ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΕΠΙΚΟΙΝΩΝΙΩΝ, ΗΛΕΚΤΡΟΝΙΚΗΣ
ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ

Κατηγοριοποίηση κειμένων χρησιμοποιώντας το μοντέλο αναπαράστασης γράφων N-γραμμάτων σε υψηλής συχνότητας ροής δεδομένων και εφαρμογές σε μέσα κοινωνικής δικτύωσης

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

Ιωάννη Α. Βιόλου
Διπλωματούχου Ηλεκτρολόγου Μηχανικού &
Μηχανικού Υπολογιστών Ε.Μ.Π.

Επιβλέπουσα:
Θεοδώρα Βαρβαρίγου
Καθηγήτρια Ε.Μ.Π.

Αθήνα: Οκτώβριος 2018

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΕΠΙΚΟΙΝΩΝΙΩΝ, ΗΛΕΚΤΡΟΝΙΚΗΣ
ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ

Κατηγοριοποίηση κειμένων χρησιμοποιώντας το μοντέλο αναπαράστασης γράφων Ν-γραμμάτων σε υψηλής συχνότητας ροής δεδομένων και εφαρμογές σε μέσα κοινωνικής δικτύωσης

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

Ιωάννη Α. Βιόλου
Διπλωματούχου Ηλεκτρολόγου Μηχανικού &
Μηχανικού Υπολογιστών Ε.Μ.Π.

Συμβουλευτική Επιτροπή:

1. Θ. Βαρβαρίγου, Καθ. Ε.Μ.Π. (Επιβλέπουσα)
2. Α. Δουλάμης Επ. Καθ. Ε.Μ.Π.
3. Δ. Ασκούνης Καθ. Ε.Μ.Π.

Εγκρίθηκε από την επταμελή εξεταστική επιτροπή στις 3 Οκτωβρίου 2018.

.....	.....	.....
Θ. Βαρβαρίγου	Α. Δουλάμης	Δ. Ασκούνης
Καθ. Ε.Μ.Π.	Επ. Καθ. Ε.Μ.Π.	Καθ. Ε.Μ.Π.

.....	.....	.....	.....
Ι. Ψαρράς	Σ. Παπαβασιλείου	Α. Σταφυλοπάτης	Ν. Δουλάμης
Καθ. Ε.Μ.Π.	Καθ. Ε.Μ.Π.	Καθ. Ε.Μ.Π.	Επ. Καθ. Ε.Μ.Π.

.....
Ιωάννης Α. Βιόλος

Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Μια σημαντική πρόκληση στην εποχή μας είναι η ταξινόμηση κειμένων σε ροές δεδομένων υψηλής συχνότητας. Σε αυτήν την έρευνα, προτείνουμε ένα καινοτόμο και υψηλής ακρίβειας μοντέλο ταξινόμησης ροής κειμένου, που σχεδιάστηκε με έναν ελαστικό κατανεμημένο τρόπο και είναι ικανό να εξυπηρετεί έναν όγκο ροής δεδομένων που παρουσιάζει διακυμάνσεις συχνότητας. Σε αυτό το μοντέλο ταξινόμησης, τα κείμενα αναπαριστώνται ως γράφοι N-γραμμάτων και η διαδικασία ταξινόμησης πραγματοποιείται χρησιμοποιώντας τεχνικές προεπεξεργασίας κειμένων, μετρικές ομοιότητας γράφων και τεχνικές κατηγοριοποίησης διανυσμάτων, ακολουθώντας το μοντέλο επιβλεπόμενης μηχανικής μάθησης.

Η έρευνα μας περιλαμβάνει την ανάλυση πολλών παραλλαγών του προτεινόμενου μοντέλου και των παραμέτρων του, όπως διαφορετικές αναπαραστάσεις των κειμένων ως γράφοι N-γραμμάτων, μετρήσεις ομοιότητας γράφων και μέθοδοι κατηγοριοποίησης, ούτως ώστε στο τέλος να καταλήξουμε σε ένα μοντέλο που παράγει προβλέψεις με υψηλή ακρίβεια. Δώσαμε ιδιαίτερη σημασία στην αντιμετώπιση της κλιμάκωσης και αποκλιμάκωσης του φόρτου εισροής των κειμένων, της διαθεσιμότητας της υπηρεσίας που παράγει τις προβλέψεις και της έγκαιρης απόκρισης των προβλέψεων για αυτό χρησιμοποιήσαμε το μοντέλο προγραμματισμού Beam. Στο μοντέλο προγραμματισμού Beam, η διαδικασία κατηγοριοποίησης εμφανίζεται ως μια ακολουθία ξεχωριστών εργασιών και διευκολύνει την κατανεμημένη υλοποίηση των πιο απαιτητικών εργασιών. Το προτεινόμενο μοντέλο και οι διάφορες παράμετροι που το συνθέτουν αξιολογούνται πειραματικά και η ροή υψηλής συχνότητας εξομοιώνεται με τη χρήση διαδεδομένων συνόλων δεδομένων, που χρησιμοποιούνται στη βιβλιογραφία για εφαρμογές ταξινόμησης κειμένων.

Το μοντέλο που προτείνουμε εκτείνεται σε πολλά ερευνητικά πεδία και αξίζει να αναφέρουμε επιγραμματικά το κάθε ένα, πώς σχετίζονται με την εργασία μας. Η κατηγοριοποίηση κειμένων είναι ένα ερευνητικό θέμα που έγκειται στα επιστημονικά πεδία της μηχανικής μάθησης και της φυσικής επεξεργασίας γλώσσας, η ροή κειμένων κυμαινόμενης υψηλής συχνότητας ανήκει στο πεδίο των μεγάλων δεδομένων. Τα μεγάλα δεδομένα για να εξυπηρετηθούν χρειάζονται υπολογιστικές υποδομές που προτείνονται από το επιστημονικό πεδίο των υπολογιστικών νεφών. Τέλος, οι εφαρμογές της κατηγοριοποίησης κειμένων στην παρούσα έρευνα θα χρησιμοποιηθεί για να επιλύσουν προβλήματα του πεδίου των μέσων κοινωνικής δικτύωσης.

Θα ξεκινήσουμε με το να παρουσιάσουμε πώς οι τεχνικές επεξεργασίας φυσικής γλώσσας χρησιμοποιούνται για την κατηγοριοποίηση, την συσταδοποίηση και την ανάκτηση κειμένων. Οι τεχνικές θα παρουσιαστούν με χρονολογική σειρά με σκοπό να φανεί η εξέλιξη της σκέψης των ερευνητών και πώς η κάθε τεχνική που προτείνεται έρχεται να επιλύσει προβλήματα ή να βελτιώσει τις προηγούμενες. Θα συνεχίσουμε με το να παρουσιάσουμε τις ιδιότητες που πρέπει να πληροί μια κατηγοριοποίηση ή συσταδοποίηση για να θεωρείται καλή, καθώς και ένα σύνολο από μετρικές που ποσοτικοποιούν την ακρίβεια μιας κατηγοριοποίησης σύμφωνα με αυτές τις ιδιότητες. Θα παρουσιαστεί η μέθοδος διεξαγωγής πειραμάτων κατηγοριοποίησης, που εφαρμόζουν αυτές τις μετρικές, η οποία θα είναι η μέθοδος αξιολόγησης που θα χρησιμοποιηθεί σε όλα τα πειραματικά σύνολα που θα παρουσιάσουμε στις επόμενες ενότητες.

Θα παρουσιαστούν σε δύο διαφορετικές ενότητες, μια μέθοδος κατηγοριοποίησης κειμένων και μια συσταδοποίησης, που κάνουν χρήση του μοντέλου αναπαράστασης γράφων N-γραμμάτων. Μια σειρά από

προβλήματα του χώρου των μέσων κοινωνικών δικτύων, θα παρουσιαστούν σε συνδυασμό με αντιπροσωπευτικές μεθόδους που χρησιμοποιούνται για την επίλυσή τους. Θα προτείνουμε την μέθοδο με την οποία το μοντέλο κατηγοριοποίησης κειμένων εφαρμόζεται, θα το επιβεβαιώσουμε και θα το αξιολογήσουμε πειραματικά και θα δούμε πως πολλές φορές ξεπερνάει σε ακρίβεια άλλες μεθόδους που χρησιμοποιούνται. Οι εφαρμογές του χώρου των μέσων κοινωνικών δικτύων όπου θα εφαρμοστεί το μοντέλο που προτείνουμε είναι η αναγνώριση κοινοτήτων, αναγνώριση γεγονότων, συναισθηματική ανάλυση και τα συστήματα συστάσεων.

Λέξεις κλειδιά: κατηγοριοποίηση κειμένων, συσταδοποίηση κειμένων, ροή κειμένων, γράφοι N-γραμμάτων, BEAM, υπολογιστικό νέφος, ανάλυση κοινωνικών δικτύων, συναισθηματική ανάλυση, αναγνώριση κοινοτήτων, αναγνώριση γεγονότων, συστήματα συστάσεων

Abstract

A prominent challenge in our information age is the classification over high frequency data streams. In this research, we propose an innovative and high-accurate text stream classification model that is designed in an elastic distributed way and is capable to service text load with fluctuated frequency. In this classification model, text is represented as N-Gram Graphs and the classification process takes place using text preprocessing, graph similarity and feature classification techniques following the supervised machine learning approach.

The work involves the analysis of many variations of the proposed model and its parameters, such as various representations of text as N-Gram Graphs, graph comparisons metrics and classification methods in order to conclude to the most accurate setup. To deal with the scalability, the availability and the timely response in case of high frequency text we employ the Beam programming model. Using the Beam programming model the classification process occurs as a sequence of distinct tasks and facilitates the distributed implementation of the most computational demanding tasks of the inference stage. The proposed model and the various parameters that constitute it are evaluated experimentally and the high frequency stream emulated using many datasets that are commonly used in the literature for text classification.

The model we propose extends to many research fields and it is worth mentioning each of them how they relate to our work. Text categorisation is a research topic that lies in the scientific fields of machine learning and natural language processing, high frequency data streams belongs to the field of big data. To service big data in an efficient and efficacy way we need computer infrastructures proposed by the scientific field of cloud computing. Finally, the text categorisation applications will be used to solve challenges in the discipline of social network analysis.

We discuss how natural language processing techniques are used to categorise, cluster and retrieve texts. The techniques will be presented in chronological order in order to show the evolution of researchers' approaches and how each technique proposed comes to solve problems or improve the previous ones. We present the properties that a categorisation or clustering must meet to be considered good as well as a set of metrics that quantify the accuracy of a categorisation according to these properties. We also present a method of conducting categorisation experiments applying these metrics. This method will be the evaluation method to be used in all the experimental sets that we will present in the following sections.

A method of text categorisation and a text clustering that use the N-Gram graph representation model are presented in two different sections. A number of social networking topics are presented and we propose that the text categorisation model which use the representation model of N-Gram graph provides efficient solutions. We evaluate our model experimentally and we see that many times it overcomes other state of the art methods. The social networking applications where the proposed model is applied are topics community detection, event detection, sentiment analysis, and recommendation systems.

Keywords: Text classification, Text clustering Text streaming, N-gram graph, Beam, Cloud Computing, social media analytics, sentiment analysis, communities detection, event detection, recommendation system

Περιεχόμενα

Ευρετήριο Σχημάτων

Ευρετήριο Πινάκων

1. Εισαγωγή

Η ταξινόμηση κειμένων είναι μια εποπτευόμενη τεχνική μηχανικής εκμάθησης που χρησιμοποιείται συχνά στο πλαίσιο πολλών εφαρμογών, όπως η ανίχνευση γεγονότων [1] και η συναισθηματική ανάλυση [2]. Τα μοντέλα γνώσης βασίζονται σε δεδομένα με τα οποία εκπαιδεύτηκαν για τη λήψη αποφάσεων σχετικά με την κατηγοριοποίηση νεοαφιχθέντων κειμένων. Οι ροές κειμένου συνήθως περιλαμβάνουν συνεχόμενα κείμενα μικρού μεγέθους, τα οποία μπορούν να αποστέλλονται ταυτόχρονα ή με υψηλή συχνότητα σε μια υπηρεσία η οποία εκτελεί συνεχή επεξεργασία σε μικρό χρόνο απόκρισης. Σε αυτό το πλαίσιο, η κατηγοριοποίηση κειμένων από ένα σημείο εξυπηρέτησης υπό πραγματικές απαιτήσεις χρόνου, μπορεί να προκαλέσει προβλήματα συμφόρησης, που θα έχουν συνέπειες στην διαθεσιμότητα της υπηρεσίας και την ακρίβεια των προβλέψεων. Για να ξεπεράσουμε αυτές τις δυσκολίες χρησιμοποιούμε ελαστικά κατανεμημένες υπολογιστικές υποδομές, νέα μοντέλα αναπαράστασης δεδομένων και εξετάζουμε καινοτόμους αλγόριθμους κατηγοριοποίησης [3] σε αντίθεση με τις παραδοσιακές προσεγγίσεις που βασίζονται σε σύνολα δεδομένων σταθερού μεγέθους.

Η πλειονότητα των εφαρμογών που επεξεργάζονται ροές κειμένου υπόκεινται στους ακόλουθους τέσσερις βασικούς περιορισμούς: τα δεδομένα επεξεργάζονται μία φορά, η απόκριση γίνεται σε πραγματικό χρόνο, υπάρχει αυστηρός περιορισμός στους διαθέσιμους υπολογιστικούς πόρους και τα θέματα που διαπραγματεύονται τα κείμενα εξελίσσονται συναρτήσει του χρόνου [4]. Η θεματική εξέλιξη των κειμένων που επεξεργαζόμαστε έχει τραβήξει την προσοχή των ερευνητών και πολλά μοντέλα έχουν προταθεί για την αντιμετώπισή τους [5]. Σε αυτή την έρευνα προτείνουμε μια μέθοδο κατηγοριοποίησης ροής κειμένων που χρησιμοποιεί το μοντέλο αναπαράστασης γράφων Ν-γραμμάτων και σχεδιάστηκε με ένα κλιμακούμενο, ασύγχρονο και αξιόπιστο τρόπο χρησιμοποιώντας το προγραμματιστικό μοντέλο Beam [6] και παρέχει προβλέψεις υψηλής ακρίβειας, σε πραγματικό χρόνο.

Οι περισσότεροι αλγόριθμοι ταξινόμησης κειμένου χρησιμοποιούν το μοντέλο του σάκου λέξεων σε συνδυασμό με μοντέλα πιθανοτήτων Bayes [7], νευρωνικά δίκτυα [8] και τεχνικές που περιλαμβάνουν πολλών διαστάσεων υπερεπίπεδα [9]. Ένα διαφορετικό μοντέλο αναπαράστασης για τις ανάγκες ταξινόμησης κειμένου είναι οι γράφοι Ν-γραμμάτων (ΓΝΓ). Σε αυτές τις αναπαραστάσεις με γράφους, ένας κόμβος αντιπροσωπεύει ένα Ν-γραμμα που υπάρχει στο αρχικό κείμενο και μια ακμή συνδέει τα γειτονικά Ν-γράμματα. Η συχνότητα γειτνίασης των Ν-γραμμάτων μπορεί να αναπαρίσταται ως βάρη στις ακμές των γράφων. Η απεικόνιση κειμένων με το προτεινόμενο μοντέλο αναπαράστασης επιτρέπει την ανάδειξη σημαντικών χαρακτηριστικών του κειμένου που υπερβαίνουν τα χαρακτηριστικά που προσφέρει ένα σύνολο λέξεων και γραμματικών κανόνων. Συγκρίνοντας γράφους Ν-γραμμάτων είμαστε σε θέση να εντοπίσουμε παρόμοια κείμενα ή την θεματική κατηγορία στην οποία ανήκουν.

Για να βρούμε το μοντέλο που παράγει προβλέψεις με την καλύτερη ακρίβεια εξετάζουμε πολλές παραμέτρους και παραλλαγές όπως την χρήση γράφων με βάρη στης ακμές ή χωρίς βάρη, διάφορες τάξεις Ν-γραμμάτων που κυμαίνονται από δύο έως δέκα, τεχνικές προεπεξεργασίας κειμένου, μετρικές ομοιότητας γράφων και κατηγοριοποίησης διανυσμάτων. Επιπλέον, προκειμένου να διασφαλιστεί ότι ελαχιστοποιούνται οι εξαρτήσεις μεταξύ των διαφόρων σταδίων του μοντέλου πρόβλεψης έχει γίνει ένας σχεδιασμός, έτσι ώστε τα στάδια επεξεργασίας να είναι διακριτά και να λειτουργούν με έναν κατανεμημένο, ασύγχρονο και κλιμακούμενο τρόπο.

Το μοντέλο κατηγοριοποίησης κειμένων με γράφους Ν-γραμμάτων συνδυάζει τα οφέλη της ευελιξίας των Ν-γραμμάτων με την καλά δομημένη αναπαράσταση των κατευθυνόμενων γράφων. Κάθε εξαγόμενη αλληλουχία γραμμάτων από ένα κείμενο μπορεί να αναπαρασταθεί ως ένα Ν-γραμμα και η σχέση αυτών των Ν-γραμμάτων μπορεί να αποδοθεί χρησιμοποιώντας ένα γράφο. Το πρόβλημα κατάταξης κειμένων μπορεί να αναχθεί σε μια θεωρία γράφων και ένα πρόβλημα αντιστοίχισης μοτίβων. Η χρήση του μοντέλου ΓΝΓ ξεπερνά κάποιους σημαντικούς περιορισμούς του μοντέλου σάκου λέξεων, όπως η διάταξη λέξεων [10] και η ορθογραφία [11]. Η αναπαράσταση κειμένων με γράφους Ν-γραμμάτων και η σύγκριση μεταξύ τους

μειώνει τις διαστάσεις του χώρου του προβλήματος κατηγοριοποίησης και ως εκ τούτου μειώνει και την πολυπλοκότητα της μεθόδου.

Η έρευνα αυτή έχει δύο βασικούς στόχους: αφενός να προτείνει μια καινοτόμο και υψηλής ακρίβειας τεχνική κατηγοριοποίησης κειμένων και αφετέρου να προτείνει μια κατανεμημένη σχεδίαση του μοντέλου που να προσφέρει την δυνατότητα επεξεργασίας και πρόβλεψης ροής κειμένων σε πραγματικό χρόνο. Πραγματοποιήσαμε την πειραματική αξιολόγηση στα 20NewsGroup και Reuters-21578, τα οποία είναι δύο από τα πιο ευρέως χρησιμοποιούμενα σύνολα δεδομένων κατηγοριοποίησης κειμένων. Αυτά τα σύνολα δεδομένων μετατράπηκαν σε ροή κειμένου χρησιμοποιώντας το μοντέλο pub / sub [12] και δόθηκαν ως είσοδο στο μοντέλο μας. Τα πειραματικά αποτελέσματα επιβεβαίωσαν την αποτελεσματικότητα του θεωρητικά προτεινόμενου μοντέλου. Χρησιμοποιήσαμε την δεκαπλή-αναδιπλώσεων διασταυρούμενη αξιολόγηση (10-fold cross validation) σε συνδυασμό με τις Μίκρο και Μάκρο μετρήσεις αξιολόγησης (Micro and Macro evaluation metrics) για να δείξουμε την ακρίβεια της πρόβλεψης των κατηγοριών. Τα αποτελέσματα δείχνουν ότι το προτεινόμενο μοντέλο είναι πρακτικό, αποτελεσματικό και σε πολλές περιπτώσεις υπερβαίνει τις άλλες μεθόδους κατηγοριοποίησης κειμένων.

2. Τεχνικές Φυσικής Επεξεργασίας Κειμένων

Η έρευνά μας έγκειται στο επιστημονικό πεδίο της φυσικής επεξεργασίας κειμένων. Θεωρούμε σημαντικό πριν παρουσιάσουμε την κατηγοριοποίηση (classification) και συσταδοποίηση (clustering) κειμένων με το μοντέλο αναπαράστασης γράφων N-γραμμάτων, για λόγους πληρότητας και καλύτερης κατανόησης να γίνει μια ανασκόπηση στις βασικές μεθόδους που έχουν προταθεί για την αυτοματοποιημένη κατηγοριοποίηση και ανάκτηση κειμένων. Η παρουσίαση ακολουθεί μια χρονολογική σειρά και έχει σκοπό να δείξει όχι μόνο τα χαρακτηριστικά κάθε μοντέλου αλλά και τον τρόπο που εξελίχθηκε η σκέψη των ερευνητών αυτού του συγκεκριμένου πεδίου.

2.1.Εισαγωγή στις Τεχνικές Φυσικής Επεξεργασίας Κειμένων

Με την εξάπλωση του Internet και την δυνατότητα αποθήκευσης τεράστιου πλήθους κειμένων σε κατανεμημένα υπολογιστικά συστήματα, έχει προκύψει επιτακτική η ανάγκη εύρεσης μεθόδων που μπορούν να διαχειρίζονται την μεγάλη ποσότητα πληροφορίας που παράγεται συνεχώς.

Η αυτοματοποιημένη κατανόηση κειμένων αφορά ένα μεγάλο εύρος εφαρμογών από ψηφιακές συλλογές κειμένων, που κατηγοριοποιούν τα αρχεία τους, έως τις πιο σύγχρονες εφαρμογές των μέσων κοινωνικών δικτύων που ανιχνεύουν τις θεματικές κοινότητες των μελών τους με βάση τα κείμενα που γράφουν και διαβάζουν.

Οι επιστημονικοί κλάδοι της ανάκτησης πληροφορίας και της φυσικής επεξεργασίας γλώσσας έρχονται να μας δώσουν συγκεκριμένα εργαλεία που χρησιμοποιούνται για να οργανώσουμε το μεγάλο πλήθος κειμένων που παράγονται συνεχώς.

Στην παρούσα μελέτη θα δούμε τέσσερις μεθόδους που χρησιμοποιούν μαθηματικά μοντέλα, στατιστικά ή γραμμικής άλγεβρας, για να επεξεργάζονται συλλογές κειμένων συναρτήσει των λέξεων που περιέχει κάθε κείμενο.

Ο πρώτος στόχος που έχουμε είναι να μπορέσουμε κάθε κείμενο να το αναγάγουμε σε κάποιες μεταβλητές που συσχετίζονται έμμεσα ή άμεσα με τις λέξεις που περιέχει. Αυτές οι μεταβλητές θα μπορούσαν επίσης να χρησιμοποιηθούν για να δεικτοδοτήσουν τα κείμενα.

Η κατηγοριοποίηση ή συσταδοποίηση των κειμένων είναι ο δεύτερος μας στόχος. Κείμενα που διαπραγματεύονται παρόμοια θέματα θα πρέπει να βρίσκονται μαζί στην ίδια κλάση ενώ κείμενα που διαπραγματεύονται διαφορετικά θέματα να είναι σε διαφορετικές κλάσεις. Αναλόγως με τις ανάγκες της εφαρμογής που έχουμε να υλοποιήσουμε, μπορούμε να έχουμε κείμενα που ανήκουν σε μία μόνο κατηγορία, σκληρή κατηγοριοποίηση (hard classification), ή κείμενα που ανήκουν σε περισσότερες από μία κατηγορίες, με ένα ποσοστό συμμετοχής σε κάθε μία, μαλακή κατηγοριοποίηση (soft classification).

Τελευταίος μας στόχος είναι, θέτοντας μια λεκτική ερώτηση, το σύστημα μας να είναι σε θέση να μπορεί να ανακτά τα κείμενα, που σε μεγαλύτερο ποσοστό ταιριάζουν και απαντούν σωστά σε αυτή την ερώτηση.

Παρακάτω παρουσιάζουμε αναλυτικά και με χρονολογική σειρά τα μοντέλα που αποτέλεσαν σταθμό στην επίτευξη των παραπάνω στόχων.

2.2.Μοντέλο Διανυσματικού Χώρου

Το μοντέλο διανυσματικού χώρου (Vector Space Model VSM) [13] παρουσιάστηκε το 1975 και είναι ένα αλγεβρικό μοντέλο που αναπαριστά κείμενα ή γενικά οπουδήποτε είδους αντικείμενα με την μορφή διανυσμάτων. Το VSM μπορεί να χρησιμοποιηθεί για να καλύψει ένα μεγάλο εύρος αναγκών, όπως την συσταδοποίηση κειμένων (document clustering), το φιλτράρισμα πληροφορίας (information filtering), την ανάκτηση πληροφοριών (information retrieval) και τα αυτοματοποιημένα ευρετήρια (automatic indexing).

Η κεντρική ιδέα είναι η ακόλουθη. Αρχικά αναπαριστούμε κάθε κείμενο ως ένα διάνυσμα μέσα σ' ένα διανυσματικό χώρο. Έπειτα συγκρίνοντας δύο διανύσματα είμαστε σε θέση να ξέρουμε κατά πόσο τα δυο αρχικά κείμενα που ανταποκρίνονται σ' αυτά τα διανύσματα είναι όμοια ή όχι. Όταν λέμε όμοια, εννοούμε σε τι ποσοστό διαπραγματεύονται παρόμοια θέματα. Αν έχουμε μια συλλογή από n κείμενα που αναπαρίστανται με τα αντίστοιχα τους διανύσματα και μια αίτηση αναζήτησης, το διάνυσμα κείμενου που έχει υψηλότερη ομοιότητα με το διάνυσμα της αίτησης, αναμένουμε να είναι η απάντηση στην αναζήτηση.

2.2.1.Απλή διανυσματική αναπαράσταση κειμένων

Κάθε κείμενο μπορεί να αναπαρασταθεί με την μορφή ενός διανύσματος με την εξής διαδικασία. Αν το πλήθος των όρων $w_{j,i}$ που μας ενδιαφέρει σ' ένα κείμενο d_i είναι t, τότε κάθε κείμενο θα είναι ένα διάνυσμα, που σε κάθε διάσταση του βάζουμε το πλήθος των φορών που υπάρχει ο όρος αυτός μέσα στο αρχικό κείμενο, όπως φαίνεται στην εξίσωση 2.1

$$d_i = \left(w_{1,i}, w_{2,i}, \ldots, w_{t,i}\right) \quad (2.1)$$

Συνήθως ως όρος χρησιμοποιείται μια λέξη αλλά θα μπορούσε να είναι και μια ολόκληρη φράση. Ο τρόπος κατασκευής των διανυσμάτων που περιγράψαμε δεν είναι ο καλύτερος, αλλά είναι ικανοποιητικός για να κατανοήσουμε αρχικά την διαδικασία. Πιο αποδοτικά διανυσματικά μοντέλα αναπαράστασης, βασίζονται στην αρχή ότι η απόσταση όλων των διανυσμάτων που δεν μας ενδιαφέρουν, είναι όσο το δυνατό πιο μακριά από το διάνυσμα που μας ενδιαφέρει. Τέτοιες μεθόδους θα περιγράψουμε στην παράγραφο 2.3

2.2.2.Σύγκριση διανυσμάτων

Μετά την αναπαράσταση των κειμένων σε διανυσματική μορφή, μας ενδιαφέρει να βρούμε μια μετρική σχέση που να μπορεί να ποσοτικοποιήσει την ομοιότητα μεταξύ δύο διανυσμάτων και των αντίστοιχων κειμένων τους. Η εξίσωση 2.2 που ακολουθεί δηλώνει τον βαθμό ομοιότητας δύο διανυσμάτων.

$$cos\theta = \frac{d_2 \cdot q}{\left\|d_2\right\| \cdot \|q\|} \quad (2.2)$$

Όπου $d_2 \cdot q$ είναι το εσωτερικό γινόμενο των διανυσμάτων d_2 και q. $\|q\|$ είναι η νόρμα του διανύσματος q. Ως γνωστό το συνημίτονο παίρνει τιμές μεταξύ του -1 και του 1. Λόγω του ότι όλες οι διαστάσεις των διανυσμάτων είναι θετικές, αφού δηλώνουν πλήθος φορών που εμφανίστηκε ένα όρος, το αποτέλεσμα θα κυμαίνεται μεταξύ του 0 και του 1 με 0 να δηλώνει ότι τα δύο διανύσματα δεν έχουν τίποτα κοινό και 1 ότι τα δύο διανύσματα συμπίπτουν.

2.2.3.Σύνθετες διανυσματικές αναπαραστάσεις κειμένων

Στην παράγραφό 2.1 παρουσιάσαμε μια μέθοδο για να αναπαριστούμε ένα κείμενο με την μορφή διανύσματος, απλώς εκχωρώντας σε κάθε διάσταση του διανύσματος το πλήθος των φορών που

συναντήσαμε έναν όρο. Σε αυτή την μέθοδο μπορούμε να εντοπίσουμε αδυναμίες, όπως του ότι είναι αναμενόμενο, κείμενα μεγαλύτερου μεγέθους να επαναλαμβάνουν περισσότερες φορές έναν όρο, από ότι κείμενα μικρότερου μεγέθους. Ή ότι μπορεί κάποιος όρος να υπάρχει σε όλα τα κείμενα οπότε η συνεισφορά του στην μετρική σύγκρισης να μην προσφέρει χρήσιμη πληροφορία. Καθώς θα προσπαθούμε να λύσουμε αυτά τα προβλήματα στις επόμενες παραγράφους, θα διαπιστώσουμε ότι βελτιώνοντας τον συντελεστή σωστής ανάκτησης κειμένων, μειώνουμε την πυκνότητα απεικόνισης των διανυσμάτων των κειμένων στον διανυσματικό χώρο.

Ξεκινάμε με την άποψη ότι δύο κείμενα με παρόμοιους όρους αναπαρίστανται από διανύσματα πολύ κοντά το ένα στο άλλο και ότι η απόσταση μεταξύ δύο κειμένων στον διανυσματικό χώρο είναι αντιστρόφως ανάλογη με την ομοιότητα των διανυσμάτων τους. Όπως καταλαβαίνουμε, ένας ιδανικός διανυσματικός χώρος κειμένων θα έχει κοντά ως συστάδες τα κείμενα που διαπραγματεύονται παρόμοια θέματα και μακριά αναμεταξύ τους τα κείμενα που διαπραγματεύονται διαφορετικά θέματα. Από την άλλη πλευρά, θα προτιμούσαμε τα κείμενα της συλλογής να αναπαρίστανται όσο το δυνατό πιο μακριά το ένα από το άλλο, ούτως ώστε όταν θα γίνει κάποιο ερώτημα να ανακτηθεί το σωστό κείμενο και να μειωθεί η πιθανότητα ανάκτησης κάποιου άλλου γειτονικού κειμένου. Αυτή η έννοια μπορεί να ποσοτικοποιηθεί μέσω της εξίσωσης 2.3

$$F = \frac{\sum_{i=1}^{n} \sum_{\substack{j=1 \\ i \neq j}}^{n} S(D_i, D_j)}{} \quad (2.3)$$

Όπου το $S\left(D_i, D_j\right)$ δηλώνει την ομοιότητα μεταξύ των κειμένων i και j. Ο υπολογισμός της εξίσωσης 2.3 έχει πολυπλοκότητα της τάξης του n^2 όπου n είναι το πλήθος των κειμένων. Μια μικρότερης πολυπλοκότητας εξίσωση είναι η 2.4

$$Q = \sum_{i=1}^{n} S\left(C^*, D_i\right) \quad (2.4)$$

Όπου το C^* είναι το κεντρικό centroid του διανυσματικού χώρου των κειμένων και υπολογίζεται από τα centroid κάθε cluster των διανυσμάτων που προκύπτουν από την εξίσωση 2.5

$$c_j = \frac{1}{m} \sum_{\substack{i=1 \\ D_i \in K}}^{m} d_{ij} \quad (2.5)$$

Η εξίσωση 2.4 έχει πολυπλοκότητα τάξης n.

Η πρώτη βελτίωση για την αναπαράσταση των διανυσμάτων των κειμένων που μπορεί να γίνει, είναι αν σε κάθε διάσταση του διανύσματος βάλουμε αντί για το πόσες φορές συναντήθηκε ο όρος $w_{k,i}$ το πόσες φορές συναντήθηκε ο όρος δια του πόσους όρους είχε το κείμενο. Αυτό τον όρο θα τον συμβολίσουμε με f_i^k όπου k ο όρος και i το κείμενο

Μια δεύτερη βελτίωση της διανυσματικής αναπαράστασης των κειμένων, μπορεί να πραγματοποιηθεί, αν σε κάθε διάσταση του διανύσματος κειμένου, βάλουμε το γινόμενο του f_i^k με έναν όρο που να δηλώνει τον αντίστροφο του σε πόσα κείμενα της συλλογής συναντήθηκε ο όρος αυτός. Ο όρος αυτός συμβολίζεται με IDF_k και δίνεται από την εξίσωση 2.6

$$IDF_k = log \frac{|D|}{|\{d' \in D \,|\, k \in d'\}|} \quad (2.6)$$

Όπου |D| είναι το πλήθος των κειμένων και $|\{d' \in D \,|\, k \in d'\}$ είναι ο αριθμός των κειμένων που περιέχουν τον όρο k.

Εν τέλει σε κάθε διάσταση του διανύσματος κειμένου βάζουμε το γινόμενο $f_i^k \cdot IDF_k$. Η ερμηνεία αυτού του γινομένου είναι ότι οι διαστάσεις του διανύσματος κειμένου θα έχουν μεγάλη τιμή όταν ο όρος στον οποίον αντιστοιχούν υπάρχει πολλές φορές σε αυτό το κείμενο, αλλά υπάρχουν λίγα κείμενα στην συλλογή κειμένων που τον έχουν. Παράδειγμα αν η λέξη "μήλο" υπάρχει 10 φορές στο κείμενο i και κανένα άλλο κείμενο πλην του i δεν έχει την λέξη "μήλο", τότε θα έχουμε υψηλό συντελεστή βάρους στην διάσταση αυτή, αν από την άλλη η λέξη "και" υπάρχει 50 φορές στο κείμενο i και υπάρχει σε άλλα 200 κείμενα από τα 205 της συλλογής κειμένων θα έχει μικρότερο συντελεστή βάρους.

