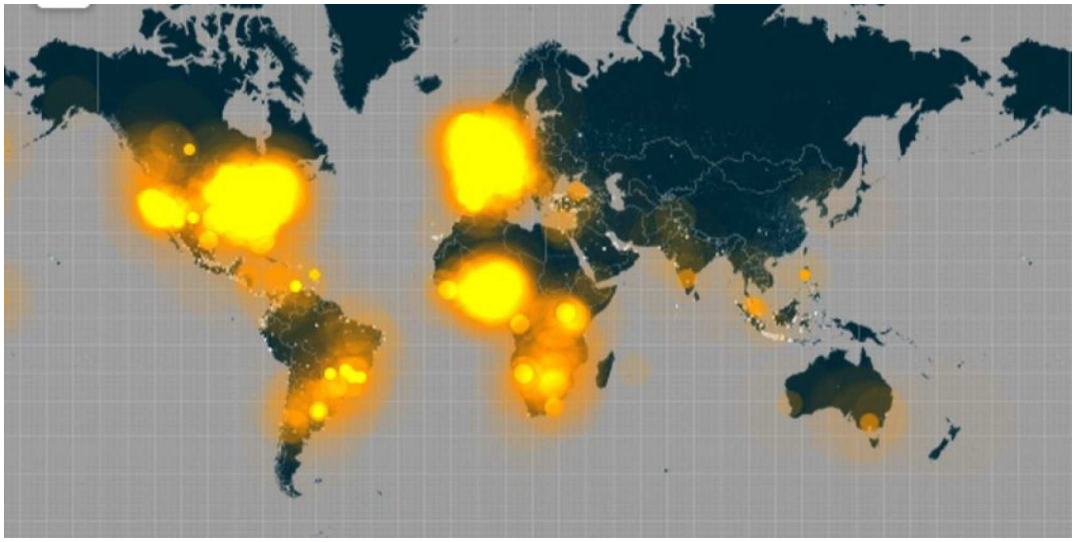
$$A = \{g_i \in Q : \exists g_j, adj(g_i, g_j) = true\}$$

ή αποτελείται από ένα μεμονωμένο κελί:

$$A = \{g_i \in Q : \forall g_j, adj(g_i, g_j) = false\}$$

όπου adj είναι μια λογική συνάρτηση που αληθεύει εφόσον τα δύο εξεταζόμενα κελιά g_i, g_j είναι γειτονικά. Άρα μια περιοχή κάλυψης A ορίζεται ως ένα σύνολο απο κελιά του καννάβου που είναι όλα γειτονικά μεταξύ τους ή από μεμονωμένα κελιά που εντοπίζονται (δεν υπάρχουν γειτονικά), καθένα από τα οποία αποτελεί απο μόνο του μία περιοχή κάλυψης.

Τέλος, θα ορίσουμε την έννοια της εξάπλωσης και συρρίκνωσης μιας δημοφιλούς συζήτησης. Εστω σύνολο κελιών Q που ανιχνεύθηκαν την χρονική στιγμή τ για μία δημοφιλή



Σχήμα 5.2: Εξάπλωση συζήτησης στο Twitter που αφορά τον ιό έμπολα (*http://www.reddit.com*)

συζήτηση T_k. Στην αμέσως επόμενη μετακύλιση του παραθύρου $\tau' = \tau + \beta$, έστω ότι η ίδια δημοφιλής συζήτηση καλύπτει μια περιοχή Q'. Ο εντοπισμός διαφορών στην χωρική κάλυψη επηρεάζει δύο διαφορετικά σύνολα αποτελούμενα από κελιά του καννάβου:

- Το σύνολο $D^+ = \{g \in G : g \in Q' \bigwedge g \notin Q\}$ περιέχει όλα τα κελιά g του κάνναβου G όπου το θέμα συζήτησης T_k δεν ήταν δημοφιλές την χρονική στιγμή τ ενώ έγινε δημοφιλές την χρονική στιγμή τ'.

- Αντίστοιχα, το σύνολο $D^- = \{g \in G : g \in Q \bigwedge g \notin Q'\}$ περιλαμβάνει τα κελιά g του καννάβου G που την χρονική στιγμή τ η συζήτηση T_k ηταν δημοφιλές, ενώ την χρονική στιγμή τ' δεν ήταν δημοφιλές.

Με βάση τα παραπάνω υποσύνολα κελιών είναι δυνατόν να εξακριβωθεί η χωρική *εξάπλωση ή συρρίκνωση* μιας δημοφιλούς συζήτησης στην περιοχή παρακολούθησης. Μια συγκεκριμενη δημοφιλής συζήτηση T_k θεωρείται ότι εξαπλώθηκε μεταξύ δύο διαδοχικών κύκλων εκτέλεσης εφόσον ικανοποείται η σχέση: $|D^+| - |D^-| > 0$, ενώ συρρικνώθηκε εφόσον ισχύει: $|D^+| - |D^-| < 0$. Ουσιαστικά η εξάπλωση ή συρρίκνωση της χωρικής κάλυψης ενός συγκεκριμένου θέματος συζήτησης T_k καθορίζεται από το πρόσημο της παράστασης $|D^+| - |D^-|$.

5.3 Διατύπωση του προβλήματος

Όπως αναφέραμε και στην εισαγωγή του κεφαλαίου, στόχος της παρούσας εργασίας είναι η ανάπτυξη μιας μεθόδου που θα εντοπίζει *δημοφιλείς συζητήσεις* στο Twitter και οι οποίες εκτυλίσσονται σε μια περιοχή ενδιαφέροντος, ενώ επίσης θα δίνει την δυνατότητα μελέτης τέτοιων συζητήσεων ως προς την εξέλιξη τους στον χώρο και τον χρόνο.

Δοθέντος ενός συνόλου S από δημοσιεύσεις $s \in S$ που έχουν καταγραφεί στο χρονικό παράθυρο $(\tau_{now} - \omega, \tau_{now}]$, σκοπός είναι η ανίχνευση δημοφιλών συζητήσεων μεταξύ των

δημοσιεύσεων, η εύρεση της γεωγραφικής τους εξάπλωσης σε έναν ομοιόμορφα διακριτοποιημένο χώρο G και τέλος η εποπτεία τους στον χωρό και τον χρόνο. Όπως προαναφέραμε σε προηγούμενη ενότητα, οι δημοφιλείς συζητήσεις εντοπίζονται σε περιοχές κάλυψης. Στην παρούσα διπλωματική εργασία, περιοχές κάλυψης θεωρήσαμε τα επιμέρους ομοιόμορφα μη επικαλυπτόμενα κελιά g του χώρου G. Επομένως για το πρόβλημα μας, ο εντοπισμός της εξάπλωσης μιας δημοφιλούς συζήτησης ανάγεται στην εύρεση του συνόλου των κελιών του G όπου αυτή η συζήτηση παρατηρείται. Ο λόγος που επιλέχθηκε να διαμερίσουμε τον χώρο σε ομοιόμορφα κελιά δεν είναι τυχαίος. Ένας λόγος είναι ότι λειτουργεί ως ευρετήριο και επομένως αυξάνεται η απόδοση του αλγορίθμου και ο άλλος λόγος είναι ότι μας εξυπηρετεί προκειμένου να εντοπίζουμε τις χωρικές διακυμάνσεις των συζητήσεων με την πάροδο του χρόνου. Η δομή του κυλιόμενου παραθύρου σε μη επικαλυπτόμενα χρονικά πλαίσια (panes) και η διακριτοποίηση του χώρου μας επιτρέπουν να κάνουμε επαυξητικά την ενημέρωση των δεδομένων, καθώς μετακυλίεται το παράθυρο, χωρίς να επαναλαμβάνουμε την ίδια διαδικασία εξ΄ αρχής κάθε φορά. Ο μηχανισμός ενημέρωσης περιλαμβάνει την διαγραφή δημοφιλών συζητήσεων που βρέθηκαν σε κάθε κελί για χρονικές στιγμές $\tau_i \leq \tau_{now} - \omega$, όπως επίσης και η αναθεωρήση των συζητήσεων για το τρέχον χρονικό πλαίσιο $(\tau_{now} - \beta, \tau_{now}]$ σε κάθε κελί.

Ουσιαστικά, η παρακολούθηση της εξάπλωσης των συζητήσεων θα είναι μια ποιοτική και ποσοτική σύγκριση της χωρικής έκτασης του φαινομένου, καθώς σε κάποιες περιοχές η ίδια συζήτηση μπορεί να είναι πιο έντονη (δηλ. περισσότερα μηνύματα) συγκριτικά με κάποιες άλλες. Αυτό θα μας δίνει την δυνατότητα να ανιχνεύσουμε αν το φαινόμενο εξαπλώθηκε στον χώρο ή αν σύρρικνώθηκε με την προηγούμενη μετακύλιση του παραθύρου. Τέλος θα είμαστε σε θέση να εκτιμήσουμε και σε ποιά ένταση έγινε η μεταβολή αυτή.

G	Κάνναβος
g	Κελί καννάβου
ω	Εύρος χρονικού παραθύρου
β	Βήμα κύλισης χρονικού παραθύρου (pane)
τ_{now}	Τρέχουσα χρονική στιγμή
S	Ρεύμα μηνυμάτων (Twitter Stream)
s	Μήνυμα (tweet) $< \tau, uid, loc, H >$
τ	Χρονόσημο
uid	Αναγνωριστικό του χρήστη
loc	Γεωγραφικό στίγμα μηνύματος
H	*ετικέτες* (hashtags) του μηνύματος
N_ω	Τρέχων αριθμός μηνυμάτων εντός παραθύρου
θ	Κατώφλι ομοιότητας συζήτησης
ϕ	Ελάχιστη τιμή δημοφιλίας
$\mathcal{P}$	Ευρετήριο με τις δημοφιλείς συζητήσεις σε όλη την έκταση της περιοχής μελέτης
$\mathcal{T}^g$	Ευρετήριο συζητήσεων που εντοπίζονται εντός ενός μεμονωμένου κελιού
$\mathcal{T}_t^g$	Σύνολο συζητήσεων που εντοπίζονται στο πλαίσιο $(t-\beta, t]$
Q	Σύνολο κελιών όπου εντοπίζεται μια δημοφιλής συζήτηση
A	Συμπαγής περιοχή κάλυψης για μια συγκεκριμένη συζήτηση
$\mathcal{A}[T_i]$	Λίστα με τις περιοχές κάλυψης μιας δημοφιλούς συζήτησης T_i

Πίνακας 5.1: Συμβολισμοί

Κεφάλαιο 6

Εποπτεία γεωγραφικής κάλυψης συζητήσεων σε κοινωνικά δίκτυα

6.1 Εισαγωγή

Στο κεφάλαιο αυτό παρουσιάζεται ο αλγόριθμος που αναπτύχθηκε με σκοπό την επεξεργασία των συζητήσεων που εκτυλίσσονται στο κοινωνικό δίκτυο *Twitter* και την εποπτεία τους στον χώρο και στον χρόνο. Η διαδικασία αυτή επαναλαμβάνεται για κάθε κύκλο εκτέλεσης της μεθοδολογίας. Η χρονική διάρκεια του κύκλου εκτέλεσης καθορίζεται απο το εύρος του χρονικού παραθυρου που έχει οριστεί ως παράμετρος εισόδου στον αλγόριθμο. Ουσιαστικά, κάθε μετακύλιση παραθύρου ορίζει και έναν κύκλο εκτέλεσης. Ο αλγόριθμος αποτελείται απο τρείς φάσεις επεξεργασίας δεδομένων. Η πρώτη φάση επεξεργασίας (ανίχνευση συζητήσεων) δέχεται τις δημοσιεύσεις των χρηστών από το ρεύμα δημοσιευμένων μηνυμάτων (Twitter Stream) εντός μιας χρονικής περιόδου (λ.χ παράθυρο 2 ωρών) και με μία διαδικασία ομαδοποίησης (συσταδοποίησης) των μηνυμάτων,την οποία θα περιγράψουμε εκτενώς στη συνέχεια του κεφαλαίου, εντοπίζονται τα θέματα που εκείνη την χρονική περίοδο οι χρήστες συζητούν.

Στην συνέχεια πραγματοποιείται ένα φιλτράρισμα των συζητήσεων προκειμένου να εντοπιστούν οι πιο *δημοφιλείς*. Τέλος, ακολουθεί το στάδιο της εύρεσης *χωρικής κάλυψης* των δημοφιλών συζητήσεων και της μελέτης της εξέλιξης τους στον χώρο και στον χρόνο, δηλαδή κατα πόσο μεταβλήθηκαν οι περιοχές κάλυψης μιας συγκεκριμένης δημοφιλούς συζήτησης σε σχέση με τον αμέσως προηγούμενο κύκλο εκτέλεσης. Στη συνέχεια του κεφαλαίου γίνεται εκτενής ανάλυση της μεθοδολογίας, η οποία απαρτίζεται από τα εξής διαδοχικά στάδια:

1. *Ανίχνευση συζητήσεων (Topic Clustering)*: Στο στάδιο αυτό εντοπίζονται ομάδες μηνυμάτων, των οποίων το γεωγραφικό στίγμα αντιστοιχεί σε ένα συγκεκριμένο κελί και ανήκουν σε μια συγκεκριμένη συζήτηση.

2. *Ανίχνευση δημοφιλών συζητήσεων (PopularityFiltering)*: Οι συζητήσεις που έχουν εντοπιστεί σε κάθε κελί του καννάβου φιλτράρονται με μία προκαθορισμένη τιμή δημοφιλίας και ανιχνεύονται οι δημοφιλείς συζητήσεις σε όλη την έκταση της περιοχής παρακολούθησης.

Algorithm 1 Monitor spatial coverage of topics in Twitter Stream

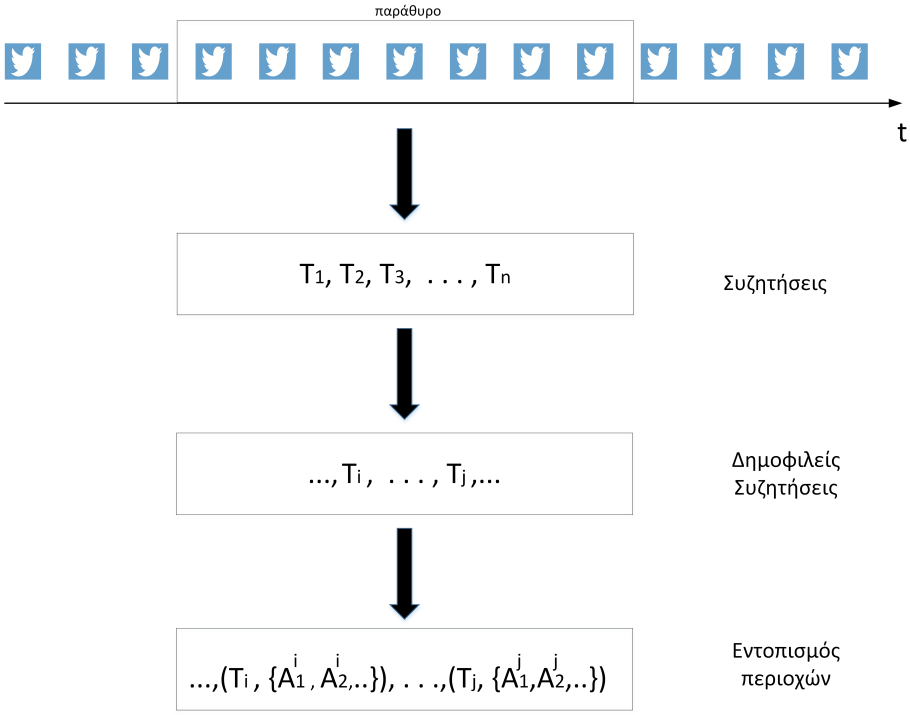
Procedure TwitterSpatialMonitoring
Input: Twitter Stream $\mathcal{S}$ of posts as tuples $\langle \tau, uid, loc, H \rangle$;
Parameter: Grid partitioning G of area of study into $g \times g$ cells;
Parameter: Window specification ⟨range ω, slide β⟩;
Parameter: Similarity threshold θ amongst hashtags in tweets;
Parameter: Minimum popularity ϕ of a topic per cell;
Output: Popular topics $\mathcal{P}$; their spatial coverage $\mathcal{A}$ per cycle;
$\tau_{now} \leftarrow \tau_0$; *//Initiation time of monitoring*
for each execution cycle spanning $(\tau_{now} - \beta, \tau_{now}]$ **do**
  $N_\omega \leftarrow |\{s \in \mathcal{S} : s.\tau \in (\tau_{now} - \omega, \tau_{now} - \beta]\}|$;
  for each tweet s with $s.\tau \in (\tau_{now} - \beta, \tau_{now}]$ **do**
    TopicClustering(s, τ_{now}, θ, β); *//s affects topics in a cell?*
    $N_\omega \leftarrow N_\omega + 1$; *//Update tweet count in window range*
  end for
  $\mathcal{P} \leftarrow$ PopularityFiltering(τ_{now}, ϕ, ω, N_ω); *//Popular topics*
  $\mathcal{A} \leftarrow$ CoverageDiscovery($\mathcal{P}$); *//Coverage per popular topic*
  $\tau_{now} \leftarrow \tau_{now} + \beta$; *//Time of the next window slide*
end for
End Procedure

3. *Εντοπισμός περιοχών κάλυψης δημοφιλών συζητήσεων και μελέτη της εξέλιξης τους (Coverage Discovery)*: Εντοπίζονται οι περιοχές κάλυψης, δηλαδή τα σύνολα των κελιών του καννάβου που είναι γειτονικά μεταξύ τους, όπου εμφανίζεται η δημοφιλής συζήτηση και τέλος συγκρίνονται τα αποτελέσματα με εκείνα που προέκυψαν από την αμέσως προηγούμενη μετακύλιση του παραθύρου, προκειμένου να εντοπιστεί πιθανή εξάπλωση ή συρρίκνωση του φαινομένου.