Όταν κάνουμε χρήση του όρου $f_i^k \cdot IDF_k$, παρατηρούμε ότι γενικώς η πυκνότητα των διανυσμάτων μειώνεται σε σχέση με το αν κάναμε χρήση μόνο του όρου f_i^k, αλλά παρατηρείται μεγαλύτερη μείωση της πυκνότητας μεταξύ των κέντρων των συστάδων και μικρότερη μείωση της πυκνότητας των διανυσμάτων κειμένων μέσα στις συστάδες. Οπότε αναλογικά έχουμε μεγαλύτερη διαφορά μεταξύ της πυκνότητας των διανυσμάτων κειμένων που διαπραγματεύονται παρόμοια θέματα, από τα διανύσματα πυκνότητας θεμάτων που διαπραγματεύονται διαφορετικά θέματα και αυτό αντικατοπτρίζεται στις τιμές ακρίβειας (precision) και ανάκλησης (recall) [14] σε πειράματα που έχουνε γίνει [13].

Τα κείμενα μέσα στις συστάδες μοιάζουν λιγότερο αναμεταξύ τους. Το ίδιο και τα κέντρα των συστάδων μοιάζουν λιγότερο αναμεταξύ τους. Ωστόσο η διασπορά των συστάδων αναμεταξύ τους είναι αναλογικά μεγαλύτερη από την διασπορά των κειμένων μέσα στις συστάδες και αυτό αιτιολογεί την βελτίωση του precision και του recall.

Από την άλλη αν δεν πολλαπλασιάσαμε με τον όρο IDF_k αλλά διαιρούσαμε, η πυκνότητα των διανυσμάτων στον χώρο θα μεγάλωνε και το precision και το recall θα μειωνόταν.

Προτάθηκαν και άλλες μέθοδοι για να απεικονίσουμε τα διανύσματα στον χώρο, όπου όλες επιβεβαίωναν ότι όσο μικραίνει η πυκνότητα των διανυσμάτων των κειμένων, τόσο αυξάνει το precision και το recall. Μια από αυτές είναι να πολλαπλασιάσουμε το βάρος κάθε όρου k μέσα σε κάθε συστάδα j με ένα παράγοντα $F_1 \cdot F_2$ όπου $F_2 = 1/NC(k)$ και

$$F_1 = \left| < CF(k) > - CF(k,j) \right| \quad (2.7)$$

$$<\text{CF(k)}> \; = \; 1/p \sum_{j=1}^{p} CF(k,j) \quad (2.8)$$

Το NC(k) δηλώνει τον αριθμό των συστάδων όπου υπάρχει ο όρος k και το CF(k,j) δηλώνει τον αριθμό των κειμένων στο cluster j που υπάρχει ο όρος k και p ο αριθμός των cluster. Να διευκρινίσουμε ότι κάθε cluster έχει άλλη τιμή για F_1, οπότε είναι διαφορετικός ο παράγοντας $F_1 \cdot F_2$. Αν από την άλλη δεν πολλαπλασιάζαμε με το $F_1 \cdot F_2$, αλλά διαιρούσαμε, το precision και το recall θα μειωνόταν.

Η τελευταία βελτίωση της μεθόδου που θα διαπραγματευθούμε, βασίζεται στην ιδέα, ότι κάποιοι όροι μέσα στα κείμενα, έχουν την δυνατότητα να αυξάνουν την ανομοιότητα μεταξύ των κειμένων, οπότε τα διανύσματα των κειμένων, βρίσκονται πιο μακριά αναμεταξύ τους. Από την άλλη, κάποιοι όροι έχουν την ιδιότητα να μειώνουν την ανομοιότητα των κειμένων, με αποτέλεσμα να αυξάνουν την πυκνότητα

διασποράς των διανυσμάτων κειμένων μέσα στον χώρο. Αφαιρώντας τους όρους που αυξάνουν την πυκνότητα διασποράς από τα κείμενα αναμένουμε να αυξήσουμε το Precision και το Recall της μεθόδου.

Εισαγάγουμε τον όρο βαθμό διακριτικότητας (Discrimination Value DV) που ορίζεται από την εξίσωση 2.9

$$DV_k = Q_k - Q \quad (2.9)$$

Όπου Q_k είναι η πυκνότητα του χώρου που καταλαμβάνουν τα διανύσματα κειμένου αφαιρώντας τον όρο k. Q είναι η πυκνότητα του χώρου διανυσμάτων. Δοκιμάζουμε να αφαιρέσουμε διάφορους όρους και τέλος επιλέγουμε να παραλείψουμε τελικά αυτούς που $DV_k > 0$.

2.2.4.Μια κριτική στο μοντέλο διανυσματικού χώρου

Τα πλεονεκτήματα του μοντέλου διανυσματικού χώρου είναι ότι το μοντέλο είναι απλό να κατανοηθεί, να υλοποιηθεί και βασίζεται σε απλές μεθόδους και πράξεις γραμμικής άλγεβρας. Οι όροι κάθε διάστασης δεν είναι απλά δυαδικοί, αλλά παίρνουν μια βαρύτητα που ανταποκρίνεται στην σημασία τους μέσα στο αρχικό κείμενο. Επιτρέπει την εύρεση ομοιότητας μεταξύ ερωτήσεων και κειμένων που μπορεί να έχουν την απάντηση. Επίσης, επιτρέπει να έχουμε ως απάντηση, όχι μόνο το κείμενο που ταιριάζει περισσότερο, αλλά μια σειρά κειμένων, ταξινομημένων με έναν συντελεστή, που να υποδηλώνει, το κάθε κείμενο σε τι ποσοστό ταιριάζει με αυτό το ερώτημα. Αντίστοιχα, επιτρέπει ως απάντηση να ανακτήσει κείμενα, που δεν ταιριάζουν πλήρως στο ερώτημα που τίθεται, αλλά σ' ένα ποσοστό.

Από την άλλη πλευρά, τα μειονεκτήματα της μεθόδου που μπορούν να αναφερθούν, είναι ότι τα μεγάλα κείμενα, δεν αναπαρίστανται καλά διότι έχουν μικρές τιμές ομοιότητας. Αυτό αιτιολογείται στο ότι έχουν μικρό εσωτερικό γινόμενο μεταξύ των διανυσμάτων τους, εξαιτίας του ότι ο χώρος είναι πολυδιάστατος.

Ένα πρόσθετο μειονέκτημα είναι ότι οι λέξεις κλειδιά και οι όροι που χρησιμοποιούμε πρέπει να ταιριάζουν ακριβώς. Παράγωγα λέξεων, οι διαφορετικοί χρόνοι κλήσης του ίδιου ρήματος, θα θεωρηθούν διαφορετικοί όροι, αν και σημασιολογικά συμπίπτουν. Αυτό το πρόβλημα έρχεται να λυθεί μέσω των διάφορων τεχνικών λημματοποίησης (stemming) [15] . Το πρόβλημα όμως που δεν μπορεί να λυθεί ούτε με την τεχνική του stemming, είναι όταν υπάρχουν λέξεις συνώνυμες, όπου η μία μπορεί να χρησιμοποιείται αντί της άλλης. Αυτές οι λέξεις θα δηλωθούν ως δυο διαφορετικοί όροι στα διανύσματα κειμένων αν και σημασιολογικά ταυτίζονται.

Το μοντέλο Vector Space λειτουργεί ως ένα μοντέλο σάκου λέξεων (Bag of Words BoW) [16]. Αυτό σημαίνει ότι για κάθε κείμενο διαχειριζόμαστε την πληροφορία, του ποιες λέξεις περιλαμβάνει και με ποια συχνότητα, αλλά αγνοούμε εντελώς με ποια σειρά εμφανίστηκαν αυτές οι λέξεις. Η σειρά εμφάνισης των λέξεων μέσα σ' ένα κείμενο, είναι μια πληροφορία που δυστυχώς χάνεται.

2.2.5.Γενικευμένο μοντέλο διανυσματικού χώρου

Το γενικευμένο μοντέλο διανυσματικού χώρου (Generalized vector space model GVSM) [17] προτάθηκε το 1985 και είναι μια γενίκευση του VSM. Η μεγάλη βελτίωση που πραγματοποιεί είναι ότι παίρνει υπόψη του και την σχέση όρου με όρο. Σε αυτή την περίπτωση η ομοιότητα ενός κειμένου d_k και μια ερώτησης q δίνεται από την εξίσωση 2.10

$$sim(d_k, q) = \frac{\sum_{j=1}^{n}\sum_{i=1}^{n} w_{i,k} * w_{j,q} * t_i \cdot t_j}{\sqrt{\sum_{i=1}^{n} w_{i,k}^2} * \sqrt{\sum_{i=1}^{n} w_{i,q}^2}} (2.10)$$

Τα $w_{i,k}$ και $w_{j,q}$ όπως και πριν δηλώνουν τα βάρη κάθε όρου και εξαρτώνται από το πόσες φορές υπάρχει αυτός ο όρος στο κείμενο και στην ερώτηση αντίστοιχα. Τα t_i και t_j δίνονται από τον τύπο 2.11

$$\vec{t_i} = \frac{\sum_{k=1}^{r} c_k(t_i)\vec{m_{ik}}}{\sqrt{\sum_{k=1}^{r} c_k^2(t_i)}} \quad (2.11)$$

Το $c_k(t_i)$ δίνεται από τον τύπο 2.12

$$c_k(t_i) = \sum_{a \in I(t_i,k)} a_{ai} \quad (2.12)$$

Όπου $a_{ai} = \begin{cases} 1 \text{ } Αν \text{ } το \text{ } κείμενο \text{ } d_a συσχετίζεται \text{ } με \text{ } τον \text{ } όρο \text{ } t_i \\ 0 \text{ } Αν \text{ } το \text{ } κείμενο \text{ } d_a δεν \text{ } συσχετίζεται \text{ } με \text{ } τον \text{ } όρο \text{ } t_i \end{cases}$

Η ιδέα που εισαγάγει η μέθοδος GVSM, είναι ότι δεν παίρνουμε μόνο υπόψη μας τα βάρη που έχουν να κάνουν με το πόσες φορές βρέθηκε μια συγκεκριμένη λέξη σ' ένα κείμενο και σε μια ερώτηση, αγνοώντας την σχέση διαφορετικών λέξεων. Αλλά μέσω του γινομένου $t_i \cdot t_j$ παίρνουμε υπόψη μας, κατά πόσο δύο όροι ακόμη και διαφορετικοί να είναι, συσχετίζονται αναμεταξύ τους. Αυτή είναι μια έννοια που θα την συναντήσουμε συχνά στις επόμενες ενότητες.

2.3.Λανθάνουσα Σημασιολογική Ανάλυση

Οι χρήστες όταν θέλουν να αναζητήσουν πληροφορίες σε μια συλλογή από κείμενα, ενδιαφέρονται με βάση την εννοιολογική συγγένεια του ερωτήματος, που έχουν σε σχέση με τα κείμενα που υπάρχουν και όχι απλώς τα κείμενα και η ερώτηση τους να μοιράζονται κοινές λέξεις.

Οι κοινές λέξεις είναι κάποιες φορές μια ένδειξη του ότι υπάρχει εννοιολογική συνάφεια μεταξύ κειμένων, αλλά υπάρχει το ενδεχόμενο της πολυσημίας και της συνωνυμίας μια λέξης. Η διαχείριση της συνωνυμίας και της πολυσημίας δεν είναι εφικτή απ' ένα μοντέλο που στηρίζεται μόνο στην συχνότητα λέξεων. Θα ορίσουμε την έννοια της συνωνυμίας και της πολυσημίας και παρακάτω θα δούμε πως η μέθοδος Λανθάνουσα Σημασιολογική Ανάλυση (Latent Semantic Analysis LSA) [18], που προτάθηκε το 1988, μέσω άδηλων (latent) μεταβλητών διαχειρίζεται καλά την συνωνυμία, εν μέρει την πολυσημία και μπορεί με έναν αποτελεσματικό τρόπο να κάνει ανάκτηση πληροφορίας.

Συνωνυμία: Το φαινόμενο κατά το οποίο υπάρχουν πολλοί τρόποι για να αναφερθούμε στο ίδιο αντικείμενο ή έννοια. Γενικώς οι άνθρωποι μέσα σε διαφορετικά πλαίσια, υπό διαφορετικές ανάγκες, κατανόηση του θέματος που τους απασχολεί και γλωσσικές συνήθειες, τείνουν να περιγράψουν την ίδια πληροφορία χρησιμοποιώντας διαφορετικούς όρους. Μάλιστα το ποσοστό που δύο άνθρωποι θα χρησιμοποιήσουν κοινές λέξεις-κλειδιά για να περιγράψουν το ίδιο αντικείμενο ή θέμα είναι λιγότερο από 20% [19]. Το φαινόμενο της συνωνυμίας μειώνει τον συντελεστή recall στις information retrieval μεθόδους.

Πολυσημία: Το φαινόμενο κατά τα οποίο κάποιες λέξεις έχουν περισσότερες από μία εννοιολογικές σημασίες. Σε διαφορετικά συμφραζόμενα ή από διαφορετικούς ανθρώπους η ίδια λέξη π.χ. "Λόγος" έχει διαφορετική σημασία που μπορεί να σημαίνει αιτιολογία, κλασματική σχέση και μέθοδος επικοινωνίας. Οπότε υπάρχει περίπτωση κάποιο κείμενο να περιέχει μια λέξη που περιέχεται σε μια ερώτηση χωρίς να διαπραγματεύεται το ίδιο θέμα με την ερώτηση.

2.3.1.Τρόποι επίλυσης του προβλήματος που εισαγάγει η πολυσημία και η συνωνυμία

Έχουν προταθεί τρόποι για να αντιμετωπιστεί το πρόβλημα της πολυσημίας, αλλά δυστυχώς όχι ικανοποιητικά. Μια κοινή προσέγγιση είναι να χρησιμοποιηθούν λεξικά μ' ένα συγκεκριμένο ελεγχόμενο πλήθος λέξεων. Μόνο από τις λέξεις που έχει το λεξικό θα σχηματίζονται τα ερωτήματα και αυτό το λεξικό θα χρησιμοποιείται ως βάση για την επιλογή των λέξεων που θα χρησιμοποιηθούν έπειτα στα διανύσματα που θα αντιστοιχούν στα κείμενα. Μια άλλη μέθοδος επίλυσης του θέματος είναι να υπάρχουν κάποιοι άνθρωποι ως ενδιάμεσοι για να αποσαφηνίζουν και να αναγάγουν τον κάθε όρο σε άλλους, που το πρόβλημα της πολυσημίας δεν θα υπάρχει. Αμέσως καταλαβαίνουμε ότι η δεύτερη μέθοδος χάνει την έννοια της αυτοματοποιημένης μεθόδου και ότι η πρώτη περιορίζει αρκετά την ευελιξία του ανθρώπινου λόγου χωρίς να είναι απαραίτητα πιο αποδοτική.

Ως επακόλουθο της συνωνυμίας διαπιστώνουμε ότι οι όροι που χρησιμοποιούμε για να δεικτοδοτήσουμε (indexing) και περιγράψουμε ένα κείμενο δεν είναι πλήρης και είναι απλώς ένα μέρος των όρων που μπορούν να χρησιμοποιηθούν. Αυτό οφείλεται στο ότι τα ίδια τα κείμενα δεν περιέχουν όλους τους όρους που θα μπορούσαν οι διάφοροι χρήστες να χρησιμοποιήσουν.

2.3.2.Η έννοια της άδηλης μεταβλητής

Αυτό στο οποίο πρέπει να δοθεί σημασία και στο οποίο βασίζεται η μέθοδος LSA, για να διαφοροποιηθεί από προηγούμενες μεθόδους, ούτως ώστε να αυξήσει την ακρίβεια της, είναι ότι υπάρχουν πολλές λέξεις που ακόμη και αν είναι διαφορετικές, συσχετίζονται αναμεταξύ τους, καθώς βρέθηκαν πολλές φορές μέσα στα ίδια κείμενα. Μπορεί ένα ερώτημα να επιφέρει ως αποτέλεσμα ένα κείμενο που δεν έχει ακριβώς τις ίδιες λέξεις με το ερώτημα αλλά λόγω του ότι οι λέξεις του ερωτήματος θεωρήθηκαν συγγενικές με τις λέξεις του κειμένου επιλέχτηκαν. Η συγγένεια, ο βαθμός σχέσεις μεταξύ των λέξεων παράγεται έμμεσα από το ότι οι λέξεις αυτές βρίσκονται συχνά μαζί στα ίδια κείμενα.

Εδώ θα προσπαθήσουμε να ερμηνεύσουμε την έννοια της άδηλης μεταβλητής (latent variable) που έρχεται να φέρει την λύση στα προηγούμενα προβλήματα.

Πίσω από μια σειρά συγγενικών λέξεων θα προσεγγίσουμε τι πραγματικά λέγεται και θα αντικαταστήσουμε μια σειρά μεταβλητών - λέξεων με μία μεταβλητή, μια άδηλη μεταβλητή που τις συνοψίζει. Αυτή η άδηλη μεταβλητή δεν μπορεί να ερμηνευθεί και να κατανοηθεί άμεσα, απλώς αποδίδει την ύπαρξη κάποιων μοτίβων (patterns) λέξεων και μας δίνει έναν υπαινιγμό ότι αν συναντηθούν κάποιες λέξεις τότε υπονοούνται και κάποιες άλλες.

{(bike), (bicycle), (road)}→{(1.276·bike+0.479·bicycle), (road)}

Παράδειγμα Α΄ Οι λέξεις bike και bicycle πολλές φορές περιλαμβάνονται στα ίδια κείμενα οπότε αντιλαμβανόμαστε ότι σχετίζονται. Αυτή την σχέση που έχουν την αντιστοιχούμε μέσω μιας άδηλης μεταβλητής που είναι ίση με το γραμμικό τους άθροισμα.

Θα χρησιμοποιήσουμε έναν πίνακα που θα δηλώνει την ύπαρξη των όρων που συνήθως είναι λέξεις σε σχέση με μια συλλογή κειμένων που έχουμε. Κάθε γραμμή του πίνακα θα αντιστοιχεί σ' ένα κείμενο και κάθε στήλη σε έναν όρο, αναλόγως με το πόσες φορές θα βρίσκεται αυτός ο όρος στο κείμενο θα αποδίδουμε τον αντίστοιχο αριθμό βαρύτητας. Αν δεν υπάρχει καμία φορά θα βάζουμε μηδέν. Σε αυτό το στάδιο η μέθοδος μας μοιάζει στην VSM που περιγράψαμε πιο πάνω. Διότι φαίνεται ότι κάθε κείμενο αντιπροσωπεύεται από όρους και συγκρίνοντας τους όρους βρίσκουμε την σχέση μεταξύ κειμένων.

Η διαφοροποίηση όπως θα δείξουμε στις επόμενες παραγράφους έρχεται στο ότι ο αρχικός πίνακας θα προσεγγιστεί από έναν μικρότερο όπου οι πολλές άμεσες μεταβλητές που αντιστοιχούν στους όρους θα δώσουν την θέση τους στις άδηλες μεταβλητές. Κατά τα άλλα μπορούμε να συστηματοποιήσουμε τα κείμενα με το να βάζουμε το ένα δίπλα στο άλλο στην δομή του πίνακα με κριτήριο τα κείμενα που μοιάζουν στους όρους τους να είναι κοντά στις αντίστοιχες γραμμές του πίνακα. Εννοείται ότι όπως και στην

μέθοδο VSM μπορούμε να απεικονίσουμε κάθε κείμενο και κάθε ερώτημα σε έναν n-διάστατο χώρο άδηλων μεταβλητών και με μια μέθοδος ομοιότητας όπως του εσωτερικού γινομένου ή του συνημιτόνου να βρίσκουμε όμοια κείμενα.

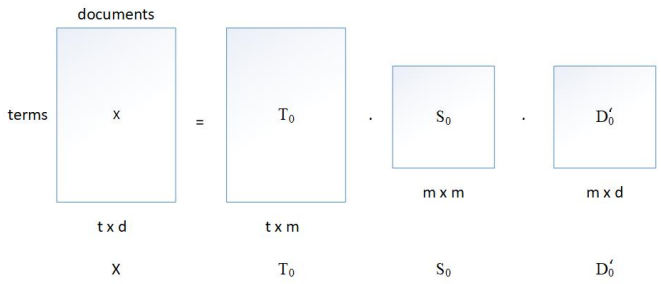
2.3.3.Εφαρμογή της μεθόδου ανάλυση πίνακα σε ιδιάζουσες τιμές

Μετά την απλή κατασκευή του πίνακα κειμένων X σε σχέση με τους όρους που περιλαμβάνουν τον αναλύουμε σε γινόμενο τριών άλλων πινάκων όπως δείχνει η εξίσωση 2.13

$$X = T_0 \cdot S_0 \cdot D_0' \quad (2.13)$$

Αυτοί οι τρεις πίνακες αναλύουν τον πίνακα X σε γραμμικά ανεξάρτητα στοιχεία. Όσα από αυτά είναι πολύ μικρά μπορούν να αφαιρεθούν και έτσι οδηγούμαστε σ' ένα προσεγγιστικό μοντέλο για τον πίνακα X και τα αντίστοιχα κείμενα και όρους που περιέχει.

Αυτή η διαδικασία ανάλυσης ενός πίνακα στο γινόμενο τριών άλλων πινάκων βασίζεται στην μέθοδο γραμμικής άλγεβρας, ανάλυση πίνακα σε ιδιάζουσες τιμές (singular-value-decomposition SVD) [20] και απεικονίζεται στο σχήμα 2.1



Σχήμα 2.1 Ανάλυση πίνακα σε m ιδιάζουσες τιμές

Στο Σχήμα 2.1 βλέπουμε την αναγωγή του ορθογώνιου πίνακα X που συσχετίζει όρους με κείμενα στο γινόμενο των τριών πινάκων T_0 S_0 και D_0 όπου t οι όροι που υπάρχουν στα κείμενα, d το πλήθος των κειμένων και m το πλήθος των singular values. Οι singular values είναι διατεταγμένες κατά φθίνουσα σειρά στον διαγώνιο πίνακα S_0. Οι πίνακες T_0 και D_0 αποτελούνται από τα αριστερά και δεξιά singular vectors αντίστοιχα που προκύπτουν από τον πίνακα X.

Η μέθοδος SVD από την πλευρά των αναγκών ανάκτησης πληροφορίας θεωρείται ως μία τεχνική για να παράγει ένα σύνολο ασυσχέτιστων μεταβλητών από ένα σύνολο συσχετισμένων μεταβλητών. Αυτές οι ασυσχέτιστες μεταβλητές, οι άδηλες μεταβλητές, μπορούν να προσδιορίσουν καλύτερα ένα κείμενο και ταυτόχρονα καταλαμβάνουν λιγότερο χώρο επιλύοντας σε ένα μεγάλο βαθμό το πρόβλημα της συνωνυμίας και σ' ένα μικρότερο της πολυσημίας.

Ο καινούριος πίνακας $\hat{X}$ που προσεγγίζει τον πίνακα X είναι τάξης k<n όπου n η τάξη του αρχικού X. Τα στοιχεία του πίνακα $\hat{X}$ δεν είναι εύκολο να τα ερμηνεύσουμε νοηματικά. Η διαδικασία να πάμε πίσω από τον πίνακα $\hat{X}$ στον X είναι εφικτή αλλά σίγουρα δεν θα γίνει με μεγάλη ακρίβεια.

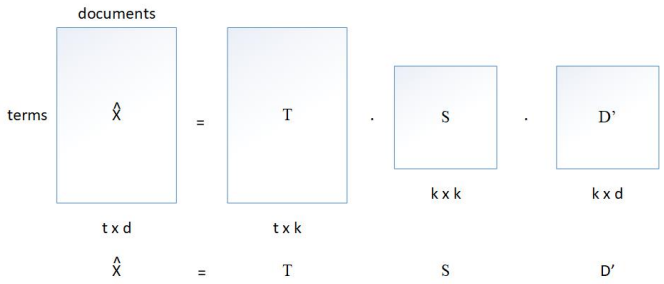
Ο αρχικός πίνακας X είναι διαστάσεων t × d όπου t είναι οι όροι των κειμένων και αντιστοιχούν σε κάθε γραμμή του πίνακα και d είναι το πλήθος των κείμενων και αντιστοιχούν στις στήλες. Οι πίνακες T_0 και D_0 έχουν ορθομοναδιαίες στήλες και ο πίνακας S_0 είναι διαγώνιος. Οι πίνακες T_0 και D_0 αποτελούνται από τα

αριστερά και δεξιά singular vectors αντίστοιχα και ο διαγώνιος S_0 σε κάθε $s_{i,i}$ στοιχείο παίρνει μια singular value. Οι singular values είναι θετικές και διατεταγμένες σε φθίνουσα σειρά σε σχέση με τις γραμμές του πίνακα.

Τα πρώτα k $s_{i,i}$ στοιχεία τα κρατάμε όπως έχουν και τα υπόλοιπα τα θέτουμε ίσον με μηδέν. Το γινόμενο αυτού του καινούριου S πίνακα με τους T και D μας δίνει τον προσεγγιστικό πίνακα $\hat{X}$ όπως φαίνεται από την εξίσωση 2.14

$$X \approx \hat{X} = TSD' \quad (2.14)$$

Ο πίνακας X θα είναι τάξης k και βρίσκεται αφού διαγράψουμε τις γραμμές και τις στήλες που είναι μηδενικές από τους T_0 και D_0 και αποκτήσουμε τους T και D αντίστοιχα. Τα γινόμενα των πινάκων φαίνονται στο σχήμα 2.2.



Σχήμα 2.2 Ανάλυση πίνακα σε k ιδιάζουσες τιμές

Στο σχήμα 2.2 εικονίζεται πως ο διαγώνιος πίνακας S αποτελείται από τις k μεγαλύτερες ιδιάζουσες τιμές του πίνακα S_0. Οι T και D είναι ορθογώνιοι πίνακες που έχουν ως στήλες μοναδιαίου μεγέθους διανύσματα ($T'T = I$) ($D'D = I$) και προκύπτουν μετά την αφαίρεση των γραμμών και στηλών που θα μηδενιστούν αν πολλαπλασιαζόντουσαν από τον S_0 όπου $s_{i,i} = 0 \ \forall \ i > k$

Η τιμή του k έχει μεγάλη σημασία. Πρέπει να είναι αρκετά μεγάλη ούτως ώστε να αντιστοιχεί στην δομή των δεδομένων αλλά και αρκετά μικρή για να μην επηρεάζεται από μη σημαντικές λεπτομέρειες.

2.3.4.Απεικόνιση του μοντέλου LSA σε k διαστάσεων χώρο

Οι k γραμμές που παράχθηκαν από τον πίνακα του μειωμένου SVD μοντέλου μπορούν να θεωρηθούν ως οι συντεταγμένες ενός k-διάστατου χώρου οπότε και κάθε κείμενο να απεικονιστεί σε αυτόν τον χώρο. Αντίστοιχα μπορούμε να υπολογίσουμε το εσωτερικό γινόμενο μεταξύ δυο διανυσμάτων-κειμένων για να βρούμε πόσο ταιριάζουν ή να τα συσταδοποιήσουμε χρησιμοποιώντας έναν αλγόριθμο όπως ο η συσταδοποίηση k-μέσων [21].

2.3.5.Υπολογισμός σχέσεων του μοντέλου LSA

Από τους πίνακες T, S, D μπορούν να υπολογιστούν συντελεστές που συγκρίνουν και δείχνουν κατά πόσο είναι όμοια δύο κείμενα, κατά πόσο συνδέονται δύο όροι και κατά πόσο συνδέεται ένας όρος μ' ένα κείμενο. Προηγούμενες μέθοδοι το πετύχαιναν αυτό συγκρίνοντας γραμμές και στήλες του πίνακα X. Από την στιγμή που ο βαθμός του πίνακα $\hat{X}$ μέσω των πινάκων T, S, D είναι μικρότερος έχουμε το όφελος ότι θα κάνουμε λιγότερους αριθμητικούς υπολογισμούς.

Σύγκριση δύο όρων

Το εσωτερικό γινόμενο μεταξύ των γραμμών του πίνακα $\hat{X}$ δηλώνει το μέγεθος που δύο όροι έχουν παρόμοιο μοτίβο εμφάνισης στο σύνολο κειμένων που εξετάζουμε. Ο πίνακας S είναι διαγώνιος και ο D ορθομοναδιαίος οπότε προκύπτει η σχέση 2.15

$$\hat{X} \cdot \hat{X}^{'} = TS^2T' \text{ (2.15)}$$

Η τιμή που βρίσκεται στην θέση (i,j) του πίνακα $\hat{X} \cdot \hat{X}^{'}$ δηλώνει το πόσο συσχετίζονται οι όροι i και j του πίνακα X. Να θυμίσουμε ότι οι όροι i,j δεν αντιστοιχούν σε λέξεις αλλά είναι άδηλες μεταβλητές που αντιστοιχούν σ' ένα σύνολο λέξεων.

Σύγκριση δύο κειμένων

Αντίστοιχη είναι και η μέθοδος που συγκρίνουμε δύο κείμενα. Εδώ το εσωτερικό γινόμενο το υπολογίζουμε μεταξύ δύο στηλών του πίνακα $\hat{X}$ και προσδιορίζει κατά πόσο δύο κείμενα έχουν τους ίδιους όρους. Κάθε κελί i, j του πίνακας $\hat{X}' \cdot \hat{X}^{'}$ ισούται με το εσωτερικό γινόμενο του κειμένου i με το κείμενο j και δηλώνει το κατά πόσο αυτά τα δύο κείμενα έχουν στοιχεία ομοιότητας. Ο πίνακας $\hat{X}' \cdot \hat{X}^{'}$ δίνεται από τον τύπο 2.16 που ακολουθεί.

$$\hat{X} \cdot \hat{X}^{'} = DS^2D' \text{ (2.16)}$$

Σχέση ενός όρου με ένα κείμενο

Η σχέση ενός όρου μ' ένα κείμενο βρίσκεται ακόμη πιο εύκολα από ότι οι προηγούμενοι συντελεστές, γιατί υπάρχει εγγενώς στον πίνακα $\hat{X}$ όπως τον είδαμε στην σχέση 2.17 και τον παραθέτουμε παρακάτω για λόγους πληρότητας.

$$\hat{X} = TSD' \text{ (2.17)}$$

Κάθε κελί (i,j) δηλώνει κατά πόσο ο όρος i σχετίζεται με το κείμενο j.

2.3.6.Αναπαράσταση ενός ερωτήματος στο μοντέλο LSA

Αφού έχουμε κατασκευάσει τον πίνακα $\hat{X}$ με βάση την συλλογή κειμένων που έχουμε είναι πολύ πιθανόν να μας δοθεί κάποιο καινούριο ερώτημα q το οποίο φυσικά και δεν το πήραμε υπόψη μας όταν φτιάχναμε τον πίνακα $\hat{X}$. Με βάση αυτό το καινούριο ερώτημα q θα πρέπει να βρούμε ποια κείμενα ταιριάζουν πιο πολύ ως απάντηση. Το πρόβλημα είναι ότι το ερώτημα q δεν είναι εκφρασμένο με τους ίδιους όρους που είναι εκφρασμένα όλα τα άλλα κείμενα στον πίνακα $\hat{X}$. Θα πρέπει να το εκφράσουμε πρώτα συναρτήσει των ίδιων άδηλων μεταβλητών και μετά να το συγκρίνουμε με βάση με μια από όλες τις μεθόδους που έχουμε αναφέρει. Το πώς θα εκφράσουμε το ερώτημα q στον καινούριο χώρο των άδηλων μεταβλητών επιτυγχάνεται από την σχέση 2.18

$$D_q = X_q'TS^{-1} \text{ (2.18)}$$

Το D_q είναι ο πίνακας στήλη η αλλιώς διάνυσμα k διαστάσεων που αντιστοιχεί στο ερώτημα q και μπορούμε να το διαχειριστούμε όπως οποιαδήποτε άλλη γραμμή του πίνακα $\hat{X}$. Το X_q είναι ο αρχικός πίνακας γραμμή που αντιστοιχεί στο ερώτημα q και T και S είναι οι πίνακες που προέκυψαν από την SVD ανάλυση και χρησιμοποιήσαμε στον τύπο 3.3.2.

2.3.7.Μια κριτική στην μέθοδο LSA

Η κατανόηση και η προγραμματιστική υλοποίηση της αλγεβρικής μεθόδου SVD είναι η πρώτη δυσκολία που καλείται να αντιμετωπίσει όποιος θέλει να την υλοποιήσει και αν δεν έχει τις κατάλληλη εξοικείωση με τα μαθηματικά μοντέλα που χρειάζονται, θα είναι ένα δύσκολο εγχείρημα. Παρόλα αυτά έχουν προταθεί μέθοδοι όπως η Lanzcos που απλοποιεί την διαδικασία [22].

Στο παράδειγμα Α΄ είδαμε πως οι δύο παρατηρούμενες (observed) μεταβλητές bike και bicycle συνδυάστηκαν σε μία. Από την άλλη περιπτώσεις όπως αυτή του παραδείγματος Α΄ μπορεί να συμβούν.

{(bike), (angle), (road)})}→{(1.276•bike+0.479•angle), (road)}

Παράδειγμα Β΄ Οι λέξεις bike και angle δεν έχουν κάποια σημασιολογική σχέση παρόλα αυτά ενοποιήθηκαν σε μια άδηλη μεταβλητή.

Αστοχίες σαν αυτή του παραδείγματος Β. είναι πιθανόν να συμβούν γιατί η μέθοδος LSA δεν μπορεί να καταλάβει το πραγματικό νόημα των λέξεων αλλά στηρίζεται σε μαθηματικά μοντέλα.

Η μέθοδος LSA αν και έρχεται να δώσει μια λύση στο πρόβλημα της συνωνυμίας που αναφέραμε στην αρχή της ενότητας δυσκολεύεται να διαχειριστεί την πολυσημία. Μια λέξη που έχει πολλά διαφορετικά νοήματα και σε κείμενα που διαπραγματεύονται διαφορετικά θέματα το πιο πιθανόν είναι να ερμηνευθεί ως ένας μέσος όρος των διαφορετικών χρίσεων και ερμηνειών που έχει.

Η μέθοδος LSA όπως και η VSM που περιγράψαμε διαχειρίζεται τις λέξεις σύμφωνα με το μοντέλο BoW και δεν αντιλαμβάνεται την θέση που μπορεί να βρίσκεται μια λέξη στο κείμενο.