6.2 Γενική περιγραφή της μεθόδου

Είναι σαφές πως ο αλγόριθμος θα πρέπει να ανανεώνει τα αποτελέσματα ανά τακτά χρονικά διαστήματα. Τα χρονικά διαστήματα αυτά ορίζουν τον *κύκλο εκτελέσης* του αλγορίθμου, ο οποίος ισοδυναμεί με την μετακύλιση του χρονικού παραθύρου. Το χρονικό παραθύρο, ορίζεται από το εύρος του ω, δηλαδή το χρονικό διάστημα για το οποίο ο αλγόριθμος επεξεργάζεται τα δεδομένα, και τη χρονική διάρκεια του βήματος β (slide).

Ο χώρος διακριτοποιείται σε ομοιόμορφα κελιά g με την χρήση ενός ευρετηρίου χωρικού κάνναβου G (uniform grid partitioning). Σε κάθε κύκλο εκτέλεσης, η εξαγωγή των αποτελεσμάτων ακολουθεί την εκτέλεση των παραπάνω σταδίων, δηλαδή την *ανίχνευση των συζητήσεων*, στο στάδιο αυτό δεικτοδοτούνται τα νέα μηνύματα στον κάνναβο G και έπειτα από επεξεργασία σχηματίζονται οι συζητήσεις. Έπειτα εντοπίζονται οι *δημοφιλείς συζητήσεις*, φιλτράροντας τις συζητήσεις που ανιχνεύτηκαν προηγουμένως και τέλος ακολουθεί η *εποπτεία και ο εντοπισμός της γεωγραφικής κάλυψης των συζητήσεων*. Τα βήματα της υλοποίησης φαίνονται συνοπτικά στο σχήμα 6.1 ενώ ο αλγόριθμος 1 παραθέτει τα στάδια επεξεργασίας που περιγράψαμε παρακάτω και θα αναλυθούν στην συνέχεια του κεφαλαίου.



Σχήμα 6.1: Διάγραμμα υλοποίησης

1. *Ανίχνευση συζητήσεων*: Σε αυτή τη διαδικασία (*TopicClustering*) πραγματοποιείται μία επαυξητική (bottom-up) συσταδοποίηση, παρόμοια της μεθόδου που αναπτύχθηκε στην εργασία [4], των μηνυμάτων όπου το γεωγραφικό τους στίγμα αντιστοιχεί στο ίδιο κελί g του καννάβου G, με βάση τις ετικέτες (hashtags) τους. Με αυτό τον τρόπο εντοπίζουμε μηνύματα των οποίων οι ετικέτες έχουν ομοιότητα μεταξύ τους. Η διαμέριση του χωρου σε ομοιόμορφα κελιά μάς επιτρέπει η συσταδοποίηση να γίνεται για μηνύματα που βρίσκονται σε κοντινή απόσταση μεταξύ τους εξασφαλίζοντας την χωρική εγγύτητα. Αυτό συμβαίνει διότι η συσταδοποίηση γίνεται για μηνυμάτα που ανήκουν στο ίδιο κελί του καννάβου, επομένως μεταβάλλοντας το πλήθος των κελιών ως παράμετρο στη μεθοδολογία μεταβάλλεται και η χωρική απόσταση που απαιτείται να έχουν τα μηνύματα μεταξύ τους.

   Ο βαθμός ομοιότητας των μηνυμάτων καθορίζεται απο μια *συνάρτηση ομοιότητας (similarity metric of function)* και ένα *κατώφλι* θ που ορίζέται ως παραμέτρο στην είσοδο. Στην προκειμένη περίπτωση έχουμε να αντιμετωπίσουμε κειμενικά δεδομένα (τις ετικέτες των μηνυμάτων). Στην παρούσα εργασία χρησιμοποιήθηκε η συνάρτηση ομοιότητας *Jaccard* η οποία ορίζεται ώς εξής: Για δύο σύνολα (συστάδες A και B) η απόσταση του A από το B είναι:

$$Jaccard(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

   Η συνάρτηση ομοιότητας *Jaccard* εκφράζει το ποσοστό ομοιότητας δύο συνόλων A, B. Στην εργασία αυτή, το σύνολο A είναι οι ετικέτες μιας προϋπάρχουσας συζήτησης, που εντοπίστηκε σε προηγούμενη εκτέλεση του αλγορίθμου, και το σύνολο B

είναι οι ετικέτες ενός νεοεισερχόμενου μηνύματος. Δηλαδή για μία προϋπάρχουσα συζήτηση T_k και ένα νεοεισερχόμενο μήνυμα s, καλείται η συνάρτηση $Jaccard$ ως εξής: $Jaccard(T_k, s.H)$. Για κειμενικά δεδομένα, παρατίθενται στην βιβλιογραφία και άλλες διαθέσιμες συναρτήσεις ομοιότητας [8] (λ.χ cosine, DamerauLevenshteinDistance κ.α). Οι συστάδες των μηνυμάτων (συζητήσεις) ουσιαστικά αποτελούν δομές σύνοψης των δεδομένων. Με αυτόν τον τρόπο μειώνουμε δραστικά των όγκο της πληροφορίας που έχουμε να επεξεργαστούμε, κάνοντας τον αλγόριθμο πιο αποδοτικό, διότι στα επόμενα στάδια επεξεργασίας διαχειριζόμαστε συστάδες που προφανώς είναι πολύ λιγότερες σε πλήθος απο ότι το πλήθος των δημοσιευμένων μηνυμάτων.

2. *Ανίχνευση δημοφιλών συζητήσεων*: Έχοντας εντοπίσει τις συζητήσεις T_i σε κάθε κελί g του ομοιόμορφου καννάβου G, είμαστε σε θέση να εντοπίσουμε δημοφιλείς συζητήσεις σε όλη την περιοχή παρακολούθησης, δηλαδή όσες εμπλέκουν σημαντικό πλήθος πρόσφατων μηνυμάτων. Θέτοντας μια έλαχιστη τιμή δημοφιλίας ϕ, της οποίας τον ορισμό και την φυσική του σημασία θα την εξηγήσουμε στην συνέχεια του κεφαλαίου, μπορούμε να φιλτράρουμε τις συζητήσεις που εντοπίστηκαν στο εκάστοτε κελί g του καννάβου G και να εξάγουμε τις δημοφιλέστερες κατα τον τρέχοντα κύκλο εκτέλεσης.

3. *Εντοπισμός περιοχών κάλυψης δημοφιλών συζητήσεων και μελέτη εξέλιξης τους*: Κάθε δημοφιλής συζήτηση που ανιχνεύεται, εισάγεται σε ένα ευρετήριο $\mathcal{P}$ στο οποίο τηρούνται οι συζητήσεις T_i που έχουν βρεθεί καθώς και το πλήθος των κελιών του καννάβου Q στα οποία εντοπίστηκαν. Για την ανίχνευση περιοχών κάλυψης, ουσιαστικά εντοπίζουμε υποσύνολα γειτονικών κελιών που σχηματίζουν μια περιοχή κάλυψης A_j^k για τη συζήτηση T_k. Για την μελέτη της εξέλιξης τους με την πάροδο του χρόνου, ουσιαστικά συγκρίνεται η χωρική κάλυψη της εκάστοτε δημοφιλούς συζήτησης με την κάλυψη που είχε στην ακριβώς προηγούμενη μετακύληση του παραθύρου, υπολογίζοντας το πρόσημο της διαφοράς των δύο ξένων συνόλων $|D^+| - |D^-|$ και εξάγοντας κάποια ποιοτικά και ποσοτικά στοιχεία. Επιπλέον, σε κάθε κύκλο εκτέλεσης, η έξοδος περιλαμβάνει τις δημοφιλείς συζητήσεις με τις περιοχές κελιών του καννάβου που εντοπίστηκαν Q, μαζί με κάποια στατιστικά στοιχεία για την εκτίμηση της εντάσης τους (απήχησης).

6.3 Τήρηση ρεύματος μηνυμάτων

Στην επεξεργασία των δεδομένων σημαντικό ρόλο διαδραματίζουν οι τεχνικές μείωσης του κόστους επεξεργασίας ώστε ο αλγόριθμος να είναι όσο το δυνατόν αποδοτικός, τόσο από άποψη χρόνου εκτέλεσης όσο και από άποψη ποιότητας αποτελεσμάτων. Στο πρώτο στάδιο ακολουθείται η τεχνική της ομοιόμορφης κατάτμησης του χώρου σε κάνναβο, έτσι ώστε να δεικτοδοτηθούν τα μηνύματα των χρηστών εντός των αντίστοιχων κελιών. Αμέσως μετά ακολουθείται η διαδικασία συσταδοποίησης των μηνυμάτων, στην συνέχεια η εύρεση της γεωγραφικής τους έκτασης και τέλος εξάγουμε τα αποτελέσματα της μεθόδου για κάθε κύκλο εκτέλεσης.

6.3.1 Δομές δεδομένων

Στο σημείο αυτό γίνεται αναφορά των κύριων δομών δεδομένων που χρησιμοποιούνται στον αλγόριθμο:

- *TopicsInfo*: Για κάθε κελί του καννάβου και για κάθε πλαίσιο $(t-\beta, t]$ τηρείται η λίστα με τις συζητήσεις T_i^t που εντοπίστηκαν σε αυτό το πλαίσιο : $\{T_1^t, T_2^t, \ldots\}$, όπου $t = \tau - \omega + \beta, \ldots, \tau - \beta, \tau$. Οι συζητήσεις T_i^t που ανιχνεύονται στο εκάστοτε πλαίσιο δεν ταυτίζονται κατ' ανάγκη.

- T_τ^g: Το σύνολο των συστάδων που σχηματίστηκαν στο τρέχον χρονικό πλαίσιο $(\tau-\beta, \tau]$, στο κελί g του καννάβου.

- $\mathcal{T}^g$: Ευρετήριο σε κάθε κελί g του καννάβου όπου τηρούνται οι συγχωνευμένες συζητήσεις (συστάδες) με την πληθυκότητα τους σε λεξικογραφική διάταξη.

- $\mathcal{P}$: Ευρετήριο όπου τηρούνται πλειάδες που περιέχουν τις δημοφιλείς συζητήσεις που εντοπίζονται και το πλήθος των κελιών του καννάβου στα οποία παρουσιάζουν δημοφιλία. Το ευρετήριο τηρεί τις πλειάδες σε λεξικογραφική σειρά και είναι της μορφής: $\mathcal{P} = \cup_i \{T_i, \{g_j \in G : T_i \text{ is popular in } g_j\}\}$.

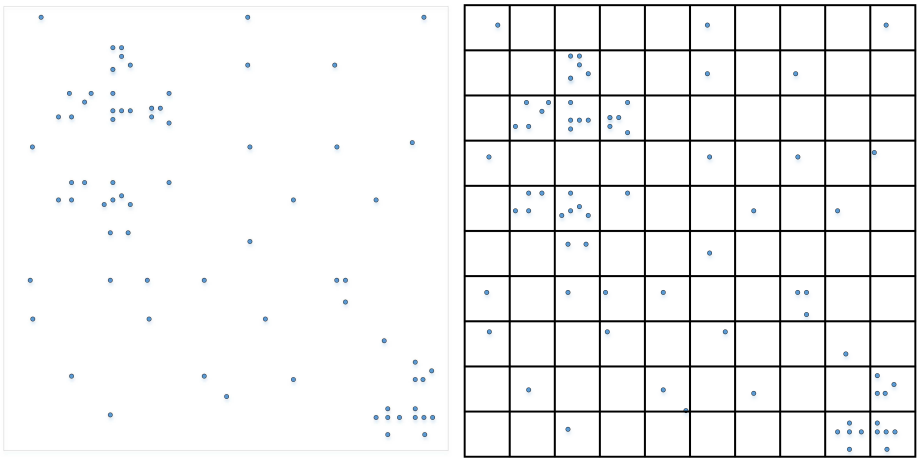
- Q: Λίστα απο κελιά που μια συγκεκριμένη συζήτηση είναι δημοφιλής.

- $\mathcal{A}$: Λίστα που περιέχει σύνολα κελιών τα οποία αντιπροσωπεύουν περιοχές κάλυψης A^i της δημοφιλής συζητής T_i: $\langle T_i, \{A_1^i,\ A_2^i,\ \ldots\}\rangle$

6.3.2 Ευρετήριο Καννάβου

Η τεχνική κατάτμησης του χώρου σε κάνναβο χρησιμοποιείται ευρέως για την επεξεργασία χωροχρονικών δεδομένων [25]. Χρήση του καννάβου είδαμε και σε εργασίες σχετικά με ρεύματα δεδομένων απο κοινωνικά δίκτυα [3, 12, 26]. Κάθε κελί του καννάβου τηρεί τις συζητήσεις (συστάδες) που έχουν σχηματιστεί για κάθε χρονόσημο εντός του χρονικού παραθύρου. Το γεωγραφικό στίγμα των μηνυμάτων αντιστοιχίζεται σε ένα μοναδικό κελί g του κάνναβου G και παύει πλέον να διαδραματίζει κάποιο ρόλο στην επεξεργασία των δεδομένων. Όπως είναι φυσικό, κάθε δημοσίευση δεν είναι δυνατόν να ανήκει σε πολλαπλά κελιά. Η ανανέωση των περιεχομένων του καννάβου γίνεται περιοδικά ανά κύκλο εκτέλεσης. Η τεχνική αυτή μας επιτρέπει τον σχηματισμό συστάδων (συζητήσεων) λαμβάνοντας υπόψη την κειμενική πληροφορία αλλά και την χωρική.

Αν δεν υπήρχε ο κάνναβος, τότε θα έπρεπε να πραγματοποιηθεί μία διαδικασία συσταδοποίησης χρησιμοποιώντας μια συνάρτηση ομοιότητας που θα λάμβανε υπόψη δύο διαστάσεις, την κειμενική και την χωρική. Σαφώς λοιπόν ο κάνναβος απλουστεύει κατα πολύ την διαδικασία της συσταδοποίησης. Το κόστος δημιουργίας του καννάβου είναι αμελητέο σε σύγκριση με το κόστος της επεξεργασίας των δεδομένων χωρίς αυτόν. Η περιοχή παρακολούθησης, όπως φαίνεται στο Σχήμα 7.1, όπου τα μηνύματα των χρηστών αναπαρίστανται ώς σημεία στο χώρο και η τοποθέτηση τους γίνεται σύμφωνα με το γεωγραφικό τους στίγμα, τεμαχίζεται



(α′) Μηνύματα (β′) Κάνναβος

Σχήμα 6.2: Περιοχή μελέτης

σε ομοιόμορφα κελιά g σταθερού μεγέθους (βλ. Σχήμα 6.2). Ο τεμαχισμός του χώρου θα μπορούσε να γίνει και με μη ομοιόμορφα κελιά. Σε αυτή την περίπτωση, δεν θα υπήρχε δυνατότητα εκτίμησης της χωρικής κάλυψης μιας συζήτησης ούτε να ανιχνεύσουμε κατά πόσο μια συζήτηση εξαπλώθηκε ή συρρικνώθηκε αφού τα κελιά δεν θα είχαν σταθερό μέγεθος.

Έτσι για τον σκοπό της εργασίας μας θεωρήσαμε ότι ο κάνναβος περιέχει τετραγωνικά κελιά (λ.χ $1km \times 1km$) και είναι τετραγωνικών διαστάσεων (λ.χ $50km \times 50km$). Οι διαστάσεις αυτές καθώς και ο αριθμός των κελιών που το χωρίζουν δίνονται ως παράμετροι στην εισόδο του αλγορίθμου. Ο αριθμός τους επηρεάζει τις επιδόσεις και το κόστος επεξεργασίας του αλγορίθμου, όπως θα φανεί αργότερα στην πειραματική μελέτη.

6.4 Ανίχνευση συζητήσεων

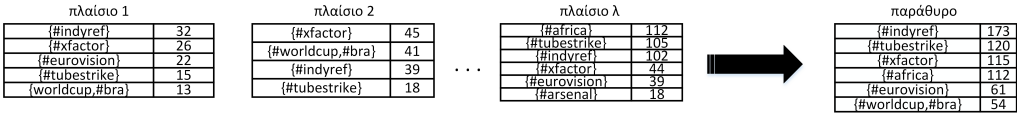
Σε κάθε μετακύλιση του χρονικού παραθύρου (βλ. σχήμα 5.1), ξεκινά ένας νέος κύκλος εκτέλεσης και εκτελείται ο αλγόριθμος 2 ο οποίος τοποθετεί τις δημοσιεύσεις στα κελιά του καννάβου. Στην συνέχεια για κάθε νεοεισερχόμενο μήνυμα ανανεώνονται οι συστάδες των δημοσιεύσεων που τηρούνται στο αντίστοιχο κελί που δεικτοδοτείται το μήνυμα. Πιο συγκεκριμένα, για την επεξεργασία των δεδομένων, χρησιμοποιείται ένας ομοιόμορφος χωρικός κάνναβος που τοποθετείται πάνω απο την περιοχή παρακολούθησης. Η τοποθέτηση των tweets σε κελιά πραγματοποιείται με hashing σε πραγματικό χρόνο. Επομένως, το κόστος αντιστοίχησης για κάθε μήνυμα (tweet) είναι σταθερό $O(1)$.