Η αδυναμία που οι μέθοδοι LSA και VSM έχουν είναι ότι τα μοντέλα τους θεωρούν ότι οι λέξεις έχουν μια συχνότητα εμφάνισης γκαουσιανής κατανομής. Οι λέξεις όμως στα περισσότερα κείμενα παρουσιάζουν μια εκθετική κατανομή (power law) [23]. Μοντέλα όπως το PLSA που θα μελετήσουμε στην επόμενη ενότητα χρησιμοποιούν πολυωνυμικές κατανομές για να μοντελοποιήσουν την συχνότητα εμφάνισης κάθε λέξης σ' ένα συγκεκριμένο σύνολο λέξεων που εξετάζουμε.

2.4. Πιθανολογική Λανθάνουσα Σημασιολογική Ανάλυση

Η μέθοδος πιθανολογικής λανθάνουσα σημασιολογική ανάλυση (Probabilistic latent sematic analysis PLSA) [24], που προτάθηκε το 1999, έρχεται να διαδεχθεί την LSA μέθοδο διαφοροποιούμενη από την δεύτερη ως το ότι χρησιμοποιεί στατιστικά μοντέλα για να αναλύσει τους παράγοντες – μεταβλητές που προέρχονται άδηλα από τα δεδομένα του συνόλου των κειμένων.

Η PLSA χρησιμοποιεί μια γενικευμένη μορφή του αλγορίθμου Expectation Maximization [25] για να βρει κάθε λέξη με τι πιθανότητα αποδίδεται σε κάθε θέμα και κάθε κείμενο που περιέχει λέξεις με τι πιθανότητα αποδίδεται σε κάθε θέμα.

Η μέθοδος PLSA ταυτόχρονα ορίζει και ένα generative data model όπου έχοντας κατασκευάσει την κατανομή πιθανότητας κάθε λέξης μπορεί πιθανότητα να αναπαράγει κείμενα.

2.4.1.Ορισμός του στατιστικού μοντέλου PLSA

Η μέθοδος PLSA βασίζεται στην ιδέα ότι υπάρχουν άδηλες (latent) μεταβλητές που παράγονται από την συνύπαρξη δεδομένων, στην συγκεκριμένη περίπτωση λέξεων, που βρίσκονται μαζί σε ίδια κείμενα. Οπότε έχουμε παρατηρούμενες μεταβλητές, τις λέξεις που τις συμβολίζουμε με τον τύπο 2.19, τα κείμενα σύμφωνα με τον τύπο 2.20 και τα θέματα που είναι άδηλες μεταβλητές που αποδίδονται από τον τύπο 2.19

$$w \in W = \{w_1,\ w_2, \ldots, w_M\} \text{ (2.19)}$$

Όπου w είναι μία λέξη από το σύνολο όλων των λέξεων και M είναι το πλήθος των λέξεων που συναντήσαμε σε όλα τα κείμενα

$$d \in D = \{d_1,\ d_2, \ldots, d_N\} \quad (2.20)$$

Όπου d είναι ένα κείμενο από το σύνολο όλων των κειμένων και N είναι το πλήθος των κειμένων που βρίσκονται στην συλλογή κειμένων.

$$z \in Z = \{z_1,\ z_2,\ \ldots, z_K\} \quad (2.21)$$

Όπου z είναι μια άδηλη μεταβλητή που αντιστοιχεί σ' ένα πλήθος παρατηρούμενων μεταβλητών w και K είναι το πλήθος των z μεταβλητών που εξάγουμε. Θα μπορούσαμε να πούμε ότι η κάθε μεταβλητή z αντιστοιχεί σ' ένα θέμα που ορίζεται από ένα πλήθος λέξεων w.

Το generative model της μεθόδου PLSA μπορεί να δημιουργηθεί όπως περιγράφουμε παρακάτω

- Επιλέγουμε ένα κείμενο d με πιθανότητα P(d).
- Διαλέγουμε μια άδηλη μεταβλητή z με πιθανότητα P(z|d).
- Παράγουμε μια λέξη w με πιθανότητα P(w|z).

Ως αποτέλεσμα θα έχουμε ένα ζευγάρι (d,w). Η άδηλη μεταβλητή αν και θα έχει χρησιμοποιηθεί δεν θα φαίνεται κάπου στο κείμενο που θα έχει παραχθεί.

Αυτή η διαδικασία φαίνεται στον τύπο 2.22 εκφραζόμενη στην από κοινού πιθανότητα.

$$P(d,w) = P(d) \cdot P(w \mid d) \quad (2.22)$$

Όπου

$$P(w \mid d) = \sum_{z \in Z} P(w \mid z) \cdot P(z \mid d) \quad (2.23)$$

Ουσιαστικά για να παραχθεί η εξίσωση 2.23 πρέπει να αθροίσουμε για όλες τις πιθανές επιλογές της μεταβλητής z. Το μοντέλο που περιγράφουμε είναι ένα στατιστικής ανάμιξης μοντέλο (statistical mixture model) [26] και βασίζεται σε δύο υποθέσεις. Πρώτον ότι τα ζευγάρια (d,w) που υπάρχουν δημιουργούνται ανεξάρτητα πράγμα που αντιστοιχεί στο BoW μοντέλο. Έπειτα ότι οι λέξεις w, που εξαρτώνται από τις άδηλες μεταβλητές z δημιουργούνται ανεξάρτητα για κάθε κείμενο d.

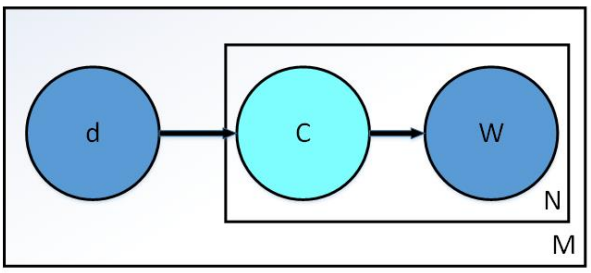
Μια συμμετρική ισοδύναμη εκδοχή του τύπου 2.22 μπορεί να δοθεί στον τύπο 2.24 όπου αντικαταστήσαμε το P(z|d) με την βοήθεια του κανόνα του Μπέυζ.

$$P(d,w) = \sum_{z \in Z} P(z) \cdot P(w \mid z) \cdot P(d \mid z) \quad (2.24)$$

Για να αποφασίσουμε τις πιθανότητες P(d), P(z|d) και P(w|z) στόχος μας είναι να μεγιστοποιήσουμε την λογαριθμική συνάρτηση πιθανότητας L που δίνεται από τον τύπο 2.25

$$L = \sum_{d \in D} \sum\nolimits_{w \in W} n(d,w) \cdot log P(d,w) \quad (2.25)$$

Όπου το n(d,w) δηλώνει την συχνότητα των λέξεων w που συναρτήθηκαν στο κείμενο d.



Σχήμα 2.3 Πιθανολογική Λανθάνουσα Σημασιολογική Ανάλυση

Στο σχήμα 2.3 βλέπουμε την σχέση μεταξύ κειμένων, λέξεων και θεμάτων. Συμβολίζουμε με M το πλήθος κειμένων που υπάρχουν στην συλλογή, d ένα κείμενο, w μια λέξη μέσα στο κείμενο d, N το πλήθος των λέξεων μέσα στο κείμενο d και c το θέμα που αντιστοιχεί στην λέξη αλλά και το πλήθος θεμάτων που αντιστοιχούν σ' ένα κείμενο με βάση της λέξεις που έχει.

2.4.2.Εκπαίδευση του PLSA μοντέλου μέσω του expectation maximization αλγορίθμου

Η διαδικασία που ακολουθείται για να μεγιστοποιήσουμε την πιθανότητα οι άδηλες μεταβλητές z να αντιπροσωπεύουν σωστά την δομή του μοντέλου PLSA είναι να εφαρμόσουμε τον Expectation Maximization (Μεγιστοποίηση Αναμενόμενης Τιμής) αλγόριθμο.

Expectation Maximization αλγόριθμος

Ο Expectation Maximization αλγόριθμος εναλλάσσει δύο βήματα με σκοπό μετά από κάθε επανάληψη οι μεταβλητές του μοντέλου να συγκλίνουν. Τα δύο βήματα είναι:

Expectation (E) βήμα

Η εκ των υστέρων πιθανότητα (posterior probability) υπολογίζεται για τις άδηλες μεταβλητές z σύμφωνα με τις τρέχουσες τιμές των παραμέτρων P(w|z), P(d|z), P(z). Αυτή δίνεται από την εξίσωση 2.26

$$P(z \mid d, w) = \frac{P(z) \cdot P(d \mid z) \cdot P(w \mid z)}{\sum_{z'} P(z') \cdot P(d \mid z') \cdot P(w \mid z')} \quad (2.26)$$

Ο τύπος 4.2.1.1 εκφράζει την πιθανότητα μια λέξη w σ' ένα συγκεκριμένο κείμενο d να προκύπτει λόγο της άδηλης μεταβλητής z.

Maximization (M) βήμα

Εδώ οι παράμετροι ανανεώνονται σύμφωνα με την εκ τον υστέρων πιθανότητα που υπολογίστηκε στο προηγούμενο βήμα και οι τύποι για το P(w|z), P(d|z) και P(z) δίνονται από τις σχέσεις 2.27, 2.28, 2.29 αντίστοιχα.

$$P(w \mid z) = \frac{\sum_{d} n(d, w) \cdot P(z \mid d, w)}{\sum_{d, w'} n(d, w') \cdot P(z \mid d, w')} \quad (2.27)$$

$$P(d|z) = \frac{\sum_w n(d,w) \cdot P(z|d,w)}{\sum_{d',w} n(d',w) \cdot P(z|d',w)} \quad (2.28)$$

$$P(z) = \frac{\sum_{d,w} n(d,w) \cdot P(z|d,w)}{\sum_{d,w} n(d,w)} \quad (2.29)$$

Εναλλάσσοντας την εξίσωση 2.26 με τις 2.27 – 2.29 έχουμε μια διαδικασία που συγκλίνει στην τοπικά μέγιστη τιμή της 2.25.

Tempered Expectation Maximization αλγόριθμος

Ο απλός expectation maximization αλγόριθμος που περιγράψαμε πιο πάνω καταφέρνει να βρει την μέγιστη τοπική πιθανότητα για όλες τις παραμέτρους μειώνοντας την περιπλοκή μεταξύ των λέξεων που υπάρχει.

Σε αυτό το σημείο θα πρέπει να ξεχωρίσουμε το κατά πόσο το μοντέλο μας μπορεί να κάνει σωστές προβλέψεις στα δεδομένα στα οποία το εκπαιδεύομε σε σχέση με το τι απόδοση θα έχει σε καινούρια δεδομένα που θα του δοθούν. Γενικώς σε οποιοδήποτε μοντέλο κατασκευάζουμε δεν θα ήταν σωστό να θεωρήσουμε ότι θα λειτουργήσει απαραίτητα σωστά σε καινούρια δεδομένα απλώς και μόνο γιατί το εκπαιδεύσαμε να έχει καλή ακρίβεια σε κάποιο εκπαιδευτικό σύνολο δεδομένων.

Το να βρούμε συνθήκες υπό τις οποίες η γενίκευση σε καινούρια δεδομένα θα είναι επιτυχημένη είναι το βασικό πρόβλημα που προσπαθεί να επιλύσει η θεωρία στατιστικής μάθησης (Statistical learning theory). Σε αυτή την ενότητα θα παρουσιάσουμε μια γενίκευση του προηγούμενου αλγορίθμου τον Tempered Expectation Maximization (TEM) που συσχετίζεται με την μέθοδο της αιτιοκρατικής ανόπτησης (deterministic annealing) [27]. Το πιο σημαντικό είναι ότι θα εισαγάγουμε την παράμετρο ελέγχου β που τροποποιεί το expectation βήμα, όπως φαίνεται και στην εξίσωση 2.30

$$P_\beta(z|d,w) = \frac{P(z) \cdot [P(d|z) \cdot P(w|z)]^\beta}{\sum_{z'} P(z') \cdot [P(d|z) \cdot P(w|z)]^\beta} \quad (2.30)$$

Παρατηρούμε ότι αν $\beta = 1$ τότε ο TEM αλγόριθμος αναγάγεται στην απλή περίπτωση του EM αλγορίθμου.

Ο αλγόριθμος TEM συγκλίνει σε κάποιες τιμές και λόγω του ότι το β μεταβάλλεται και αποφεύγει τα τοπικά μέγιστα ενώ ταυτόχρονα δεν υπόκειται σε overfitting λάθη. Στα λάθη δηλαδή που προκύπτουν στα στατιστικά μοντέλα που έχουν πολλές μεταβλητές και αντί να περιγράφουν πραγματικές σχέσεις περιγράφουν τυχαία περιστατικά και θόρυβο. Ο αλγόριθμος 2.1 περιγράφει πλήρως την διαδικασία.

	Αλγόριθμος 2.1
1:	Θέτουμε $\beta \leftarrow 1$ και εκτελούμε τον αλγόριθμο EM έως ότου η απόδοση του στα δεδομένα που έχουμε για να το δοκιμάσουμε χειροτερέψει.
2:	Μειώνουμε το β θέτοντας $\beta \leftarrow \eta\beta$ όπου η κάποια παράμετρος $\eta < 1$.
3:	Επαναλαμβάνουμε όσο η απόδοση στα δεδομένα που έχουμε κρατήσει για δοκιμή βελτιώνεται.
4:	Σταματάμε όταν η περεταίρω μείωση του β δεν βελτιώνει το αποτέλεσμα αλλιώς πηγαίνουμε στο βήμα 2.

5 Εκτελούμε κάποιες παραπάνω επαναλήψεις χρησιμοποιώντας αμφότερα τα δεδομένα που είναι για εκπαίδευση και για δοκιμή.

2.4.3.Η γεωμετρία του PLSA μοντέλου

Όπως μπορούσαμε να κάνουμε γεωμετρική απεικόνιση των κειμένων σ' έναν k διαστάσεων χώρο με τις προηγούμενες μεθόδους, έτσι αντίστοιχα μπορούμε να απεικονίσουμε τα κείμενα σε συνάρτηση με τα ποσοστά που ανήκουν σε κάθε θέμα.

Μια ακόμη γραφική απεικόνιση που μπορεί να προκύψει από το μοντέλο PLSA είναι να εικονίσουμε σ' έναν M-1 διαστάσεων χώρο κάθε θέμα $z_1,z_2,\ldots,z_K$ όπου M το πλήθος των λέξεων και K το πλήθος των θεμάτων. Από αυτά τα K σημεία μπορούμε να ορίσουμε το στερεό K-1 διαστάσεων (simplex). Τα βάρη που υπάρχουν για ένα κείμενο P(z|d) αντιστοιχούν στις συντεταγμένες του κειμένου και μπορούν αντίστοιχα να ορίσουν ένα simplex

2.4.4.Μαθηματική σχέση μεταξύ LSA και PLSA μοντέλου

Όπως έχουμε καταλάβει οι μέθοδοι LSA και PLSA βασίζονται στο ίδιο σκεπτικό να μπορέσουμε να περιγράψουμε ένα κείμενο όχι άμεσα από τις λέξεις του αλλά από κάποιες άδηλες μεταβλητές που παράγονται συναρτήσει των λέξεων του. Για αυτή την διαδικασία η μέθοδος LSA χρησιμοποιεί την μέθοδο γραμμικής άλγεβρας SVD ενώ η PLSA ένα πιθανοτικό μοντέλο. Το ερώτημα που τίθεται είναι αν αυτά τα δύο μαθηματικά μοντέλα συνδέονται αναμεταξύ τους. Η απάντηση είναι πως ναι και παρακάτω θα περιγράψουμε πως αυτά τα δύο μοντέλα μπορούν να συσχετιστούν.

Στο μοντέλο LSA είδαμε ότι ο πίνακας που συσχετίζει όρους με κείμενα μπορεί να εκφραστεί ως το διάνυσμα τριών πινάκων αντίστοιχα και στο μοντέλο PLSA μπορούμε να ορίσουμε τους ακόλουθους τρεις πίνακες

$$\hat{U} = (P(d_i \mid z_k))_{i,k} \quad (2.31)$$

$$\hat{V} = (P(w_j \mid z_k))_{j,k} \quad (2.32)$$

$$\hat{\Sigma} = diag(P(z_k))_k \quad (2.33)$$

Τότε το μοντέλο συνδυαστικής πιθανότητας (joint probability model) P(d,w) όπως δόθηκε από την εξίσωση 4.1.6 μπορεί να γραφτεί ως γινόμενο των τριών παραπάνω πινάκων.

$$P = \hat{U} \cdot \hat{\Sigma} \cdot \hat{V}^t \quad (2.34)$$

Παρατηρούμε ότι τα αριστερά και δεξιά ιδιοδιανύσματα του πίνακα από την μέθοδο SVD αντιστοιχούν στους παράγοντες P(w|z) και P(d|z) αντίστοιχα και τα singular values αντιστοιχούν στις τιμές του P(z)

2.4.5.Ανάκτηση πληροφορίας μέσω εφαρμογής ερωτήσεων

Σε όλη την προηγούμενη παρουσίαση της μεθόδου θεωρήσαμε μια συλλογή κειμένων όπου βρίσκουμε κάθε κείμενο σε τι ποσοστό αντιστοιχεί σε κάθε ένα από τα άδηλα θέματα που σχηματίστηκαν, με βάση τις λέξεις που υπάρχουν σε κάθε κείμενο. Ο επόμενος στόχος που θα επιδιώξουμε είναι θέτοντας ένα ερώτημα q που προφανώς δεν είναι ένα από τα κείμενα της συλλογής να ανακτηθούν τα κείμενα που εννοιολογικά θα ταιριάζουν περισσότερο ως απάντηση του. Οπότε σε αυτή την ενότητα θα διαπραγματευθούμε το πώς απεικονίζουμε ένα ερώτημα ή ακόμη και ένα κείμενο που δεν το είχαμε περιλάβει στο σύνολο κειμένων που εκπαίδευσαν την μέθοδο καθώς και το πώς να βρούμε την ομοιότητα του με τα υπόλοιπα κείμενα.

Η απεικόνιση ενός καινούριου κειμένου - ερωτήματος μπορεί να επιτευχθεί με το να εκτελέσουμε κάποιες επαναλήψεις της μεθόδου ΕΜ στις μεταβλητές z ώστε οι πιθανότητες P(z|q) να προσαρμόζονται σε κάθε Maximization βήμα.

Η μέθοδος αυτή θα εκτελεί επαναληπτικά τα βήματα του αλγορίθμου Expectation Maximization και περιμένουμε να μας δώσει ως αποτέλεσμα έναν πίνακα διάνυσμα που θα δηλώνει σε τι ποσοστό το κείμενο θα αναλογεί σε κάθε θέμα.

Αφού έχουμε τον πίνακα διάνυσμα που αντιστοιχεί στο καινούριο κείμενο ή ερώτημα μπορούμε πάλι με μια συνάρτηση ομοιότητας εσωτερικού γινομένου ή συνημιτόνου όπως κάναμε στις προηγούμενες μεθόδους να βρούμε ποιο κείμενο ταιριάζει εννοιολογικά καλύτερα.

2.5.Λανθάνουσα Κατανομή Ντίριχλετ

Εξέλιξη της μεθόδου PLSA είναι η μέθοδος Λανθάνουσα Κατανομή του Ντίριχλετ (Latent Dirichlet Allocation LDA) που προτάθηκε το 2002 και υποθέτει ότι κάθε κείμενο παράγεται ως μια μίξη θεμάτων. Η συνεχής τιμή που αναλογεί για κάθε θέμα δίνεται βάση της Dirichlet συνάρτησης.

2.5.1.Εισαγωγή στην λανθάνουσα κατανομή Ντίριχλετ

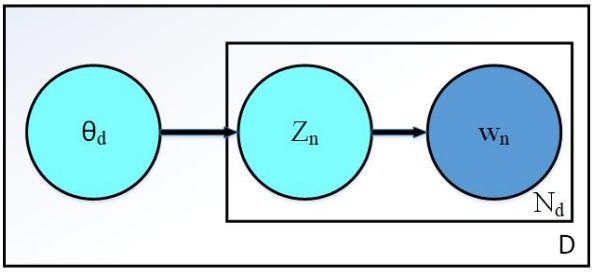
Στην μέθοδο LDA [28] θεωρούμε ότι έχουμε k άδηλα θέματα σύμφωνα με τα οποία τα κείμενα παράγονται. Κάθε θέμα αναπαρίσταται ως μια πολυωνυμική κατανομή πάνω στο |V| πλήθος διαφορετικών λέξεων όλων τον κειμένων. Κάθε κείμενο παράγεται μέσω μια διαδικασίας δειγματοληψίας σ' ένα συνδυασμό από θέματα και έπειτα για κάθε θέμα παράγονται οι λέξεις που πιθανοτικά προκύπτουν.

Πιο συγκεκριμένα ένα κείμενο από N λέξεις w = $<w_1, w_2,\ldots,w_N>$ παράγεται με την ακόλουθη διαδικασία. Αρχικά ο βαθμός που κάθε θέμα i σχετίζεται μ' ένα κείμενο αναφέρεται ως θ_i. Το θ_i προκύπτει από την Dirichlet $(\alpha_1, \alpha_2,\ldots,\alpha_k)$ κατανομή όπου $\theta_i \geq 0$ και $\sum_i \theta_i = 1$. Έπειτα για κάθε λέξη από τις N λέξεις που υπάρχουν στο κείμενο ένα θέμα $z_n \in \{1,\ldots k\}$ παράγεται από την Mult(θ) κατανομή με τέτοιο τρόπο ώστε $p(z_n=i|\theta) = \theta_i$. Στο τέλος κάθε λέξη w_n παράγεται συναρτήσει του z_n θέματος από την πολυωνυμική κατανομή $p(w|z_n)$. Συνοψίζοντας τα παραπάνω σε μια μαθηματική έκφραση έχουμε την σχέση 2.35

$$p(w) = \int_\theta \left(\prod_{n=1}^{N} \sum_{z_n=1}^{k} p\left(w_n \middle| z_n;\beta\right) \cdot p\left(z_n \middle| \theta\right)\right) \cdot p(\theta;\alpha) d\theta \quad (2.35)$$

Όπου p(θ;α) είναι η Dirichlet κατανομή, $p\left(z_n \middle| \theta\right)$ είναι η πολυωνυμική κατανομή παραμετροποιημένη από το θ και $p\left(w_n \middle| z_n;\beta\right)$ είναι η πολυωνυμική κατανομή πάνω στις λέξεις. Το μοντέλο αυτό παραμετροποιείται από τις k διαστάσεων παραμέτρους α = $<\alpha_1, \alpha_2,\ldots,\alpha_k>$ και τον πίνακα β διαστάσεων $k \times |V|$. Ο πίνακας β δηλώνει τις k πολυωνυμικές κατανομές πάνω στις λέξεις.

Στο σχήμα 2.4 βλέπουμε ότι στην LDA μέθοδο, η Dirichlet κατανομή δειγματοληπτείται για κάθε κείμενο για να παράγει το θ και έπειτα για κάθε θέμα του κειμένου η πολυωνυμική κατανομή z δειγματοληπτείται για να παράγει λέξεις.



Σχήμα 2.4 Λανθάνουσα Κατανομή του Ντίριχλετ

Αναλύοντας παραπάνω την σχέση 2.35 θα μπορούσαμε να πούμε ότι οι παράμετροι θ_i προκύπτουν από την Dirichlet κατανομή μέσω του α και οι λέξεις παράγονται από την κατανομή $p(w|\theta)=\sum_{z=1}^{k} p(w \mid z) \cdot p(z \mid \theta)$.

Έτσι το LDA είναι μοντέλο ανάμιξης (mixture model) όπου τα p(w|θ) είναι τα mixture components και τα p(θ;α) δίνουν τα βάρη.

2.5.2.Εκτίμηση παραμέτρων στο μοντέλο LDA

Το πιο σημαντικό πρόβλημα που πρέπει να λύσουμε για να χρησιμοποιήσουμε το LDA μοντέλο είναι αυτό που βρίσκει την εκ των υστέρων κατανομή των άδηλων μεταβλητών που δίνεται από ένα κείμενο και εκφράζονται στη σχέση 2.36

$$p\left(\theta, z \middle| w, \alpha, \beta\right) = \frac{p(\theta, z, w \mid \alpha, \beta)}{p(w \mid \alpha, \beta)} \quad (2.36)$$

Δυστυχώς αυτή η κατανομή στις περισσότερες περιπτώσεις είναι αδύνατο να υπολογιστεί από έναν υπολογιστή. Παρόλα αυτά υπάρχουν προσεγγιστικοί αλγόριθμοι που μπορούν να χρησιμοποιηθούν όπως οι Laplace approximation, variational approximation, και Markov chain Monte Carlo. Παρακάτω θα περιγράψουμε έναν απλό convexity-based variational αλγόριθμο.

Η βασική ιδέα σε αυτή την μέθοδο είναι να κάνουμε χρήση της ανισότητας του Jensen [29] για να αποχτήσουμε ένα κάτω φράγμα των παραμέτρων που μας ενδιαφέρουν. Έτσι θα έχουμε μια οικογένεια από κατανομές με άδηλες μεταβλητές όπως φαίνονται στην εξίσωση 2.37

$$q\left(\theta, z \middle| \gamma, \varphi\right) = q(\theta \mid \gamma) \cdot \prod_{n=1}^{N} q(z_n \mid \varphi_\nu) \quad (2.37)$$

Όπου οι Dirichlet παράμετροι γ και οι πολυγαμικές παράμετροι $(\varphi_1, \ldots, \varphi_N)$ είναι ελεύθερες variational παράμετροι.

Αφού έχουμε παρουσιάσει μια απλοποιημένη μορφή των πιθανοτικών κατανομών, το επόμενο βήμα είναι να εκφράσουμε μια μέθοδο βελτιστοποίησης του προβλήματος που βρίσκει τις παραμέτρους γ και φ. Αυτή η μέθοδος φαίνεται στην σχέση 2.38

$$\left(\gamma^*, \varphi^*\right) = \underset{(\gamma, \varphi)}{argmin}\, D(q(\theta, z \mid \gamma, \varphi) \,||\, p\left(\theta, z \middle| w, a, \beta\right)) \quad (2.38)$$

Οι βέλτιστες τιμές της εξίσωσης 2.38 μπορούν να βρεθούν με το να ελαχιστοποιήσουμε την Kullback-Leibler απόκλιση [30] και να εφαρμόσουμε επαναληπτικά τους τύπους 2.39 και 2.40

$$\varphi_{ni} \propto \beta_{iw_n} \exp\{E_q[\log(\theta_i)]\gamma\} \quad (2.39)$$

$$\gamma_i = \alpha_i + \sum_{n=1}^{N} \varphi_{ni} \quad (2.40)$$

Ο αλγόριθμος 2.2 μπορεί να χρησιμοποιηθεί για να υπολογιστούν αριθμητικά τα γ και φ

	Αλγόριθμος 2.2
1:	Αρχικοποιούμε φ_{ni}^{0}:= 1/k για όλα τα i και n
2:	Αρχικοποιούμε γ_i:= a_i+N/k για όλα τα i
3:	Επανέλαβε
4:	Για n = 1 έως N
5	Για i =1 έως k
6	$\varphi_{ni}^{t+1} := \beta_{iw_n} \cdot \exp(\Psi(\gamma_i^t))$
7	Κανονικοποιούμε φ_{ni}^{t+1} να έχει άθροισμα 1
8	$\gamma^{t+1} := a + \sum_{n=1}^{N} \varphi_n^{t+1}$
9	Έως ότου υπάρχει σύγκλιση

Οι παράμετροι α και β που θέλουμε να βρούμε μεγιστοποιούν την λογαριθμική πιθανότητα που εκφράζεται από την σχέση 2.41

$$l(\alpha, \beta) = \sum_{d=1}^{M} logp(w_d|\alpha, \beta) \quad (2.41)$$

Χρησιμοποιώντας την variational Expectation Maximization διαδικασία που ακολουθεί μπορούμε να βρούμε ένα κάτω φράγμα για τις παραμέτρους γ και φ και έπειτα τα α και β.

Ε βήμα

Για κάθε κείμενο βρίσκουμε της βέλτιστες τιμές των παραμέτρων $\{\gamma_d^*, \varphi_d^* : d \in D\}$. Όπως περιγράψαμε σε προηγούμενη παράγραφο.

Μ βήμα

Μεγιστοποιούμε το χαμηλό φράγμα της λογαριθμικής πιθανότητας όσον αφορά της παραμέτρους του μοντέλου α και β. Αυτό γίνεται με το να βρούμε την μέγιστη εκτίμηση πιθανότητας για κάθε κείμενο με τις κατάλληλες τιμές που προέκυψαν από το Ε βήμα.

2.6.Τεχνικές Κατηγοριοποίησης Ροής Κειμένων

Οι τεχνικές κατηγοριοποίησης κειμένων συνεχούς ροής είναι κυρίως παραλλαγές ή επεκτάσεις των κλασικών μοντέλων κατηγοριοποίησης κειμένων στις οποίες εφαρμόζονται επιπλέον μέθοδοι διαχείρισης της συνεχούς ροής δεδομένων. Η μπεϋζιανή ταξινόμηση κειμένων σε συνδυασμό με μια τεχνική συνεχούς εκμάθησης [7] έχει εφαρμοστεί για την εξαγωγή γνώσης από νεοαφιχθέντα κείμενα. Αντίστοιχα μια συνεχής κατηγοριοποίηση ροής κειμένου που χρησιμοποιεί έναν ταξινομητή SVM έχει προταθεί από [9].

Οι συνδυαστικοί μέθοδοι (ensemble methods) ταξινόμησης ροής κειμένων των συναρτήσεων χρησιμοποιούνται ευρέως λόγω του ότι συνδυάζουν πλεονεκτήματα από διαφορετικές τεχνικές ταξινόμησης [31]. Η Iterative Boosting Streaming μέθοδος [32] προσαρμόζει αυτόματα το μοντέλο ταξινόμησης της σε νέες έννοιες που δημιουργούνται δυναμικά, και χρησιμοποιεί ένα σύνολο από βασικές διαδικασίες μάθησης σύμφωνα με την τρέχουσα ακρίβεια που έχει η κάθε μία. Η εκμάθηση και η ταξινόμηση ροών δεδομένων πολλαπλών ετικετών βασίζεται σε έναν k-nn αλγόριθμο πολλών ετικετών και μια λειτουργία προσαρμογής βάρους [33], η οποία συνδυάζει τις προβλέψεις με έναν αποτελεσματικό και υψηλής ταχύτητας τρόπο. Ένα σύνολο μοντέλων ταξινόμησης για κάθε κατηγορία έχει προταθεί από τους [34] ως μέθοδος αντιμετώπισης του φαινομένου, κατηγορίες που εξαφανίζονται από την είσοδο των δεδομένων για πολύ καιρό και ξανά εμφανίζονται να κατηγοριοποιούνται σωστά.

Τα δέντρα αποφάσεων (Decision trees) έχουν χρησιμοποιηθεί επίσης στην κατηγοριοποίηση ροής κειμένων. Συγκεκριμένα το Hoeffding bound [35] επιλέγει ένα βέλτιστο χαρακτηριστικό διάσπασης στο εύρος ροής δεδομένων. Τα Hoeffding trees δεσμεύουν έναν αριθμό παρατηρήσεων που απαιτούνται για να μετρηθεί η καταλληλότητα ενός χαρακτηριστικού που υπάρχει στη ροή κειμένων και παρουσιάζει καλύτερα αποτελέσματα από τις μη σταδιακές τεχνικές μάθησης που χρησιμοποιούν ένα απεριόριστο πλήθος παρατηρήσεων. Με τον ίδιο τρόπο, τεχνικές Random Forests έχουν ενσωματώσει έναν προσαρμοστικό αλγόριθμο [36] για να αναλύσουν αποτελεσματικά την ροή κειμένων και να αντιμετωπίσουν το ταχέως εξελισσόμενο περιεχόμενο της. Για να το πετύχουν αυτό χρησιμοποιούν μια διαδικασία συσσωρευμένης εκκίνησης και περιορίζουν τις αποφάσεις διαχωρισμού των φύλλων σε ένα υποσύνολο χαρακτηριστικών. Πλατφόρμες λογισμικού που επεξεργάζονται ροές κειμένου είναι επίσης διαθέσιμες όπως το RapidMiner (2018) [37] και το Massive Online Analysis MOA [38] και παρέχουν ένα σύνολο τεχνικών ταξινόμησης, ομαδοποίησης και παλινδρόμησης (Regression) κατά τη ροή δεδομένων. Το MOA παρέχει επίσης δύο ταξινομητές που διαχειρίζονται συγκεκριμένα το φαινόμενο αλλαγής θεμάτων (drifting concept) σε μια ροή κειμένων: Probabilistic Adaptive Windowing και Self-Adjusting Memory.

Η βαθιά εκμάθηση έχει προσελκύσει σημαντική προσοχή [39] τα τελευταία χρόνια ως εξέχουσα τεχνική μάθησης σε ένα σύνολο στατικών δεδομένων και ροής δεδομένων [40]. Τα βαθιά συνελικτικά δίκτυα (Convolution Networks) [41] παρουσιάζουν σημαντικά αποτελέσματα, καθώς το βάθος των νευρωνικών δικτύων αυξάνεται και χρησιμοποιούν μικρές συνελίξεις με χαρακτήρες ως είσοδο. Τα συνελικτικά νευρωνικά δίκτυα που έχουν αμφίδρομα επαναλαμβανόμενα επίπεδα στα οποία χρησιμοποιούνται διανύσματα από λέξεις μειώνουν την απώλεια τοπικής πληροφορίας και αποδίδουν τις εξαρτήσεις που υπάρχουν μεταξύ των ακολουθιών λέξεων [42]. Επίσης τα Deep Belief Networks έχουν συνδυαστεί με επιτυχία με την τεχνική regression softmax [8] για την επίλυση προβλημάτων υπολογισμού μεγάλου όγκου δεδομένων κειμένου. Αυτές οι τεχνικές βαθιάς εκμάθησης έχουν εφαρμοστεί στα πλαίσια προγραμματιστικών μοντέλων όπως το Apache Spark [43]. Επιπλέον, στην έρευνα [44] έχει εφαρμοστεί μια προσαρμογή της ταξινόμησης συνεχούς ροής κειμένων χρησιμοποιώντας τα δίκτυα Boltzman Machines και Deep Belief.