Στην συνέχεια ακολουθεί η διαδικασία της συσταδοποίησης για κάθε νεοεισερχόμενο μήνυμα s από το ρεύμα δημοσιεύμενων μηνυμάτων S (Twitter Stream). Η διαδικασία της συσταδοποίησης αποτελεί μια παραλλαγή της διαδικασίας που πραγματοποιείται στην εργασία [5] και περιγράφεται απο τον αλγόριθμο 2. Επεξηγηματικά, ο αλγόριθμος 2 δέχεται ώς είσοδο (γρ. 1) το νεοεισερχόμενο μήνυμα s, το χρονόσημο τ, που αντιστοιχεί στο πέρας του χρονικού πλαισίου $(\tau - \omega,\ \tau]$, ένα κατώφλι ομοιότητας θ και το βήμα κύλισης β. Η διαδικασία αυτή δεν πραγματοποιείται για τις συστάδες που εντοπίστηκαν σε όλο το εύρος

Algorithm 2 Refresh topics for each incoming tweet

1: **Procedure** TopicClustering (tweet s, timestamp τ, threshold θ, window slide β)
2: $g \leftarrow$ HashLocation($s.loc$); *//Grid cell where this tweet appears*
3: $\mathcal{T}_\tau^g \leftarrow G[g].TopicsInfo((\tau-\beta,\tau])$; *//Topics at latest pane*
4: **if** $\mathcal{T}_\tau^g = \emptyset$ **then**
5: $T_{new} \leftarrow \{s.H\}$; *//From hashtags of tweet s, create...*
6: $\mathcal{T}_\tau^g \leftarrow \langle \tau, T_{new}, 1\rangle$; *//... a new topic with cardinality 1*
7: **else**
8: $n \leftarrow |\mathcal{T}_\tau^g|$; *//Number of topics in latest pane at cell g*
9: **for** $i \leftarrow 1$ **to** n **do**
10: $\sigma[i] \leftarrow$ similarity($\mathcal{T}_\tau^g[i], s$); *//Text similarity with i^{th} topic*
11: **end for**
12: $i_{max} \leftarrow \text{argmax}_{i\in\{1\ldots n\}}\sigma$; *//Pointer to most similar topic*
13: **if** $\sigma[i_{max}] \geq \theta$ **then** { *//Update most similar topic*}
14: $\mathcal{T}_\tau^g[i_{max}].topic \leftarrow \mathcal{T}_\tau^g[i_{max}].topic \cup \{s.H\}$;
15: $\mathcal{T}_\tau^g[i_{max}].count \leftarrow \mathcal{T}_\tau^g[i_{max}].count + 1$;
16: **else** { *//Create a new topic only from this tweet*}
17: $T_{new} \leftarrow \{s.H\}$;
18: $\mathcal{T}_\tau^g \leftarrow \mathcal{T}_\tau^g \cup \langle \tau, T_{new}, 1\rangle$;
19: **end if**
20: **end if**
21: $G[g].TopicsInfo((\tau-\beta,\tau]) \leftarrow \mathcal{T}_\tau^g$; *//Retain topics in cell*
22: **End Procedure**

του παραθύρου αλλά μόνο για αυτές που σχηματίστηκαν στο πιο πρόσφατο χρονικό πλαίσιο $(\tau_{now}-\beta,\ \tau_{now}]$ του παραθύρου, στο κελί g που αντιστοιχεί το γεωγραφικό στίγμα $s.loc$ του νεοεισερχόμενου μήνυματος. Η λίστα των συζητήσεων (συστάδων) που τηρούνται σε κάθε κελί είναι ταξινομημένες με βάση το πλαίσιο στο οποίο εντοπίστηκαν (pane), από το πιο παλιό στο πιο πρόσφατο. Στην συνέχεια συγκρίνουμε το νεοεισερχόμενο μηνυμα s με τις προϋπάρχουσες συστάδες που έχουν σχηματιστεί στο τρέχον χρονικό πλαίσιο και τηρούνται στη δομή $TopicsInfo((\tau-\beta,\tau])$ (γρ. 3). Παραδείγματος χάρη, για έυρος κυλιόμενου παραθύρου ω και βήμα κύλισης β, απο προηγούμενες εκτελέσεις έχουν εντοπιστεί οι εξής συζητήσεις που τηρούνται στη δομή: $TopicsInfo = \langle\ \{\{\#indyref, \#eurovision\}, (\tau_{now}-\omega,\ \tau_{now}-\omega+\beta]\}, \{\{\#eurovision\}, (\tau_{now}-\omega+\beta,\ \tau_{now}-\omega+2\beta]\}\ldots\{\{\#elections\}, (\tau_{now}-\beta,\ \tau_{now}]\}\ \rangle$. Στην περίπτωση που δεν υπάρχουν σχηματισμένες συστάδες στο τρέχον χρονικό πλαίσιο, τότε το μήνυμα s σχηματίζει από μόνο του μία συστάδα (γρ. 4-6). Η σύγκριση του μηνύματος s με τις προϋπάρχουσες συστάδες γίνεται με βάση μια συνάρτηση ομοιότητας $similarity$. Στην εργασία αυτή χρησιμοποιήθηκε η συνάρτηση Jaccard. Συγκρίνεται κάθε μία προϋπάρχουσα συστάδα $T_\tau^g[i]$,που ανήκει στο πλαίσιο $(\tau-\beta,\tau]$, και έχει εντοπιστεί στο κελί g, με το νεοεισερχόμενο μήνυμα s. Έπειτα, επιλέγεται η συζήτηση που ξεπερνά το κατώφλι ομοιότητας θ και παρουσιάζει τον μεγαλύτερο βαθμό ομοιότητας με το μήνυμα s (γρ. 8-15), σε σύγκριση με τις υπόλοιπες συζητήσεις. Στην συνέχεια, ανανεώνεται η ταυτότητα της συζήτησης (ετικέτες) και η πληθυκότητα της (γρ. 14-15). Στην εργασία μας, δεν εντοπίζουμε ομοιότητα μεταξύ δύο διαφορετικών ετικετών αλλά δύο διαφορετικών συνόλων (λ.χ οι ετικέτες $\{\#eurovision, \#eurovisionsongcontest\}$ ενδέχεται να ανήκουν σε διαφορετικό θέμα



Σχήμα 6.3: Συγχώνευση συζητήσεων βάσει των πλαισίων ενός παραθύρου

συζήτησης). Στην περίπτωση που το νεοεισερχόμενο μήνυμα s δεν ανήκει σε καμία συζήτηση, τότε και πάλι σχηματίζει μία συστάδα με το μεμονωμένο αυτό μήνυμα. (γρ. 17-18).

Παρατηρούμε ότι στα χρονικά πλαίσια $(\tau_{now}-\omega,\ \tau_{now}-\omega+\beta]$ και $(\tau_{now}-\omega+\beta,\ \tau_{now}-\omega+2\beta]$ περιέχεται η ίδια συζήτηση. Αυτό συμβάινει διότι η συζήτηση με την ετικέτα $\#eurovision$ εντοπίστηκε σε δύο διαφορετικά χρονικά πλαίσια. Επομένως το μόνο που απαιτείται προκειμένου να βρεθεί το συνολικό πλήθος των συζητήσεων εντός του παραθύρου είναι να συγχωνευτούν οι συζητήσεις που εντοπίστηκαν εντός ενός κελιού g και εκτυλίχθησαν εντός των χρονικών πλαισίων (panes) του τρέχοντος παραθύρου. Η λειτουργία της συγχώνευσης θα περιγράφεί στην συνέχεια (ενότητα 6.5).

Ειδικότερα, πραγματοποιούμε μια επαυξητική συσταδοποίηση για κειμενικά δεδομένα. Οι υπάρχουσες συστάδες δημιουργήθηκαν εντός του χρονικού ορίζοντα του παραθύρου με τον ίδιο ακριβώς τρόπο. Η συσταδοποίηση γίνεται με βάση την κειμενική πληροφορία που στην παρούσα εργασία είναι τα σύνολα των ετικετών της εκάστοτε δημοσίευσης. Οι συστάδες διαδραματίζουν τον ρόλο των συζητήσεων που εκτυλίσσονται στα κοινωνικά δίκτυα και ερευνάται η τοπικότητα μέσω της χρήσης του κάνναβου.

6.5 Ανίχνευση δημοφιλών συζητήσεων

Στο επόμενο στάδιο της μεθόδου, εντοπίζονται δημοφιλείς συζητήσεις, το πλήθος των μηνυμάτων που τις απαρτίζουν και οι περιοχές κάλυψης (κελιά καννάβου) στις οποίες εμφανίζονται. Αυτή η διαδικασία πραγματοποιείται φιλτραροντας τις συζητήσεις που έχουν εντοπιστεί στο εκτάστοτε κελί g του καννάβου, για την εκάστοτε μετακύλιση του παραθύρου, με μία ελάχιστη τιμή δημοφιλίας σύμφωνα με τον ορισμό της δημοφιλούς συζήτησης που δώσαμε στο κεφάλαιο 5.

Η ελάχιστη τιμή δημοφιλίας ϕ που ορίζεται αναφέρεται σε κάθε κελί του καννάβου ξεχωριστά. Έστω $S_\omega = \{s_i, \ldots, s_{i+N_\omega}\}$ το τμήμα του ρεύματος μηνυμάτων που επεξεργαζόμαστε το τρέχον παράθυρο. Το πλήθος των μηνυμάτων του τρέχοντος παραθύρου είναι $|S_\omega| = N_\omega$. Προφανώς ο αριθμός N_ω κυμαίνεται με την πάροδο του χρόνου, αφού δεν είναι δυνατό κάθε χρονική περιόδο να έχουμε τον ίδιο αριθμό απο μηνύματα. Επομένως, όταν θέλουμε να ανιχνεύσουμε δημοφιλείς συζητήσεις, σε ένα συγκεκριμένο κελί g του καννάβου, ελέγχουμε αν το πλήθος των μηνυμάτων N_g στο κελί g ικανοποιεί τον περιορισμό $\geqslant \phi \times N_\omega$.

Η διαδικασία εξαγωγής των δημοφιλών συζητήσεων απαιτεί την ανίχνευση των συζητήσεων εντός του χρονικού παραθύρου ω. Επομένως, απαιτείται η συγχώνευση των επιμέρους χρονικών πλαισίων $\frac{\omega}{\beta}$ που εντοπίστηκαν στην προηγούμενη διαδικασία και συνιστούν το εύρος του παραθύρου για την εύρεση του πλήθους των συζητήσεων (βλ. Σχήμα 6.3). Για την ακρίβεια, σε κάθε μετακύλιση διαγράφεται το απώτατο πλαίσιο της προηγούμενης μετακύλισης

και προστίθεται ένα καινούργιο (το πιο πρόσφατο διάστημα με τίς νέες συζητήσεις).

Για την εύρεση δημοφιλών συζητήσεων, ο αλγόριθμος 3 αρχικά καλεί την διαδικασία $dropExpiredTopics(g, \tau - -\omega)$ (γρ. 4) η οποία διαγράφει από το εκάστοτε κελί g τις συζητήσεις (συστάδες) που σχηματίστηκαν στο χρονικό πλαίσιο $(\tau - \omega - \beta,\ \tau - \omega]$, το οποίο εκπίπτει από το τρέχον παράθυρο. Έπειτα ακολουθεί η διαδικασία συγχώνευσης των συστάδων που σχηματίστηκαν σε διαφορετικά χρονικά πλαίσια και αφορούν το ίδιο θέμα συζήτησης για κάθε κελί g του καννάβου. Με αυτόν τον τρόπο βρίσκουμε την πληθυκότητα των συστάδων εντός του εύρους του παραθύρου σε κάθε κελί g του καννάβου. Η διαδικασία αυτή εκτελείται στην $MergeTopics$ του αλγορίθμου 3. Η συγχώνευση γίνεται εξάρχής σε κάθε μετακύλιση του παραθύρου. Κατά την διάρκεια εκτέλεσης της διαδικασίας, οι συγχωνευμένες συστάδες τοποθετούνται σε ένα ευρετήριο $\mathcal{T}^g$ (γρ. 6), στο οποίο διατηρούνται σε λεξικογραφική διάταξη σε κάθε κελί του καννάβου, διευκολύνοντας έτσι την διαδικασία της συγχώνευσης. Για να είναι δυνατή η λεξικογραφική διάταξη των συζητήσεων (συστάδων) θα πρέπει οι ετικέτες τους να είναι ταξινομημένες επίσης σε λεξικογραφική σειρά. Για παράδειγμα, η λεξικογραφική διάταξη της συστάδας με ετικέτες $\{\#uk, \#indyref, \#scotland\}$ είναι η $\{\#indyref, \#scotland, \#uk\}$. Επομένως για κάθε συστάδα υπο εξέταση, δεν χρείαζεται να ψάχνουμε όλες τις υπόλοιπες παρα μόνο τις συστάδες που η πρώτη ετικέτα απο το σύνολο τους ξεκινά απο το ίδιο γράμμα. Η παραπάνω δομή και λειτουργία του ευρετηρίου δεν είναι τίποτε άλλο απο ένα κοινό λεξικό. Για παράδειγμα, αν ψάχνουμε την λέξη $indyref$ δεν χρειάζεται να ψάχνουμε ένα ένα τα λήμματα του λεξικού αλλά μπορούμε να περιορίσουμε την αναζήτηση στο γράμμα i. Κατά αυτόν τον τρόπο βελτιστοποιούμε κατα πολύ την απόδοση της μεθόδου μας, αφού η εύρεση του πρώτου γράμματος γίνεται σε σταθερό χρόνο και η μετέπειτα αναζήτηση της κατάλληλης συστάδας γίνεται σε γραμμικό χρόνο.

Στην συνέχεια, φιλτράρονται οι συζητήσεις του εκάστοτε κελιού g, εντοπίζοντας τις συζητήσεις που η πληθυκότητα τους ξεπερνούν την ελάχιστη τιμή $\phi \times N_\omega$ (γρ. 8). Οι δημοφιλείς συζητήσεις που εντοπίζονται σε κάθε κελί τοποθετούνται σε ένα ευρετήριο $\mathcal{P}$ (γρ. 9) με λεξικογραφική σειρά, όπου τηρούνται όλες οι δημοφιλείς συζητήσεις και το σύνολο των κελιών που ανιχνεύθηκαν σε όλη την έκταση του καννάβου. Τέλος η συνάρτηση αυτή επιστρέφει το ευρετήριο $\mathcal{P}$.

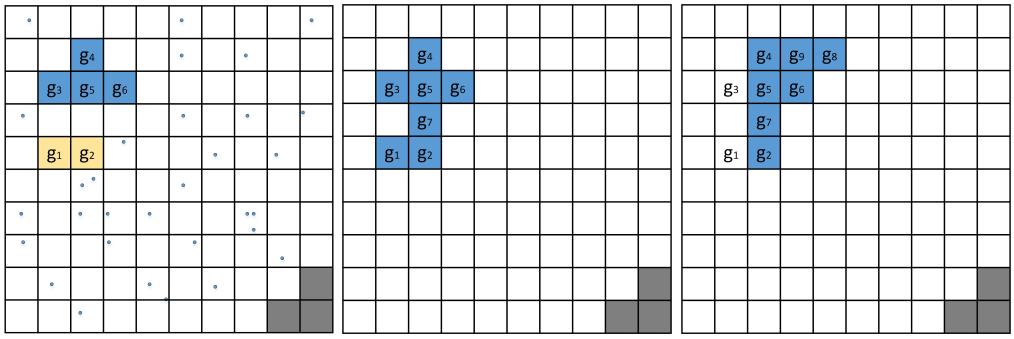
6.6 Εύρεση χωρικής κάλυψης και μελέτη εξάπλωσής τους

Στο προηγούμενο κεφάλαιο είχαμε δώσει έναν αφηρημένο ορισμό της έννοιας της χωρικής κάλυψης των δημοφιλών συζητήσεων. Συνοπτικά είχαμε ορίσει ότι χωρική κάλυψη είναι το σύνολο των γειτονικών κελιών όπου η πληθικότητα μιας συζήτησης ξεπερνά την ελάχιστη τιμή δημοφιλίας ϕ. Η μεθοδολογία που αναπτύχθηκε χρησιμοποιεί τον κάνναβο για την διακριτοποίηση του χώρου σε ομοιόμορφα κελιά g. Ο αλγόριθμος αυτός δεν περιορίζεται στον εντοπισμό των κελιών στις οποίες εμφανίζεται μια συγκεκριμένη συζήτηση, αλλά εντοπίζει και τις περιοχές κάλυψης, οι οποίες αποτελούνται απο ομάδες γειτονικών κελιών χωρίς χάσματα μεταξύ τους.

Τέτοιες περιοχές φαίνονται στο Σχήμα 6.4α΄. Παρατηρούνται 3 περιοχές κάλυψης (ομάδες

Algorithm 3 Identify popular topics

1: **Function** PopularityFiltering (timestamp τ, MinPopularity ϕ, window range ω, tweet count in current window N_ω)
2: **Output**: $\mathcal{P} = \cup_i \{T_i, \{g_j \in G : T_i \text{ is popular in } g_j\}\}$
3: $\mathcal{P} \leftarrow \emptyset$; *//Set to hold all popular topics in this execution cycle*
4: **for** each cell $g \in G$ **do**
5: dropExpiredTopics$(g, \tau{-}\omega)$; *//Topics in evicted window pane*
6: $\mathcal{T}^g \leftarrow$ MergeTopics(g, ω); *//All topics in window at this cell*
7: **for** each topic $\langle T_i, |T_i| \rangle \in \mathcal{T}^g$ **do**
8: **if** $|T_i| \geqslant \phi \times N_\omega$ **then**
9: $\mathcal{P} \leftarrow \mathcal{P} \cup \{\langle T_i, g \rangle\}$; *//$T_i$ is enough popular in this cell*
10: **end if**
11: **end for**
12: **end for**
13: **return** $\mathcal{P}$; *//For each topic, it lists the cells where it is popular*
14: **End Function**



(α′) Χρονική στιγμή τ (β′) Χρονική στιγμή $\tau + \beta$ (γ′) Χρονική στιγμή $\tau + 2\beta$

Σχήμα 6.4: Περιοχές κάλυψης για διαδοχικές μετακυλίσεις του παραθύρου

κελιών με διαφορετικό χρώμα). Η μπλέ περιοχή περιλαμβάνει μια ομάδα τεσσάρων κελιών, η γκρί περιοχή περιλαμβάνει τρία κελιά και τέλος η κίτρινη περιοχή τρία κελιά. Οι περιοχές εντοπίζονται και μεταβάλλονται δυναμικά κατά την εκτέλεση του αλγορίθμου. Υπάρχει το ενδεχόμενο την χρονική στιγμή τ να διακρίνουμε τρείς περιοχές όπως στο Σχήμα 6.4α′ και την χρονική στιγμή $\tau + \beta$, στην αμέσως επόμενη μετακύλιση του παραθύρου, να διακρίνουμε δύο περιοχές, η οποία προκύπτει από την συνένωση της μπλέ και της κίτρινης περιοχής. Αυτή η περίπτωση φαίνεται καλύτερα στο σχήμα 6.4β′.

Στον αλγόριθμο 3 εντοπίζονται δημοφιλείς συζητήσεις στο ομοιόμορφο πλέγμα, αποθηκεύονται προσωρινά σε ένα ευρετήριο(με την μορφή του λεξιλογίου όπως περιγράψαμε και σε προηγούμενη ενότητα) $\mathcal{P}$ και στην συνέχεια ανιχνεύεται το σύνολο των κελιών Q της συζήτησης αυτής. Για τον εντοπισμό των περιοχών κάλυψης A ακολουθούμε τον αλγόριθμο 4.

Η συνάρτηση *Coverage Discovery* δέχεται ως είσοδο το ευρετήριο $\mathcal{P}$ όπου τηρούνται τα δημοφιλή θέματα συζήτησης μαζί με το πλήθος των κελιών Q που εκτείνεται η εκάστοτε συζήτηση (γρ. 1). Για κάθε δημοφιλή συζήτηση T_i (γρ. 3-27) εντοπίζεται η χωρική κάλυψη

Algorithm 4 Find coverage areas in the grid for all popular topics

```
1: Function CoverageDiscovery (popular topics 𝒫)
2: Output: Coverage 𝒜 = {T_i, {A_1, A_2, ...}}, ∀ T_i ∈ 𝒫
3: for each topic T_i ∈ 𝒫 do
4:    𝒜[T_i] ← ∅;   //Initialize list of coverage areas for topic T_i
5:    Q ← 𝒫[T_i];  //Set of cells where T_i has been found popular
6:    dA ← ∅;      //Set of cells to signify changes in coverage
7:    for each g_k ∈ Q do
8:       merged ← false;  //Flag an area that needs consolidation
9:       if 𝒜[T_i] ≠ ∅ then
10:         for each A_j ∈ 𝒜[T_i] do
11:            if adjacent(g_k, A_j) then
12:               A_j ← A_j ∪ {g_k};  //Cell g_k is adjacent with...
13:               dA ← dA ∪ A_j;   //...at least one cell of area A_j
14:               𝒜[T_i] ← 𝒜[T_i] \ A_j;
15:               merged ← true;
16:            end if
17:         end for
18:      end if
19:      if merged = true then
20:         A_j ← ⋃ dA;  //A compact area with all adjacent cells
21:         𝒜[T_i] ← 𝒜[T_i] ∪ A_j;
22:      else
23:         A_j ← {g_k};   //Coverage area consists of cell g_k only
24:         𝒜[T_i] ← A_j;  //...and this is where T_i is popular
25:      end if
26:   end for
27: end for
28: return 𝒜;
29: End Function
```

(δλ. περιοχές στον κάνναβο απο γειτονικά μεταξύ τους κελιά). Οι περιοχές κάλυψης εντοπίζονται με μόνο με ένα πέρασμα ελέγχωντας πάντα αν το εξεταζόμενο κελί είναι γειτονικό με ένα τουλάχιστον κελί απο την υπάρχουσα περιοχή που έχει σχηματιστεί. Αρχικά για κάθε κελί που εντοπίζεται θεωρείται σαν μια ξεχωριστεί περιοχή εξάπλωσης (γρ. 23-25). Έφοσον εντοπίζεται κάποιο σύνολο που περιέχει ένα τουλάχιστον γειτονικό κελί (γρ. 11) με το εξεταζόμενο, είναι περιττό να ελέγχθουν επιπλέον κελιά της ίδιας περιοχής. Παραδείγματος χάρη, σε μια αυθαίρετη κατάσταση εκτέλεσης έχουμε εντοπίσει τις εξής περιοχές κάλυψης για την δημοφιλή συζήτηση $\{\#indyref\}$ όπως φαίνεται στο Σχήμα 6.4α΄: $\{\{g_1, g_2\}, \{g_3, g_4, g_5, g_6\}\}$ και στην επόμενη μετακύλιση του παραθύρου εντοπίζουμε το κελί g_7. Το εξεταζόμενο κελί g_7 είναι γειτονικό με τα κελιά g_3, g_5, g_6. Επομένως οι δύο περιοχές που εντοπίστηκαν θα πρέπει να συνενωθούν και θα σχηματιστεί μία ενιαία περιοχή κάλυψης $\{g_1, g_2, g_3, g_4, g_5, g_6, g_7\}$ (βλ. Σχήμα 6.4β΄). Έτσι δεν χρειάζεται να εξετάστούν τα κελιά απο την αρχή αλλά αρκεί να βρεθούν οι περιοχές που ανήκει το κελί g_7 (γρ. 13). Εφόσον βρεθεί παραπάνω απο μία περιοχή τότε αυτές θα πρέπει να συνενωθούν. Το σύνολο των περιοχών που πρέπει να συνενωθούν τηρούνται στο σύνολο dA (γρ. 20). Επομένως η συνένωση των περιοχών γίνεται γραμμικά.

Για την παρακολούθηση της εξάπλωσης, σύμφωνα με τον ορισμό που δόθηκε στο κεφάλαιο 5, αρκεί να συγκρίνουμε τις περιοχές κάλυψης των δημοφιλών συζητήσεων απο διαδοχικές μετακυλίσεις του παραθύρου. Η παραπάνω διαδικασία είναι τετριμμένη, αφού για να βρεθούν τα κελιά στα οποία η συζήτηση ήταν δημοφιλής στο παράθυρο $(\tau - \omega, \tau]$ και στο παράθυρο $(\tau, \tau + \beta]$ παύει να είναι, αρκεί να αναζητήσουμε το συγκεκριμένο κελί που αντιστοιχεί στην συγκεκριμένη συζήτηση. Παραδείγματος χάρη, η δημοφιλής συζήτηση $\{\#indyref\}$ τη χρονική στιγμή $\tau + \beta$ τηρείται στο ευρετήριο $\mathcal{P} = \{\{\#indyref\}, \{g_1, g_2, g_3, g_4, g_5, g_6, g_7\}\}$, ενώ στην αμέσως επόμενη χρονική στιγμή $\tau + 2\beta$ η συζήτηση διαμορφώνεται στο ευρετηριο ως εξής: $\mathcal{P} = \{\{\#indyref\}, \{g_2, g_4, g_5, g_6, g_7, g_8, g_9\}\}$. Τα κελιά στα οποία η συζήτηση δεν είναι πια δημοφιλής με την μετακύλιση είναι $D^- = \{g_1, g_3\}$, ενώ τα κελιά στα οποία επεκτάθηκε η δημοφιλία είναι τα ακόλουθα: $D^+ = \{g_8, g_9\}$. Στην προκειμένη περίπτωση, παρατηρούμε πως $|D^+| - |D^-| = 0$. Επομένως μπορούμε να πούμε πως αυτή η συζήτηση για την χρονική στιγμή . Επομένως, μπορούμε να ισχυριστούμε πως η συγκεκριμένη συζήτηση παρέμεινε στάσιμη. Η πολυπλοκότητα της διαδικασίας αυτής είναι αμελητέα καθώς το πλήθος των κελιών, στα οποία εκτείνεται η εκάστοτε συζήτηση, είναι αμελητέο.

Κεφάλαιο 7

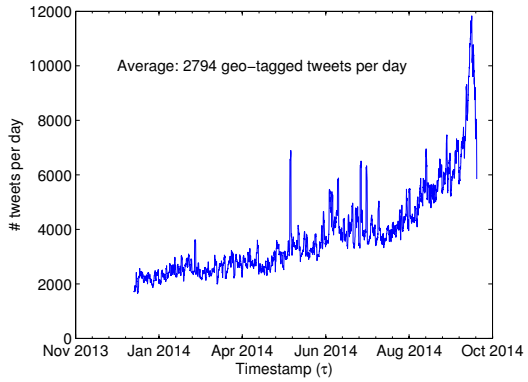
Πειραματική Αξιολόγηση

Στο κεφάλαιο αυτό παρουσιάζεται η πειραματική αξιολόγηση και ο έλεγχος ορθής λειτουργίας του αλγορίθμου. Γίνεται περιγραφή των χαρακτηριστικών των αρχείων εισόδου, των παραμέτρων που χρησιμοποιήθηκαν, καθώς και παράθεση των συγκριτικών πειραμάτων που εκτελέστηκαν, βάσει των οποίων γίνεται αξιολόγηση των επιδόσεων.

7.1 Πειραματικό Πλαίσιο

Τα πειραματικά δεδομένα που έχουμε στη διάθεσή μας αφορούν δημοσιεύσεις από το κοινωνικό δίκτυο Twitter από διάφορες τοποθεσίες του κόσμου και έχουν χρησιμοποιηθεί σε προηγούμενη διερεύνηση [9] σχετικά με την ανίχνευση χαρακτηριστικών τοποθεσιών βάσει της δραστηριότητας χρηστών του Twitter. Για την πειραματική αξιολόγηση του αλγορίθμου ελήφθησαν υπ΄ όψιν δημοσιεύσεις που προέρχονται από την περιοχή του ευρύτερου Λονδίνου, διότι εκεί παρατηρείται ο μεγαλύτερος όγκος και πυκνότητα μηνυμάτων στο συγκεκριμένο αρχείο. Τα μηνύματα που δεν περιέχουν γεωγραφικές συντεταγμένες αγνοούνται εντελώς και δεν συμμετέχουν καθόλου στην επεξεργασία. Το περιεχόμενο του μηνύματος εκκαθαρίζεται από όλες τις λέξεις και μένουν μόνο οι ετικέτες (hashtags). Το σύνολο των μηνυμάτων που προέκυψε είναι λοιπόν της μορφής $\langle \tau, uid, loc, H \rangle$, όπου τ είναι η χρονική στιγμή δημοσίευσης του μηνύματος, uid η ταυτότητα του χρήστη που δημοσίευσε το μήνυμα, loc η γεωγραφική τοποθεσία του χρήστη τη στιγμή της δημοσίευσης σε καρτεσιανές συντεταγμένες και τέλος H το σύνολο των *ετικετών* του μηνύματος.

Τα δεδομένα αποτελούνται από 807.479 εγγραφές και αφορούν την χρονική περίοδο 30-12-2013 00:00:01 έως 14-10-2014 21:47:02. Σύμφωνα με το Σχήμα 7.1, για την περίοδο που περιλαμβάνει το πειραματικό σύνολο των δεδομένων, δημοσιεύονται καθημερινά 2.794 μηνύματα. Είναι προφανές ότι τέτοιος όγκος δεδομένων είναι πολύ μικρός ώστε να μπορεί να χαρακτηριστεί ως ρεύμα, γι΄ αυτόν τον λόγο προχωρήσαμε στην πλασματική επαύξηση του αρχείου. Πιο συγκεκριμένα, κάθε πρωτογενές μήνυμα επαναλήφθηκε 9, 19, 49 και 99 φορές, οπότε προέκυψαν συνολικά 10-, 20-, 50-, 100-πλάσια μηνύματα που αποθηκεύτηκαν σε αντίστοιχα χωριστά αρχεία. Αυτό έγινε προκειμένου να αξιολογήσουμε την αποδοτικότητα του αλγορίθμου για κλιμακούμενους όγκους δεδομένων.



Σχήμα 7.1: Ρυθμός άφιξης μηνυμάτων σε κάθε κύλιση παραθύρου

Έγιναν πειράματα για τις εξής παραμέτρους:

1. Πλήθος κελιών $g \times g$ καννάβου.
2. Εύρος ω (range) χρονικού παραθύρου.
3. Βήμα β (slide) μετακύλισης του χρονικού παραθύρου (ισούται με τη χρονική διάρκεια κάθε πλαισίου).
4. Κατώφλι ομοιότητας θ για την ανίχνευση συζητήσεων, δηλαδή ο ελάχιστος βαθμός ομοιότητας των ετικετών ενός νεοεισερχόμενου μηνύματος με μια προϋπάρχουσα συζήτηση.
5. Ελάχιστη δημοφιλία ϕ μιας συζήτησης εντός μεμονωμένου κελιού. Αυτή εκφράζεται ως ποσοστό (%) επί του συνολικού πλήθους μηνυμάτων που έχουν δημοσιευθεί κατά τη διάρκεια ω ενός χρονικού παραθύρου.

Πιο συγκεκριμένα, οι τιμές που δόθηκαν για κάθε παράμετρο ήταν:

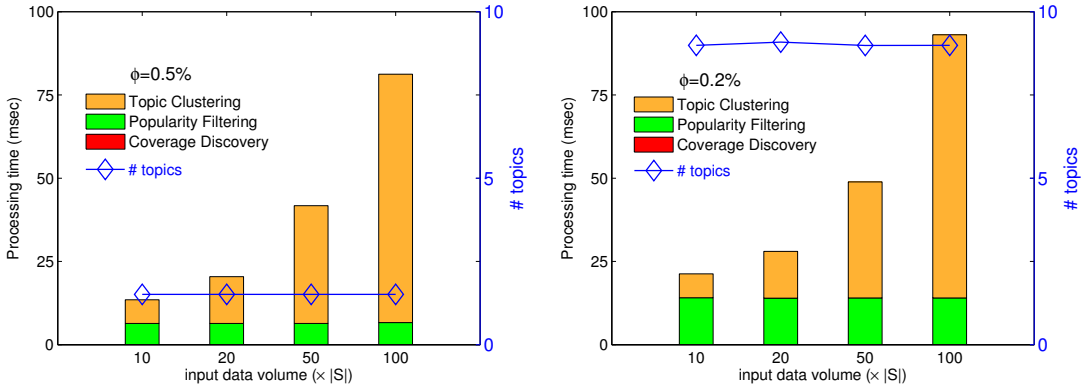
Συντελεστής κλιμάκωσης αρχικού συνόλου δεδομένων $\|S\|$	$\times$10, $\times$20, $\times$50, $\times$**100**
Πλήθος κελιών καννάβου $g \times g$	10 $\times$ 10, 20 $\times$ 20, 25 $\times$ 25, 30 $\times$ 30, **50** $\times$ **50**, 100 $\times$ 100
Εύρος ω χρονικού παραθύρου	2h, 4h, 12h, **24h**
Βήμα β μετακύλισης παραθύρου	**1h**, 2h, 4h, 8h
Κατώφλι ομοιότητας θ	0.3, **0.5**, 0.6, 0.8
Ελάχιστη δημοφιλία ϕ	0.02%, 0.03%, 0.05%, 0.1%, **0.2%**, 0.3%, 0.5%

Πίνακας 7.1: Παράμετροι πειραμάτων

Με έντονους χαρακτήρες(**bold**) δηλώνονται οι τυπικές τιμές των παραμέτρων όπως ισχύουν στα περισσότερα πειράματα.

Κατά την διάρκεια των πειραμάτων μετρήθηκαν οι χρόνοι εκτέλεσης του αλγορίθμου σε *κάθε κύκλο εκτέλεσης χωριστά για κάθε φάση*:

- Ανίχνευση συζητήσεων (*Topic Clustering*),
- Ανίχνευση δημοφιλών συζητήσεων (*Popularity Filtering*) και
- Εντοπισμός χωρικής κάλυψης (*Coverage Discovery*),



(α′) $\phi = 0.5\%$ (β′) $\phi = 0.2\%$

Σχήμα 7.2: Χρόνος εκτέλεσης ανά φάση και πλήθος δημοφιλών συζητήσεων για κλιμακούμενο όγκο δεδομένων

καθώς επίσης και το πλήθος των δημοφιλών συζητήσεων που εντοπίζονται σε όλη την περιοχή μελέτης (Λονδίνο). Στα γραφήματα που ακολουθούν απεικονίζονται μέσοι όροι των μεγεθών αυτών ανά κύκλο εκτέλεσης.