Ένα κατανεμημένο, κλιμακωτό μοντέλο με μικρό χρόνο απόκρισης μπορεί να υποστηρίξει αλγόριθμους κατηγοριοποίησης ροής κειμένων που εμφανίζονται με υψηλή συχνότητα. Τα υπολογιστικά νέφη είναι μια ευρέως χρησιμοποιούμενη λύση που ικανοποιεί τις απαιτήσεις υπολογισμού, μνήμης και απόδοσης σε οικονομικά αποδοτικές λύσεις. Μια επισκόπηση των εργαλείων λογισμικού και των τεχνολογιών για την ανάπτυξη ανάλυσης κλιμακούμενων δεδομένων σε υπολογιστικά συστήματα νέφους παρέχεται από τον Belcastro [45]. Μια ανασκόπηση σχετικά με την επεξεργασία ροών δεδομένων και τα εργαλεία που παρέχονται ως υπηρεσία περιλαμβάνει το Amazon Kinesis, το Google Dataflow και το Azure Stream Analysis [46]. Για την καλύτερη λειτουργία μια υπηρεσίας όπου κάνει αυτόματη ταξινόμηση ροής κειμένων προκύπτουν προκλήσεις όπως ο μεγάλος όγκος δεδομένων εισόδου, η άμεση διαχείριση των ροών κειμένου,

ο εντοπισμός των σημείων συμφόρησης, ο προγραμματισμός και η προσαρμογή των εφαρμογών και η κατανομή πόρων, ώστε να ανταποκρίνονται στις απαιτήσεις των υπηρεσιών επεξεργασίας ροών κειμένου με δυναμικό τρόπο.

Η μέθοδος ταξινόμησης γράφων Ν-γραμμάτων που θα χρησιμοποιηθεί στη δουλειά μας χρησιμοποίησε το μοντέλο προγραμματισμού Beam [6]. Το Beam διαθέτει σημαντικά πλεονεκτήματα σε σύγκριση με τα Spark και Hadoop καθώς ενοποιεί τα στατικά δεδομένα και τα δεδομένα ροής με τέτοιο τρόπο όπου παρέχει έναν ολοκληρωμένο τρόπο επεξεργασίας με χαρακτηριστικά όπως παράθυρα γεγονότων, υδατογραφήματα και ενεργοποιητές (triggers). Το Beam διαθέτει επίσης εξαιρετικές δυνατότητες αυτοσυσχέτισης [47] και πολύ ευέλικτες, επεκτάσιμες αλλά και ταυτόχρονα διακριτές δυνατότητες [48] για τον σχεδιασμό αγωγών επεξεργασίας δεδομένων.

2.7.Συμπεράσματα

Στο κεφάλαιο αυτό ξεκινήσαμε από την διαισθητική αντίληψη ότι κείμενα που έχουν τις ίδιες λέξεις θα διαπραγματεύονται και παρόμοια θέματα. Αρχικά θεωρήσαμε κάθε λέξη ενός κειμένου μια μεταβλητή και απεικονίσαμε τα κείμενα ως διανύσματα σε έναν ν-διάστατο χώρο σύμφωνα με το μοντέλο σάκου λέξεων. Τα κείμενα που αποτελούν πυκνώματα σε αυτόν τον ν-διάστατο χώρο, πιστεύουμε ότι αποτελούν συστάδες. Όταν θέλουμε να βρούμε την απάντηση μιας ερώτησης, απλώς απεικονίζουμε αυτή την ερώτηση ως ένα διάνυσμα στον ν-διάστατο χώρο και όσα κείμενα βρίσκονται πιο κοντά της είναι η απάντηση.

Το πρόβλημα της συνωνυμίας, της πολυσημίας των λέξεων, καθώς και το ότι οι διαστάσεις που έχουν τα διανύσματα είναι πολλές, μας οδήγησαν να προσπαθήσουμε να χρησιμοποιήσουμε μαθηματικά μοντέλα.

Η μέθοδος LSA χρησιμοποιώντας αλγεβρικά κριτήρια μπορεί να ενοποιεί συνώνυμες λέξεις ή ακόμη και λέξεις που έχουν νοηματική συγγένεια σε άδηλες μεταβλητές, οπότε το πρόβλημα της συνωνυμίας λύθηκε, καθώς επίσης και το να μειωθεί το πλήθος των διαστάσεων των διανυσμάτων.

Εξέλιξη της LSA είναι η PLSA. Εδώ χρησιμοποιούμε στατιστικά και πιθανοτικά μοντέλα για να εκπαιδεύσουμε στο σύστημα μας. Κάθε λέξη αντιστοιχεί σε κάποιο θέμα, το θέμα είναι μια άδηλη μεταβλητή, και κάθε κείμενο αναλόγως ποιες λέξεις έχει, αντιστοιχεί ποσοστιαία σε μια σειρά από θέματα.

Τέλος είδαμε πως με την χρήση της συνάρτησης κατανομής Dirichlet υλοποιήθηκε η μέθοδος LDA, που ορίζει ένα generative μοντέλο και είναι σε θέση να συσταδοποιήσει μια συλλογή κειμένων σ' ένα k πλήθος κατηγοριών. Αυτό στις περισσότερες περιπτώσεις το πετυχαίνει με καλύτερη ακρίβεια από οποιαδήποτε άλλη μέθοδο, ενώ ταυτόχρονα είναι σε θέση να βρει ποιες λέξεις είναι σημαντικές σε κάθε κατηγορία ταξινομώντας τες σε φθίνουσα σειρά σημασίας.

Είναι αξιοσημείωτο πως ενώ ως δεδομένο έχουμε μόνο την παρατήρηση λέξεων μέσα σε κείμενα ο μηχανικός και ο επιστήμονας εφαρμόζει διαδοχικά όλο και πιο ανεπτυγμένα μαθηματικά μοντέλα για να βελτιώσει την ακρίβεια των αποτελεσμάτων του. Ταυτόχρονα αποτελεί πρόκληση η ανάπτυξη νέων μαθηματικών μοντέλων που θα μπορούσαν να εφαρμοστούν για τις ανάγκες της τεχνολογίας που ολοένα αυξάνονται.

Τέλος, το μοντέλο γράφων Ν-γραμμάτων που θα αναλύσουμε στην συνέχεια είναι ένα μοντέλο αναπαράστασης κειμένων που διαφέρει από το ευρέως χρησιμοποιούμενο μοντέλο BoW. Στο μοντέλο BoW, το κείμενο αναπαρίσταται ως μια πολλαπλή ομάδα αντικειμένων, όπως λέξεις ή Ν-γράμματα, χωρίς να λαμβάνεται υπόψη η ακολουθία των στοιχείων. Οι γράφοι Ν-γραμμάτων μπορούν να διατηρήσουν την ακολουθία στοιχείων και να αναπαραστήσουν ένα κείμενο με πιο αντιπροσωπευτικό και συνεκτικό τρόπο. Το μοντέλο αναπαράστασης γράφων Ν-γραμμάτων προτάθηκε αρχικά για την αξιολόγηση περιλήψεων

κειμένων [49] και παραμένει ανοιχτό πεδίο για περαιτέρω έρευνα και εφαρμογές όπως θα δούμε στις επόμενες ενότητες.

3. Μετρικές Αξιολόγησης Κατηγοριοποιήσεων και Συσταδοποιήσεων

Στο παρόν κεφάλαιο παρουσιάζουμε τις περισσότερες μετρικές που μπορούν να χρησιμοποιηθούν για να αξιολογήσουν κατά πόσο μια κατηγοριοποίηση - συσταδοποιήση κειμένων, διανυσμάτων ή οποιοδήποτε αντικειμένων είναι καλή. Οι μετρικές μπορεί να χρησιμοποιούν ένα σημείο αναφοράς, εξωτερικές, δηλαδή μια σωστή κατηγοριοποίηση των δεδομένων ή να βασίζονται μόνο στα αποτελέσματα ενός αλγορίθμου συσταδοποίησης, εσωτερικές. Παρουσιάζουμε τις βασικές ιδιότητες που πρέπει να ικανοποιούν και τις κατατάσσουμε βάση κριτηρίων όπως το ταίριασμα συστάδων με κατηγορίες, την αναλογία μεταξύ σωστών και λάθος συσταδοποιημένων ζευγαριών, την edit distance ομοιότητα και σχέσεων από την θεωρία πληροφορίας. Επίσης αναφέρουμε τις BCubed μετρικές και τις επεκτείνουμε για επικαλυπτόμενες συσταδοποιήσεις. Τέλος παρουσιάζουμε την δεκαπλή-αναδιπλώσεων διασταυρούμενη αξιολόγηση (10-fold cross validation) σε συνδυασμό με τις Μίκρο και Μάκρο μετρήσεις αξιολόγησης που θα τις χρησιμοποιήσουμε για να αξιολογήσουμε τα περισσότερα πειραματικά αποτελέσματα.

3.1. Εισαγωγή στης Μετρικές Αξιολόγησης Κατηγοριοποιήσεων και Συσταδοποιήσεων

Συσταδοποίηση ονομάζουμε την διαδικασία κατά την οποία καταχωρούμε αντικείμενα σε λογικές ομάδες, με τέτοιο τρόπο ούτως ώστε τα αντικείμενα που παρουσιάζουν στοιχεία ομοιότητας να είναι μαζί στην ίδια λογική ομάδα ενώ τα αντικείμενα που διαφέρουν αναμεταξύ τους να βρίσκονται σε διαφορετικές λογικές ομάδες.

Τεχνικές συσταδοποίησης έχουν εφαρμογή σε πολλούς επιστημονικούς κλάδους όπως η βιολογία[50], η φαρμακευτική [51], το marketing [52], τα κοινωνικά δίκτυα [53], η επεξεργασία εικόνας [54], τα συστήματα συστάσεων [55], η ηλεκτρονική [56] και οι κοινωνικές επιστήμες [57].

Αλγόριθμοι που παράγουν συσταδοποιήσεις είναι ένα ερευνητικό θέμα που έχει απασχολήσει σε μεγάλο βαθμό την επιστημονική κοινότητα και είναι αξιοσημείωτο το μεγάλο πλήθος μεθόδων που έχουν προταθεί [58].

Το μεγάλο πλήθος αλγορίθμων συσταδοποίησης που υπάρχει, οι διαφορετικές ανάγκες τις οποίες χρειάζεται να καλύψει κάθε εφαρμογή και οι ιδιαιτερότητες κάθε data set το οποίο καλούμαστε να συσταδοποιήσουμε κάνουν μεγάλη την ανάγκη ύπαρξης μιας σειράς μετρικών σχέσεων με τα οποία θα μπορούμε να αξιολογήσουμε τις συσταδοποιήσεις. Αυτές τις μετρικές θα παρουσιάσουμε στις επόμενες ενότητες.

3.2. Ιδιότητες Αξιολόγησης Συσταδοποιήσεων

Θα περιγράψουμε μια σειρά από ιδιότητες που θα πρέπει μια μετρική σχέση να ικανοποιεί ούτως ώστε να παράγει αξιόπιστα αποτελέσματα. Οι ιδιότητες αυτές είναι η βάση για να κατανοήσουμε τα προτερήματα αλλά και τα μειονεκτήματα κάθε μετρικής σχέσης.

Αρχικά θα περιγράψουμε τις πιο σημαντικές ιδιότητες που πρέπει μια μετρική σχέση να ικανοποιεί. Έπειτα αναφέρουμε δύο ακόμη σύνολα ιδιοτήτων. Για να διατυπώσουμε τις ιδιότητες που θα δούμε χρησιμοποιούμε

την ακόλουθη ορολογία. Σε ένα υποθετικό data set θα έχουμε δύο σύνολα συσταδοποιήσεων D_1 και D_2 όπου το D_2 θα είναι πιο καλά συσταδοποιημένο από το D_1. Κάθε μετρική σχέση συμβολίζεται ως Q και το ότι μια ιδιότητα συμβάλει στο να είναι η συσταδοποίηση D_2 καλύτερη από την D_1 φαίνεται από την σχέση $Q(D_1) < Q(D_2)$. Οι σωστές κατηγορίες θα συμβολίζονται ως $L_1 \ldots L_n$.

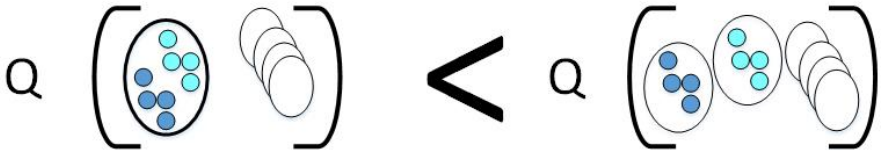
Επίσης καθ' όλη την διάρκεια της εργασίας ως συστάδες εννοούμε τα σύνολα στοιχείων που δημιουργήθηκαν από έναν αλγόριθμο συσταδοποίησης. Ως κλάσεις εννοούμε τα σύνολα στοιχείων που δημιουργήθηκαν από έναν αλγόριθμο ταξινόμησης (classification). Ως κατηγορίες εννοούμε τα σύνολα στοιχείων που είναι σωστά ταξινομημένα σύμφωνα με ένα validation data set. Οι κατηγορίες δεν προκύπτουν από ένα αλγόριθμο αλλά είναι ένα gold standard (σημείο αναφοράς) με το οποίο πρέπει να συγκρίνουμε την μέθοδο μας για να δούμε κατά πόσο μια μετρική αξιολόγησης συστάδων είναι καλή.

Τους όρους ταξινόμηση, συσταδοποίηση, classification και clustering μπορούμε να τους εναλλάσσουμε εξίσου για τις ανάγκες της εργασίας, αν και διαπραγματεύονται διαφορετικές οικογένειες αλγορίθμων. Αυτό οφείλεται στο ότι οι ίδιες μετρικές αξιολόγησης μπορούν να χρησιμοποιηθούν και για προβλήματα ταξινόμησης και για προβλήματα συσταδοποίησης.

3.2.1.Ομοιογένεια συστάδας

Έστω S ένα σύνολο από αντικείμενα που ανήκουν στις κατηγορίες $L_1 \ldots L_n$. D_1 είναι μια συσταδοποίηση όπου το cluster C αναμειγνύει αντικείμενα από δύο κατηγορίες L_i και L_j. D_2 είναι μια συσταδοποίηση πανομοιότυπη με τη D_1, με την μόνη διαφορά ότι το cluster C έχει αντικατασταθεί με δύο cluster που περιέχουν σωστά τα αντικείμενα της κατηγορίας L_i και L_j αντίστοιχα. Η μετρική αξιολόγησης Q ικανοποιεί την σχέση $Q(D_1) < Q(D_2)$.

Η ιδέα στην οποία βασίζεται η ιδιότητα της ομογένειας είναι ότι κάθε συστάδα που θα δημιουργηθεί θα πρέπει να αποτελείται από αντικείμενα μόνο μιας κατηγορίας, που είναι όμοια και να μην αναμιγνύονται ανόμοια αντικείμενα στην ίδια συστάδα. Στοιχεία από δύο διαφορετικές κατηγορίες δεν πρέπει να συνδυάζονται στην ίδια συστάδα.

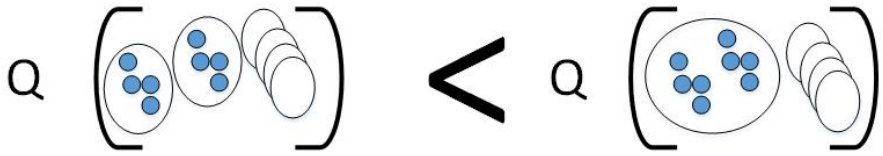


Σχήμα 3.1 Ομοιογένεια συστάδας

Στο σχήμα 3.1 βλέπουμε την Ομοιογένεια συστάδας (Cluster Homogeneity) και πώς η ένωση δύο σωστών κατηγοριών σε μία συστάδα συμβάλει στο να μειωθεί η ακρίβεια μιας μετρικής μεθόδου. Δεξιά έχουμε την σωστή συσταδοποίηση D_2 που όλα τα μαύρα αντικείμενα είναι μαζί σε μια συστάδα και όλα τα λευκά είναι μαζί σε μιαν άλλη συστάδα. Αριστερά στην συσταδοποίηση D_2, τα λευκά με τα μαύρα αντικείμενα συνδυάστηκαν οπότε $Q(D_1) < Q(D_2)$.

3.2.2.Πληρότητα συστάδας

Η συμπληρωματική ιδιότητα του cluster homogeneity είναι η Πληρότητα συστάδας (cluster completeness). Εδώ το ζητούμενο είναι τα αντικείμενα που ανήκουν στην ίδια αρχικά σωστή κατηγορία να συσταδοποιούνται μαζί στην ίδια συστάδα και να μην κατανέμονται σε πολλές συστάδες. Εκφράζοντας το με άλλα λόγια, διαφορετικές συστάδες πρέπει να περιέχουν αντικείμενα από διαφορετικές κατηγορίες. Αυτή την διαισθητική έννοια θα την διατυπώσουμε πιο επίσημα. Έστω D_1 είναι μια συσταδοποίηση τέτοια όπου δύο clusters C_1 και C_2 περιέχουν αντικείμενα από την ίδια κατηγορία L. Έστω επίσης μια άλλη συσταδοποίηση D_2 που είναι πανομοιότυπη με την D_1 εκτός από την διαφορά ότι τα C_1 και C_2 είναι ενωμένα σ' ένα cluster. Τότε η συσταδοποίηση D_2 είναι καλύτερη από την D_1 και $Q(D_1) < Q(D_2)$.



Σχήμα 3.2 Πληρότητα συστάδας

Στο σχήμα 3.2 αναπαριστάται η πληρότητα συστάδας (Cluster Completeness) και πώς τα αντικείμενα μιας αρχικής σωστής κατηγορίας θα πρέπει να διατηρηθούν μαζί και να μην διασπαστούν σε δύο διαφορετικές συστάδες. Παρατηρούμε στην αριστερή συσταδοποίηση ότι ενώ όλα τα μαύρα αντικείμενα είναι όμοια, έχουν συσταδοποιηθεί ξεχωριστά, σε δύο διαφορετικές συστάδες. Ενώ στην δεξιά έχουν συσταδοποιηθεί μαζί.

Οι δύο ιδιότητες Cluster Homogeneity και Cluster Completeness είναι οι δύο πιο βασικές που ένας αλγόριθμος συσταδοποίησης πρέπει να ικανοποιεί. Ενώνοντάς τες σε μια φράση, θα μπορούσαμε να εκφράσουμε την πρόταση ότι σκοπός μας είναι να κρατήσουμε τα αντικείμενα από την ίδια κατηγορία μαζί και να κρατήσουμε τα αντικείμενα από διαφορετικές κατηγορίες χωριστά. Παρότι αυτές οι δύο ιδιότητες φαίνονται πολύ βασικές υπάρχουν κριτήρια αξιολόγησης συσταδοποιήσεων που δεν τις ικανοποιούν.

3.2.3.Σάκος κουρελιών

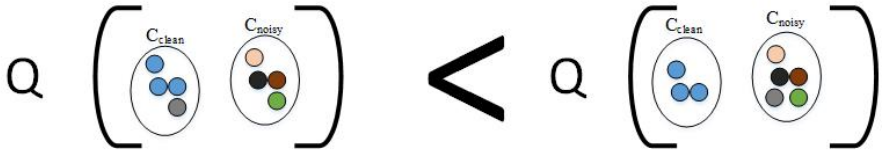
Η ιδιότητα σάκος κουρελιών (Rag Bag) δηλώνει ότι το να εισαγάγουμε διαταραχή (disorder) σε ένα ήδη διαταραγμένο cluster είναι λιγότερο επιβλαβές από το να εισαγάγουμε διαταραχή σε ένα καθαρό cluster. Με τον όρο διαταραγμένο cluster εννοούμε ένα cluster που περιέχει ανόμοια αντικείμενα.

Καθαρό cluster είναι ένα cluster που όλα του τα αντικείμενα είναι όμοια αναμεταξύ τους. Λέγοντας να εισαγάγουμε διαταραχή σε ένα καθαρό cluster, εννοούμε σε ένα καθαρό cluster να εισαγάγουμε κάποια αντικείμενα που δεν ταιριάζουν με τα υπόλοιπα που το αποτελούν.

Σίγουρα θα υπάρχουν αντικείμενα που δεν θα ταιριάζουν με κανένα από τα ήδη υπαρκτά cluster. Δεν θα ήταν καλή τακτική αυτά τα αντικείμενα να τα εισαγάγουμε σε καθαρά cluster, αλλά θα ήταν προτιμότερο να φτιάξουμε ένα ειδικό cluster, που το ονομάζουμε rag bag, ώστε να τα εισαγάγουμε σε αυτό. Το Rag Bag είναι μια συστάδα που αποτελείται από όλα τα αντικείμενα που παρουσιάζουν χαμηλή έως και καθόλου ομοιότητα με τα υπόλοιπα cluster.

Ένα ιδανικό σύστημα συσταδοποίησης θα μπορούσε όλα τα αντικείμενα που δεν ταιριάζουν με κάποιο cluster να μην τα συσταδοποιήσει και να τα απορρίψει. Από την άλλη, όταν συγκρίνουμε συστήματα συσταδοποίησης, καλό θα ήταν για λόγους πληρότητας να μην αφαιρούμε αντικείμενα, γιατί το ένα σύστημα μπορεί ένα αντικείμενο να το θεωρήσει χρήσιμη πληροφορία και το άλλο διαταραχή.

Διατυπώνοντας αυτή την ιδιότητα πιο επίσημα θεωρούμε C_{clean} ένα cluster με n όμοια αντικείμενα που ανήκουν στην ίδια κατηγορία και C_{noisy} ένα cluster που ενώνει n αντικείμενα από μοναδιαίες κατηγορίες. D_1 είναι μια συσταδοποίηση, όπου το αντικείμενο μιας μοναδιαίας κατηγορίας, συσταδοποιείται σε μια καθαρή συστάδα C_{clean} και D_2 είναι μια συσταδοποίηση, που το αντικείμενο της μοναδιαίας συσταδοποίησης συσταδοποιείται στην C_{noisy}, τότε $Q(D_1) < Q(D_2)$. Μοναδιαία κατηγορία εννοούμε μια κατηγορία που αποτελείται από ένα αντικείμενο που δεν μπορεί να περιληφθεί σε οποιαδήποτε άλλη κατηγορία.



Σχήμα 3.3 Σάκος κουρελιών

Στο σχήμα 3.3 εικονίζεται πως στην αριστερή συσταδοποίηση ένα αντικείμενο που δεν ταιριάζει με τα υπόλοιπα μαύρα αντικείμενα έχει συσταδοποιηθεί μαζί τους. Αυτό χαλάει την καθαρότητα της συστάδας με τα μαύρα αντικείμενα. Στην δεξιά συσταδοποίηση, η συστάδα με τα μαύρα αντικείμενα παραμένει καθαρή και έχουμε μια συστάδα Rag Bag που συναθροίζει όλα τα αντικείμενα που δεν ταιριάζουν να συσταδοποιηθούν κάπου αλλού.

3.2.4.Μέγεθος ενάντιων ποσότητας

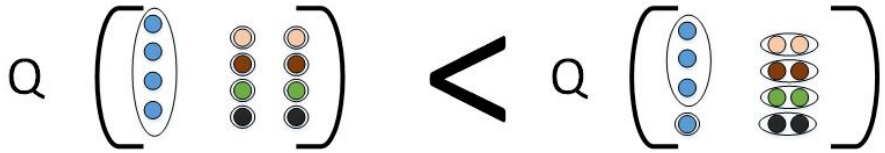
Η ιδιότητα συστάδας μέγεθος ενάντιων ποσότητας (clusters size vs. quantity) εκφράζει ότι είναι προτιμότερο ένα μικρό λάθος σ' ένα μεγάλο cluster παρά ένας μεγάλος αριθμός από μικρά λάθη σε μικρά cluster.

Μια μορφή αυτής τις ιδιότητας είναι ότι είναι προτιμότερο να διαχωρίσουμε ένα αντικείμενο από την μεγάλη συστάδα που ανήκει από το να διασπάσουμε πολλές μικρές συστάδες. Αυτό φαίνεται στο σχήμα 3.4.

Θα διατυπώσουμε καλύτερα αυτήν την ιδιότητα που περιγράψαμε. Έστω D μια κατηγοριοποίηση στην οποία υπάρχει μια κατηγορία L με N+1 αντικείμενα και n κατηγορίες $L_1 \ldots L_n$ που περιλαμβάνουν δύο αντικείμενα. Παράγουμε επίσης, την συσταδοποίηση D_1 όπου είναι όμοια με την κατηγοριοποίηση D με την διαφορά ότι κάθε κατηγορία L_i που έχει δύο αντικείμενα σπάει, ούτως ώστε κάθε αντικείμενο να είναι μια μοναδιαία συστάδα. D_2 είναι μια συσταδοποίηση όμοια με την κατηγοριοποίηση D με την διαφορά ότι η κατηγορία L που περιέχει n + 1 αντικείμενα σπάει σε μια συστάδα των n αντικειμένων και μια συστάδα του ενός αντικειμένου. Η συσταδοποίηση D_2 είναι καλύτερη από την συσταδοποίηση D_1 $Q(D_1) < Q(D_2)$.

Στο σχήμα 3.4 βλέπουμε ότι είναι προτιμότερο να έχουμε ένα μικρό λάθος σε μια μεγάλη συστάδα από ότι πολλά λάθη σε μικρές συστάδες. Η δεξιά συσταδοποίηση κάνει λάθος μόνο σ' ένα μαύρο αντικείμενο που ανήκει σε μια μεγάλη συστάδα ενώ έχει διατηρήσει σωστές όλες τις μικρές συστάδες. Η αριστερή

συσταδοποίηση έχει αφήσει ανέπαφη την συστάδα με τα μαύρα αντικείμενα ενώ έχει διασπάσει όλες τις μικρές συστάδες. Είναι προτιμότερη η δεξιά συσταδοποίηση.



Σχήμα 3.4 Μέγεθος εναντίον ποσότητας

3.2.5.Περιορισμός του Dom

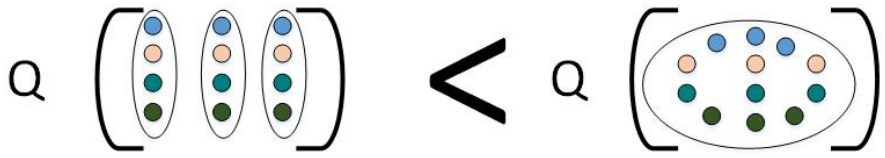
Σε αυτή την υποενότητα θα αναφέρουμε ιδιότητες που βασίζονται στην θεωρία της εντροπίας της πληροφορίας[59][60]. Η ποιότητα μιας συσταδοποίησης μπορεί να προκύψει από μετρικές που έχουν τις εξής παραμέτρους: Τον αριθμό των κατηγοριών, τον αριθμό των noise συστάδων, τον αριθμό των χρήσιμων συστάδων και τρεις ιδιότητες της μάζας πιθανότητας του λάθους που μπορεί να προκύψει.

Noise συστάδες είναι αυτές που περιλαμβάνουν αντικείμενα από όλες τις κατηγορίες, με ίση πιθανότητα από την κάθε μία. Χρήσιμες συστάδες είναι αυτές στις οποίες επικρατεί κυρίως μία κατηγορία. Η μάζα πιθανότητα λάθους δηλώνει σε τι βαθμό τα αντικείμενα δεν συσταδοποιούνται στις χρήσιμες συστάδες.

Οι ιδιότητες αυτές δηλώνουν την ιδέα ότι μια συσταδοποίηση δεν είναι καλή όταν:

- Ο αριθμός των χρήσιμων συστάδων διαφέρει κατά πολύ από τις κατηγορίες.
- Ο αριθμός των noise συστάδων είναι μεγάλος.
- Η μάζα πιθανότητα λάθους είναι μεγάλη.

Γενικώς αυτές οι ιδιότητες ανάγονται στις ιδιότητες που αναφέραμε στις προηγούμενες υποενότητες 3.2.1 - 3.2.4. Για παράδειγμα οι ιδιότητες Cluster Homogeneity και Cluster Completeness ανάγονται σε μια μείωση ή μια αύξηση των χρήσιμων cluster σε σχέση με τον αριθμό των κατηγοριών. Από την άλλη η ιδιότητα Rag Bag που διαπραγματευτήκαμε στην υποενότητα 3.2.3 δεν καλύπτεται. Τέλος αυτές οι ιδιότητες δεν είναι εύκολο να εκφραστούν καθαρά από μια μετρική σχέση.



Σχήμα 3.5 Περιορισμός του Dom

Στο σχήμα 3.5 βλέπουμε ότι ένα μεγάλο πλήθος από λάθος συστάδες παράγει μια χειρότερης ποιότητας συσταδοποίηση από ότι λίγες λάθος συστάδες.

3.2.6.Περιορισμός Meila

Στην έρευνα comparing clustering [61] προτάθηκαν δώδεκα ιδιότητες που θα πρέπει να ικανοποιεί ένα κριτήριο αξιολόγησης. Οι περισσότερες από αυτές δεν σχετίζονται άμεσα με την ποιότητα που έχει μια μετρική αλλά με τα εγγενή χαρακτηριστικά που πρέπει να έχει, όπως την δυνατότητα να εφαρμόζεται σε μεγάλο πλήθος δεδομένων και την πολυπλοκότητα του.

Οι ιδιότητες που προτείνονται δεν έχουν να προσφέρουν κάτι περισσότερο από ότι οι ιδιότητες που έχουμε ήδη περιγράψει. Το μόνο όφελος που θα μπορούσε να έχει κάποιος μελετώντας τες, είναι να δει μια διαφορετική και πιο αναλυτική διατύπωση των ιδιοτήτων 3.2.1 – 3.2.4.

3.3.Μετρικές Ομοιότητας Συστάδας - Κατηγορίας

Σε αυτή και τις επόμενες ενότητες θα παρουσιάσουμε τις πιο γνωστές αλλά και λιγότερο γνωστές μετρικές που αξιολογούν μια συσταδοποίηση.

Μια από τις πιο διαδεδομένες μετρικές σχέσεις, αν όχι η πιο διαδεδομένη, είναι η Purity και η αντιστροφή της, η Inverse Purity. Η Purity και η Inverse Purity συνδυάζονται σε μία σχέση, τον αρμονικό μέσο F.

3.3.1.Καθαρότητα

Στις μετρικές αυτής της ενότητας αντιστοιχούμε κάθε συστάδα που παράχθηκε σε μία κατηγορία. Η μετρική καθαρότητας (Purity) βασίζεται στην συχνότητα που τα αντικείμενα μιας κατηγορίας βρίσκονται μέσα σε μια συστάδα. Το ποιας κατηγορίας τα αντικείμενα κυριαρχούν, δηλαδή είναι τα πιο πολλά σε μια συστάδα, χαρακτηρίζουν την συστάδα και δηλώνουν ότι συνδέεται με την αντίστοιχη κατηγορία.

Θεωρούμε L τις αρχικές σωστές κατηγορίες, C τις συστάδες και N το πλήθος των αντικειμένων που θα συσταδοποιηθούν. Η μετρική Purity υπολογίζεται από τον τύπο 3.1.

$$Purity = \sum_i \frac{|C_i|}{N} max_j\, Precision(C_i, L_j) \quad (3.1)$$

Το precision του cluster C_i για μια κατηγορία L_j ορίζεται από την σχέση 3.2

$$Precision\left(C_i, L_j\right) = \frac{\left|C_i \bigcap L_j\right|}{\left|C_i\right|} \quad (3.2)$$

Η μετρική Purity εκφράζει ότι μειώνεται η ποιότητα μιας συσταδοποίησης όταν εισέρχεται θόρυβος σ' ένα cluster, αλλά δεν μπορεί να εκφράσει την βελτίωση της ποιότητας μιας συσταδοποίησης όταν συσταδοποιούνται αντικείμενα από την ίδια κατηγορία μαζί σε μια συστάδα.

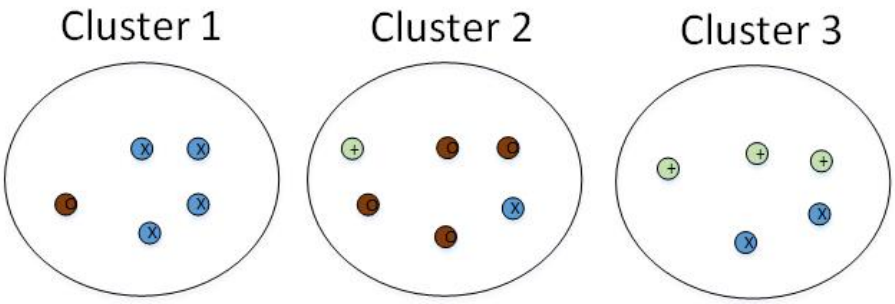
Ακραία περίπτωση της μετρικής Purity είναι να φτιάχνουμε μοναδιαία cluster που περιέχουν μόνο ένα αντικείμενο. Σε αυτή την περίπτωση η μετρική Purity θα είναι ένα. Αυτό το κατανοούμε γιατί κάθε συστάδα θα θεωρείται καθαρή και ότι δεν αναμιγνύει αντικείμενα από διαφορετικές κατηγορίες.