Όλες οι δομές και διαδικασίες του αλγορίθμου υλοποιήθηκαν στη γλώσσα προγραμματισμού C++ και τα πειράματα εκτελέστηκαν σε λειτουργικό σύστημα Ubuntu Linux σε προσωπικό υπολογιστή Intel(R) Core(TM) i7, 2.4 GHz με μνήμη RAM 8 Gb.

7.2 Αξιολόγηση πειραμάτων

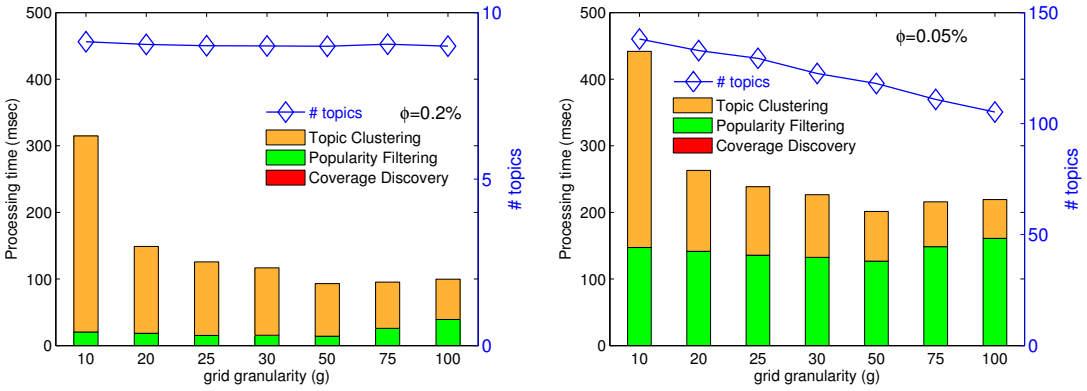
7.2.1 Κλιμακωσιμότητα του αρχικού συνόλου δεδομένων

Προκειμένου να αξιολογήσουμε καλύτερα τις επιδόσεις του αλγορίθμου, χρησιμοποιήσαμε έναν συντελεστή κλιμάκωσης του αρχικού συνόλου δεδομένων. Έτσι παρήχθησαν δεδομένα $\times 10$, $\times 20$, $\times 50$ και $\times 100$ επί του αρχικού συνόλου δεδομένων και αντίστοιχα μετρήθηκαν οι μέσοι χρόνοι των επιμέρους φάσεων του αλγορίθμου για τιμές ελάχιστης δημοφιλίας $\phi = 0.5\%$ και $\phi = 0.2\%$, όπως φαίνεται στο Σχήμα 7.2. Η φάση της ανίχνευσης συζητήσεων επηρεάζεται άμεσα απο τον όγκο των δεδομένων που έχει να επεξεργαστεί διότι αυτή η φάση δέχεται τα δεδομένα και εξάγει τις συζητήσεις (συστάδες). Κάθε νεοεισερχόμενο μήνυμα συγκρίνεται με τις προϋπάρχουσες συστάδες. Επομένως με την αύξηση του πλήθους των μηνυμάτων, αυξάνεται και το πλήθος των συγκρίσεων που εκτελεί η διαδικασία *ανίχνευση συζητήσεων* προκειμένου το νεοεισερχόμενο μήνυμα να ενταχθεί σε μία από τις συστάδες. Όπως φαίνεται από το Σχήμα 7.2, ο χρόνος εκτέλεσης της φάσης *Ανίχνευση συζητήσεων* συναρτήσει του πλήθους των μηνυμάτων είναι γραμμικός. Έτσι, είναι λογικό ότι η αύξηση των δεδομένων προκαλεί επιπλέον επιβάρυνση της φάσης *ανίχνευση συζητήσεων* (Topic Clustering).

Όσον αφορά την φάση *ανίχνευση δημοφιλών συζητήσεων* (Popularity Filtering), παρατηρούμε πως διατηρείται σχεδόν σταθερός σε καθεμία περίπτωση ξεχωριστά ($\phi = 0.2\%$, $\phi = 0.5\%$), κάτι που είναι λογικό, καθώς η φάση αυτή επηρεάζεται άμεσα απο το διαφορετικό

Συντελεστής κλιμάκωσης αρχικού συνόλου δεδομένων $\|S\|$	×10	×20	×50	**×100**
Μέσο πλήθος μηνυμάτων ανά παράθυρο	38010.85	76021.7	190054.2	380108.5
Μέσο πλήθος μηνυμάτων ανά συζήτηση	14.1	28.4	70.8	140.1

Πίνακας 7.2: Μέσο πλήθος μηνυμάτων ανά συζήτηση



(α′) $\phi = 0.2\%$ (β′) $\phi = 0.5\%$

Σχήμα 7.3: Χρόνοι εκτέλεσης ανά παράθυρο και πλήθος δημοφιλών συζητήσεων για διάφορες υποδιαιρέσεις του καννάβου

πλήθος των συζητήσεων που ανιχνεύθηκαν στην προηγούμενη φάση. Με τον πολλαπλασιασμό των μηνυμάτων δεν παρήχθησαν διαφορετικά δεδομένα, αλλά δημιουργήσαμε πολλαπλά αντίγραφα των αρχικών, επομένως ο αριθμός των διαφορετικών δημοφιλών συζητήσεων που ανιχνεύονται παραμένει σχεδόν αμετάβλητος αλλά με μεγαλύτερη ένταση (βλ. Σχήμα 7.2).

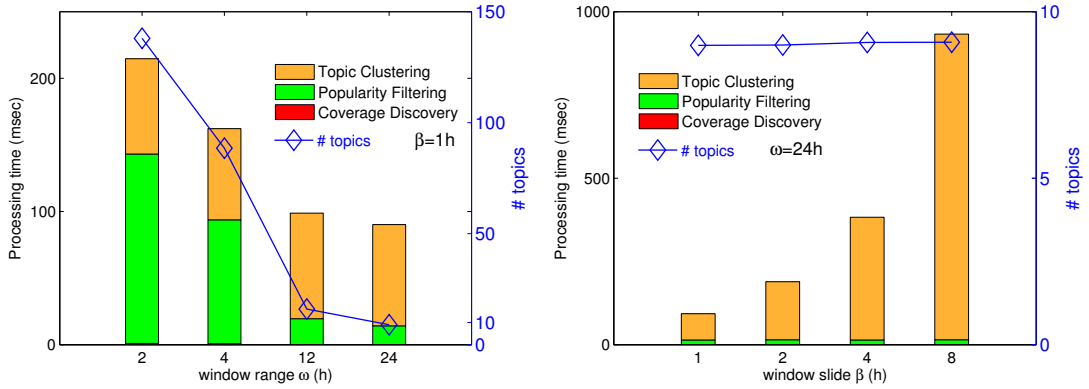
Πρακτικά, η διαδικασία του εντοπισμού χωρικής κάλυψης δεν φαίνεται στην γραφική παράσταση διότι ο χρόνος εκτέλεσής της είναι αμελητέος, αφού εμπλέκει μικρό πλήθος κελιών. Η ίδια συμπεριφορά παρατηρείται και σε όλα τα επόμενα πειράματα, καταδεικνύοντας ότι ο εντοπισμός χωρικής κάλυψης διευκολύνεται δραματικά χάρη στην αναγωγή μηνυμάτων σε δημοφιλείς συζητήσεις ανά κελί κατά τις δύο προηγούμενες φάσεις.

Τέλος, κυρίως λόγω της αύξησης του χρόνου εκτέλεσης της φάσης *Ανίχνευση συζητήσεων* παρατηρείται και αύξηση του συνολικού μέσου χρόνου εκτέλεσης του αλγορίθμου.

7.2.2 Διαστασιολόγηση Καννάβου

Για την αξιολόγηση του αλγορίθμου θεωρήσαμε τετραγωνικό κάνναβο μεγέθους $50km \times 50km$. Η επιλογή του πλήθους κελιών του καννάβου έγινε με βάση το παρακάτω πείραμα: Για σταθερή ελάχιστη δημοφιλία ϕ μεταβάλλαμε το πλήθος των κελιών του καννάβου. Οι χρόνοι εκτέλεσης των επιμέρους φάσεων ξεχωριστά και ο αριθμός των δημοφιλών συζητήσεων που εντοπίζονται για αυθαίρετες τιμές ελάχιστης δημοφιλίας φαίνονται στο Σχήμα 7.3. Πραγματοποιήθηκαν δύο πειράματα με ελάχιστες τιμές δημοφιλίας $\phi = 0.2\%$ και $\phi = 0.05\%$, όπου παρατηρήθηκαν αντίστοιχες συμπεριφορές στους χρόνους εκτέλεσης.

Για την φάση *Ανίχνευση συζητήσεων*, παρατηρούμε ότι όσο αυξάνουμε το πλήθος των κελιών, ο χρόνος εκτέλεσης της φάσης μειώνεται. Αυτό συμβαίνει επειδή η φάση *Ανίχνευση*



(α′) Μεταβαλόμενο εύρος παραθύρου (β′) Μεταβαλόμενο βήμα παραθύρου

Σχήμα 7.4: Επίδραση χρονικού παραθύρου

συζητήσεων, δημιουργεί συστάδες (συζητήσεις) εντός του εκαστοτε κελιού. Επομένως, όσο αυξάνουμε το πλήθος των κελιών, δημιουργούνται ολοένα και λιγότερες συστάδες ανά κελί, με αποτέλεσμα ο αριθμός συγκρίσεων του εκάστοτε μηνύματος με τις προϋπάρχουσες συστάδες να μειώνεται.

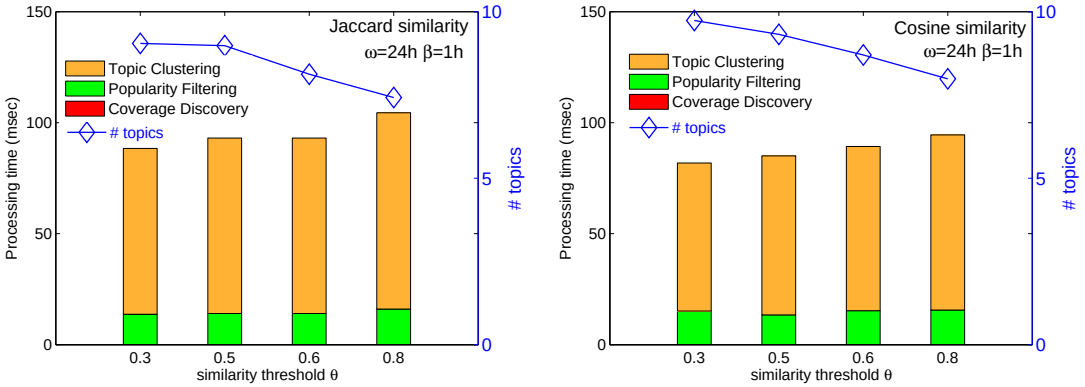
Για την φάση *Ανίχνευση δημοφιλών συζητήσεων*, παρατηρούμε ότι παραμένει σχεδόν αμετάβλητη για πλήθος κελιών 10×10 έως 50×50. Στην συνέχεια παρατηρούμε μια αισθητή αύξηση για πλήθος κελιών 75×75 και 100×100. Αυτό οφείλεται στο μεγάλο πλήθος κελιών που έχουμε να επεξεργαστούμε προκειμένου να ανιχνεύσουμε τις δημοφιλείς συζητήσεις, γεγονός που επιβαρύνει το κόστος.

Εν τέλει, παρατηρούμε πως ο συνολικός χρόνος εκτέλεσης σχηματίζει μια καμπύλη, με ελάχιστο χρόνο εκτέλεσης που αντιστοιχεί στον κάνναβο 50×50. Αυτό οφείλεται στην μείωση του χρόνου εκτέλεσης για την *Ανίχνευση συζητήσεων* και στην αύξηση του κόστους της φάσης *Ανίχνευση δημοφιλών συζητήσεων*. Ο αριθμός των συζητήσεων μειώνεται όσο αυξάνεται το πλήθος των κελιών του καννάβου, κάτι που είναι λογικό, εφόσον η ελάχιστη τιμή δημοφιλίας ϕ διατηρείται σταθερή και ο χώρος τεμαχίζεται ολοένα και περισσότερο, είναι πιθανό κάποιες συζητήσεις να παύουν να είναι δημοφιλείς σε ένα περαιτέρω τεμαχισμένο κελί του καννάβου. Επομένως είναι λογικό κάποιες συζητήσεις να μην ξεπερνούν την ελάχιστη τιμή δημοφιλίας. Επομένως η καλύτερη επιλογή καννάβου για την επεξεργασία του συγκεκριμένου συνόλου δεδομένων είναι ο κάνναβος 50×50, η οποία υιοθετείται σε όλα τα υπόλοιπα πειράματα.

7.2.3 Επίδραση χρονικού παραθύρου

Στην συνέχεια, κρατώντας σταθερό το πλήθος των κελιών του καννάβου 50×50, το κατώφλι ομοιότητας $\theta = 0.5$, την ελάχιστη δημοφιλία $\phi = 0.2\%$, δοκιμάζουμε τις επιδόσεις του αλγορίθμου για διάφορους προσδιορισμούς κυλιόμενου παραθύρου.

Στην γραφική παράσταση του Σχήματος 7.5α′, παρατηρούμε ότι όσο αυξάνουμε το εύρος ω του χρονικού παραθύρου, μειώνεται το πλήθος των δημοφιλών συζητήσεων που εντοπίζονται. Το γεγονός αυτό οφείλεται στη φύση και τον τρόπο λειτουργίας της ελάχιστης τιμής



(α′) Συνάρτηση ομοιότητας Jaccard (β′) Συνάρτηση ομοιότητας συνημιτόνου

Σχήμα 7.5: Επίδραση κατωφλίου ομοιότητας

Κατώφλι ομοιότητας θ	**0,3**	**0,5**	**0,6**	**0,8**
Μέσο πλήθος συζητήσεων	2.642,83	2.720,24	2.740,58	2.757,95
Μέσο πλήθος δημοφιλών συζητήσεων	9,04	8,98	8,12	7,42

Πίνακας 7.3: Μέσο πλήθος συζητήσεων για διάφορες τιμές του κατωφλίου ομοιότητας θ

δημοφιλίας ϕ που επιλέξαμε. Όπως έχουμε περιγράψει και στο Κεφάλαιο 6, μια συζήτηση είναι δημοφιλής εφόσον προκύπτει από τουλάχιστον $N_\omega \times \phi$ μηνύματα, επομένως η ελάχιστη τιμή δημοφιλίας εξαρτάται απο το πλήθος των μηνυμάτων ανά παράθυρο. Έτσι είναι λογικό ότι όσο μεγαλύτερο είναι το εύρος του παραθύρου τόσο μεγαλύτερη τιμή θα λαμβάνει η μεταβλητή N_ω άρα το όριο δημοφιλίας θα αυξάνεται με αποτελέσμα το πλήθος των δημοφιλών συζητήσεων να προκύπτει μικρότερο. Η ραγδαία μείωση του πλήθους των δημοφιλών συζητήσεων με την αύξηση του εύρους του παραθύρου προκαλεί και την αύξηση του χρόνου εκτέλεσης για την *ανίχνευση δημοφιλών συζητήσεων*, εφόσον η φάση αυτή εξαρτάται άμεσα από το πλήθος των συζητήσεων που ανιχνεύθηκαν εντός του παραθύρου σε όλο τον κάνναβο.

Έπειτα εκτελέσαμε ένα αντίστοιχο πείραμα διατηρώντας σταθερή την τιμή του εύρους $\omega =$ 24h του χρονικού παραθύρου και μεταβάλλοντας το βήμα μετακύλισης $\beta = \{1h, 2h, 4h, 8h\}$. Στο γράφημα του Σχήματος 7.5β′, βλέπουμε πως η αύξηση του βήματος β προκαλεί την αύξηση του χρόνου εκτέλεσης για την φάση *ανίχνευση συζητήσεων*. Αυτό δικαιολογείται επειδή η *ανίχνευση συζητήσεων* εκτελείται για κάθε χρονικό πλαίσιο διάρκειας β του παραθύρου. Όσο μεγαλύτερο είναι το βήμα μετακύλισης β (δηλαδή η χρονική περιόδος που κάλυπτει κάθε πλαίσιο), τόσο μεγαλύτερος ο όγκος εισερχόμενων δεδομένων που καλείται να επεξεργαστεί. Αντίθετα, η φάση *ανίχνευση δημοφιλών συζητήσεων* και το πλήθος των δημοφιλών συζητήσεων δεν επηρεάζεται σχεδόν καθόλου απο την μεταβολή του βήματος μετακύλισης, αφού αυτοί οι υπολογισμοί διεξάγονται σε όλο το εύρος ω του παραθύρου, το οποίο παραμένει σταθερό στις $24h$ για το πείραμα αυτό.

7.2.4 Επίδραση κατωφλίου ομοιότητας συζητήσεων (θ)

Στο πείραμα αυτό, διατηρήσαμε σταθερό το εύρος παραθύρου $\omega = 24h$, το βήμα μετακύλισης $\beta = 1h$, την ελάχιστη τιμή δημοφιλίας $\phi = 0.2\%$ και μεταβάλαμε το κατώφλι ομοιότητας $\theta = \{0.3, 0.5, 0.6, 0.8\}$ για τις δύο συναρτήσεις ομοιότητας, συνάρτηση Jaccard, συνάρτηση συνημιτόνου, που χρησιμοποίηθηκαν κατά την σύγκριση ετικετών των μηνυμάτων με προϋπάρχουσες συζητήσεις.

Απο την γραφική παράσταση του Σχήματος 7.5 παρατηρούμε πως όσο αυξάνεται το κατώφλι ομοιότητας θ, το πλήθος των δημοφιλών συζητήσεων μειώνεται, ενώ ταυτόχρονα ο μέσος χρόνος εκτέλεσης για την φάση *ανίχνευση συζητήσεων* αυξάνεται. Αυτό συμβαίνει διότι όσο μεγαλύτερο είναι το κατώφλι ομοιότητας είναι πιο δύσκολο ένα μήνυμα να ενταχθεί σε μία συστάδα (συζήτηση), με αποτέλεσμα να δημιουργούνται περισσότερες συστάδες συζητήσεων (βλ. πίνακα 7.3). Ωστόσο, το πλήθος μηνυμάτων που συγκροτούν κάθε συστάδα θα είναι μικρότερο συγκριτικά με το πλήθος μηνυμάτων μιας συστάδας για μεγαλύτερο κατώφλι. Επομένως, τελικά προκύπτουν περισσότερες αλλά λιγοτερο δημοφιλείς συζητήσεις (συστάδες) πολλές απο τις οποίες δεν μπορούν να υπερβούν την ελάχιστη τιμή δημοφιλίας ϕ κατά την επόμενη φάση $PopularityFiltering$.

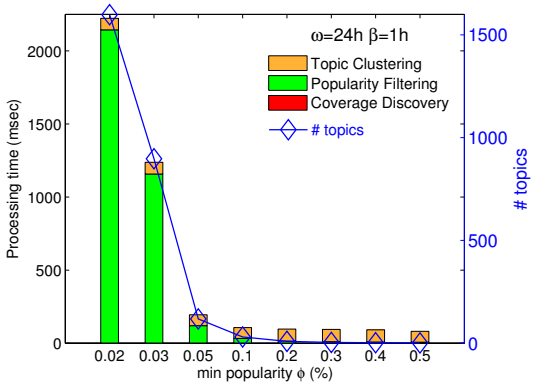
7.2.5 Επίδραση ελάχιστης τιμής δημοφιλίας συζητήσεων

Στο πείραμα αυτό μελετήθηκε η επίδραση της ελάχιστης τιμής δημοφιλίας ϕ στους μέσους χρόνους εκτέλεσης της εκάστοτε φάσης και στο πλήθος των δημοφιλών συζητήσεων που ανιχνεύθηκαν. Για την εκτέλεση του πειράματος, διατηρήσαμε σταθερές τις εξής παραμέτρους: εύρος παραθύρου $\omega = 24h$, βήμα μετακύλισης $\beta = 1h$, κατώφλι ομοιότητας $\theta = 0.5$ και πλήθος κελιών καννάβου 50×50. Μεταβάλοντας την ελάχιστη τιμή δημοφιλίας ϕ, από το γράφημα του Σχήματος 7.6 παρατηρούμε ότι ο χρόνος εκτέλεσης για την *ανίχνευση συζητήσεων* παραμένει σχεδον σταθερός. Αυτό συμβαίνει διότι η ελάχιστη τιμή δημοφιλίας ϕ εμπλέκεται μόνο στην φάση της ανίχνευσης των δημοφιλών συζητήσεων ($PopularityFiltering$). Πράγματι, όσο αυξάνεται η ελάχιστη τιμή δημοφιλίας ϕ, τόσο λιγότερες δημοφιλείς συζητήσεις εντοπίζονται. Ταυτόχρονα, παρατηρούμε αντίστοιχη μείωση του χρόνου εκτέλεσης της φάσης *Ανίχνευση δημοφιλών συζητήσεων* σε σχέση με το πλήθος τους. Πράγματι, το πλήθος των δημοφιλών συζητήσεων που τηρούνται στο ευρετήριο $\mathcal{P}$ είναι μεγαλύτερο όσο ελλατώνεται η ελάχιστη τιμή δημοφιλίας ϕ, επομένως οι αναζητήσεις και ανανεώσεις στο ευρετήριο $\mathcal{P}$ κοστίζουν περισσότερο χρόνο.

7.3 Σύνοψη συμπερασμάτων αξιολόγησης

Τέλος, από την πειραματική αξιλόγηση του αλγορίθμου προκύπτουν τα εξής συμπεράσματα:

- Η κλιμάκωση του όγκου των πρωτογενών δεδομένων επιβαρύνουν κυρίως την φάση της *ανίχνευσης συζητήσεων*, αφού κάθε νεοεισρχόμενο μήνυμα συγκρίνεται με τις προϋπάρχουσες συζητήσεις. Επομένως ο χρόνος εκτέλεσης της φάσης αυτής είναι γραμμικός ως



Σχήμα 7.6: Ελάχιστη τιμή δημοφιλίας ϕ

προς το πλήθος των δεδομένων που καλείται να διαχειριστεί. Η επιβάρυνση της φάσης *ανίχνευση δημοφιλών συζητήσεων* είναι ανεπαίσθητη, διότι με την αύξηση του όγκου των δεδομένων, δεν αυξάνεται το πλήθος των διαφορετικών συζητήσεων που εντοπίζονται εντός των κελιών g αλλά το πλήθος των μηνυμάτων που απαρτίζουν την εκάστοτε συζήτηση.

- Ο χρόνος εκτέλεσης της φάσης *ανίχνευση δημοφιλών συζητήσεων*, όπως επιβεβαιώνεται και από την πειραματική αξιολόγηση, εξαρτάται από το πλήθος των κελιών του καννάβου G, αφού χρειάζεται να διατρέξει ολά τα κελιά προκειμένου να εξάγει τις δημοφιλείς συζητήσεις σε όλη την έκταση της περιοχής μελέτης, από το πλήθος των διαφορετικών συζητήσεων που εντοπίζονται στα κελιά του καννάβου και τέλος από το πλήθος των διαφορετικών συζητήσεων που εντοπίζονται σε όλη την έκταση του καννάβου.

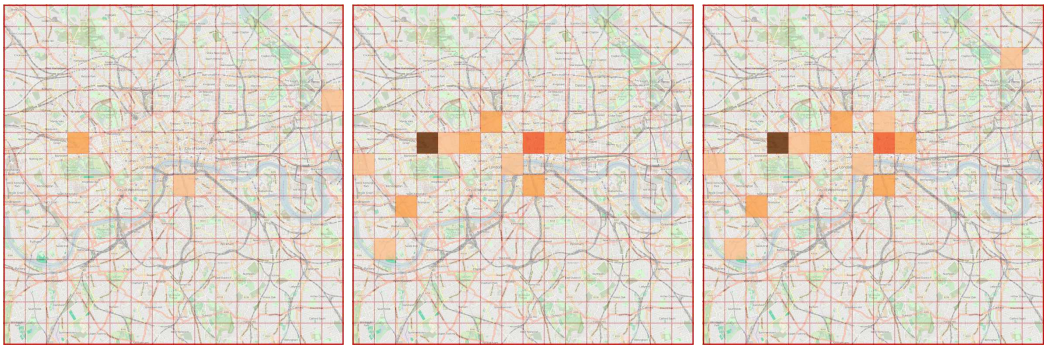
- Ο χρόνος εκτέλεσης της φάσης *ανίχνευση συζητήσεων* επιβαρύνεται από τον όγκο των νεοεισερχόμενων δεδομένων και επιπλέον από το πλήθος των συζητήσεων που εντοπίζονται στο εκάστοτε κελί του καννάβου.

- Η επίβαρυνση του χρόνου εκτέλεσης της φάσης *εντοπισμός χωρικής κάλυψης* είναι γενικά ανεπαίσθητος, αφού επεξεργάζεται ένα μικρό σε σύγκριση με το πλήθος των δημοσιευμένων μηνυμάτων, αριθμό κελιών που αντιστοιχούν στην χωρική κάλυψη της εκάστοτε δημοφιλούς συζήτησης.

Γενικά, κάθε κύκλος εκτέλεσης γενικά ολοκληρώνεται εντός δευτερολέπτων (μόνο για χαμήλη ελάχιστη τιμή δημοφιλίας ϕ παρατηρούνται χρόνοι περίπου 2 sec) και στις περισσότερες περιπτώσεις σε δέκατα του δευτερολέπτου για μεγάλο όγκο νεοεισερχόμενων μηνύματων. Επομένως, μπορούμε να ισχυριστούμε πως ο αλγόριθμος που αναπτύχθηκε εκτελείται και εξάγει αποτελέσματα σε πραγματικό χρόνο.

7.3.1 Ποιοτικά αποτελέσματα

Κατά την πειραματική μελέτη του αλγόριθμου που αναπτύχθηκε, εξετάστηκαν επιπλέον τα ποιοτικά αποτελέσματα που εξάγονται ανά τακτά χρονικά διαστήματα. Τα ποιοτικά αποτελε-



(α΄) 29 Απριλίου, 9:00 π.μ (β΄) 29 Απριλίου, 10:00 π.μ (γ΄) 29 Απριλίου, 11:00 π.μ

Σχήμα 7.7: Χωρική κάλυψη δημοφιλούς συζήτησης *#tubestrike* με την πάροδο του χρόνου

σμάτα περιλαμβάνουν το σύνολο των ετικετών (hashtags) που συνιστούν την ταυτότητα των συζητήσεων, τις περιοχές χωρικής κάλυψης των συζητήσεων, το πλήθος των μηνυμάτων σε κάθε κελί (ένταση) και τις μεταβολές στην εκτάση τους με την πάροδο του χρόνου.

Ενδεικτικά, στο Σχήμα 7.7 απεικονίζεται η χωρική κάλυψη, για διαδοχικές μετακυλίσεις του παραθύρου, της δημοφιλούς συζήτησης με την ετικέτα *#tubestrike*. Η συζήτηση αυτή έχει θέμα την απεργία στο μετρό του Λονδίνου. Αρχικά, η συζήτηση εκτυλίσσεται σε ένα μόνο κελί του καννάβου και στην συνέχεια σταδιακά εξαπλώνεται στα γύρω κελιά με κυμαινόμενη ένταση σε κάθε κελί του καννάβου.

Κεφάλαιο 8

Επίλογος

8.1 Συμπεράσματα

Ο αλγόριθμος *εποπτεία γεωγραφικής εξάπλωσης συζητήσεων στα κοινωνικά δίκτυα* που αναπτύχθηκε επικεντρώθηκε στην μελέτη συζητήσεων που εκτυλίσσονται στο κοινωνικό δίκτυο Twitter. Ο ρύθμος των μηνυμάτων που δημοσιεύονται στο Twitter είναι ραγδαίος. Γι΄ αυτό τον λόγο, το πλήθος των δημοσιευμένων μηνυμάτων θεωρείται ρεύμα δεδομένων.

Η μεθοδολογία που σχεδιάστηκε, επεξεργάζεται τα νεοεισερχόμενα μηνύματα σε πραγματικό χρόνο προκειμένου να εντοπίσει δημοφιλή θέματα που συζητιούνται την τρέχουσα χρονική στιγμή. Επιπλέον, για κάθε δημοφιλή συζήτηση που ανιχνεύεται, εντοπίζεται και παρακολουθείται η χωρική εξάπλωση της στον χώρο και στον χρόνο. Η διαδικασία για την εύρεση και την εποπτεία της χωρικής κάλυψης πραγματοποιείται σε πραγματικό χρόνο και τα αποτελεσμάτα που εξάγονται ανά τακτά χρονικά διαστήματα είναι προσεγγιστικά, αλλά ικανοποιητικά ποιοτικά. Πιο συγκεκριμένα, προκύπτουν τα εξής συμπεράσματα:

- Ο κάνναβος G που χρησιμοποιήθηκε στην περιοχή μελέτης αποδείχθηκε ικανοποιητικός για την ανίχνευση δημοφιλών συζητήσεων και την εποπτεία τους στο χώρο και στο χρόνο. Παρ΄ όλα αυτά, η χωρική κάλυψη των συζητήσεων είναι προσεγγιστική και εντοπίζεται σε επίπεδο κελιών. Επιπλέον, η διακριτοποίηση του καννάβου διαδραματίζει σημαντικό ρόλο στην εύρεση συζητήσεων, καθώς μεταβάλλοντας το πλήθος των κελιών είναι πιθανόν να χάνονται ορισμένες συζητήσεις, εξαιτίας της φύσης της ελάχιστης τιμής δημοφιλίας που έχει οριστεί.

- Η συνάρτηση ομοιότητας για την ανίχνευση συζητήσεων, ουσιαστικά συγκρίνει την ομοιότητα δύο συνόλων (σύνολο ετικετών). Επομένως, κατά την σύγκριση, πραγματοποιείται ταύτιση των ετικετών και όχι σύγκριση χαρακτήρων. Για παράδειγμα, όπως φάνηκε και από τα πειραματικά αποτελέσματα, οι δύο ετικέτες $H_1 = \{\#tubestrike\}$, $H_2 = \{\#tubestrikes\}$ δημιουργούν δύο διαφορετικές συζητήσεις. Κάτι τέτοιο επηρεάζει τόσο το πλήθος, όσο και την ένταση των ανιχνευμένων δημοφιλών συζητήσεων.

- Οι επιδόσεις που μετρήθηκαν πειραματικά, αποδείχθηκαν ιδιαίτερα επαρκείς για την επεξεργασία μεγάλου όγκου δεδομένων σε πραγματικό χρόνο. Επιπλέον, τα ποιοτικά

αποτελέσματα που προέκυψαν, εκτιμάται ότι δεν απέχουν από την πραγματικότητα.

8.2 Μελλοντικές επεκτάσεις

Από την μελέτη του συγκεκριμένου προβλήματος προκύπτουν ενθαρρυντικές προοπτικές επέκτασής του. Συγκεκριμένα:

- Στην εργασία αυτή ελήφθησαν υπ΄ όψιν μόνο τα μηνύματα που διαθέτουν γεωγραφικό στίγμα. Όμως, μόνο το 2% του συνολικού πλήθους μηνυμάτων που δημοσιεύονται καθημερινά, διαθέτουν γεωγραφικές συντεταγμένες. Επομένως, η προτεινόμενη μέθοδος αγνοεί σημαντικό όγκο πληροφορίας. Θα ήταν λοιπόν δυνατόν, για τα μηνύματα που δεν διαθέτουν γεωγραφικό στίγμα, να πραγματοποιείται κειμενική ανάλυση (text analysis) όλου του περιεχομένου τους, προκειμένου να γίνεται μία εκτίμηση για την τοποθεσία που δημοσιεύθηκε το μήνυμα (π.χ. #Syntagma).

- Επιπλέον, οι τιμές των παραμέτρων θα ήταν δυνατόν να επιλέγονται με βάση τεχνικές μηχανικής μάθησης για αποτελεσματικότερη εποπτεία αντί να δίνονται αυθαίρετες τιμές.

- Τέλος, η διαδικασία αυτή θα μπορούσε να παραλληλοποιηθεί διαμοιράζοντας σε κάθε επεξεργαστικό κόμβο ένα υποσύνολο των κελιών του καννάβου, εκτελώντας τοπικά την ανίχνευση συζητήσεων και ανταλλάσσοντας τα ενδιάμεσα αποτελέσματα για τον σχηματισμό του συνολικού ευρετηρίου.

Βιβλιογραφία

[1] H.Abdelhaq and M.Gertz. On the Locality of Keywords in Twitter Streams. In *Proceedings of the 5th ACM SIGSPATIAL International Workshop on GeoStreaming (IWGS)*, pp. 12-20, Dallas, Texas, USA, November 2014.

[2] H.Abdelhaq, M.Gertz, and C.Sengstock. Spatio-temporal Characteristics of Bursty Words in Twitter Streams. In *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL)*, pp. 194-203, Orlando, Florida, USA, November 2013.

[3] H.Abdelhaq, C.Sengstock, and M.Gertz. EvenTweet: Online Localized Event Detection from Twitter. *PVLDB*, 6(12): 1326-1329, August 2013.

[4] C. C.Aggarwal, J.Han. J.Wang, P.Yu. A Framework for Clustering Evolving Data Streams. In *Proceedings of 29th International Conference on Very Large Data Bases (VLDB)*, pp. 81-92, Berlin, Germany, September 2003.

[5] C. C. Aggarwal, and K Subbian. Event Detection in Social Streams. In*Proceedings of the Twelfth SIAM International Conference on Data Mining*, pp. 624-635, Anaheim, California, USA, April 2012.

[6] B. Babcock, S. Babu, M. Datar, R. Motwani, and J.Widom. Models and Issues in Data Stream Systems. In *Proceedings of the 21stACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS'02)*, pp.1-16, Madison, Wisconsin, May 2002.

[7] L. Chen, G. Cong, C.S. Jensen, and D. Wu. Spatial Keyword Query Processing: An Experimental Evaluation. *PVLDB*, 6(3): 217-228, 2013.