3.3.2.Αντίστροφη καθαρότητα

Καταλαβαίνουμε ότι η μετρική Purity δεν μπορεί από μόνη της να εκφράσει κατά πόσο μια συσταδοποίηση είναι καλή. Θα εισαγάγουμε την μετρική Inverse Purity που είναι συμμετρική της Purity και μπορεί να εκφράσει σε τι βαθμό μια συστάδα περιέχει όλα τα αντικείμενα μιας αρχικής κατηγορίας. Η Inverse Purity εκφράζεται από τον τύπο 3.3.

$$Inverse\,Purity = \sum_i \frac{|L_i|}{N} max_j\, Precision(L_i, C_j) \quad (3.3)$$

Η μετρική Inverse Purity εκφράζει σε τι βαθμό τα αντικείμενα μιας συστάδας περιέχουν όλα τα αντικείμενα μιας αρχικής κατηγορίας. Δεν μπορεί όμως να εκφράσει αν μέσα στην ίδια συστάδα περιέχονται και αντικείμενα από άλλες κατηγορίες. Τα αντικείμενα από άλλες κατηγορίες σε αυτήν την περίπτωση θεωρούνται θόρυβος.



Σχήμα 3.6 Αντίστροφη καθαρότητα

Στο σχήμα 3.6 παρατηρούμε πως το cluster 1 συμβάλει στο να έχουμε υψηλό purity γιατί έχει μέσα του πέντε αντικείμενα (x) από μια κατηγορία και μόνο ένα ως θόρυβο. Το cluster 2 συμβάλει στο να έχουμε υψηλό inverse Purity γιατί περιέχει τέσσερα από τα συνολικά πέντε αντικείμενα της κατηγορίας (ο). Το Purity αυτής της συσταδοποίησης που έχει 3 cluster είναι Purity=(1/17) · (5+4+3) ≈ 0.71

3.3.3.Φ-μέτρο

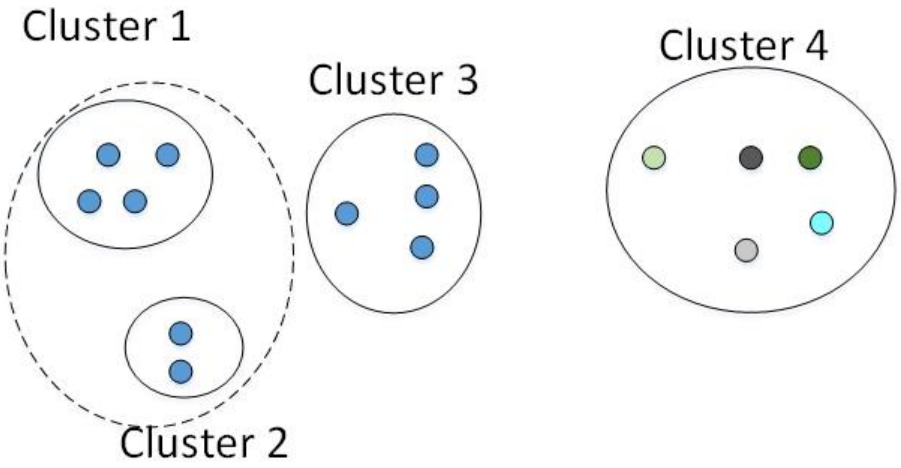
Μπορούμε να συνδυάσουμε την μετρική Purity και Inverse Purity στην μετρική F measure όπως φαίνεται στην σχέση 3.4.

$$F = \sum_i \frac{|L_i|}{N} max_j\{F(L_i, C_j)\} \quad (3.4)$$

Το $F(L_i, C_j)$ δίνεται από τον τύπο 3.5

$$F\left(L_i, C_j\right) = \frac{2 \cdot Recall(L_i, C_j) \cdot Precision(L_i, C_j)}{Recall\left(L_i, C_j\right) + Precision(L_i, C_j)} \quad (3.5)$$

Το μειονέκτημα με τις μετρικές Purity και Inverse Purity είναι ότι δεν ικανοποιούν την ιδιότητα Cluster Completeness και Rag Bag. Κάθε κατηγορία κρίνεται με βάση την συστάδα που έχει πιο πολλά αντικείμενα της και οι αλλαγές σε άλλες συστάδες δεν αναγνωρίζονται.



Σχήμα 3.7 Φ-μέτρο

Στο σχήμα 3.7 παρατηρούμε ότι τα cluster C1 και cluster C2 περιέχουν αντικείμενα από την ίδια κατηγορία. Αν τα ενώσουμε θα βελτιωθεί η ποιότητα της συσταδοποίησης, η ιδιότητα Cluster Completeness. Αλλά η μετρική Purity και Inverse Purity δεν ικανοποιούν αυτή την ιδιότητα.

3.4.Μετρικές Σχέσεις Ζευγαριών Αντικειμένων

Μια διαφορετική κατηγορία μετρικών σχέσεων βασίζεται στην καταμέτρηση των ζευγαριών αντικειμένων που βρίσκονται μέσα στις συστάδες και τις αρχικές κατηγορίες εξίσου[62].

Έστω SS είναι το πλήθος των ζευγαριών αντικειμένων που ανήκουν στην ίδια κατηγορία και συστάδα. SD είναι το πλήθος των ζευγαριών που ανήκουν στην ίδια συστάδα και διαφορετική κατηγορία. DS είναι το πλήθος των ζευγαριών που ανήκουν σε διαφορετική συστάδα και στην ίδια κατηγορία. DD είναι ο αριθμός των ζευγαριών που ανήκουν σε διαφορετική κατηγορία και συστάδα. Αυτό που θέλουμε από μια συσταδοποίηση είναι να έχουμε μεγάλα SS και DD. Υψηλές τιμές στα DS και SD δηλώνουν ότι η συσταδοποίηση που έχει γίνει δεν είναι καλή.

Στις υποενότητες 3.4.1 – 3.4.3 παρουσιάζουμε τις μετρικές Rand measure [63], Jaccard Coeficient[64] και Folkes and Mallows[65]. Το μειονέκτημα σε αυτές τις μετρικές είναι ότι δεν ικανοποιούν τις ιδιότητες Rag Bag και Clusters size vs. quantity. Υπάρχει περίπτωση κάποια αντικείμενα από μία κατηγορία να διαχωριστούν σε διαφορετικές συστάδες και και το SS να μειωθεί και το DS να αυξηθεί.

3.4.1.Rand μέτρο

Το Rand μέτρο υπολογίζει κατά πόσο είναι όμοιες οι συστάδες σε σχέση με τις κατηγορίες. Αυτό το εκφράζει ως ένα ποσοστό των σωστών συσταδοποιήσεων που έχουν γίνει προς όλες τις συσταδοποιήσεις. Η εξίσωση δίνεται από τον τύπο 3.6.

$$Rand\,measure\,R = \frac{(SS + DD)}{SS + SD + DS + DD} \quad (3.6)$$

Η εξίσωση 4.1.1 εκφράζει το άθροισμα των ζευγαριών των αντικειμένων που έχουν συσταδοποιηθεί σωστά μαζί συν το άθροισμα των ζευγαριών των αντικειμένων που σωστά έχουν συσταδοποιηθεί σε διαφορετικές συστάδες δια το άθροισμα όλων των ζευγαριών των αντικειμένων που έχουν συσταδοποιηθεί σωστά ή λάθος, μαζί ή χώρια.

3.4.2.Jaccard συντελεστής

Ο συντελεστής Jaccard αρχικά χρησιμοποιήθηκε για να ποσοτικοποιήσει την ομοιότητα μεταξύ δύο data set και μετράει την ομοιότητα μεταξύ πεπερασμένων συνόλων. Ο Jaccard συντελεστής εκφράζει τον λόγο της τομής δυο συνόλων δια της ένωσης τους όπως φαίνεται από τον τύπο 3.7.

$$J(A,B) = \frac{\left|A \cap B\right|}{\left|A \cup B\right|} \quad (3.7)$$

Ο Συντελεστής Jaccard κυμαίνεται μεταξύ μηδέν και ένα $0 \leq J(A,B) \leq 1$. Στην ακραία περίπτωση που τα Α και Β είναι άδεια ορίζεται J(A,B)=1.

Ακλουθώντας την ορολογία που εκφράσαμε στην αρχή της υποενότητας 3.4 μπορούμε να εκφράσουμε τον συντελεστή Jaccard σύμφωνα με την σχέση 3.8

$$Jaccard\,Coefficient\,J = \frac{SS}{SS + SD + DS} \quad (3.8)$$

Ο συντελεστής Jaccard ισούται με τον λόγο των ζευγαριών των αντικειμένων που είναι σωστά συστηματοποιημένα στο ίδιο cluster όπως είναι και στην ίδια κατηγορία δια του αθροίσματος των ζευγαριών που βρίσκονται στην ίδια συστάδα αλλά σε διαφορετικές κατηγορίες, στην ίδια κατηγορία αλλά σε διαφορετικές συστάδες και στην ίδια συστάδα και κατηγορία.

Ο συντελεστής Jaccard έχει έναν συμμετρικό συντελεστή, την απόσταση Jaccard (Jaccard distance) η οποία μετράει την ανομοιότητα μεταξύ των συστάδων που δημιουργήθηκαν. Η απόσταση Jaccard ισούται με τον λόγο της ένωσης δύο συνόλων πλην την τομή τους δια την ένωση τους που εκφράζεται από τον τύπο 3.9 ή πιο αναλυτικά τον 3.10.

$$d_J(A,B) = 1 - J(A,B) = \frac{\left|A \bigcup B\right| - \left|A \bigcap B\right|}{\left|A \bigcup B\right|} \quad (3.9)$$

$$Jaccard\,Simillarity\,JS = \frac{SD + DS}{SS + SD + DS} \quad (3.10)$$

Μι ακόμη παραλλαγή του Jaccard συντελεστή είναι η συμμετρική διαφορά $A \triangle B = \left(A \bigcup B\right) - (A \bigcap B)$.

3.4.3.Folkes και Mallows

Ο συντελεστής Folkes και Mallows είναι ένα συντελεστής που είναι απευθείας ανάλογος με το πλήθος των ζευγαριών των αντικειμένων που είναι σωστά συσταδοποιημένα σύμφωνα με τις αρχικές κατηγορίες. Ο συντελεστής Fokes and Mallows δίνεται από τον τύπο 3.11

$$Folkes\ and\ Mallows\ FM = \sqrt{\frac{SS}{SS+SD} \cdot \frac{SS}{SS+DS}} \quad (3.11)$$

Αν έχουμε μια συσταδοποίηση που οι συστάδες δεν σχετίζονται με τις αρχικές κατηγορίες, η τιμή αυτού του συντελεστή προσεγγίζει το μηδέν και δείχνει καλύτερα αποτελέσματα[65] από άλλους συντελεστές, όπως ο Rand measure. Επίσης η μετρική Folkes και Mallows έχει καλά αποτελέσματα αν θόρυβος προστεθεί στο data set και την συταδοποίηση, καθώς επίσης και στην πιο ακραία περίπτωση που έχουν σχηματιστεί διαφορετικό πλήθος συστάδες από ότι αρχικές κατηγορίες.

3.5.Μετρικές που Βασίζονται στην Θεωρία Πληροφορίας

Η Θεωρία πληροφορίας είναι τμήμα των εφαρμοσμένων μαθηματικών και βασίζεται στην θεωρία πιθανοτήτων και στατιστικής. Ένας από τους κύριους σκοπούς της είναι η ποσοτικοποίηση της πληροφορίας. Δύο από τις βασικότερες ποσότητες τις πληροφορίας είναι η εντροπία και η κοινή πληροφορία (mutual information). Με βάση αυτές τις δύο θα παρουσιάσουμε δύο μετρικές αξιολόγησης συσταδοποιήσεων.

3.5.1.Μετρικές βασισμένες στην εντροπία

Η μετρική που βασίζεται στην εντροπία[66] δείχνει πως τα μέλη από k κατηγορίες κατανέμονται μέσα σε κάθε cluster. Η γενική ποιότητα της συσταδοποίησης υπολογίζεται από την μέση εντροπία από όλα τα cluster και δίνεται από τον τύπο 3.12

$$Entropy = -\sum_j \frac{n_j}{n} \sum_i P(i,j) \cdot log_2 P(i,j) \quad (3.12)$$

Το P(i,j) είναι η πιθανότητα του να βρεθεί ένα αντικείμενο από την κατηγορία i στην συστάδα j, n_j είναι το πλήθος των αντικειμένων στο cluster j και n είναι ο συνολικός αριθμός των αντικειμένων που συσταδοποιούνται.

Όλες οι μετρικές που βασίζονται στην εντροπία δεν ικανοποιούν την ιδιότητα Rag Bag. Η αύξηση της εντροπίας όταν ένα αντικείμενο προστίθεται είναι ανεξάρτητη από το πόσο διαταραγμένο ήταν το cluster πριν και είναι αντιστοιχεί μόνο με το κατά πόσο εισήχθηκε λάθος ένα αντικείμενο σ' ένα καθαρό ή disordered cluster.

3.5.2.Μετρικές βασισμένες στην κοινή πληροφορία

Η κοινή πληροφορία (Mutual information) στην θεωρία πληροφορίας μετρά το ποσό της πληροφορίας που μπορεί να αποκτηθεί για μια τυχαία μεταβλητή παρατηρώντας μια άλλη.

Για τις ανάγκες της αξιολόγησης συσταδοποιήσεων, μπορεί να χρησιμοποιηθεί η κοινή πληροφορία για να δηλώσει πόση πληροφορία μοιράζεται μεταξύ μιας συσταδοποίησης σε σχέση με τις αρχικές τις κατηγορίες. Η μετρική αυτή δυστυχώς δεν έχει καλά αποτελέσματα[67] και δεν χρησιμοποιείται συχνά για αυτό θα παρουσιαστεί συνοπτικά.

Η μετρική Mutual information δύο μεταβλητών μετρά κατά πόσο δύο μεταβλητές, στην περίπτωση μας δύο σύνολα στοιχείων, εξαρτώνται το ένα από το άλλο. Δηλαδή αναμένουμε ότι αν μια συσταδοποίηση και μια κατηγοριοποίηση έχουν υψηλή Mutual information τιμή τότε η συσταδοποίηση εξαρτάται σε μεγάλο βαθμό από την κατηγοροποίηση και άρα θα είναι καλή.

Η μετρική Mutual information για δύο διακριτές μεταβλητές X και Y δίνεται από την εξίσωση 3.13

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log\left(\frac{p(x,y)}{p(x) \cdot p(y)}\right) \quad (3.13)$$

Η συνάρτηση αυτή μετρά σε τι βαθμό η πληροφορία που υπάρχει στο X υπάρχει και στο Y. Αν το X και Y είναι ανεξάρτητα τότε το X δεν προσφέρει κάποια πληροφορία για το Y. Στις δικές μας ανάγκες αυτό μεταφράζεται ότι αν τα X και Y είναι ανεξάρτητα τότε τα αντικείμενα που ανήκουν στην κατηγορία A θα συσταδοποιηθούν τυχαία σε άλλα cluster και δεν θα είναι μαζί σε ένα αντιπροσωπευτικό cluster. Η άλλη περίπτωση είναι αν η εξίσωση 3.13 έχει υψηλή τιμή τότε αναμένουμε τα αντικείμενα που είναι μαζί στις ίδιες κατηγορίες να είναι μαζί και στις αντίστοιχες συστάδες.

3.6.Μετρικές που Βασίζονται στο Edit Distance

Μια διαφορετική μετρική μέθοδος για να αξιολογήσει του πόσο καλή είναι μια συσταδοποιήση προκύπτει με βάση του πόσο απέχει μια συσταδοποίηση από την αρχική σωστή κατηγοριοποίηση[68].

Η μετρικές αυτής της κατηγορίας βασίζονται στον αριθμό των μετασχηματισμών που πρέπει να γίνουν ούτως ώστε ξεκινώντας από μια συσταδοποίηση να καταλήξουμε στην αρχική κατηγορία. Όσο λιγότεροι μετασχηματισμοί χρειαστούν τόσο πιο κοντά θεωρούμε ότι βρίσκονται η συστάδα με την κατηγορία. Όσο λιγότεροι μετασχηματιστούν χρειαστούν τόσο πιο καλή είναι μια συσταδοποίηση.

Η μετρική που βασίζεται στο πλήθος των μετασχηματισμών δηλώνει έμμεσα τι απόσταση υπάρχει μεταξύ μιας συσταδοποίησης και των αρχικών κατηγοριών.

Οι μετασχηματισμοί που μπορούν να γίνουν είναι η ένωση δύο συστάδων και μετακίνηση ενός αντικειμένου από μια συστάδα σε μιαν άλλη.

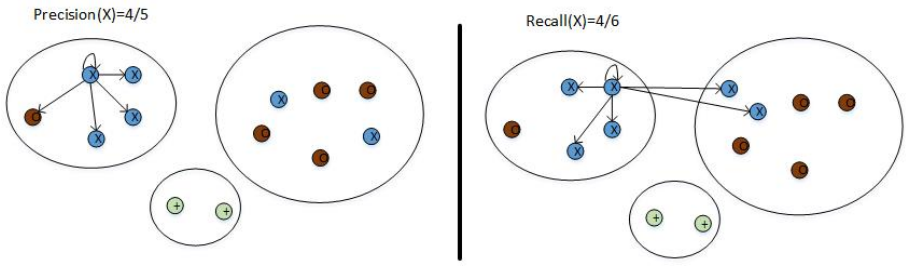
Οι ιδιότητες Cluster Homogeneity και Rag Bag δεν ικανοποιούνται από την μετρική που βασίζεται στην απόσταση μεταξύ συστάδων με κατηγοριών. Ανεξάρτητα με το πού έχει εισαχθεί ο θόρυβος είτε σε ένα καθαρό cluster είτε σ' ένα disordered το μόνο που μετράει είναι το πλήθος των μετασχηματισμών.

3.7.Μετρικές Ομοιότητας Αντικειμένων (BCubed)

Οι μετρικές σχέσεις που μελετήσαμε στις προηγούμενες ενότητες δεν μπορούν να ικανοποιήσουν όλες τις ιδιότητες που αναφέραμε στις υποενότητες 3.2.1 – 3.2.4. Ειδικά την ιδιότητα Rag Bag δεν την ικανοποιεί καμία μετρική σχέση.

Οι μετρικές BCubed precision και recall[69] είναι μετρικές που μπορούν να ικανοποιήσουν και τις τέσσερις ιδιότητες. Σε αντίθεση με τις μετρικές Purity και Entropy οι οποίες υπολογίζουν ανεξάρτητα την ποιότητα του κάθε cluster οι μετρικές BCubed αναγάγουν την αξιολόγηση στο να υπολογίσουν το precision και το recall για κάθε αντικείμενο.

Το precision κάθε αντικειμένου δηλώνει πόσα αντικείμενα από την ίδια συστάδα ανήκουν στην κατηγορία του. Αντίστοιχα το recall δηλώνει πόσα αντικείμενα από την ίδια κατηγορία εμφανίζονται στο cluster που βρίσκεται το αντικείμενο.



Σχήμα 3.8 Precision και Recall

Στο σχήμα 3.8 παρατηρούμε ότι το precision για το αντικείμενο e είναι precision(e)=4/5 γιατί στην κάτω αριστερά συστάδα υπάρχουν σύνολο πέντε αντικείμενα από τα οποία τα τέσσερα ανήκουν στην ίδια κατηγορία με το e. Το recall για το αντικείμενο e υπολογίζεται Recall(e) =4/6 γιατί υπάρχουν σύνολο έξι αντικείμενα που ανήκουν στην κατηγορία του e, τέσσερα εκ των οποίον στην ίδια συστάδα με το e.

Από την πλευρά ενός χρήστη οι BCubed μετρικές αντιπροσωπεύουν την ακρίβεια της συσταδοποίησης. Δηλαδή όταν ο χρήστης ανατρέξει σε ένα αντικείμενο, κατά πόσο μόνο παρόμοια αντικείμενα υπάρχουν με αυτό στην ίδια συστάδα και κατά πόσο όλα τα παρόμοια αντικείμενα βρίσκονται στην ίδια συστάδα και όχι διάσπαρτα σε άλλες συστάδες.

Αν το αντικείμενο που ανάκτησε ο χρήστης έχει υψηλό BCubed recall, τότε ο χρήστης θα βρει πολλά αντικείμενα που είναι όμοια με αυτό μέσα στην ίδια συστάδα.

Αν από την άλλη το αντικείμενο που ανακτά ένας χρήστης από μια συστάδα έχει υψηλό BCubed precision, τότε ο χρήστης δεν θα βρει μέσα σε αυτό το cluster αντικείμενα που δεν ταιριάζουν με αυτό.

Η βασική διαφορά μεταξύ των μετρικών Purity, Entropy με τις BCubed είναι ότι το κατά πόσο τα αντικείμενα είναι καλά συσταδοποιημένα εξαρτάται από το αντικείμενο αναφοράς και όχι από την κυρίαρχη κατηγορία σ’ ένα cluster.

Το L(e) συμβολίζει την κατηγορία του αντικειμένου e και το C(e) την συστάδα του. Η ορθότητα της σχέσης μεταξύ του e και του e΄ σε μια συσταδοποίηση είναι

$$Correctness(e, e') = \begin{cases} 1 \; \alpha\nu\nu \; L(e) = L(e') \leftrightarrow C(e) = C(e') \\ 0 \; \alpha\lambda\lambda\iota\acute{\omega}\varsigma \end{cases} \quad (3.14)$$

Δύο αντικείμενα που προέρχονται από την ίδια κατηγορία είναι σωστά συσχετισμένα, αν βρίσκονται και στην ίδια συστάδα.

Το BCubed precision ενός αντικειμένου αντιστοιχεί στην αναλογία αντικειμένων τα οποία προέρχονται από την ίδια κατηγορία και βρίσκονται στο ίδιο cluster με αυτό.

Αντίστοιχα υπάρχει και το ολικό BCubed precision πού είναι ο μέσος όρος του precision από όλα τα αντικείμενα στην συσταδοποίηση. Εφόσον ο μέσος όρος υπολογίζεται με βάση όλα τα αντικείμενα, δεν χρειάζεται να πάρουμε υπόψη μας το μέγεθος της συστάδας και της κατηγορίας. Το BCubed recall είναι αντίστοιχο, μόνο που αντί να διαχειριστούμε συστάδες χρησιμοποιούμε κατηγορίες. Οι δύο BCubed μετρικές δίνονται από τους τύπους 3.15 και 3.16.

$$\text{Precision BCubed} = Avg_e[Avg_{e'.C(e)=C(e')}[Correctness(e, e')]] \quad (3.15)$$

$$\text{Recall BCubed} = Avg_e[Avg_{e'.L(e)=L(e')}[Correctness(e, e')]] \quad (3.16)$$

Οι BCubed μετρικές συνδυάζουν τα καλύτερα στοιχεία από όλες τις άλλες μετρικές σχέσεις. Αντίστοιχα με την Purity και την Inverse Purity βασίζεται στο precision και στο recall. Εν συνεχεία, αντίστοιχα με την μετρική που βασίζεται στην εντροπία οι BCubed μετρικές παίρνουν υπόψη τους το disorder από κάθε cluster και όχι μόνο από την κυρίαρχη κατηγορία.

Στο σημείο αυτό θα θέλαμε να ξανά αναφέρουμε συνοπτικά τις τέσσερις ιδιότητες που πρέπει να ικανοποιεί μία μετρική και να αιτιολογήσουμε γιατί οι BCubed τις ικανοποιούν.

Cluster Homogeneity: Αν διαχωρίσουμε μια συστάδα που αναμιγνύει δύο κατηγορίες σε δύο καθαρές συστάδες τότε αυξάνεται το BCubed precision ενώ δεν επηρεάζεται το recall.

Cluster Completeness: Ενώνοντας δύο συστάδες που περιέχουν μόνο αντικείμενα από την ίδια κατηγορία αυξάνεται το BCubed recall ενώ το precision των αντικειμένων παραμένει το ίδιο.

Rag Bag: Ας υποθέσουμε δύο περιπτώσεις. Στην μία περίπτωση έχουμε ένα αντικείμενο σ' ένα cluster μόνο του και το εισαγάγουμε σ' ένα καθαρό cluster από n αντικείμενα τότε μειώνεται το precision κάθε αντικειμένου από 1 σε $\frac{n}{n+1}$

Στην δεύτερη περίπτωση το αντικείμενο που είναι μόνο του το εισαγάγουμε σ' ένα Rag Bag, τότε μειώνεται το precision κάθε αντικειμένου από 1 σε $\frac{1}{n+1}$. Το recall δεν επηρεάζεται σε καμία από τις δύο περιπτώσεις.

Η ολική μείωση του precision στην πρώτη περίπτωση εκφράζεται από την εξίσωση 3.17

$$DEC_{D_1} = \frac{1 + n \cdot 1}{N_{tot}} - \frac{\frac{1}{n+1} + n \cdot \frac{n}{n+1}}{N_{tot}} = \frac{\frac{2n}{n+1}}{N_{tot}} \simeq \frac{2}{N_{tot}} \quad (3.17)$$

Το πρώτο κλάσμα αντιπροσωπεύει το precision πριν την εισαγωγή του αντικειμένου από το μοναδιαίο cluster στο καθαρό cluster. Το δεύτερο κλάσμα αντιπροσωπεύει την συσταδοποίηση που έχει το αντικείμενο από την μοναδιαία κατηγορία μέσα στο καθαρό cluster. N_{tot} είναι ο συνολικός αριθμός αντικειμένων που συσταδοποιείται.

Η ολική μείωση του precision στην δεύτερη περίπτωση εκφράζεται από την εξίσωση 3.18

$$DEC_{D_2} = \frac{1 + n \cdot \frac{1}{n}}{N_{tot}} - \frac{\frac{1}{n+1} + n \cdot \frac{1}{n+1}}{N_{tot}} = \frac{1}{N_{tot}} < DEC_{D_1} \quad (3.18)$$

Το πρώτο κλάσμα αντιπροσωπεύει το precision πριν την εισαγωγή του αντικειμένου από το μοναδιαίο cluster στο Rag Bag. Το δεύτερο κλάσμα αντιπροσωπεύει την συσταδοποίηση που έχει το αντικείμενο από την μοναδιαία κατηγορία μέσα στο Rag Bag.

Παρατηρούμε ότι $DEC_{D_2} < DEC_{D_1}$, δηλαδή το precision μειώνεται περισσότερο αν εισαγάγουμε το αντικείμενο από ένα μοναδιαίο cluster σε ένα καθαρό από ότι σε ένα Rag Bag.

Cluster Sive vs. Quantity: Αν ξανά ανατρέξουμε στο σχήμα 3.4 στην πρώτη συσταδοποίηση τα 2n αντικείμενα μειώνουν το recall κατά 50%. Αυτό οδηγεί σε μια ολική μείωση που φαίνεται από την σχέση 3.19.

$$DEC_{D_1} = \frac{2n}{N_{tot}} - \frac{2n\frac{1}{2}}{N_{tot}} = \frac{n}{N_{tot}} \quad (3.19)$$

Στην δεύτερη συσταδοποίηση το recall των n αντικειμένων μειώνεται από το 1 στο $\frac{n}{n+1}$ και το recall ενός αντικειμένου μειώνεται από το 1 στο $\frac{1}{n+1}$. Έτσι η ολική μείωση στην δεύτερη συσταδοποίηση φαίνεται στην σχέση 3.20

$$DEC_{D_2} = \frac{n+1}{N_{tot}} - \frac{n\frac{n}{n+1} + \frac{1}{n+1}}{N_{tot}} = \frac{\frac{2n}{n+1}}{N_{tot}} \simeq \frac{2}{N_{tot}} \quad (3.20)$$

Παρατηρούμε ότι $DEC_{D_2} < DEC_{D_1}$.

Από ότι βλέπουμε οι BCubed μετρικές ικανοποιούν όλες τις ιδιότητες που αναφέραμε στις υποενότητες 3.2.1 – 3.2.4. Πιο συγκεκριμένα η μετρική BCubed precision ικανοποιεί την Rag Bag και την Cluster Homogeneity. Ενώ η BCubed recall την Cluster Completeness και την Clusters size vs. quantity.

Οι BCubed μετρικές precision και recall συνδυάζονται σε μια μετρική που με έναν αριθμό συνοψίζει το πόσο καλή είναι μια συσταδοποίηση. Μια μετρική που συνδυάζει τις BCubed precision και recall ικανοποιεί όλες τις ιδιότητες 3.2.1 – 3.2.4

Η μετρική F [70] συνδυάζει τις μετρικές BCubed precision και recall και δίνεται από την σχέση 3.21

$$F(R,P) = \frac{1}{a\left(\frac{1}{p}\right) + (1-a)\frac{1}{R}} \quad (3.21)$$

Στην σχέση 7.8 το R αντιστοιχεί στο BCubed recall και το P στο BCubed precision. Μέσω του α και του (1-α) δηλώνουμε τι βαρύτητα δίνουμε στο precision και στο recall αντίστοιχα. Αν πάρουμε α = 0,5 τότε έχουμε τον αρμονικό μέσο του P και R.

3.8.Μετρικές Επικαλυπτόμενης Συσταδοποίησης

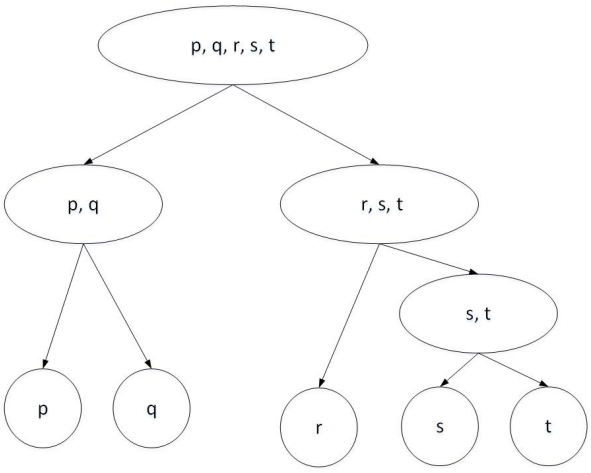
Στην ενότητα 3.7 είδαμε ότι οι μετρικές BCubed precision και recall ικανοποιούν όλα τις ιδιότητες 3.2.1 – 3.2.4 και θεωρούνται τα καλύτερα κριτήρια, μαζί με την F μετρική που τις συνδυάζει, για να κρίνουν κατά πόσο μια συσταδοποίηση είναι καλή.

Οι μετρικές συσταδοποιήσης όμως έτσι όπως διατυπωθήκαν στην ενότητα 3.7, καθώς και στις προηγούμενες 3.3, 3.4, 3.5, 3.6 περιγράφουν την περίπτωση που η συσταδοποίηση έγινε με τέτοιο τρόπο, που κάθε αντικείμενο να ανήκει σε μία και μόνο μια συστάδα. Υπάρχει πολύ συχνά η περίπτωση τα αντικείμενα να ανήκουν ταυτόχρονα σε περισσότερες από μία συστάδες.

Ας θεωρήσουμε την περίπτωση που έχουμε να συσταδοποιήσουμε κείμενα σε ένα πλήθος θεμάτων. Κάποια κείμενα μπορεί να διαπραγματεύονται ένα μόνο θέμα όπως τα αθλητικά, οπότε να συσταδοποιηθούν σε ένα cluster και κάποια κείμενα να ανήκουν σε περισσότερα θέματα, όπως ένα κείμενο που διαπραγματεύεται πολιτική και οικονομικά ταυτόχρονα. Υπάρχουν αρκετοί αλγόριθμοι συσταδοποίησης που είναι σε θέση να συσταδοποιήσουν αντικείμενα σε επικαλυπτόμενες συστάδες. Τέτοιες συσταδοποιήσεις χρειάζονται μια διαφορετική μετρική που να τις αξιολογεί.

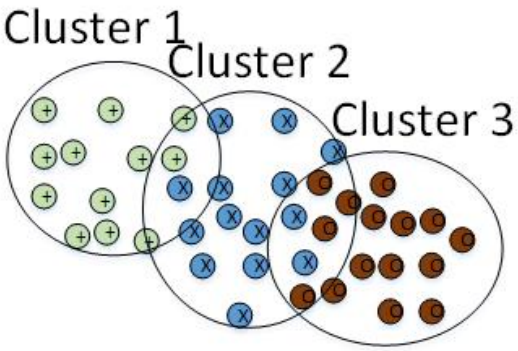
Θα διατυπώσουμε μια επέκταση των BCubed precision και recall μετρικών, που είδαμε στην ενότητα 3.7, για να είναι σε θέση να αξιολογούν συσταδοποιήσεις πού τα αντικείμενα μπορούν να ανήκουν σε περισσότερες από μια συστάδες ταυτόχρονα.

Μια επικαλυπτόμενη συσταδοποίηση μπορεί να έχει δύο μορφές. Στην πρώτη μορφή είναι η ιεραρχική συσταδοποίηση όπου έχουμε βασικές συστάδες που εξειδικεύονται σε επιμέρους συστάδες έχοντας μια δενδροειδή μορφή όπως φαίνεται στο σχήμα 3.9 είτε σε απλές επικαλυπτόμενες συστάδες που δεν παρουσιάζουν κάποια σχέση κληρονομικότητας αναμεταξύ τους και φαίνονται στο σχήμα 3.10.



Σχήμα 3.9 Ιεραρχική συσταδοποίηση

Στο σχήμα 3.9 παρατηρούμε πως κάθε κόμβος του δέντρου είναι μια συστάδα η οποία εξειδικεύεται καθώς κατεβαίνουμε σε συστάδες που είναι υποσύνολα του αρχικού κόμβου. Οι Ιεραρχικές συσταδοποιήσεις μπορούν να γίνουν είτε από πάνω προς τα πάνω όπου τις γενικές συστάδες τις εξειδικεύουμε σε μικρότερες είτε από κάτω προς τα πάνω όπου ενώνουμε σε κάθε επίπεδο τις πιο όμοιες συστάδες αναμεταξύ τους.



Σχήμα 3.10 Επικαλυπτόμενες συστάδες

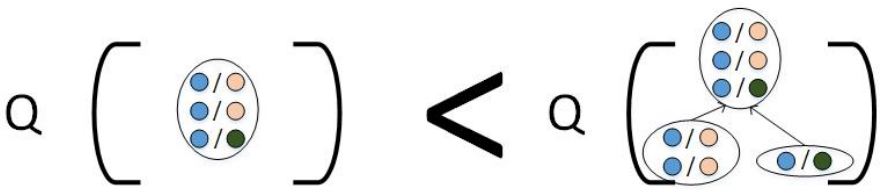
Στο σχήμα 3.10 παρατηρούμε την Επικαλυπτόμενη συσταδοποίηση χωρίς να έχει ιεραρχική δομή. Παρατηρούμε ότι κάποιοι κόμβοι ανήκουν σε μια συστάδα κάποιοι σε δύο και κάποιοι σε τρεις.