[8] W. Cohen, P.D. Ravikumar and S.E. Fienberg. A Comparison of String Distance Metrics for Name-Matching Tasks. In*IIWeb, pp. 73-78,2003*

[9] H. Efstathiades, D. Antoniades, G. Pallis, and M.D. Dikaiakos. Identification of Key Locations based on Online Social Network Activity. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 218-225, Paris, France, August 2015.

[10] M. Ester, H. Kriegel, J. Sander, and X. Xu. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD'96)*, pp. 226-231, Portland, Oregon, USA, August 1996.

[11] Facebook Inc. http://www.facebook.com

[12] W.Feng, C.Zhang, W.Zhang, J.Han, J.Wang, C.Aggarwal, and J. Huang. StreamCube: Hierarchical Spatio-temporal Hashtag Clustering for Event Exploration over the Twitter Stream. In *Proceedings of the 31st International Conference on Data Engineering (ICDE)*, pp. 1561-1572, Seoul, Korea, April 2015.

[13] Foursquare Inc. http://foursquare.com

[14] V. Gaede, and O. Gunther. MultidimensionalAccess Methods. *ACM Computing Surveys*, 30 : 170-231, 1998.

[15] L. Golab, and M. Tamer Ozsu. Issues in Data Stream Management.*ACM SIGMOD Record*, 32(2):5-14,June 2003.

[16] M. Hadjieleftheriou, G. Kollios, D. Gunopulos, and V. J. Tsotras. On-Line Discovery of Dense Areas in Spatio-temporal Databases. In *Proceedings of the 8th International Symposium on Spatial and Temporal Databases (SSTD'03)*, pp.306-324, Santorini Island, Greece, July 2003.

[17] C.S. Jensen, D. Lin, B.C. Ooi. Continuous Clustering of Moving Objects. *IEEE Transactions on Knowledge and Data Engineering*, 19(9): 1161-1174, 2007.

[18] C. S. Jensen, D. Lin, B. Chin Ooi, and R. Zhang. Effective Density Queries on Continuously Moving Objects. In *Proceedings of the 22nd International Conference on Data Engineering (ICDE'06)*, pp. 71-81, Atlanta, Georgia, USA, April 2006.

[19] J. Jiang, H. Lu, B. Yang, and B. Cui. Finding Top-k Local Users in Geo-tagged Social Media Data. In *Proceedings of the 31st International Conference on Data Engineering (ICDE)*, pp. 267-278, Seoul, Korea, April 2015.

[20] J. Li, D. Maier, K. Tufte, V. Papadimos, and P. A. Tucker. No Pane, No Gain: Efficient Evaluation of Sliding-Window Aggregates over Data Streams.*ACM SIGMOD Record*, 34(1): 39-44, 2005.

[21] T. Lappas, M.R. Vieira, D. Gunopulos, and V.J. Tsotras. On the Spatiotemporal Burstiness of Terms. *PVLDB*, 5(9): 836-847, May 2012.

[22] A. Magdy, L. Alarabi, S. Al-Harthi, M. Musleh, T.M. Ghanem, S. Ghani, and M.F. Mokbel. Taghreed: a System for Querying, Analyzing, and Visualizing Geotagged Microblogs. In *Proceedings of the 22nd ACM SIGSPATIAL International Conference*

on Advances in Geographic Information Systems (ACM GIS), pp. 163-172, Dallas, Texas, USA, November 2014.

[23] T. Sakaki, M. Okazaki, Y. Matsuo. Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. *International World Wide Web Conference Committee (IW3C2)*, pp. 851-860, Raleigh, North Carolina, April 2010.

[24] J. Sander, M. Ester, H. Kriegel, and X. Xu. Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications. *Data Mining and Knowledge Discovery*, 2:169-194, 1998.

[25] J. Shi, N. Mamoulis, D. Wu, D.W. Cheung. Density-based Place Clustering in Geo-Social Networks. In *Proceedings of ACM SIGMOD International Conference on Management of Data*, pp. 99-110, Snowbird, Utah, USA, 2014.

[26] A. Skovsgaard,D. Sidlauskas, and C. Jensen. Scalable Top-k Spatio-Temporal Term Querying. In *Proceedings of 30th IEEE International Conference on Data Engineering*, pp. 148-159, Chicago, Illinois, USA, April 2014.

[27] M. Stonebraker, U. Cetintemel, and S. Zdonik. The 8 Requirements of Real-Time Stream Processing. *ACM SIGMOD Record*, 34(4):42-47, December 2005.

[28] Twitter Inc. http://www.twitter.com

[29] W. Wang, J. Yang, and R. Muntz. STING: A Statistical Information Grid Approach to Spatial Data Mining. In *Proceedings of the 23rd International Conference on Very Large Data Bases (VLDB)*, pp. 186-195, Athens, Greece, 1997.

[30] K.Watanabe, M.Ochi, M.Okabe, and R.Onai. Jasmine: A Realtime Local-event Detection System based on Geolocation Information Propagated to Microblogs. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM)*, pp. 2541-2544, Glasgow, Scotland, UK, 2011.

[31] T.Zhang, R.Ramakrishan, M.Livny: BIRCH An Efficient Data Clustering Method for Very Large Databases. In *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data*, pp. 103-114, Montreal, Quebec, Canada, June 1996.

Γλωσσάριο

Ελληνικός όρος	Αγγλικός όρος
αποτίμηση ερωτημάτων	query evaluation
δεικτοδότηση	indexing
δημοφιλείς συζητήσεις	popular topics
ένταση	intensity
ερώτημα διαρκείας	continuous query
ερώτημα εγγύτερου γείτονα	nearest-neighbor query
εξάπλωση	expansion
επαυξητική	bottom-up
ιεραρχική	hierarchical
κατηγορικά δεδομένα	categorical data
κατώφλι	threshold
κάνναβος	grid
κινούμενο αντικείμενο	moving object
μήνυμα	tweet
παράθυρο	window
ρεύμα δεδομένων	data stream
συνάρτηση ομοιότητας	similarity function
συστάδα	cluster
συνάθροιση	aggregation
συρρίκνωση	shrinking
φιλτράρισμα	filtering
χρονόσημο	timestamp
χρονικό πλαίσιο παραθύρου	pane
χωρική κάλυψη	spatial coverage

Εποπτεία γεωγραφικής κάλυψης συζητήσεων σε κοινωνικά δίκτυα

Εμμανουήλ Λουκαδάκης

manos.loyk@gmail.com

Διπλωματική εργασία στο Εργαστήριο Συστημάτων Βάσεων Γνώσεων και Δεδομένων
Επιβλέπων: Καθηγητής ***Ι. Βασιλείου***

1 Γενικό πλαίσιο

Αντικείμενο της παρούσας διπλωματικής εργασίας αποτελεί ο σχεδιασμός και η υλοποίηση ενός αλγορίθμου για την ανίχνευση δημοφιλών συζητήσεων που εκτυλίσσονται στα κοινωνικά δίκτυα, αλλά και την διαρκή εποπτεία τέτοιων συζητήσεων με σκοπό την μελέτη της εξέλιξής τους στον χώρο και στον χρόνο.

Σήμερα, γνωρίζουν άνθηση πολλά κοινωνικά δίκτυα όπως το Twitter, το Facebook, το Foursquare, κ.ά. Από αυτά επιλέξαμε να μελετήσουμε συζητήσεις που εκτυλίσσονται στο Twitter, διότι ο τρόπος λειτουργίας και η δομή του εξυπηρετούν καλύτερα την παρούσα μελέτη. Πιο συγκεκριμένα, το Twitter αποτελεί ένα ρεύμα δεδομένων (δεν είναι τυχαίο που το αποκαλούν Twitter Stream), καθώς παγκοσμίως δημοσιεύονται μηνύματα με πολύ γρήγορο ρυθμό. Ενδεικτικά, κατά το 2015, οι χρήστες δημοσίευαν μηνύματα με ρυθμό 500.000.000 ανά ημέρα. Επιπλέον, κάθε μήνυμα (tweet) των χρηστών περιορίζεται σε 140 χαρακτήρες το πολύ, κάνοντας έτσι την επεξεργασία της πληροφορίας πιο εύκολη σε σχέση με τα άλλα κοινωνικά δίκτυα στα οποία αυτός ο περιορισμός δεν υπάρχει.

Ο αλγόριθμος που υλοποιήθηκε δίνει την δυνατότητα μελέτης του φαινομένου αυτού θέτοντας διαφορετικές παραμέτρους στην είσοδο. Τα δεδομένα που επεξεργαζόμαστε είναι τα μηνύματα των χρηστών. Για τον σκοπό της μελέτης λάβαμε υπ΄ όψιν μόνο τα μηνύματα που περιλαμβάνουν γεωγραφικό στίγμα (geo-tagged tweets), προκειμένου να υπάρχει η δυνατότητα μελέτης της εξέλιξής τους στο χώρο. Τα μηνύματα θεωρούνται ότι δημοσιεύονται σε πραγματικό χρόνο, οπότε είναι ανάγκη η επεξεργασία των δεδομένων να είναι αποδοτική, αποδεχόμενοι ότι τα αποτελέσματα θα είναι κατ΄ ανάγκην προσεγγιστικά. Επομένως, για την σχεδίαση της μεθοδολογίας επιδιώκουμε ελαχιστοποίηση της πολυπλοκότητας. Ο αλγόριθμος εξάγει ενημερωμένα αποτελέσματα ανά τακτά χρονικά διαστήματα που ανταποκρίνονται στην συνεχή ροή μηνυμάτων.

2 Επεξεργασία δεδομένων σε κοινωνικά δίκτυα

Η επεξεργασία της πληροφορίας που παράγεται στα κοινωνικά δίκτυα και η εφαρμογή ερωτημάτων σε τέτοια δεδομένα μπορεί να μας δώσει κάποια πολύ σημαντικά συμπεράσματα. Χαρακτηριστικά ερωτήματα και αναλύσεις που έχουν προταθεί κατά καιρούς στην βιβλιογραφία είναι τα ακόλουθα:

- *Ανίχνευση τοπικών γεγονότων*: Όταν ένα γεγονός εξελίσσεται, τότε οι χρήστες που συμμετέχουν ή το παρατηρούν, έχουν την τάση να το δημοσιοποιούν και να παράγουν μηνύματα σχετικά με αυτό μέσω του κινητού τους. Πρακτικά, οι χρήστες λειτουργούν ως αισθητήρες που παρατηρούν ένα γεγονός πλησίον τους και το καταγράφουν όπως το αντιλαμβάνονται. Επομένως, εκμεταλλευόμενοι την πληροφορία των κοινωνικών δικτύων θα μπορούσαμε να εντοπίσουμε και να μάθουμε για γεγονότα πολύ πιο γρήγορα από τα μέσα μαζικής ενημέρωσης.
- *Ερωτήματα top-k χωροχρονικών λεκτικών όρων*: Τέτοια ερωτήματα χρησιμοποιούνται για να εντοπίσουν τους k επικρατέστερους λεκτικούς όρους που εμφανίζονται πιο συχνά σε μια περιοχή ενδιάφεροντος. Αποτελούν ερωτήματα διαρκείας, επομένως είναι αναγκαίο η επεξεργασία να γίνεται σε πραγματικό χρόνο. Με

αυτό τον τρόπο μπορεί να εντοπιστεί δυναμικά οποιοδήποτε τοπικό, επίκαιρο θέμα συζήτησης ανάλογα με το πλήθος και το περιεχόμενο των εκάστοτε μηνυμάτων.

Διερευνώντας μία άλλη κατεύθυνση, σε αυτήν την εργασία επιχειρείται μεν ο εντοπισμός δημοφιλών συζητήσεων, αλλά μόνο όπου υπάρχει ισχυρή συγκέντρωση πολλών σχετικών μηνυμάτων. Επομένως, εντοπίζονται *συμπαγείς* περιοχές όπου μία συζήτηση είναι δημοφιλής και επιπλέον ανιχνεύονται *μεταβολές* στην γεωγραφική της έκταση. Προφανώς, η ανακάλυψη τέτοιων περιοχών μπορεί να γίνει μέσω συσταδοποίησης (π.χ. DBSCAN), όμως κάτι τέτοιο θα ήταν ασύμφορο από πλευράς κόστους.

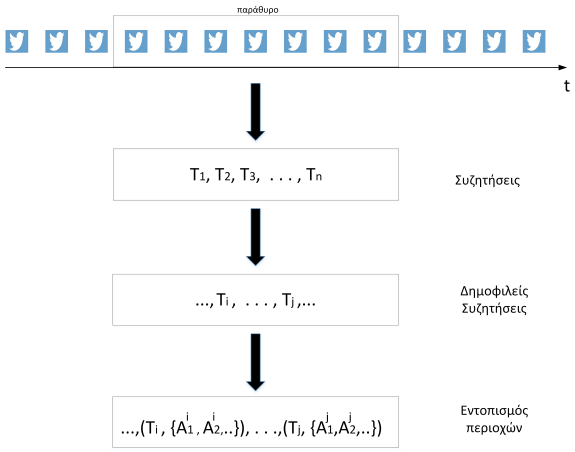
3 Μοντέλο συστήματος

Τα δημοσιευμένα μηνύματα των χρηστών στο Twitter σχηματίζουν ένα ρεύμα δεδομένων της μορφής S=$(..., s_i, ..., s_j, ...)$. Κάθε μήνυμα (tweet) ακολουθεί την μορφή: $s = \langle \tau, uid, loc, H \rangle$, όπου τ είναι η χρονική στιγμή της δημοσίευσης, uid είναι ένας μοναδικός κωδικός ταυτοποίησης κάθε χρήστη, το γεωγραφικό στίγμα loc του χρήστη κατά τη στιγμή της δημοσίευσης και τέλος το σύνολο *ετικετών* (hashtags) $H = \{H_1, H_2, ...\}$ του μηνύματος. Στην παρούσα μελέτη όλα τα μηνύματα χωρίς γεωγραφικό στίγμα αγνοούνται κατά την επεξεργασία, καθώς επίσης και το υπόλοιπο κείμενο (πλήν ετικετών).

Για την αποδοτικότερη επεξεργασία των δεδομένων, χρησιμοποιήθηκε χρονικά *κυλιόμενο παράθυρο* (sliding window) με *χρονικό εύρος* ω και *χρονικό βήμα* β. Το παράθυρο διακριτοποιείται σε $\frac{\omega}{\beta} = \lambda$ χρονικά μη επικαλυπτόμενα *πλαίσια (panes)* προκαθορισμένου ορίζοντα, όπου $\omega, \beta, \lambda \in \mathbb{Z}$.

Για το πρόβλημά μας, μία *συζήτηση* T_k αποτελείται από το σύνολο των ετικετών των σχετικών μηνυμάτων, τα οποία παρουσιάζουν τον μεγαλύτερο βαθμό ομοιότητας μεταξύ τους. Αυτή η ομοιότητα καθορίζεται από ένα *κατώφλι* $0 \leq \theta \leq 1$ και υπολογίζεται από μία *συνάρτηση ομοιότητας*, λ.χ. Jaccard ή cosine.

Μία τέτοια συζήτηση χαρακτηρίζεται ως *δημοφιλής* όταν τα μηνύματα που την απαρτίζουν βρίσκονται εντός μιας δεδομένης *περιοχής κάλυψης* A και το πλήθος τους ξεπερνά μία *ελάχιστη τιμή δημοφιλίας* $\phi \in \mathbb{R}^+$ εντός του χρονικού εύρους ω του παραθύρου. Το ζήτημα είναι ότι η περιοχή κάλυψης A δεν είναι γνωστή εκ των προτέρων, αλλά πρέπει να ανακαλύπτεται δυναμικά βάσει των εκάστοτε δημοσιευμένων μηνυμάτων.



Σχήμα 1: *Στάδια επεξεργασίας.*

Αποφεύγοντας την online συσταδοποίηση, επιλέγουμε να ανασυγκροτούμε κάθε περιοχή A ως σύνολο από προκαθορισμένες ψηφίδες σταθερού μεγέθους, οι οποίες είναι γειτονικές μεταξύ τους. Συγκεκριμένα, με χρήση ενός *καννάβου* G, διαμερίζουμε όλη την περιοχή παρακολούθησης σε $g \times g$ ομοιόμορφα κελιά. Συνεπώς αναζητούμε *συμπαγείς περιοχές κάλυψης*, καθεμιά από τις οποίες αποτελείται είτε από γειτονικά είτε μεμονωμένα κελιά του καννάβου (διακρίνονται με διαφορετικά χρώματα στο Σχήμα 2α΄). Το γενικό μοντέλο της μεθοδολογίας προβλέπει την συνεχή ανανέωση των αποτελεσμάτων ανά *κύκλο εκτέλεσης* (μετακύλιση παραθύρου) σε πραγματικό χρόνο.

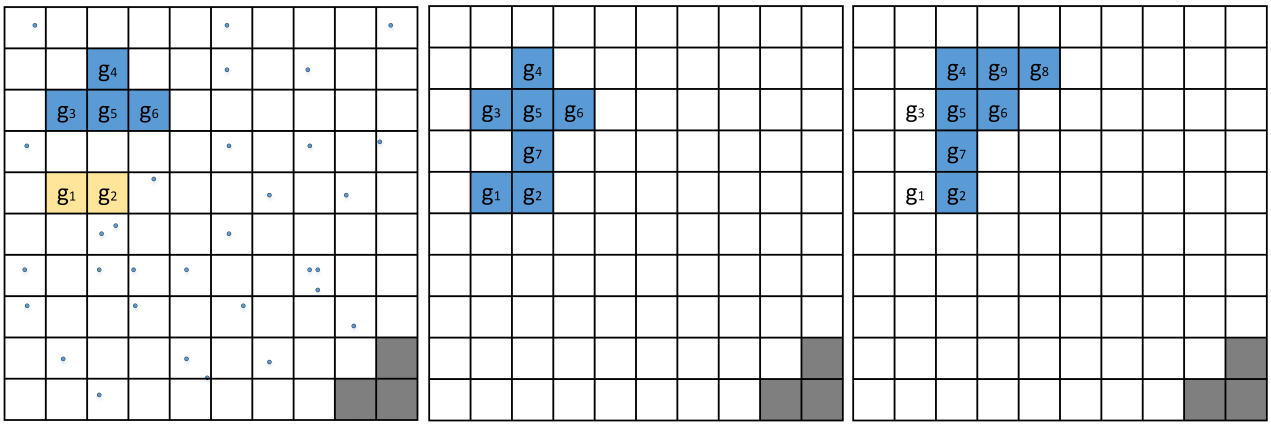
Επιπλέον, ενδιαφερόμαστε να ανακαλύψουμε πιθανές μεταβολές σε κάθε περιοχή κάλυψης A μεταξύ διαδοχικών κύκλων εκτέλεσης. Πιο συγκεκριμένα, μία περιοχή κάλυψης θεωρείται ότι *εξαπλώθηκε* όταν το συνολικό πλήθος κελιών που εντοπίστηκαν στον τρέχοντα κύκλο εκτέλεσης είναι μεγαλύτερο από τον προηγούμενο. Αντίστοιχα, μια περιοχή κάλυψης *συρρικνώθηκε* εφόσον το πλήθος των κελιών έχει ελαττωθεί σε σύγκριση με την προηγούμενη μετακύλιση του παραθύρου.

Σημειώνεται πως ο αλγόριθμος δεν παρέχει ακριβή αποτελέσματα, αλλά *προσεγγιστικά*. Ωστόσο, όπως θα φανεί στην πειραματική μελέτη, τα αποτελέσματα είναι ποιοτικά αξιόπιστα και ανταποκρίνονται στην πραγματικότητα.

4 Αλγόριθμος online επεξεργασίας

Η μεθοδολογία που προτείνεται για την επίλυση του προβλήματος, αποτελείται από τρία στάδια επεξεργασίας (βλ. Σχήμα 1):

- Ανίχνευση συζητήσεων (*TopicClustering*).



(α′) Χρονική στιγμή τ (β′) Χρονική στιγμή $\tau + \beta$ (γ′) Χρονική στιγμή $\tau + 2\beta$

Σχήμα 2: Περιοχές κάλυψης για διαδοχικές μετακυλίσεις του παραθύρου

- Ανίχνευση δημοφιλών συζητήσεων ($PopularityFiltering$).
- Εντοπισμός περιοχών κάλυψης δημοφιλών συζητήσεων και μελέτη της εξέλιξής τους ($CoverageDiscovery$).

4.1 Ανίχνευση συζητήσεων

Στο στάδιο αυτό εντοπίζονται ομάδες πρόσφατων μηνυμάτων, των οποίων το γεωγραφικό στίγμα αντιστοιχεί σε ένα συγκεκριμένο κελί και όλα εντάσσονται στην ίδια συζήτηση. Για κάθε νεοεισερχόμενο μήνυμα s πραγματοποιείται μία επαυξητική (bottom-up) συσταδοποίηση των μηνυμάτων σε κάθε κελί g του καννάβου G βάσει ομοιότητας των ετικετών (hashtags) πάνω από το κατώφλι θ. Η διαδικασία πραγματοποιείται μόνο για τις συστάδες που εντοπίστηκαν στο τρέχον χρονικό πλαίσιο, και για το κελί όπου το νεοεισερχόμενο μήνυμα δεικτοδοτήθηκε. Επομένως, σε κάθε κελί τηρούνται οι συστάδες (συζητήσεις) που εντοπίστηκαν σε κάθε πλαίσιο του παραθύρου. Η διαμέριση του χώρου σε ομοιόμορφα κελιά μάς επιτρέπει η συσταδοποίηση να γίνεται για μηνύματα που βρίσκονται σε κοντινή απόσταση μεταξύ τους εξασφαλίζοντας την χωρική εγγύτητα.

4.2 Ανίχνευση δημοφιλών συζητήσεων

Οι συζητήσεις που έχουν εντοπιστεί σε κάθε κελί g του καννάβου φιλτράρονται με μία προκαθορισμένη τιμή δημοφιλίας ϕ ώστε να ανιχνεύονται οι δημοφιλείς συζητήσεις σε όλη την έκταση του καννάβου. Αρχικά, θα πρέπει να ενημερωθεί η πληθικότητα των συζητήσεων στο εκάστοτε κελί όπου ανήκουν, για όλο το εύρος του παραθύρου. Επομένως, διαγράφονται οι συζητήσεις που εντοπίστηκαν στο παλαιότερο πλαίσιο που εκπίπτει από το παράθυρο και στην συνέχεια συγχωνεύονται οι ταυτόσημες συζητήσεις που βρέθηκαν στα επόμενα πλαίσια, προκειμένου να υπολογιστεί η συνολική πληθικότητά τους σε όλο το εύρος ω του παραθύρου.

Στην συνέχεια, οι συζητήσεις που ενημερώθηκαν, φιλτράρονται σύμφωνα με την *ελάχιστη τιμή δημοφιλίας* ϕ. Η ελάχιστη τιμή δημοφιλίας εκφράζει το ποσοστό επί του συνολικού αριθμού μηνυμάτων του παραθύρου ο οποίος απαιτείται για να χαρακτηριστεί μια συζήτηση δημοφιλής εντός ενός μεμονωμένου κελιού g. Οι δημοφιλείς συζητήσεις που εντοπίζονται σε κάθε κελί, εισάγονται σε ένα γενικό ευρετήριο κατά λεξικογραφική σειρά, αναγράφοντας τα κελιά στα οποία κάθε τέτοια συζήτηση είναι δημοφιλής, για την διευκόλυνση στο επόμενο στάδιο ανανεώσεων και αναζητήσεων.

4.3 Εντοπισμός περιοχών κάλυψης δημοφιλών συζητήσεων και μελέτη της εξέλιξής τους

Για κάθε δημοφιλή συζήτηση που τηρείται στο γενικό ευρετήριο, υπολογίζονται υποσύνολα κελιών τα οποία σχηματίζουν περιοχές κάλυψης (λ.χ. η μπλέ περιοχή στο Σχήμα 2β′). Οι περιοχές κάλυψης υπολογίζονται αποδοτικά, διότι αναμένεται να είναι μικρό το πλήθος των κελιών που τηρούνται για κάθε συζήτηση.

Για την μελέτη της εξέλιξής τους (εξάπλωση ή συρρίκνωση) με την πάροδο του χρόνου, ουσιαστικά συγκρίνουμε το πλήθος των κελιών της εκάστοτε δημοφιλούς συζήτησης με εκείνο που είχε στην ακριβώς προηγούμενη μετακύλιση του παραθύρου.

Έτσι, στο Σχήμα 2γ΄, στην αμέσως επόμενη χρονική στιγμή $\tau + 2\beta$, η συζήτηση στην μπλέ περιοχή έγινε δημοφιλής στα καινούργια κελιά $\{g_8, g_9\}$, ενώ έπαψε να είναι δημοφιλής στα κελιά $\{g_1, g_3\}$. Τέλος, σε κάθε κύκλο εκτέλεσης, η έξοδος περιλαμβάνει τις δημοφιλείς συζητήσεις με τις περιοχές κελιών όπου εντοπίστηκαν, μαζί με στατιστικά στοιχεία για την εκτίμηση της *απήχησής* τους (πλήθος σχετικών μηνυμάτων).

5 Πειραματική αξιολόγηση

Ο αλγόριθμος υλοποιήθηκε στην γλώσσα προγραμματισμού C++ και δοκιμάστηκε πειραματικά. Τα δεδομένα αποτελού-νται από 807.479 εγγραφές και αφορούν χρονική περίοδο δέκα μηνών στην ευρύτερη περιοχή του Λονδίνου, με μέσο αριθμό 2.794 δημοσιεύσεων ανά ημέρα. Είναι προφανές ότι τέτοιος όγκος δεδομένων είναι πολύ μικρός ώστε να μπορεί να χαρακτηριστεί ως ρεύμα, γι΄ αυτόν τον λόγο προχωρήσαμε στην πλασματική επαύξηση του αρχείου. Πιο συγκεκριμένα, κάθε πρωτογενές μήνυμα επαναλήφθηκε 9, 19, 49 και 99 φορές, οπότε προέκυψαν συνολικά 10-, 20-, 50-, 100-πλάσια μηνύματα που αποθηκεύτηκαν σε αντίστοιχα χωριστά αρχεία. Αυτό έγινε προκειμένου να αξιολογήσουμε την αποδοτικότητα του αλγορίθμου για κλιμακούμενους όγκους δεδομένων. Μελετήθηκαν οι επιδόσεις στον χρόνο επεξεργασίας για τις εξής παραμέτρους:

- Πλήθος κελιών $g \times g$ καννάβου: 10×10, 20×20, 30×30, 50×50, 75×75, 100×100.
- Εύρος ω χρονικού παραθύρου: 2h, 4h, 12h, 24h.
- Βήμα β μετακύλισης του χρονικού παραθύρου: 1h, 2h, 4h, 8h.
- Κατώφλι ομοιότητας θ για την ανίχνευση συζητήσεων: 0.3, 0.5, 0.6, 0.8.
- Ελάχιστη δημοφιλία ϕ μιας συζήτησης εντός μεμονωμένου κελιού: 0.02%, 0.03%, 0.05%, 0.1%, 0.2%, 0.3%, 0.5%.

Από την πειραματική αξιολόγηση προέκυψαν τα εξής συμπεράσματα:

– Η κλιμάκωση του όγκου των πρωτογενών δεδομένων επιβαρύνει κυρίως την φάση της *ανίχνευσης συζητήσεων*, αφού κάθε νεοεισερχόμενο μήνυμα συγκρίνεται με τις προϋπάρχουσες συζητήσεις. Πράγματι, το κόστος εκτέλεσης της φάσης αυτής είναι γραμμικό ως προς το πλήθος των δεδομένων ανά παράθυρο. Η επιβάρυνση από την φάση *ανίχνευσης δημοφιλών συζητήσεων* είναι ανεπαίσθητη, διότι με την αύξηση του όγκου των δεδομένων, δεν αυξάνεται το πλήθος των διαφορετικών συζητήσεων εντός κελιών, αλλά μόνο το πλήθος των μηνυμάτων που απαρτίζουν την εκάστοτε συζήτηση.

– Ο χρόνος εκτέλεσης για την *ανίχνευση δημοφιλών συζητήσεων* εξαρτάται από το πλήθος των κελιών του καννάβου G, αφού χρειάζεται να διατρέξει όλα τα κελιά του καννάβου, για καθεμιά από τις διαφορετικές συζητήσεις που εντοπίζονται σε όλη την περιοχή παρακολούθησης.

– Η επιβάρυνση λόγω της φάσης *εντοπισμός χωρικής κάλυψης* είναι εντελώς αμελητέα, αφού για κάθε συζήτηση αρκεί να εξετάσει ένα μικρό πλήθος κελιών σε σύγκριση με τον όγκο των δημοσιευμένων μηνυμάτων.

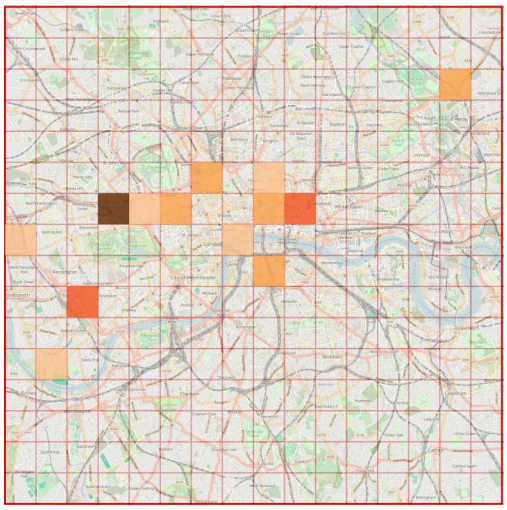
– Κάθε κύκλος εκτέλεσης ολοκληρώνεται εντός δευτερολέπτων (μόνο για χαμηλό ϕ παρατηρούνται χρόνοι περίπου 2 sec) και στις περισσότερες περιπτώσεις σε δέκατα του δευτερολέπτου για μεγάλο όγκο νεοεισερχόμενων μηνυμάτων. Επομένως, μπορούμε να ισχυριστούμε πως ο αλγόριθμος που αναπτύχθηκε εκτελείται και εξάγει αποτελέσματα σε πραγματικό χρόνο.

– Επίσης, τα ποιοτικά αποτελέσματα που προκύπτουν θεωρούνται ρεαλιστικά. Ενδεικτικά, στο Σχήμα 3 απεικονίζεται η χωρική έκταση (στο κέντρο του Λονδίνου) της δημοφιλούς συζήτησης #tubestrike που αναφέρεται σε απεργία του μετρό. Οι διαφορετικές αποχρώσεις των κελιών που διακρίνονται, αναπαριστούν την διαφορετική ένταση (πλήθος μηνυμάτων) της συζήτησης αυτής σε κάθε κελί.

6 Συμπεράσματα – Προοπτικές

Από την υλοποίηση του αλγορίθμου προκύπτουν τα εξής συμπεράσματα:

- Ο κάνναβος G που χρησιμοποιήθηκε στην περιοχή μελέτης αποδείχθηκε ικανοποιητικός για την ανίχνευση δημοφιλών συζητήσεων και την εποπτεία τους στο χώρο και στο χρόνο. Παρ΄ όλα αυτά, η χωρική κάλυψη των συζητήσεων είναι προσεγγιστική και εντοπίζεται σε επίπεδο κελιών. Επιπλέον, η διακριτοποίηση του



Σχήμα 3: *Συζήτηση #tubestrike στις 30 Απριλίου 2014 ώρα 22:00.*

καννάβου διαδραματίζει σημαντικό ρόλο στην εύρεση συζητήσεων, καθώς μεταβάλλοντας το πλήθος των κελιών είναι πιθανόν να χάνονται ορισμένες συζητήσεις, εξαιτίας της φύσης της ελάχιστης τιμής δημοφιλίας που έχει οριστεί.

- Η συνάρτηση ομοιότητας για την ανίχνευση συζητήσεων, ουσιαστικά συγκρίνει την ομοιότητα δύο συνόλων (σύνολο ετικετών). Επομένως, κατά την σύγκριση, πραγματοποιείται ταύτιση των ετικετών και όχι σύγκριση χαρακτήρων. Για παράδειγμα, όπως φάνηκε και από τα πειραματικά αποτελέσματα, οι δύο ετικέτες $H_1 = \{\#tubestrike\}$, $H_2 = \{\#tubestrikes\}$ δημιουργούν δύο διαφορετικές συζητήσεις. Κάτι τέτοιο επηρεάζει τόσο το πλήθος, όσο και την ένταση των ανιχνευμένων δημοφιλών συζητήσεων.

- Οι επιδόσεις που μετρήθηκαν πειραματικά, αποδείχθηκαν ιδιαίτερα επαρκείς για την επεξεργασία μεγάλου όγκου δεδομένων σε πραγματικό χρόνο. Επιπλέον, τα ποιοτικά αποτελέσματα που προέκυψαν, εκτιμάται ότι δεν απέχουν από την πραγματικότητα.

Από την μελέτη του συγκεκριμένου προβλήματος προκύπτουν ενθαρρυντικές προοπτικές επέκτασής του. Συγκεκριμένα, στην εργασία αυτή ελήφθησαν υπ΄ όψιν μόνο τα μηνύματα που διαθέτουν γεωγραφικό στίγμα. Όμως, μόνο το 2% του συνολικού πλήθους μηνυμάτων που δημοσιεύονται καθημερινά, διαθέτουν γεωγραφικές συντεταγμένες. Επομένως, η προτεινόμενη μέθοδος αγνοεί σημαντικό όγκο πληροφορίας. Θα ήταν λοιπόν δυνατόν, για τα μηνύματα που δεν διαθέτουν γεωγραφικό στίγμα, να πραγματοποιείται κειμενική ανάλυση (text analysis) όλου του περιεχομένου τους, προκειμένου να γίνεται μία εκτίμηση για την τοποθεσία που δημοσιεύθηκε το μήνυμα (π.χ. #Syntagma). Επιπλέον, οι τιμές των παραμέτρων θα ήταν δυνατόν να επιλέγονται με βάση τεχνικές μηχανικής μάθησης για αποτελεσματικότερη εποπτεία αντί να δίνονται αυθαίρετες τιμές. Τέλος, η διαδικασία αυτή θα μπορούσε να παραλληλοποιηθεί διαμοιράζοντας σε κάθε επεξεργαστικό κόμβο ένα υποσύνολο των κελιών του καννάβου, εκτελώντας τοπικά την ανίχνευση συζητήσεων και ανταλλάσσοντας τα ενδιάμεσα αποτελέσματα για τον σχηματισμό του συνολικού ευρετηρίου.