Η ιεραρχική συσταδοποίηση έχει σχέσεις γονέα – παιδί. Γενικώς όμως μια οποιαδήποτε ιεραρχική συσταδοποίηση μπορεί να θεωρηθεί επικαλυπτόμενη συσταδοποίηση χωρίς να ισχύει το αντίστροφο.

Μια μετρική που χρησιμοποιείται για να αξιολογήσει επικαλυπτόμενες συστάδες θα πρέπει να λαμβάνει υπόψη της την απαίτηση ότι αν δύο αντικείμενα συνυπάρχουν μαζί στις n τον αριθμό ίδιες κατηγορίες θα πρέπει να συνυπάρχουν μαζί και στις n συστάδες. Αν και αυτή η απαίτηση φαίνεται τετριμμένη οι περισσότερες μετρικές όπως αυτές που βασίζονται στην εντροπία και η purity δεν την ικανοποιούν. Ο λόγος είναι ότι αξιολογούν μια συσταδοποίηση είτε βασιζόμενες στις συστάδες (purity) είτε στις κατηγορίες (inverse purity). Παρακάτω θα δώσουμε ένα σύντομο παράδειγμα όπου φαίνεται γιατί δεν μπορούν να αξιολογήσουν σωστά μια ιεραρχική συσταδοποίηση.

Έστω ότι έχουμε μια τετριμμένη συσταδοποίηση με τρία αντικείμενα, όπως φαίνεται στο σχήμα 3.11 δεξιά, που συσταδοποιούνται ιεραρχικά. Η αριστερή συσταδοποίηση εισαγάγει όλα τα αντικείμενα σ' ένα cluster.

Όλες οι μετρικές που βασίζονται στις κατηγορίες ή τις συστάδες αποτυγχάνουν να αξιολογήσουν την αριστερή συσταδοποίηση χαμηλότερα από την δεξιά. Αυτό οφείλεται στο ότι στην αριστερή μοναδιαία συστάδα όλα τα αντικείμενα προέρχονται από μια κατηγορία, την γκρι και ταυτόχρονα όλα τα αντικείμενα από την κάθε κατηγορία βρίσκονται στο ίδιο cluster.



Σχήμα 3.11 Ιεραρχική συσταδοποίηση έναντι απλής συσταδοποίησης

Σχήμα 3.11 Η δεξιά, ιεραρχική συσταδοποίηση πρέπει να αξιολογηθεί υψηλότερα από την αριστερή. Παρόλα αυτά μετρικές σχέσεις όπως η Purity και Inverse Purity δεν το επιτυγχάνουν.

Οι BCubed μετρικές υπολογίζουν ανεξάρτητα το precision και το recall για κάθε αντικείμενο της συσταδοποίησης. Το precision ενός αντικειμένου ισούται με το πλήθος των αντικειμένων, της συστάδας που βρίσκεται το αντικείμενο που ανήκουν στην ίδια κατηγορία με το αντικείμενο διά το πλήθος των αντικειμένων μέσα στο cluster.

Αντίστοιχα το recall ενός αντικειμένου ισούται με το πλήθος των αντικειμένων που βρίσκονται στην ίδια συστάδα και προέρχονται από την ίδια κατηγορία διά το πλήθος των αντικειμένων της κατηγορίας.

Το αν δυο αντικείμενα συσχετίζονται σωστά σε μια συσταδοποίηση με βάση τις αρχικές, σωστές τους κατηγορίες σε μία μη επικαλυπτόμενη συσταδοποίηση δίνεται από την σχέση 3.15.

Στην περίπτωση μιας επικαλυπτόμενης συσταδοποιήσης η σχέση μεταξύ δύο αντικειμένων δεν μπορεί να αναπαρασταθεί ως μια δυαδική συνάρτηση, όπως της σχέσης 3.15. Αυτό οφείλεται στο ότι στα επικαλυπτόμενα cluster πρέπει να λάβουμε υπόψη μας ότι κάθε αντικείμενο μπορεί να ανήκει σε πολλές κατηγορίες και συστάδες. Για παράδειγμα δύο αντικείμενα μπορεί να μοιράζονται δύο κατηγορίες και στην συσταδοποίηση που έγινε να μοιράζονται μία ή τρεις συστάδες. Αν μοιράζονται μία συστάδα τότε έχουμε απώλεια πληροφορίας. Αν μοιράζονται τρεις συστάδες τότε εισάγεται λάθος πληροφορία.

Οι μετρικές Multiplicity Precision και Multiplicity Recall μεταξύ δύο αντικειμένων μπορούν να ικανοποιήσουν τις απαιτήσεις της αξιολόγησης μιας επικαλυπτόμενης συσταδοποίησης. Οι μετρικές Multiplicity Precision και Multiplicity Recall δίνονται από τις σχέσεις 3.21 και 3.22

$$Multiplicity\,Precision(e,e') = \frac{\min(\left|C(e)\bigcap C(e')\right|, \left|L(e)\bigcap L(e')\right|)}{\left|C(e)\bigcap C(e')\right|} \quad (3.21)$$

$$Multiplicity\,Recall(e,e') = \frac{\min(\left|C(e)\bigcap C(e')\right|, \left|L(e)\bigcap L(e')\right|)}{|L(e)\bigcap L(e')|} \quad (3.22)$$

Όπου e και e΄ είναι ένα ζευγάρι αντικειμένων, L(e) είναι το σύνολο των κατηγοριών και C(e) το σύνολο των συστάδων στις οποίες ανήκει ο κόμβος e.

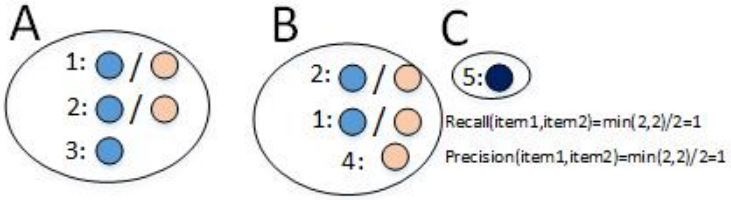
Η Multiplicity Precision μετρική ορίζεται μόνο όταν οι κόμβοι e και e΄ βρίσκονται μαζί σε κάποια συστάδα και η Multiplicity Recall μετρική ορίζεται μόνο όταν οι κόμβοι e και e΄ βρίσκονται μαζί σε κάποια κατηγορία.

Η μετρική Multiplicity Precision είναι μέγιστη, ίση με ένα, όταν ο αριθμός των συστάδων που βρίσκονται δύο κόμβοι είναι μικρότερος ή ίσος με τον αριθμό των αρχικών κατηγοριών που βρίσκονται μαζί. Η μετρική είναι ίση με μηδέν όταν δύο αντικείμενα δεν προέρχονται από καμία κοινή κατηγορία.

Αντίστοιχα η μετρική Multiplicity Recall χρησιμοποιείται όταν δύο αντικείμενα προέρχονται από μία ή περισσότερες κατηγορίες, είναι μέγιστη όταν το πλήθος των κατηγοριών από τις οποίες προέρχονται τα αντικείμενα είναι λιγότερες ή ίσες από το πλήθος των cluster που συσταδοποιήθηκαν και είναι ελάχιστη όταν δύο αντικείμενα δεν βρίσκονται σε καμία συστάδα μαζί.

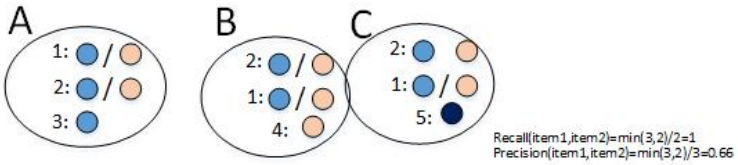
Η μετρική Multiplicity Precision μεγαλώνει αν ταιριάζουν οι κατηγορίες για κάθε συστάδα όπου δύο αντικείμενα συνυπάρχουν μαζί. Η μετρική Multiplicity Recall μεγαλώνει όταν προσθέτουμε μια συστάδα μέσα στην οποία είναι δύο αντικείμενα τα οποία βρίσκονται μαζί και σε μια κατηγορία. Αν έχουμε λιγότερες συστάδες από ότι χρειάζεται μειώνεται το Recall. Αν έχουμε λιγότερες κατηγορίες από όσες συστάδες

έχουμε τότε μειώνεται το Precision. Τα σχήματα 3.12 – 3.14 δείχνουν ένα παράδειγμα πως υπολογίζονται οι μετρικές.



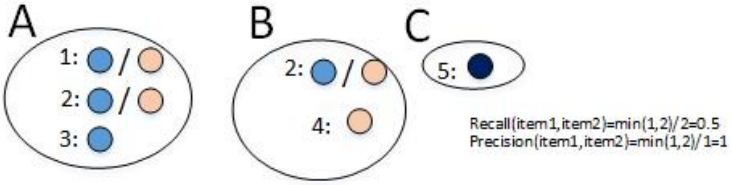
Σχήμα 3.12 Συσταδοποίηση με Precision και Recall ένα

Στο σχήμα 3.12 βλέπουμε πως τα αντικείμενα 1 και 2 που προέρχονται από δύο κατηγορίες συσταδοποιούνται μαζί σε δύο κατηγορίες. Η συσταδοποίηση συμφωνεί πλήρως με τις αρχικές κατηγορίες για αυτό το Precision και το Recall είναι ένα.



Σχήμα 3.13 Συσταδοποίηση με μείωση της μετρικής Recall

Στο σχήμα 3.13 Τα αντικείμενα 1 και 2 προέρχονται από δύο κατηγορίες αλλά συσταδοποιούνται μαζί σ' ένα cluster. Στην περίπτωση αυτή έχουμε χάσει πληροφορία. Η μετρική Recall αντιλαμβάνεται και ερμηνεύει την απώλεια πληροφορίας με μία μειωμένη τιμή.



Σχήμα 3.14 Συσταδοποίηση με μείωση της μετρικής Precision

Στο σχήμα 3.14 Τα αντικείμενα 1 και 2 προέρχονται από δύο κατηγορίες, αλλά συσταδοποιούνται μαζί σε τρία cluster. Στην περίπτωση αυτή προστίθεται πλεονάζουσα πληροφορία. Η μετρική Precision αντιλαμβάνεται και ερμηνεύει την εισαγωγή λάθος πληροφορίας με μία μειωμένη τιμή.

Την μετρική Multiplicity Recall και Multiplicity Precision τις προσαρμόζουμε στις BCubed μετρικές. Για αυτό θα χρησιμοποιήσουμε τους αρχικούς BCubed ορισμούς, αλλά θα αντικαταστήσουμε την Correctness συνάρτηση με την Multiplicity για το Precision και Recall αντίστοιχα.

Η καινούρια BCubed precision μετρική που συσχετίζεται με ένα αντικείμενο θα είναι ίση με τον μέσο όρο του Multiplicity Precision με όλα τα αντικείμενα που υπάρχουν μαζί σε κάποιες κατηγορίες. Το συνολικό BCubed Precision για επικαλυπτόμενες συστάδες είναι ίσο με τον μέσο όρο του Precision για όλα τα αντικείμενα. Με τον ίδιο τρόπο υπολογίζεται και το συνολικό BCubed Recall για επικαλυπτόμενες συστάδες. Οι δύο μετρικές αποδίδονται στις εξισώσεις 3.23 και 3.24.

$$Precision\,BCubed = Avg_e[Avg_{e'.C(e)\cap C(e')\neq 0}[Multiplicity\,precision(e,e')]] \quad (3.23)$$

$$Recall\,BCubed = Avg_e[Avg_{e'.C(e)\cap C(e')\neq 0}[Multiplicity\,recall(e,e')]] \quad (3.24)$$

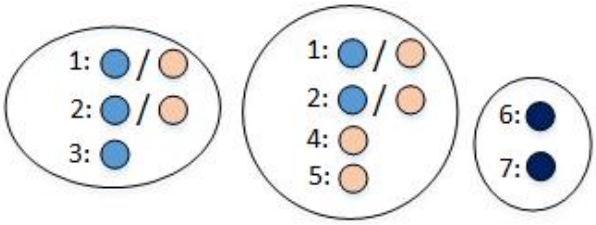
Οι μετρικές 3.23 και 3.24 περιλαμβάνουν στον υπολογισμό τους την σχέση κάθε αντικειμένου με τον εαυτό του. Αυτό μειώνει τον συντελεστή της μετρικής, αν διαγραφεί ή γραφτεί δύο φορές από μια συσταδοποίηση, μια συστάδα με ένα στοιχείο μόνο, ενώ δεν θα έπρεπε.

Στην περίπτωση που οι συστάδες δεν είναι επικαλυπτόμενες οι μετρικές 3.23 και 3.24 μπορούν να συμπεριφερθούν όπως οι 3.15 και 3.16 αποδίδοντας το Precision και το Recall σαν να ήταν απλή συσταδοποίηση.

Θα περιγράψουμε στις επόμενες παραγράφους μια εφαρμογή των μετρικών για μια επικαλυπτόμενη συσταδοποίηση με σκοπό να φανεί πως οι μετρικές επηρεάζονται από τον τρόπο που συσταδοποιούνται τα αντικείμενα.

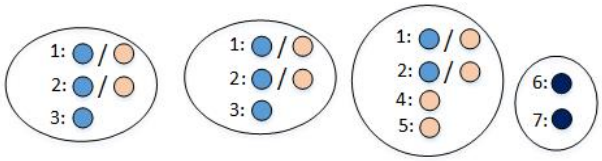
Έχουμε αρχικά τις κατηγορίες που φαίνεται στο σχήμα 3.15. Αν μια συσταδοποίηση έχει θόρυβο, όπως το να έχει αναπαράγει δύο φορές την ίδια κατηγορία (σχήμα 3.16), τότε το recall παραμένει μέγιστο αλλά μειώνεται το precision. Αν η συσταδοποίηση έχει λιγότερη πληροφορία από ότι θα έπρεπε μειώνεται το recall.

Στην περίπτωση που μια κατηγορία είναι διαιρεμένη σε δύο συστάδες (σχήμα 3.19) κάποια ζευγάρια αντικειμένων δεν θα είναι μαζί για αυτό το recall μειώνεται. Από την άλλη, αν δύο κατηγορίες ενωθούν σε μία συστάδα (σχήμα 3.20) τότε κάποιες από τις νέες συνδέσεις δεν θα είναι σωστές με αποτέλεσμα αμφότερα το precision και το recall να μειωθούν.



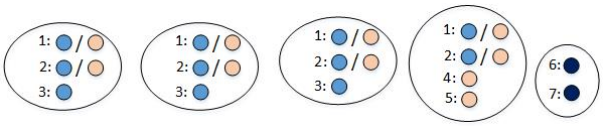
Σχήμα 3.15 Ορθή συσταδοποίηση

Στο σχήμα 3.15 βλέπουμε την σωστή συσταδοποίηση που έχει φτιάξει τα cluster να συμφωνούν απόλυτα με τις κατηγορίες σε αυτή την περίπτωση Precision = 1 και Recall = 1



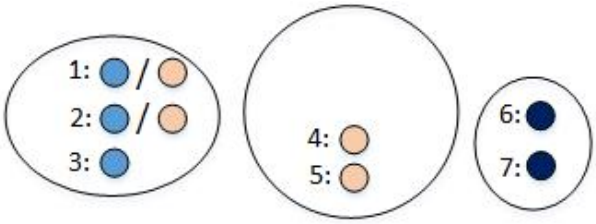
Σχήμα 3.16 Μείωση του Precision

Στο σχήμα 3.16 βλέπουμε πως όταν μία κατηγορία έχει αναπαραχθεί δύο φορές το Recall παραμένει το ίδιο αλλά το Precision μειώνεται. Precision =0,86 και Recall = 1



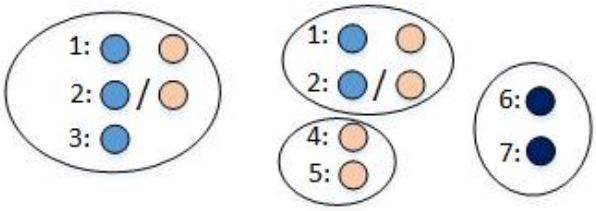
Σχήμα 3.17 Περαιτέρω μείωση του Precision

Στο σχήμα 3.17 βλέπουμε ότι όταν μία κατηγορία έχει αναπαραχθεί τρεις φορές το Recall παραμένει το ίδιο αλλά το Precision μειώνεται ακόμη περισσότερο. Precision =0,80 και Recall = 1



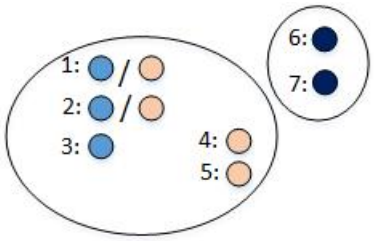
Σχήμα 3.18 παράληψη παρατηρήσεων και μείωση του Recall

Στο σχήμα 3.18 Δύο αντικείμενα από την μεσαία κατηγορία δεν περιελήφθησαν όπως θα έπρεπε στην μία από τις δύο συστάδες όπως θα αναλογούσε. Το precision παραμένει ίδιο ενώ το Recall μειώνεται. Precision = 1 και Recall = 0.68



Σχήμα 3.19 Μείωση του Recall μέσω διάσπασης μιας συστάδας σε δύο

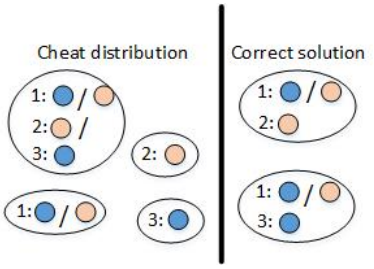
Στο σχήμα 3.19 παρατηρούμε πως δυο μικρότερες συστάδες όταν μοιράζονται τα αντικείμενα που θα έπρεπε να περιέχονται στην αρχική μεσαία κατηγορία το Precision παραμένει ίδιο ενώ το Recall μειώνεται. Precision = 1 και Recall = 0.74



Σχήμα 3.20 Μείωση του Precision και του Recall μέσω ένωσης συστάδων

Στο σχήμα 3.20 βλέπουμε πως μια συστάδα περιέχει δύο κατηγορίες, ενώνοντας την αριστερή και την μεσαία σε μία μειώνονται και το Precision και το Recall. Precision = 0.88 και Recall = 0.94

Τέλος παρουσιάζουμε μια περίπτωση όπου η συσταδοποίηση είναι πολύ διαφορετική από τις αρχικές κατηγορίες, οι BCubed μετρικές το αντιλαμβάνονται αλλά μετρικές όπως η purity και η inverse purity όχι.



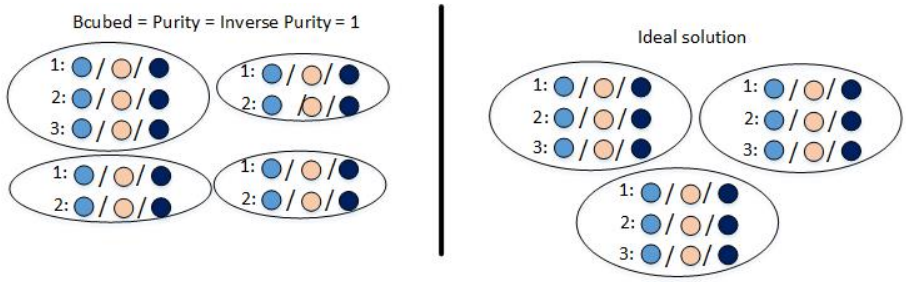
Σχήμα 3.21 Κάθε αντικείμενο αποτελεί δική του συστάδα

Στο σχήμα 3.21 βλέπουμε την δεξιά σωστή κατηγοριοποίηση και αριστερά φτιάξαμε μία συστάδα που περιέχει όλα τα αντικείμενα από τις κατηγορίες και μία συστάδα για κάθε αντικείμενο που είναι μόνο του.

Η αριστερή συστάδα του σχήματος 3.21 παίρνει μέγιστη Inverse Purity τιμή γιατί όλα τα αντικείμενα είναι συσταδοποιημένα μαζί στο μεγάλο cluster και μέγιστη Purity τιμή γιατί όλα τα αντικείμενα είναι μόνα τους σε μια συστάδα το κάθε ένα. Λόγω του ότι οι BCubed μετρικές υπολογίζουν το Precision με βάση τα αντικείμενα και όχι τις συστάδες που είναι σε θέση να εκφράσει την διαφορά.

Είδαμε ότι οι μετρικές BCubed ικανοποιούν όλες τις ιδιότητες 3.2.1- 3.2.4 και αποδίδουν καλύτερα από οποιεσδήποτε άλλες μετρικές και για επικαλυπτόμενα cluster και για απλά. Παρόλα αυτά υπάρχει μια αδυναμία. Μια μεγάλη BCubed τιμή δεν σημαίνει απαραίτητα μια σωστή συσταδοποίηση. Αυτό ισχύει για τις επικαλυπτόμενες συσταδοποιησεις και θα εξηγήσουμε που οφείλεται με βάση το σχήμα 3.22.

Στην επικαλυπτόμενη συσταδοποίηση του αριστερού σχήματος 3.22 οι BCubed μετρικές είναι μέγιστες γιατί οι τρεις κόμβοι εμφανίζονται τρεις φορές ο κάθε ένας όπως στις κατηγορίες και κάθε ζευγάρι κόμβων βρίσκεται σε τρεις κατηγορίες και συστάδες.



Σχήμα 3.22 Περίπτωση αποτυχίας των BCubed μετρικών

Στο σχήμα 3.22 βλέπουμε την δεξιά σωστή κατηγοριοποίηση των αντικειμένων. Η αριστερή συσταδοποίηση έχει τις BCubed μετρικές ίσες με ένα χωρίς να είναι η σωστή συσταδοποίηση.

Ευτυχώς η περίπτωση του σχήματος 3.22 σε πρακτικά προβλήματα συσταδοποίησης είναι αμελητέα και δεν θα πρέπει να μας αποθαρρύνει από το να χρησιμοποιούμε και να εμπιστευόμαστε τις BCubed μετρικές.

3.9.Πίνακες Σύγχυσης

Οι πίνακες σύγχυσης (Confusion πίνακες) [71], πολλές φορές τους συναντάμε και με την ονομασία πίνακες συνάφειας και πίνακες λάθους, είναι πίνακες που με μια εποπτική εικόνα παρουσιάζουν συνοπτικά σε τι βαθμό είναι μια συσταδοποίηση καλή σε σχέση με τις αρχικά σωστές κατηγορίες.

Κάθε στήλη του Confusion πίνακα αντιπροσωπεύει αντικείμενα σε μια συστάδα και κάθε γραμμή αντιπροσωπεύει σε ποια κατηγορία ανήκουν τα αντικείμενα πραγματικά.

Ένα από τα βασικά πλεονεκτήματα του Confusion πίνακα είναι ότι μπορεί πολύ γρήγορα ο ενδιαφερόμενος να αντιληφθεί τα συστηματικά λάθη που μπορεί να γίνονται σε μια συσταδοποίηση, βλέποντας από ποια κατηγορία προέρχονται τα αντικείμενα που συσταδοποιούνται λάθος σε ένα cluster.

	Συστάδες				
Κατηγορίες		Α	Β	Γ	Δ
	Α	8	1	1	0
	Β	0	9	0	1
	Γ	4	2	4	0
	Δ	4	0	0	6

Πίνακας 3.1 Συσταδοποίηση 40 αντικείμενων

Στον Πίνακας 3.1 βλέπουμε την συσταδοποιήση 40 αντικειμένων. Κάθετα έχουμε τις συστάδες οριζόντια έχουμε τις σωστές κατηγορίες. Κάθε κελί δηλώνει τον αριθμό των αντικειμένων που προέρχονται από την κάθε κατηγορία (γραμμή) σε ποια cluster συσταδοποιούνται (στήλη). Ιδανικά θα έπρεπε όλα τα στοιχεία εκτός από την κύρια διαγώνια να είναι μηδενικά.

Μια παραλλαγή του Confusion πίνακα (table of confusion), αποτελείται από δύο γραμμές και δύο στήλες και αναφέρει το πλήθος των λάθος θετικών συσταδοποιήσεων, λάθος αρνητικών συσταδοποιήσεων, σωστών θετικών συσταδοποιήσεων και σωστών αρνητικών συσταδοποιήσεων.

Οι λάθος θετικές συσταδοποιήσεις είναι όταν ένα αντικείμενο καταχωρήθηκε σε μια συστάδα ενώ δεν έπρεπε να καταχωρηθεί. Οι λάθος αρνητικές συσταδοποιήσεις είναι όταν ένα αντικείμενο έπρεπε να καταχωρηθεί σε μια συστάδα αλλά δεν καταχωρήθηκε. Οι σωστές θετικές συσταδοποιήσεις είναι όταν ένα αντικείμενο έπρεπε να καταχωρηθεί σε μια συστάδα και καταχωρήθηκε. Τέλος οι σωστές αρνητικές συσταδοποιήσεις είναι όταν ένα αντικείμενο δεν έπρεπε να καταχωρηθεί σε μια συστάδα και δεν καταχωρήθηκε.

Αυτή η παρουσίαση είναι πιο λεπτομερής από το να αναφέρουμε απλά το ποσοστό των σωστών συταδοποιήσεων. Βέβαια για κάθε κατηγορία θα πρέπει να σχηματίσουμε έναν πίνακα και δεν μπορούμε με έναν πίνακα να αναπαραστήσουμε συνοπτικά όλη την πληροφορία. Στον πίνακα 3.2 παρουσιάζουμε τον table of confusion για την κατηγορία Α του πίνακα 3.1

8 true positives	2 false negatives
8 false positives	22 true negatives

Πινάκας 3.2 Table of Confusion για την κατηγορία Α

Στον πίνακα 3.2 έχουμε το Table of Confusion για την κατηγορία Α. Έχουμε 8 true positives γιατί συσταδοποίησαμε ορθώς 8 αντικείμενα στην συστάδα Α. 2 false negatives γιατί αν και η κατηγορία Α είχε 8 αντικείμενα συσταδοποίησαμε στο αντίστοιχο cluster μόνο τα 8 και τα υπόλοιπα 2 μπήκαν σε λάθος cluster (Β και Γ). 8 false positives γιατί στην συστάδα Α μπήκαν 8 αντικείμενα 4 από την Γ και 4 από την Δ που δεν θα έπρεπε. 22 true negatives γιατί 22 αντικείμενα ορθώς δεν συσταδοποιήθηκαν στην συστάδα Α και εισήχθηκαν σε άλλες.

3.10.Εσωτερικά Κριτήρια

Όλες οι μετρικές συσταδοποίησης που περιγράψαμε στις προηγούμενες ενότητες βασίζονται στο ότι έχουμε ένα data set το οποίο είναι ήδη κατηγοριοποιημένο και εφαρμόζουμε σε αυτό έναν αλγόριθμο συσταδοποίησης. Οι μετρικές αυτές μας λένε σε τι βαθμό μια συσταδοποιήση είναι αντίστοιχη με την αρχική κατηγοριοποίηση. Αυτές οι μετρικές χαρακτηρίζονται ως εξωτερικές μετρικές.

Στις επόμενες ενότητες θα περιγράψουμε μια κατηγορία μετρικών σχέσεων που δεν χρησιμοποιεί ένα ήδη κατηγοριοποιημένο data set για να αξιολογήσει μια συσταδοποίηση. Αυτές οι μετρικές σχέσεις λέμε ότι χρησιμοποιούν εσωτερικά κριτήρια αξιολόγησης. Η αξιολόγηση αυτή γίνεται στις συστάδες που έχουν παραχθεί από τον αλγόριθμο συσταδοποίησης αγνοώντας την σωστή κατηγοριοποίηση των δεδομένων.

Τα εσωτερικά κριτήρια που χρησιμοποιούνται βαθμολογούν υψηλά τις συσταδοποιήσεις που παράγουν clusters τα οποία έχουν την ακόλουθη ιδιότητα. Τα αντικείμενα μέσα σε κάθε cluster έχουν υψηλή ομοιότητα ενώ τα αντικείμενα μεταξύ διαφορετικών cluster έχουν από ελάχιστη έως καθόλου ομοιότητα.

Οι μετρικές που χρησιμοποιούν εσωτερικά κριτήρια πολλές φορές δεν είναι αξιόπιστες για κάποιους αλγόριθμους συσταδοποίησης. Αυτό οφείλεται στο ότι μπορεί το κριτήριο που συσταδοποιεί ένας αλγόριθμος και το κριτήριο που μια μετρική αξιολογεί τις συστάδες που παρήχθησαν να είναι το ίδιο ή παραπλήσιο.

3.11.Δείκτης Davies–Douldin

Ο δείκτης Davies – Bouldin[72] εφαρμόζεται σε αντικείμενα που έχουν μια διανυσματική αναπαράσταση και έχουν συσταδοποιηθεί σε έναν n-διάστατο χώρο. Ο δείκτης αυτός βασίζεται στην διασπορά που υπάρχει στα αντικείμενα μέσα σ' ένα cluster καθώς και σε τι βαθμό είναι διαχωρισμένα αναμεταξύ τους τα cluster. Η Davies–Bouldin index τιμή που προκύπτει εξαρτάται και από τον αλγόριθμο συσταδοποίησης και από την μορφή που έχουν τα δεδομένα.

Ο τύπος που δίνει την διασπορά των διανυσμάτων μέσα σε μια συστάδα δίνεται από την σχέση 3.25. C_i είναι μια συστάδα από διανύσματα. A_i είναι το κέντρο (centroid) της συστάδας C_i και T_i είναι το πλήθος των κόμβων που έχει η συστάδα. X_j είναι ο j-στος κόμβος που ανήκει στην συστάδα C_i.

$$S_i = \left(\frac{1}{T_i} \sum_{j=1}^{T_i} \left| X_j - A_i \right|^q \right)^{\frac{1}{q}} \quad (3.25)$$

Το q συνήθως το παίρνουμε ίσο με 2 και δηλώνει την ευκλείδεια απόσταση μεταξύ του centroid της i συστάδας και του X_j κόμβου. Για να έχουμε σωστά αποτελέσματα, η επιλογή της μετρικής απόστασης θα πρέπει να γίνει με βάση την μετρική που έχει χρησιμοποιηθεί στην συσταδοποίηση.

Η σχέση $M_{i,j}$ δηλώνει σε τι βαθμό είναι διαχωρισμένες δύο συστάδες C_i, C_j αναμεταξύ τους και δίνεται από τον τύπο 3.26

$$M_{i,j} = \left\| A_i - A_j \right\|_p = \left(\sum_{k=1}^{n} \left| a_{k,i} - a_{k,j} \right|^p \right)^{\frac{1}{p}} \quad (3.26)$$

$\alpha_{k,i}$ είναι η k-στη συνιστώσα του centroid του cluster C_i. Θεωρούμε ότι κάθε διάνυσμα έχει n διαστάσεις οπότε το ίδιο ισχύει και για τα centroid. Με την επιλογή p = 2 η εξίσωση 3.26 χρησιμοποιεί την ευκλείδεια απόσταση αντίστοιχα με την 3.25.

Τις τιμές S_i και $M_{i,j}$ τις συνδυάζουμε σε μία σχέση την $R_{i,j}$. Η σχέση $R_{i,j}$ δηλώνει σε τι βαθμό η συστάδα i είναι μακριά από την συστάδα j και κατά πόσο τα αντικείμενα μιας συστάδας είναι κοντά ή διάσπαρτα αναμεταξύ τους. Το ιδανικό θα ήταν τα αντικείμενα μέσα σε μια συστάδα να είναι κοντά αναμεταξύ τους ενώ οι συστάδες μακριά. Το $R_{i,j}$ δίνεται από την σχέση 3.27.

$$R_{i,j} = \frac{S_i + S_j}{M_{i,j}} \quad (3.27)$$

Η σχέση 3.27 έχει τις ακόλουθες ιδιότητες.

$R_{i,j} \geq 0$ (3.28)

$R_{i,j} = R_{j,i}$ (3.29)

Αν $S_j \geq S_k \, και \, M_{i,j} = M_{i,k} \, τότε \, R_{i,j} > R_{i,k}$ (3.30)

Αν $S_j = S_k \, και \, M_{i,j} \leq M_{i,k} \, τότε \, R_{i,j} > R_{i,k}$ (3.31)

Οι συσταδοποιήσεις που έχουν την χαμηλότερη τιμή $R_{i,j}$ είναι οι καλύτερες. Από τα $R_{i,j}$ υπολογίζουμε το D_i να είναι το μεγαλύτερο, σχέση 3.32 και το DB να είναι ο μέσος όρος τον D_i σχέση 3.33.

$$D_i \equiv \max_{j:i \neq j} R_{i,j} \text{ (3.32)}$$

$$DB \equiv \frac{1}{N} \sum_{i=1}^{N} D_i \text{ (3.33)}$$

Το DB της σχέσης 3.33 είναι η τιμή του Davies Bouldin index. Η τιμή DB βασίζεται στην σχέση 3.32 όπου πήραμε την χειρότερη πιθανή τιμή θα μπορούσαμε αντί για το μέγιστο $R_{i,j}$ να είχαμε τον μέσο όρο.

3.12.Δείκτης Dunn

Η μετρική Dunn index[73] βασίζεται στην ίδια αρχή με την Davies – Bouldin index, προσπαθεί να αναγνωρίσει κατά πόσο οι συστάδες που υπάρχουν είναι πυκνές και έχουν μικρή διασπορά μεταξύ των μελών τους καθώς επίσης κατά πόσο οι συστάδες αναμεταξύ τους είναι καλά διαχωρισμένες.

Για μια δεδομένη συσταδοποίηση όσο πιο μεγάλος είναι ο συντελεστής Dunn index τόσο καλύτερα έχει σχηματιστεί.

Στην Dunn index μετρική χρησιμοποιείται η διάμετρος μια συστάδας. Η διάμετρος μιας συστάδας μπορεί να οριστεί με πολλούς τρόπους όπως φαίνεται στις επόμενες τρεις παραγράφους. C_i ορίζεται μια συστάδα διανυσμάτων, x και y δύο n διαστάσεων διανύσματα μέσα στο cluster C_i.

Η διάμετρος μιας συστάδας μπορεί να οριστεί ως η απόσταση μεταξύ των δύο πιο απομακρυσμένων σημείων μέσα στο cluster όπως δείχνει η σχέση 3.34.

$$\Delta_i = \max_{x,y \in C_i} d(x,y) \text{ (3.34)}$$

Η διάμετρος μιας συστάδας μπορεί να οριστεί ως η μέση απόσταση μεταξύ όλων των ζευγαριών των σημείων μέσα σε ένα cluster, όπως φαίνεται από την σχέση 3.35.

$$\Delta_i = \frac{1}{\left|C_i\right| \left|C_i - 1\right|} \sum_{x,y \in C_i, x \neq y} d(x,y) \text{ (3.35)}$$

Η διάμετρος μιας συστάδας μπορεί να οριστεί ως η μέση απόσταση οποιουδήποτε σημείου από το centroid μέσα στο cluster, όπως φαίνεται από την σχέση 3.36.

$$\Delta_i = \frac{\sum_{x \in C_i} d(x, \mu)}{\left| C_i \right|} \quad (3.36)$$

Όπου μ είναι το centroid του i cluster και δίνεται από την σχέση 3.37.

$$\mu = \frac{\sum_{x \in C_i} x}{\left| C_i \right|} \quad (3.37)$$

Έπειτα, υπολογίζουμε την απόσταση μεταξύ των cluster $\delta(C_i,C_j)$. Η απόσταση μεταξύ των cluster μπορεί να υπολογιστεί αντίστοιχα όπως κάναμε και για την απόσταση μεταξύ των κόμβων μέσα σ' ένα cluster. Οι τρεις βασικές επιλογές είναι το $\delta(C_i,C_j)$ να είναι ίσο με την απόσταση των δύο πιο απομακρυσμένων centroid ή το μέσο όρο απόστασης μεταξύ όλων των ζευγαριών των centroid ή να βρούμε το κέντρο όλων των cluster και τον μέσο όρο των αποστάσεων όλων των centroid από αυτό.

Η μετρική Dunn index για m συστάδες χρησιμοποιεί την διάμετρο μιας συστάδας Δ_i και την απόσταση μεταξύ των cluster $\delta(C_i,C_j)$ και δίνεται από την σχέση 3.38.

$$DI_m = \min_{1 \le i \le m} \left\{ \min_{1 \le j \le m, j \ne i} \left\{ \frac{\delta(C_i, C_j)}{\max\limits_{1 \le k \le m} \Delta_k} \right\} \right\} \quad (3.38)$$

Κάποιες τελευταίες παρατηρήσεις που θα πρέπει να κάνουμε είναι ότι η μετρική Dunn index μπορεί να χρησιμοποιήσει αντί της Ευκλείδειας απόστασης άλλες όπως την Manhattan απόσταση.

Σε προβλήματα όπου το πλήθος των συστάδων δεν είναι γνωστό μπορούμε να εφαρμόσουμε την Dunn index μετρική για μια σειρά συσταδοποιήσεων με διαφορετικό πλήθος συστάδων και εκεί όπου θα έχουμε το μέγιστο DI θα είναι το σωστό πλήθος συστάδων.

Τέλος μια αδυναμία αυτής της μετρικής σχέση είναι ότι αν υπάρχει μόνο ένα cluster που δεν είναι καλά συσταδοποιημένο αυτό θα επηρεάσει κατά πολύ το αποτέλεσμα, ακόμη και αν όλα τα άλλα είναι καλά συσταδοποιημένα. Αυτό οφείλεται στον παρανομαστή της σχέσης 3.38 που χρησιμοποιεί την μέγιστη διάμετρο από όλες τις συστάδες.

3.13.Συντελεστής Silhouette

Ο συντελεστής Silhouette[74] είναι και αυτός μια μετρική σχέση που χρησιμοποιεί εσωτερικά κριτήρια για να δηλώσει το κατά πόσο μια συσταδοποίηση είναι καλή. Η μετρική Silhouette βασίζεται στην ιδέα να μετρηθεί κατά πόσο ένας κόμβος ταιριάζει να βρίσκεται στην δική του συστάδα ή θα ταίριαζε περισσότερο να ήταν σε κάποια άλλη. Επεκτείνοντας για όλους τους κόμβους του data set μπορούμε να ξέρουμε κατά πόσο η συσταδοποίηση είναι καλή.

α(i) είναι η μέση ανομοιότητα του κόμβου i με όλα τα άλλα αντικείμενα του cluster στο οποίο βρίσκεται. Ως μέτρο ανομοιότητας μπορεί να χρησιμοποιηθεί μια συνάρτηση απόστασης όπως η ευκλείδεια ή η Manhattan απόσταση. Το a(i) ερμηνεύεται ως το κατά πόσο ταιριάζει ο κόμβος i να είναι μέσα στο cluster που βρίσκεται. Αυτό προκύπτει από την μέση απόσταση του κόμβου i με οποιονδήποτε άλλον κόμβο της συστάδας. Όσο πιο μικρή είναι η τιμή a(i) τόσο πιο καλά ταιριάζει.

Στην συνέχεια βρίσκουμε την ανομοιότητα του κόμβου i με όλες τις άλλες συστάδες. Η μικρότερη ανομοιότητα με κάποια άλλη συστάδα του κόμβου i είναι το b(i) και η συστάδα για την οποία προκύπτει αυτή η τιμή λέγεται γειτονική συστάδα.

Η μετρική Silhouette ορίζεται από τον τύπο 3.39.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (3.39)$$

Είναι φανερό ότι η τιμή της μετρικής Silhouette κυμαίνεται $-1 \leq s(i) \leq 1$. Μια τιμή του s(i) κοντά στο 1 δηλώνει ότι a(i) $\ll$ b(i) οπότε ο κόμβος i βρίσκεται σωστά στο cluster του. Μια τιμή του s(i) κοντά στο -1 δηλώνει ότι a(i) $\gg$ b(i) οπότε ο κόμβος i θα έπρεπε να βρίσκεται στην γειτονική του συστάδα. Μια τιμή του s(i) κοντά στο 0 δηλώνει ότι a(i) $\approx$ b(i) οπότε ο κόμβος i βρίσκεται ακριβώς στην μέση των δύο συστάδων.

Υπολογίζοντας την μέση τιμή του s(i) για όλους τους κόμβους i ξέρουμε κατά πόσο είναι σωστά συσταδοποιημένοι όλοι οι κόμβοι μέσα στα cluster τους.

3.14.Δεκαπλή-αναδίπλωση Διασταυρούμενης Αξιολόγησης

Για να αξιολογηθεί η ακρίβεια της διαδικασίας ταξινόμησης κειμένων χρησιμοποιούνται οι Μίκρο μετρικές, Μίκρο ακρίβειας (Micro precision), Μίκρο Ανάκλησης (Micro Recall), και Μίκρο Φ-Μέτρου (Micro F-Measure) και οι αντίστοιχες Μάκρο μετρικές, Μάκρο ακρίβειας (Macro Precision), Μάκρο Ανάκλησης (Macro Recall), και Μάκρο Φ-Μέτρου(Macro F-Measure) σε συνδυασμό με την δεκαπλή-αναδίπλωση διασταυρούμενης αξιολόγησης (Ten-fold cross-validation) [75].

Με την δεκαπλή-αναδίπλωση διασταυρούμενης αξιολόγησης, το σύνολο δεδομένων χωρίζεται σε δέκα ισάριθμα υποσύνολα, τα εννέα από αυτά χρησιμοποιούνται ως δεδομένα εκπαίδευσης και το ένα ως δεδομένα ελέγχου. Η μέθοδος επαναλαμβάνεται δέκα φορές. Κάθε φορά χρησιμοποιείται ένα διαφορετικό υποσύνολο κειμένων για έλεγχο και τα υπόλοιπα εννέα υποσύνολα ως δεδομένα εκπαίδευσης.

Χρησιμοποιώντας την δεκαπλή-αναδίπλωση διασταυρούμενης αξιολόγησης, όλα τα κείμενα του συνόλου δεδομένων θα δοκιμαστούν και θα ταξινομηθούν μία φορά και εννέα φορές θα χρησιμοποιηθούν για την εκπαίδευση του μοντέλου. Τα πλεονεκτήματα αυτής της μεθόδου αξιολόγησης είναι ότι είναι διεξοδική και παράγει αξιόπιστα αποτελέσματα, αλλά έχει το μειονέκτημα ότι η όλη διαδικασία της ταξινόμησης πρέπει εξολοκλήρου να λάβει μέρος δέκα φορές.

Η ακρίβεια, η ανάκληση και ο αρμονικός μέσος τους (Φ-Μέτρο) μπορούν να εκφράσουν την ποιότητα μιας μεθόδου κατηγοριοποίησης σε ένα σύνολο δεδομένων γνωρίζοντας τις σωστές κατηγορίες που αντιστοιχούν σε κάθε κείμενο. Η ακρίβεια για μια κλάση υποδηλώνει το ποσοστό των σωστά ανακτηθέντων κειμένων που είναι σχετικά με την κατηγορία. Η ανάκληση υποδηλώνει το ποσοστό των σχετικών κειμένων που ανακτώνται. Η ακρίβεια, η ανάκληση και το Φ-μέτρο δίδονται στις εξισώσεις 3.40, 3.41 και 3.42. Στον πίνακα 3.3 παρέχονται οι όροι που χρησιμοποιούνται.

	Κατηγορία Α	όχι Κατηγορία Α
Κλάση Α	Αληθώς Θετικά (TP)	Ψευδείς Θετικά(FP)
όχι Κλάση Α	Ψευδείς Αρνητικά(FN)	Αληθώς Αρνητικά(TN)

Πίνακας 3.3. Ταξινόμηση των εγγράφων σε όλες τις πιθανές περιπτώσεις

Ο όρος Κατηγορία χρησιμοποιείται για μια κατηγορία κειμένων όπως δίδεται στο σύνολο δεδομένων.

Ο όρος κλάση χρησιμοποιείται για την αντίστοιχη πρόβλεψη της κατηγορίας Α όπως εκτιμάται με μια μέθοδο κατηγοριοποίησης.

Ο όρος Αληθώς Θετικά (True Positives TP) είναι ο συνολικός αριθμός των κειμένων που προβλέφθηκαν σωστά στην κατηγορία Α και πραγματικά ανήκουν στην κατηγορία Α.

Ο όρος Αληθώς Αρνητικά (True Negatives TN) είναι ο συνολικός αριθμός των κειμένων που δεν ανήκουν στην κατηγορία Α και σωστά δεν προβλέφθηκαν στην κατηγορία Α.

Ο όρος Ψευδείς Θετικά (False Positives FP) είναι ο συνολικός αριθμός των κειμένων που δεν ανήκουν στην κατηγορία Α, αλλά προβλέφθηκαν ότι ανήκουν εσφαλμένα στην κλάση Α.

Ο όρος Ψευδείς Αρνητικά (False Negatives FN) είναι ο συνολικός αριθμός των κειμένων που αν και ανήκουν στην κατηγορία Α, εσφαλμένα δεν προβλέφθηκαν να ανήκουν στην κατηγορία Α.

$$Precision = \frac{TP}{TP + FP} (3.40)$$

Η ακρίβεια εκφράζει το βαθμό στον οποίο έχουν ανακτηθεί σχετικά κείμενα σε σχέση με όλα τα κείμενα (σχετικά και μη σχετικά) που ανακτήθηκαν στην κλάση. Η ακρίβεια είναι μια μέτρηση που περιγράφει την καθαρότητα μιας κλάσης.

$$Recall = \frac{TP}{TP + FN} (3.41)$$

Η Ανάκληση εκφράζει το βαθμό στον οποίο έχουν ληφθεί σχετικά κείμενα σε σχέση με όλα τα σχετικά κείμενα που περιλαμβάνει η κατηγορία. Εάν παραλείπονται ορισμένα κείμενα, ο βαθμός ανάκλησης μειώνεται. Η Ανάκληση εμφανίζει την πληρότητα κειμένων μιας κλάσης.

$$FMeasure = 2\frac{Precision \cdot Recall}{Precision + Recall} (3.42)$$

Η ακρίβεια και η ανάκληση, εκφράζουν το βαθμό επιτυχίας μιας κατηγοριοποίησης από διαφορετικές οπτικές γωνίες. Μια μετρική που τα συνδυάζει εξίσου είναι ο αρμονικός τους μέσος Φ-μετρική.

Η Φ-μετρική, όπως θα την χρησιμοποιήσουμε, είναι ο αρμονικός μέσος της ακρίβειας και της ανάκλησης ούτως ώστε να εκφράζει την ακρίβεια μιας κατηγοριοποίησης λαμβάνοντας υπόψη τόσο την καθαρότητα όσο και την πληρότητα κειμένων που κατηγοριοποιήθηκαν σε μία κλάση.

Η ακρίβεια, η ανάκληση και η Φ-μετρική όπως δίνονται στις εξισώσεις 3.40, 3.41 και 3.42 αναφέρονται σε μία κατηγορία. Ένας υπολογισμός πρέπει να γίνει για όλες τις κατηγορίες. Επιπλέον, όλα τα αποτελέσματα θα πρέπει να υπολογίζονται κατά μέσο όρο για όλες τις αναδιπλώσεις της δεκαπλής-αναδίπλωσης διασταυρούμενης αξιολόγησης. Στις επόμενες ενότητες περιγράφονται οι μέσες Μάκρο και Μίκρο σχέσεις και πως εξάγονται.

Μάκρο μετρικές

Οι μέσες τιμές ακρίβειας, ανάκλησης και Φ-μετρικής υπολογίζονται με τα ακόλουθα τρία βήματα.

1. Οι μετρήσεις υπολογίζονται για κάθε κλάση σε μία αναδίπλωση με βάση τις εξισώσεις 3.40, 3.41 και 3.42.

2. Οι μέσες μετρήσεις για όλες τις κλάσεις σε μία αναδίπλωση εκτιμώνται από το βήμα 1.

3. Οι μετρήσεις του βήματος 2 υπολογίζονται κατά μέσο όρο για όλες τις αναδιπλώσεις.

Μίκρο μετρικές

Οι μετρήσεις ακριβείας, ανάκλησης και Φ-μετρικής υπολογίζονται με τα ακόλουθα τρία βήματα.

1. Οι μετρήσεις υπολογίζονται για όλα τα έγγραφα με βάση τις εξισώσεις 3.40, 3.41 και 3.42 σε μία αναδίπλωση.

2. Η μέση ακρίβεια, η ανάκληση και η Φ-μετρική υπολογίζονται για όλα τα κείμενα με βάση το βήμα 1 ανεξάρτητα από την κλάση που κάθε κείμενο ανήκει σε μία αναδίπλωση.

3. Οι μετρήσεις του βήματος 2 υπολογίζονται κατά μέσο όρο για όλες τις αναδιπλώσεις.

Σύγκριση μεταξύ μάκρο και μίκρο μετρικών.

Οι μέσες μάκρο μετρικές υποδηλώνουν τον βαθμό που οι κλάσεις έχουν προβλεφθεί καλά, ενώ οι μίκρο μετρικές δείχνουν τον βαθμό που τα κείμενα κατηγοριοποιούνται σωστά στις κλάσεις. Οι μέσες τιμές των μακρο μετρικών αξιολογούν την κατηγοριοποίηση από την πλευρά των κατηγοριών, ενώ οι μέσες τιμές των μίκρο μέτρησης από την προοπτική των κειμένων.

Επιπλέον, είναι σημαντικό να αναφέρουμε ότι η μέση ακρίβεια, των μακρό μετρικών ακρίβειας, ανάκλησης και Φ-μετρικής είναι τρεις διαφορετικοί αριθμοί, ενώ αποδεικνύεται ότι οι αντίστοιχες μίκρο τιμές ανάγονται στο ίδιο αποτέλεσμα. Έτσι στους πίνακες των πειραματικών αποτελεσμάτων κάθε αναφορά σε μάκρο μετρικές μετρήσεις συνεπάγεται τρία αποτελέσματα ενώ οι αναφορές σε μίκρο μετρικές αντιστοιχούν σε ένα αποτέλεσμα, συχνά της Φ-μετρικής.

3.15.Συμπεράσματα

Στην παρούσα εργασία είδαμε μια σειρά εξωτερικών και εσωτερικών μετρικών σχέσεων που μπορούν να αξιολογήσουν σε τι βαθμό είναι καλή μια συσταδοποίηση. Είδαμε τις ιδιότητες που πρέπει να ικανοποιεί μία μετρική σχέση και διαπιστώσαμε ότι μόνο οι BCubed σχέσεις τις ικανοποιούν. Εν συνεχεία επεκτείναμε τις BCubed μετρικές για να αξιολογούν επικαλυπτόμενες συσταδοποιήσεις.

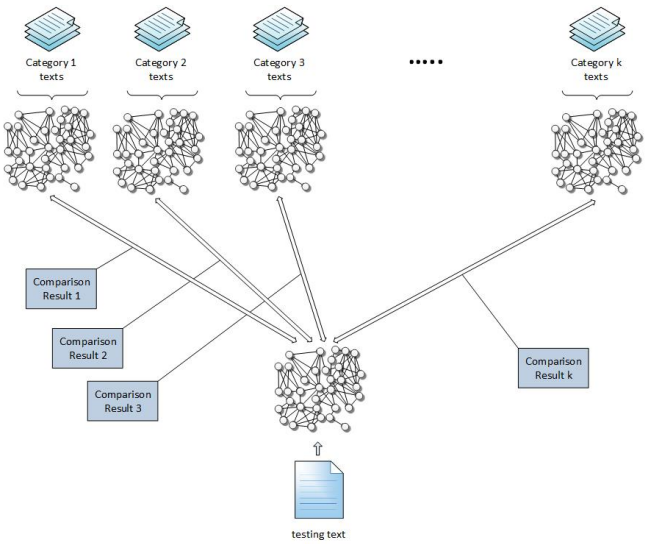
Καταλήγουμε ότι γνωρίζοντας ένα σωστά κατηγοριοποιημένο data set η καλύτερη μέθοδος για να αξιολογήσουμε μια συσταδοποίηση είναι να χρησιμοποιήσουμε της BCubed μετρικές. Αν δεν υπάρχει ένα τέτοιο data set μπορούμε να χρησιμοποιήσουμε κάποιο εσωτερικό κριτήριο επιλέγοντας με μεγάλη προσοχή ποια συνάρτηση απόστασης θα εφαρμόσουμε.

Το πιο σημαντικό όμως είναι ότι μια βαθιά γνώση των μετρικών αξιολόγησης συσταδοποιήσεων μπορεί να προσφέρει στον ερευνητή μηχανικό τον τρόπο σκέψης για το πώς θα σχεδιάσει έναν καινοτόμο αλγόριθμο συσταδοποίησης που να είναι πιο αποδοτικός από αυτούς που ήδη υπάρχουν.

4. Το μοντέλο κατηγοριοποίησης κειμένων με Γράφους Ν-γραμμάτων

Το μοντέλο κατηγοριοποίησης κειμένων που προτείνουμε έγκειται στις μεθόδους κατηγοριοποίησης πολλαπλών θεματικών κατηγοριών κειμένου. Το μοντέλο αποτελείται από πέντε στάδια: προεπεξεργασία κειμένων, κατασκευή γράφων, μέτρηση ομοιότητας γράφων, διανυσματική αναπαράσταση και κατηγοριοποίηση. Αυτά τα στάδια έχουν σχεδιαστεί ως μια ακολουθία ξεχωριστών εργασιών και επιμέρους λειτουργίες όπου οι πιο απαιτητικές εργασίες να μπορούν να κλιμακωθούν ή να αποκλιμακωθούν σε σχέση με το φόρτο εργασίας των εισερχόμενων ροών κειμένου. Όλα αυτά τα στάδια περιγράφονται στις επόμενες παραγράφους και αναδεικνύονται τα εγγενή πλεονεκτήματα του μοντέλου αναπαράστασης γράφων Ν-γραμμάτων για τις ανάγκες κατηγοριοποίησης κειμένων και την ικανότητά τους να υλοποιηθούν σε ένα κατανεμημένο σύστημα.

Οι κατηγορίες κειμένου και θέματος μπορούν να εκπροσωπούνται με έναν συνεκτικό τρόπο ως γράφοι Ν-γραμμάτων. Η σύγκριση αυτών των γράφων μπορεί να ποσοτικοποιηθεί ώστε να παρέχει ένα μέτρο ομοιότητας μεταξύ ενός κειμένου με τις αντίστοιχες κατηγορίες θεμάτων όπως απεικονίζεται στο Σχήμα 4.1. Το προτεινόμενο μοντέλο ταξινόμησης ακολουθεί μια κλιμακούμενη κατανεμημένη υπολογιστική προσέγγιση για τους μετασχηματισμούς κειμένων σε γράφους και τις συγκρίσεις γράφων ενισχύοντας έτσι την ακρίβεια των προβλέψεων και μειώνοντας την χρονική απόκριση του συστήματος.



Σχήμα 4.1 Γραφική αναπαράσταση και σύγκριση μεταξύ κειμένων και θεματικών κατηγοριών

Το μοντέλο κατηγοριοποίησης κειμένων που προτείνουμε είναι ανεξάρτητο από την γλώσσα εφαρμογής. Ο ΓΝΓ μπορεί να κατασκευαστεί από κείμενα γραμμένα σε οποιαδήποτε γλώσσα. Επιπλέον, μπορεί να χρησιμοποιηθεί ακόμη και εάν ένα κείμενο συνδυάζει λέξεις από δύο ή περισσότερες γλώσσες. Το φαινόμενο να χρησιμοποιούνται αγγλικές λέξεις σε έγγραφα που είναι γραμμένα σε διαφορετικές γλώσσες, είναι αρκετά συνηθισμένο.

Το προτεινόμενο μοντέλο επιτυγχάνει μια αρκετά καλή υπολογιστική πολυπλοκότητα για τις εργασίες κατασκευής και σύγκρισης γράφων. Η πολυπλοκότητα για την κατασκευή ενός ΓΝΓ είναι O (l) όπου l είναι ο συνολικός αριθμός των χαρακτήρων των κειμένων και η πολυπλοκότητα της μεθόδου ομοιότητας γραφήματος είναι $O(e_1 \cdot e_2)$ όπου e_1 και e_2 είναι ο αριθμός των ακμών σε δύο διαγράμματα σύγκρισης

Το αποτέλεσμα σύγκρισης μεταξύ ενός γράφου κειμένου με τους γράφους που αντιπροσωπεύουν τις θεματικές κατηγορίες παράγει βαθμούς ομοιότητας που μπορούν να σχηματίσουν ένα διάνυσμα k διαστάσεων, όπου το k υποδηλώνει τον αριθμό των κατηγοριών. Αυτά τα διανύσματα k-διαστάσεων περνάνε σε έναν ταξινομητή διανυσμάτων όπου υπολογίζεται η καταλληλότερη κατηγορία στην οποία θα ταξινομηθούν τα αρχικά κείμενα. Ο αγωγός (pipeline) ροής κειμένων για τη στήριξη του μοντέλου ταξινόμησης ΓΝΓ χωρίζεται σε παράλληλες εργασίες (embarrassingly parallel tasks), δηλαδή εργασίες εξαγωγής, μετασχηματισμού και φόρτωσης. Η διαδικασία ταξινόμησης καταμερίζεται σε μικρότερες εργασίες που εκτελούνται ως λειτουργίες μετασχηματισμού και κατευθύνουν επεξεργασμένα δεδομένα με τη μορφή PCollections (Π-Συλλογών), από στάδιο σε στάδιο. Στο πιο απαιτητικό υπολογιστικό στάδιο, που είναι η εργασία σύγκρισης γράφων, τα δεδομένα και η επεξεργασία ανατίθενται σε υπολογιστικές μονάδες κατά παράλληλο τρόπο. Όλες οι εργασίες περιλαμβάνουν τη δυνατότητα αυτόματης κλιμάκωσης όπως παρέχεται από το Dataflow στις υποδομές του Google Cloud, το οποίο εγγυάται την έγκαιρη απόκριση, τη διαθεσιμότητα και την ανοχή σε σφάλματα στο απεριόριστο μήκος και την τυχαία αύξηση ή μείωση της ροής κειμένων.

Στόχος μας είναι να προτείνουμε μια μέθοδο που έχει γενική μορφή και μπορεί να χρησιμοποιηθεί σε διαφορετικές ανάγκες ταξινόμησης κειμένων. Η μέθοδος αξιολογείται τόσο σε σχέση με ένα ίσο κατανεμημένο σύνολο δεδομένων όσο και σε ένα σύνολο δεδομένων που είναι ανόμοια κατανεμημένο. Στο πρώτο σύνολο δεδομένων, τα κείμενα διανέμονται εξίσου σε όλες τις κατηγορίες. Στο δεύτερο ένα σύνολο δεδομένων εμπεριέχει ορισμένες κατηγορίες όπου έχουν πολύ περισσότερα κείμενα από άλλες. Στο ανόμοια κατανεμημένο σύνολο δεδομένων, πολλοί αλγόριθμοι ταξινόμησης τείνουν να αναθέτουν κείμενα που ανήκουν σε μικρές κατηγορίες σε μεγαλύτερες. Ακόμη και αν ορισμένες μέθοδοι προσπάθησαν να αντιμετωπίσουν αυτό το ζήτημα [76] παραμένει μια από τις κύριες προκλήσεις στην ταξινόμηση κειμένου. Οι Μίκρο μετρήσεις αξιολόγησης δεν μπορούν να εκφράσουν αυτήν την εσφαλμένη ταξινόμηση σε αντίθεση με τις Μάκρο μετρήσεις αξιολόγησης, οι οποίες είναι σε θέση να το κάνουν. Η ανομία κατανομή των κειμένων σε θεματικές κατηγορίες είναι ένα ζήτημα που λαμβάνεται υπόψη σε αυτή την έρευνα και προσπαθούμε να το αντιμετωπίσουμε.

Στη συνέχεια παρέχουμε τις λεπτομέρειες σχετικά με τις κύριες λειτουργίες στον αγωγό ταξινόμησης ΓΝΓ, δηλαδή την κατασκευή γράφων, τον υπολογισμό ομοιότητας γράφων, την προεπεξεργασία, την αντιπροσώπευση και ταξινόμηση των φορέων. Αναφερόμαστε επίσης στον τρόπο με τον οποίο οι λειτουργίες αυτές τοποθετούνται σε ένα κατανεμημένο περιβάλλον.

4.1. Κατασκευή Γράφων

Η διαδικασία κατασκευής γράφων μπορεί να χωριστεί σε δύο βασικά βήματα. Πρώτον, όλα τα Ν-γράμματα εξάγονται μετατοπίζοντας ένα παράθυρο Ν χαρακτήρων στο αρχικό κείμενο. Έπειτα αυτά τα Ν-γράμματα χρησιμοποιούνται ως κόμβοι στον ΓΝΓ. Οι κόμβοι συνδέονται μέσω μιας ακμής μόνο αν τα αντίστοιχα Ν-γράμματα πληρούν κάποιο κριτήριο εγγύτητας (π.χ. ακολουθίας ή περιεχομένου). Το κριτήριο εγγύτητας καθορίζεται από το μέγεθος του πλαισίου ολίσθησης Ν-γραμμάτων, οι λεπτομέρειες του οποίου περιγράφονται σε μια επόμενη υποενότητα και εξηγείται η διαδικασία κατασκευής των ακμών. Οι ακμές των γράφων που παράγονται είναι μονής κατεύθυνσης (έτσι αναφερόμαστε σε ένα κατευθυνόμενο γράφημα) προκειμένου να καταγράψουμε την ακολουθία των Ν-γραμμάτων όπως υπάρχουν στο αρχικό κείμενο. Θα εξετάσουμε γράφους με βάρη και χωρίς βάρη. Τα βάρη υποδηλώνουν πόσο συχνά πληρούται το προαναφερθέν κριτήριο γειτνίασης για ένα ζεύγος Ν-γραμμάτων. Ο γράφος Ν-γραμμάτων μιας κατηγορίας δημιουργείται από τη συνάθροιση των γράφων που ανήκουν στην ίδια κατηγορία από το στάδιο εκπαίδευσης του μοντέλου.

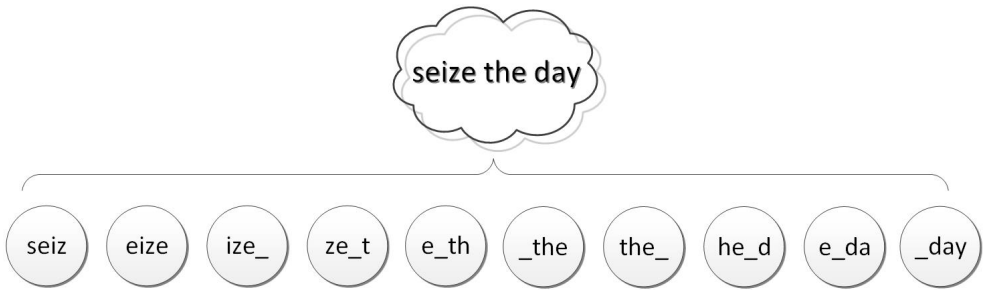
Ένα κείμενο μπορεί να μετατραπεί σε ΓΝΓ, αλλά το αντίστροφο δεν είναι δυνατό. Ένας γράφος μπορεί να υποδεικνύει εάν δύο Ν-γράμματα είναι κοντά μεταξύ τους καθώς και την ακολουθία τους. Αλλά δεν μπορούμε να ακολουθήσουμε μια διαδρομή ακμών ούτως ώστε να ανασυγκροτήσουμε το αρχικό κείμενο.

Αυτό οφείλεται στο ότι ένας κόμβος Ν-γραμμάτων μπορεί να έχει πολλές εξερχόμενες ακμές. Επιπλέον δύο ή περισσότερα τμήματα κειμένου μπορούν να έχουν τον ίδιο ΓΝΓ καθιστώντας ακόμα πιο αβέβαιη την ανασύσταση του αρχικού κειμένου.

Ένας ΓΝΓ περιλαμβάνει επαναλαμβανόμενη πληροφορία λόγω του ότι οι Ν-1 χαρακτήρες ενός κόμβου Ν-γραμμάτων μπορούν να επαναληφθούν με βάση το κριτήριο της γειτνίασης στους συνδεδεμένους του κόμβους. Αυτός ο φαινομενικά πλεονασμός εξακολουθεί να είναι πολύ χρήσιμος, διότι με βάση αυτό μπορούν να ανιχνευθούν τα παρόμοια κείμενα. Σε πολλές περιπτώσεις, μια σύνθετη λέξη περικλείει τις έννοιες των απλών λέξεων. Η σχέση μεταξύ μιας απλής λέξης και μιας σύνθετης λέξης μπορεί να αναγνωριστεί με τεχνικές μερικής ομοιότητας ως ομοιότητες γράφων. Πολλές τεχνικές ομοιότητας γράφων δοκιμάζονται στο προτεινόμενο μοντέλο, όπως θα δούμε. Κάθε μία από αυτές ορίζεται με διαφορετικό τρόπο, αλλά όλες βασίζονται στον πλήθος και την διάταξη των χαρακτήρων που υπάρχουν στους γράφους που συγκρίνονται.

4.1.1. Κατασκευή κόμβων

Τα πιθανά Ν-γράμματα εξάγονται από ένα κείμενο και αντιπροσωπεύονται ως κόμβοι. Οι χαρακτήρες ενός Ν-γράμματος μπορεί να προέρχονται από μία λέξη ή από διαδοχικές λέξεις που συμπεριλαμβάνουν ακόμη και τον χαρακτήρα διαστήματος που τις χωρίζει, όπως φαίνεται στο παράδειγμα του Σχήματος 4.2.



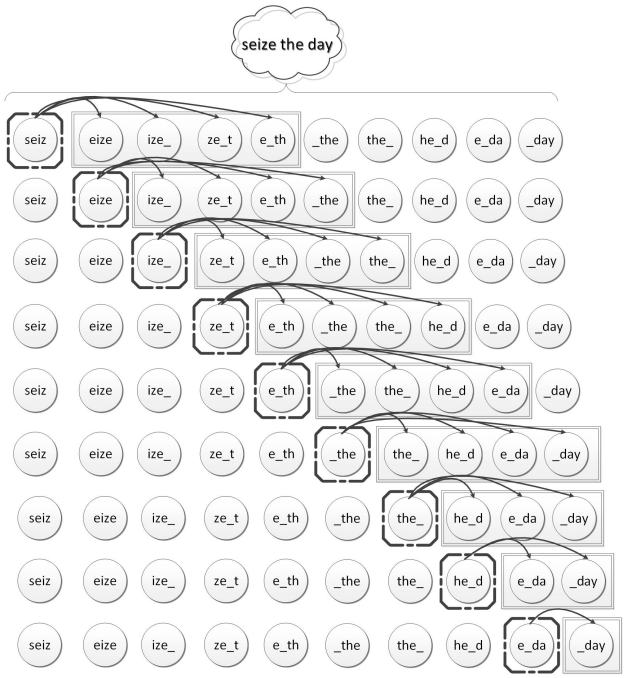
Σχήμα 4.2 Κατασκευή κόμβου 4 γραμμάτων της φράσης"seize the day"

Ανάλογα με την εφαρμογή, τα Ν-γράμματα μπορεί να διακρίνουν πεζά με κεφαλαία. Για παράδειγμα, στις εφαρμογές ανάλυσης συναισθημάτων προτιμάται τα Ν-γράμματα να διακρίνουν τα πεζά από τα κεφαλαία, διότι η χρήση κεφαλαίων και μικρών γραμμάτων να εκφράζουν τα συναισθήματα του συγγραφέα. Επιπλέον, τα σημεία στίξης είναι σημαντικά για τον ίδιο λόγο [77]. Η προτεινόμενη γενική μέθοδος θεματικής κατηγοριοποίησης κειμένων με ΓΝΓ δεν κάνει διάκριση κεφαλαίων και πεζών, διότι στις περισσότερες εφαρμογές οι λέξεις δεν διαφέρουν νοηματικά εάν είναι γραμμένες με κεφαλαία ή πεζά γράμματα. Στη χειρότερη περίπτωση, ο αριθμός των κόμβων είναι (#Chars) Ν όπου #Chars είναι ο αριθμός διακριτών χαρακτήρων και σημείων στίξης που μπορούν να χρησιμοποιηθούν. Προκειμένου να διατηρηθεί ο γράφος σε ένα μικρό, διαχειρίσιμο μέγεθος, το προτεινόμενο μοντέλο χρησιμοποιεί τα Ν-γράμματα μόνο ως συνδυασμό πεζών γραμμάτων και κενών χαρακτήρων. Ακόμη και αν ένα Ν-γραμμα υπάρχει πολλές φορές στο αρχικό κείμενο, μόνο ένας κόμβος θα δημιουργηθεί για να το αντιπροσωπεύσει. Δεδομένου ότι δεν υπάρχουν ενδείξεις σχετικά με την κατάλληλη τιμή του Ν, ερευνήσαμε επιλογές μεταξύ 2 και 10.

Στην περίπτωση ενός μεγάλου κειμένου, ένα μεγάλο Ν μπορεί να απεικονίσει με μεγαλύτερη ακρίβεια την ακολουθία των λέξεων, αλλά ο γράφος γίνεται μεγαλύτερος και η πολυπλοκότητα ψηλότερη. Από την άλλη πλευρά, εάν το Ν είναι μικρό, τότε το γράφημα είναι μικρότερο αλλά η ακρίβεια της μεθόδου επηρεάζεται αρνητικά. Σε μικρά κείμενα όπως tweets, ένα μικρό Ν μπορεί να είναι πιο αποτελεσματικό από τη χρήση ενός μεγάλου Ν. Σε αυτά τα συμπεράσματα καταλήξαμε μέσω πειραμάτων και αναφέρονται στην ενότητα αξιολόγησης.

4.1.2.Κατασκευή ακμών

Η γειτνίαση μαζί με τη σειρά εμφάνισης δύο Ν-γραμμάτων στο αρχικό κείμενο αναπαρίσταται στον γράφο από μια ακμή η οποία ενώνει τους δύο αντίστοιχους κόμβους. Έτσι, κάθε κόμβος Ν-γραμμάτων συνδέεται στον ακόλουθο κόμβο Ν-γράμματος μέσω μιας κατευθυνόμενης ακμής. Ο αριθμός των επόμενων κόμβων που θα ενωθούν καθορίζεται από ένα πλαίσιο. Για να ορίσουμε την έννοια του πλαισίου πρέπει να απαντήσουμε στην ερώτηση: Υποθέτοντας ένα Ν-γραμμα που θεωρείται πηγή, πόσες επόμενα Ν-γράμματα θα πρέπει να θεωρήσουμε επαρκώς κοντά έτσι ώστε να μπορούμε να δηλώσουμε αυτήν την εγγύτητα με μιαν ακμή στο γράφημα; Η απάντηση σε αυτή την ερώτηση ορίζει την σημασία του πλαισίου.

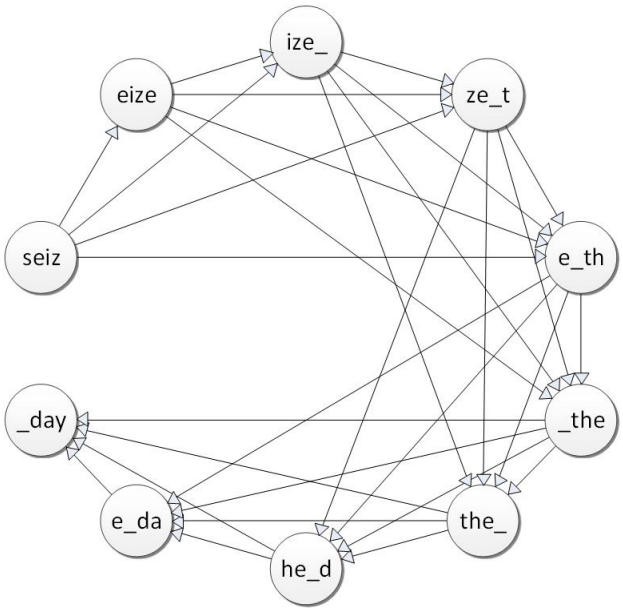


Σχήμα 4.3 Κατασκευή ακμών με κυλιόμενο πλαίσιο μήκους 4

Σε αυτή την έρευνα, πειραματιστήκαμε με μεγέθη πλαισίων μεταξύ δύο έως δέκα όπως θα δούμε στην ενότητα αξιολόγησης. Μεγαλύτερα μεγέθη πλαισίων από 10 δεν έχουν δοκιμαστεί λόγω της υψηλής υπολογιστικής επιβάρυνσης που προσθέτουν, ενώ μικρότερα μεγέθη (δηλ. Μέγεθος = 1) δεν έχουν νόημα. Το σημαντικότερο συμπέρασμα είναι ότι όταν χρησιμοποιούνται επιλογές προς το κάτω άκρο, τότε το γράφημα θα αντιπροσωπεύει αυστηρά τη σχέση μεταξύ των λέξεων. Από την άλλη πλευρά, εάν χρησιμοποιείται ένα μεγάλο πλαίσιο, τότε το γράφημα θα αντιπροσωπεύει με πιο ευέλικτο τρόπο τη σχέση μεταξύ των λέξεων. Τελικά, επιλέξαμε ένα μέγεθος πλαισίου ίσο με Ν, δηλ. τον ίδιο αριθμό Ν με τον αριθμό των χαρακτήρων στα Ν-γράμματα. Η χρήση εναλλακτικών μεγεθών πλαισίων που διαφέρουν από την τάξη Ν των Ν-γραμμάτων παραμένει μια μελλοντική εργασία.

Το πλαίσιο που περιέχει τους ακόλουθους κόμβους Ν-γραμμάτων ολισθαίνει ένα Ν-γραμμα κάθε φορά και για κάθε κόμβο Ν-γραμμα, παράγονται Ν ακμές που συνδέουν τους Ν αντίστοιχους κόμβους Ν-γραμμάτων όπως απεικονίζεται στο Σχήμα 3 και στο Σχήμα 4. Κάθε γραμμή αντιπροσωπεύει έναν κόμβο πηγής και το πλαίσιο που περιλαμβάνει τους τερματικούς κόμβους. Οι άκρες μπορούν να έχουν βάρη ή να μην έχουν. Οι ακμές χωρίς βάρη αντιπροσωπεύουν απλώς ότι μερικά Ν-γράμματα βρίσκονται κοντά σε ένα κείμενο. Οι σταθμισμένες ακμές μπορούν να αντιπροσωπεύουν τη συχνότητα της εμφάνισης γειτονικών Ν-γραμμάτων. Παρά την διαισθητική πεποίθηση ότι ο σταθμισμένος ΓΝΓ μπορεί να έχει καλύτερη ακρίβεια, τα πειράματα όπως περιγράφονται στην ενότητα αξιολόγησης απέδειξαν ότι στις περισσότερες περιπτώσεις συμβαίνει το αντίθετο. Αυτό εξηγείται από το γεγονός ότι πολλά κοινά Ν-γράμματα συνυπάρχουν συχνά σε διάφορα κείμενα ανεξαρτήτως κατηγορίας. Η κατηγοριοποίηση με βάση την συχνότητα των γειτονικών Ν-γραμμάτων ελαττώνει την ακρίβεια του μοντέλου.

Για να βελτιωθεί η χρήση σταθμισμένων γράφων, τα βάρη των ακμών μπορούν να κανονικοποιηθούν ώστε να είναι ανεξάρτητα από το μέγεθος κειμένου. Εάν μια κατηγορία αποτελείται από πολλά κείμενα, θα έχει ένα μεγάλο γράφο και τα σημαντικά ζευγάρια γειτονικών N-γραμμάτων θα υπάρχουν πολλές φορές. Από την άλλη πλευρά, αν μια κατηγορία περιλαμβάνει μόνο μερικά κείμενα, τα σημαντικά ζευγάρια γειτονικών N-γραμμάτων μπορεί να υπάρχουν μόνο δύο ή τρεις φορές. Αυτό έχει ως αποτέλεσμα, τα βάρη των ακμών να υπολογίζονται λαμβάνοντας υπόψη τόσο τον αριθμό της συχνότητας που κάθε άκρη υπάρχει όσο και τον συνολικό αριθμό ακμών. Η κανονικοποίηση των σταθμισμένων γράφων μπορεί να συμβάλει στην αντιμετώπιση της ανόμοιας κατανομής κειμένων σε θεματικές κατηγορίες.



Σχήμα 4.4. Γράφος 4-γραμμάτων της πρότασης “seize the day"

4.2.Μετρικές Ομοιότητας Γράφων

Ένα βασικό σημείο του προτεινόμενου μοντέλου ταξινόμησης με ΓΝΓ είναι η ποσοτική μέτρηση της ομοιότητας γράφων. Η ομοιότητα γράφων μπορεί να υποδεικνύει τον βαθμό που ένα κείμενο ανήκει σε μια κατηγορία. Πολλές μέθοδοι για την εκτίμηση της ομοιότητας γράφων υπάρχουν στη βιβλιογραφία. Ένα βασικό κριτήριο είναι η αναγνώριση του μέγιστου κοινού υπογράφου, όπως ο Rascal [78] που βασίζεται στη μέγιστη κλίκα μεταξύ δύο γράφων. Ένα άλλο κριτήριο βασίζεται στην ομοιότητα μεταξύ των κόμβων των γράφων [79].

Ο συνολικός αριθμός των κοινών ακμών μεταξύ δύο γράφων N-gram είναι ένα καλύτερο κριτήριο για την ομοιότητα από το πλήθος των κοινών κόμβων. Ο λόγος είναι ότι υπάρχουν διακριτά N-γράμματα που εμφανίζονται συχνά μεταξύ διαφορετικών λέξεων χωρίς να υπάρχει πραγματική συσχέτιση μεταξύ των λέξεων από όπου προέρχονται. Ενώ η γειτνίαση και η διαδοχή των N-γραμμάτων υποδηλώνουν σε μεγαλύτερο βαθμό τις σχέσεις μεταξύ των λέξεων. Το κριτήριο που πρέπει να χρησιμοποιηθεί είναι ένα κριτήριο που βασίζεται στον αριθμό των κοινών ακμών.

Η ομοιότητα περιεχομένου (Containment Similarity CS) [80] είναι μια μετρική που μπορεί να χρησιμοποιηθεί σε σταθμισμένους γράφους και μετρά τον αριθμό των ακμών που συνυπάρχουν στους δύο

γράφους σε σύγκριση με τον συνολικό αριθμό ακμών του μικρότερου εκ των δύο γράφων. Μια ακμή συνυπάρχει και στους δύο γράφους εάν οι προσκείμενοι κόμβοι του έχουν τις ίδιες ετικέτες N-γραμμάτων και στους δύο γράφους. Η εξίσωση της ομοιότητας περιεχομένου δίνεται παρακάτω.

$$\boldsymbol{CS}\left(\boldsymbol{G_T}, \boldsymbol{G_C}\right) = \frac{\sum_{e \in G_T} \boldsymbol{\mu}(\boldsymbol{e}, \boldsymbol{G_C})}{\boldsymbol{min}(|\boldsymbol{G_T}|, |\boldsymbol{G_C}|)} \quad (4.1)$$

Όπου G_T είναι ο γράφος ενός κειμένου και το G_C είναι ο γράφος μιας κατηγορίας $\left|G_T\right|$, $\left|G_C\right|$ είναι ο αριθμός των ακμών του γράφου G_T και G_C αντίστοιχα και το e είναι μια ακμή που ανήκει στο γράφημα G_T. Η συνάρτηση μ (e, G_C) είναι ίση με 1 αν ο γράφος G_C περιέχει την ακμή e και 0 διαφορετικά.

Σε όλες σχεδόν τις περιπτώσεις, ο γράφος του κειμένου είναι μικρότερος από τον γράφο της θεματικής κατηγορίας. Για αυτόν τον λόγο, είναι πιο αποδοτικό να διατρέχουμε τις ακμές του γράφου κειμένου και να ελέγχουμε αν υπάρχουν στον γράφο κατηγορίας. Σημειώνουμε ότι από τη στιγμή που οι γράφοι είναι κατευθυνόμενοι, οι ακμές π.χ. "Appl" - "pple" και "pple" - "appl" θεωρούνται διαφορετικές.

Στην περίπτωση σταθμισμένων γράφων, μπορεί να χρησιμοποιηθεί το μέτρο της Ομοιότητας Αξίας (Value Similarity VS) και της κανονικοποιημένης Ομοιότητας Αξίας (Normalized Value Similarity NVS) [80]. Η ομοιότητα αξίας εκφράζει το άθροισμα των κοινών ακμών μεταξύ δύο γράφων λαμβάνοντας υπόψη το βάρος των ακμών (εξίσωση 4.2). Η ομοιότητα αξίας είναι υψηλή εάν τα κοινά N-γράμματα είναι γειτονικά με την ίδια συχνότητα και στους δύο γράφους και χαμηλή εάν τα κοινά N-γράμματα έχουν διαφορετικές συχνότητες.

$$\boldsymbol{VS}\left(\boldsymbol{G_T}, \boldsymbol{G_C}\right) = \frac{\sum_{e \in G_T} \frac{\boldsymbol{min}(\boldsymbol{w_T}(\boldsymbol{e}), \boldsymbol{w_c}(\boldsymbol{e}))}{\boldsymbol{max}(\boldsymbol{w_T}(\boldsymbol{e}), \boldsymbol{w_c}(\boldsymbol{e}))}}{\boldsymbol{max}(|\boldsymbol{G_T}|, |\boldsymbol{G_C}|)} \quad (4.2)$$

Όπου w_T είναι το βάρος της άκρης e στο γράφημα κειμένου και το w_C είναι το βάρος της ακμής e στο γράφημα θεματικών κατηγοριών. Η ομοιότητα αξίας μεταξύ ενός μεγάλου γράφου και ενός μικρού γράφου μπορεί να μην παράγει εύλογα αποτελέσματα. Επομένως, ένας παράγοντας κανονικοποίησης θα πρέπει να χρησιμοποιηθεί για να διαιρέσει την ομοιότητα αξίας (εξίσωση 4.3).

$$\boldsymbol{SS}\left(\boldsymbol{G_T}, \boldsymbol{G_C}\right) = \frac{\boldsymbol{min}\left(\boldsymbol{G_T}, \boldsymbol{G_C}\right)}{\boldsymbol{max}\left(\boldsymbol{G_T}, \boldsymbol{G_C}\right)} \quad (4.3)$$

Ο κανονικοποιημένος παράγοντας είναι η διαίρεση μεταξύ των ελάχιστων ακμών προς τις περισσότερες ακμές των δύο γράφων.

Η ομοιομορφία κανονικοποιημένης τιμής (NVS) εξίσωση 4.4 υποδηλώνει την ομοιότητα μεταξύ δύο γράφων που ανεξαρτητοποιείται από τα μεγέθη των γράφων.

$$\boldsymbol{NVS}\left(\boldsymbol{G_T}, \boldsymbol{G_C}\right) = \frac{\boldsymbol{VS}\left(\boldsymbol{G_T}, \boldsymbol{G_C}\right)}{\boldsymbol{SS}\left(\boldsymbol{G_T}, \boldsymbol{G_C}\right)} \quad (4.4)$$

Τα CS, VS και NVS δίνουν ως αποτέλεσμα έναν αριθμό μεταξύ μηδέν και ενός. Η μηδενική ομοιότητα δηλώνει ότι οι γράφοι δεν έχουν τίποτα κοινό και η ομοιότητα μεταξύ G_D, G_C ισούται με ένα και δηλώνει ότι είναι ταυτόσημα γράφοι ή υπογράφοι.

Ομοίως, η ομοιότητα περιεχομένου και η ομοιότητα αξίας έχουν χρησιμοποιηθεί για τη σύγκριση ΓΝΓ για την συναισθηματική ανάλυση κειμένων [80] και για την αξιολόγηση αυτοματοποιημένων περιλήψεων [44]. Αυτές οι μετρικές είναι κατάλληλες για την καταγραφή της ομοιότητας κειμένου που αντιπροσωπεύονται από τους γράφους.

Έχουν εφαρμοστεί τρεις παραλλαγές μετρικών που βασίζονται στον υπολογισμό μέγιστων κοινών υπογραφών (Maximum Common Subgraph MCS). Ο υπολογισμός του μέγιστου κοινού υπογράφου χωρίς ετικέτες (labels) είναι ένα NP-πλήρες πρόβλημα. Ευτυχώς, οι γράφοι N-γραμμάτων χρησιμοποιούν κόμβους με ετικέτες και η πολυπλοκότητα τους είναι πολυωνυμική. Μετά τη μέτρηση του μέγιστου κοινού υπογράφου, η ποσοτικοποίηση της ομοιότητας μπορεί να εκτιμηθεί με βάση τον αριθμό των κοινών κόμβων ή ακμών. Η εξίσωση 4.5 χρησιμοποιεί τον αριθμό των κόμβων που περιέχει ο μέγιστος ο κοινός υπογράφος ενώ οι εξισώσεις 4.6 και 4.7 βασίζονται στον αριθμό των ακμών.

$$\boldsymbol{MCSNS} = \frac{\boldsymbol{MCSN}(\left|\boldsymbol{G_T}\right|, \left|\boldsymbol{G_C}\right|)}{\mathbf{min}(\left|\boldsymbol{G_T}\right|, \left|\boldsymbol{G_C}\right|)} \quad (4.5)$$

όπου $MCSN(\left|G_t\right|, \left|G_C\right|)$ δηλώνει τον αριθμό των κόμβων που περιέχονται στον μέγιστο κοινό υπογράφο.

$$\boldsymbol{MCSUES} = \frac{\boldsymbol{MCSUE}(\left|\boldsymbol{G_T}\right|, \left|\boldsymbol{G_C}\right|)}{\mathbf{min}(\left|\boldsymbol{G_T}\right|, \left|\boldsymbol{G_C}\right|)} \quad (4.6)$$

όπου $MCSUE\left(\left|G_T\right|, \left|G_C\right|\right)$ δηλώνει τον αριθμό των ακμών που περιέχονται στον μέγιστο κοινό υπογράφο και είναι ανεξάρτητη από την κατεύθυνσή των ακμών στους αρχικούς γράφους.

$$\boldsymbol{MCSDES} = \frac{\boldsymbol{MCSDE}(\left|\boldsymbol{G_T}\right|, \left|\boldsymbol{G_C}\right|)}{\mathbf{min}(\left|\boldsymbol{G_T}\right|, \left|\boldsymbol{G_C}\right|)} \quad (4.7)$$

όπου $MCSDES\left(\left|G_T\right|, \left|G_C\right|\right)$ δηλώνει τον αριθμό των κόμβων που περιέχονται στον μέγιστο κοινό υπογράφο και έχουν την ίδια κατεύθυνση στους γράφους του κειμένου που εξετάζουμε G_T και της θεματικής κατηγορίας G_C

Η μετρική MCSNS βασίζεται στον αριθμό των κόμβων που περιέχει ο μέγιστος κοινός υπογράφος. Αυτό σημαίνει ότι εξετάζεται το πλήθος των διαφορετικών ακολουθιών χαρακτήρων που υπάρχουν τόσο στο αρχικό κείμενο όσο και στην αντίστοιχη κατηγορία θέματος. Οι μετρήσεις MCSUES και MCSDES

βασίζονται στον αριθμό των ακμών που περιέχουν ο μέγιστος κοινός υπογράφος. Το MCSUES δεν λαμβάνει υπόψη την κατεύθυνση των ακμών, ενώ το MCSDES απαιτεί οι ακμές να γειτονεύουν στους κόμβους προέλευσης και τερματισμού με την ίδια ετικέτα. Αυτές οι μετρικές αποτυπώνουν την έννοια του πόσο ισχυρή είναι η σχέση των αλληλουχιών N-γραμμάτων μεταξύ κειμένου και κατηγορίας θέματος.

4.3.Προεπεξεργασία Κειμένων

Τόσο οι ΓΝΓ των θεματικών κατηγοριών όσο και των κειμένων περιέχουν πολλούς κόμβους και ακμές N-γραμμάτων. Μια περαιτέρω ερώτηση είναι αν μπορούμε να διακρίνουμε τους κόμβους και τις ακμές που προσφέρουν χρήση πληροφορία στην διαδικασία της κατηγοριοποίησης από τους "θορυβώδεις" κόμβους και ακμές. Οι κόμβοι και οι ακμές θεωρούνται ως θόρυβος αν υπάρχουν σε όλες τις κατηγορίες ή ακόμα χειρότερα αν συστηματικά έχουν αρνητική συμβολή κατά την ταξινόμηση των κειμένων. Υπάρχουν δύο πλεονεκτήματα από το φιλτράρισμα τμημάτων των γράφων. Πρώτον, οι γράφοι γίνονται μικρότεροι μειώνοντας τις απαιτήσεις υπολογισμού και μνήμης. Δεύτερον, η ακρίβεια της μεθόδου ταξινόμησης μπορεί να αυξηθεί.

Η επιλογή χαρακτηριστικών (Feature selection) είναι μια προσέγγιση μηχανής μάθησης που εφαρμόζεται για την ποσοτικοποίηση της σημασίας κάθε χαρακτηριστικού σε ένα συγκεκριμένο πρόβλημα. Για τις ανάγκες της επεξεργασίας φυσικής γλώσσας, είναι συνηθισμένη πρακτική να αφαιρούνται οι κοινές λέξεις (Stop Words) και να γίνεται λημματοποίηση (Stemming) του αρχικού κειμένου [81]. Ακόμα κι αν παραλείπονται ή αλλάζουν ορισμένες λέξεις σε ένα στάδιο προεπεξεργασίας, διατηρείται η ακολουθία των σημαντικών λέξεων. Οι κοινές λέξεις είναι οι πολύ συνηθισμένες λέξεις που μπορούν να βρεθούν σε ολόκληρο το σύνολο κειμένων που εξετάζονται και δεν σχετίζονται με συγκεκριμένα θέματα. Η λημματοποίηση αφορά την διαγραφή της κατάληξης μιας λέξης ή την εύρεση του λήμματος από όπου έχει παραχθεί.

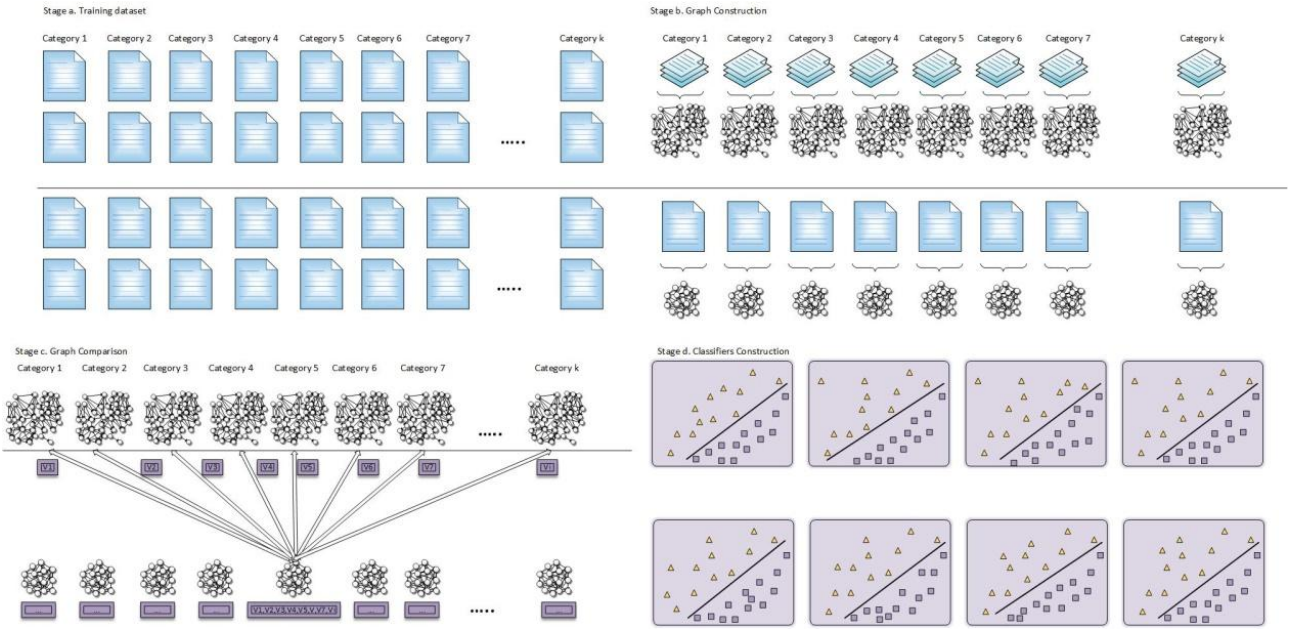
Στο στάδιο της εκπαίδευσης, τα κείμενα χωρίζονται σε δύο μέρη. Το πρώτο μέρος περιλαμβάνει τα κείμενα που χρησιμοποιούνται για την κατασκευή των k κατηγοριών ΓΝΓ, ενώ στο δεύτερο μέρος, κάθε κείμενο αντιπροσωπεύεται ως ξεχωριστό ΓΝΓ. Οι ξεχωριστοί γράφοι των μεμονωμένων κειμένων συγκρίνονται με τους γράφους των k θεματικών κατηγοριών για να παραχθούν τα διανύσματα που τους αντιπροσωπεύουν. Ο ταξινομητής διανυσμάτων εκπαιδεύεται με αυτά τα διανύσματα. Από τη συνολική διαδικασία του σταδίου εκπαίδευσης, διατηρούμε τις κατηγορίες γράφων N-γραμμάτων και το εκπαιδευμένο μοντέλο ταξινόμησης διανυσμάτων, ως δομή γνώσης τα οποία θα χρησιμοποιήσουμε στο στάδιο πρόβλεψης.

Στο στάδιο πρόβλεψης, κάθε κείμενο αντιπροσωπεύεται ως ΓΝΓ και συγκρίνεται με τους ΓΝΓ των κατηγοριών που κατασκευάστηκαν στο στάδιο εκπαίδευσης. Αυτή η σύγκριση παράγει διανύσματα που αντιπροσωπεύουν τα κείμενα και μεταφέρονται στο εκπαιδευμένο μοντέλο ταξινόμησης διανυσμάτων προκειμένου να ανατεθούν στην πλέον κατάλληλη κατηγορία. Η προεπεξεργασία πραγματοποιείται με παρόμοιο τρόπο στην αρχή της εκπαίδευσης και της πρόβλεψης, προτού τα κείμενα να χρησιμοποιηθούν για την κατασκευή των γράφων.

Για να επιλέξουμε τη μέθοδο ταξινόμησης διανύσματος, λάβαμε υπόψιν μας τη ροή και την ασύγχρονη φύση του μοντέλου ταξινόμησης ΓΝΓ και τα αξιολογήσουμε πειραματικά σε μια συγκριτική μελέτη. Τα χαρακτηριστικά των διανυσμάτων είναι ομοιογενή, καθώς όλα αντιστοιχούν σε αριθμητικές ομοιότητες γράφων που εκφράζονται στο εύρος [0,1]. Επιπλέον, κάθε μια από τις συνιστώσες του διανύσματος που έχει παραχθεί συμβάλλει στην κατηγοριοποίηση χωρίς να εξαρτάται από τις υπόλοιπες συνιστώσες. Ως αποτέλεσμα ένας ταξινομητής που βασίζεται στης αποστάσεις των διανυσμάτων ή σε πιθανότητες όπως οι (SVM) και Bayes να είναι μια καλή επιλογή.

Στο μοντέλο SVM, τα επίπεδα αποφάσεων (decision planes) παράγονται για να ορίσουν τα όρια αποφάσεων που χωρίζουν ένα σύνολο παρατηρήσεων σε ένα χώρο N-διαστάσεων. Τα επίπεδα απόφασης σχηματίζονται στο στάδιο της εκπαίδευσης και πρέπει να ικανοποιούν τον περιορισμό ότι το χάσμα μεταξύ των παρατηρήσεων που ανήκουν σε διαφορετικές κατηγορίες θα πρέπει να είναι όσο το δυνατόν ευρύτερο. Η

διαδικασία ταξινόμησης περιλαμβάνει πολλές κλάσεις έτσι χρησιμοποιήθηκε ο SVM για ταξινόμηση πολλαπλών κατηγοριών. Συγκεκριμένα, εφαρμόσαμε το γραμμικό SVM σε μια προσέγγιση ταξινόμησης μιας έναντι μίας (one vs. one). Στην μία έναντι μίας ταξινόμησης, ένας ξεχωριστός δυαδικός ταξινομητής εκπαιδεύεται για κάθε ζευγάρι κατηγοριών. Στο στάδιο προβλέψεων, το μοντέλο τροφοδοτείται ένα μη κατηγοριοποιημένο σύνολο κείμενων σε όλους τους δυαδικούς ταξινομητές και η κατηγορία που έχει τον υψηλότερο αριθμό προβλέψεων είναι η κατηγορία στην οποία εκχωρείται το κείμενο.



Σχήμα 4.5 Στάδια κατασκευής και ταξινόμησης γράφων και διανυσμάτων

Ο ταξινομητής Gaussian Bayes βασίζεται σε ένα πιθανοτικό μοντέλο στο οποίο τα κείμενα αναπαριστάνονται ως διανύσματα όπως περιγράφονται στις προηγούμενες παραγράφους. Οι k διαστάσεις των διανυσμάτων είναι ανεξάρτητες μεταβλητές και στην έρευνα μας σχετίζονται με τις εξόδους των μετρήσεων ομοιότητας των γράφων. Στην εξίσωση 4.8 η κατηγορία πρόβλεψης είναι η εξαρτημένη μεταβλητή y, t_i το μη παρατηρημένο κείμενο και το σ_y, μ_y υπολογίζονται χρησιμοποιώντας τη μέγιστη πιθανότητα.

$$P(t_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{\left(t_i-\mu_y\right)^2}{2\sigma_y^2}\right) \quad (4.8)$$

Η βαθιά μάθηση (Deep Learning) αφορά πολλαπλά επίπεδα ενός νευρωνικού δικτύου. Το διάνυσμα των ομοιοτήτων των γράφων μεταξύ ενός μη κατηγοριοποιημένου κειμένου και των κατηγοριών μπορεί να τροφοδοτηθεί στο επίπεδο εισόδου του νευρωνικού δικτύου και χρησιμοποιώντας μια τεχνική διάδοσης ενεργοποιήσεων κάθε επίπεδο μετατρέπει τα δεδομένα σε μια πιο αφηρημένη και σύνθετη μορφή μέχρι την ενεργοποίηση του τελευταίου επιπέδου που υποδηλώνει την κατηγορία στην οποία ανήκει το κείμενο. Η βαθιά εκμάθηση δείχνει πολύ καλή απόδοση σε προβλήματα που εκπαιδεύονται με μεγάλο αριθμό δεδομένων, αλλά εάν το εκπαιδευμένο σύνολο δεδομένων δεν επαρκεί, πιθανόν άλλες τεχνικές να έχουν καλύτερη απόδοση.

4.4.Κατανεμημένος Σχεδιασμός του ΓΝΓ για Ροές Κειμένων

Στην ταξινόμηση συνόλου (Batch) κειμένων, τα δεδομένων είναι στατικά και μπορεί να υποστούν επανειλημμένες επεξεργασίες χωρίς χρονικούς περιορισμούς. Η ταξινόμηση ροής (Streaming) κειμένου ΓΝΓ έχει σχεδιαστεί για να ικανοποιεί πιο συγκεκριμένους περιορισμούς, όπως η συνεχής παραγωγή κειμένων, η έγκαιρη απόκριση, η δυνατότητα επεκτάσεως του μοντέλου επεξεργασίας, η συχνή και πιθανή ταυτόχρονη άφιξη δεδομένων. Για να καλύψουμε όλες αυτές τις προκλήσεις, συνδυάσαμε την ταξινόμηση γράφων N-γραμμάτων με τα πλεονεκτήματα του αγωγού (pipieline) Apache Beam και τις δυνατότητες που προσφέρουν οι υπηρεσίες νέφους της Google.

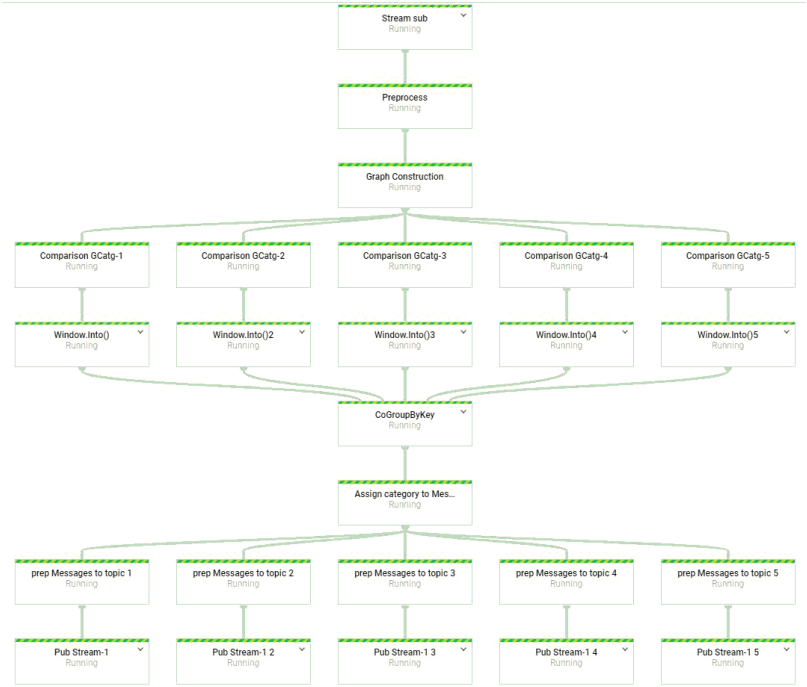
Το μοντέλο κατηγοριοποίησης ΓΝΓ εμπίπτει στην κλάση των αμήχανα παράλληλων προβλημάτων (Embarrassingly parallel problems) τα οποία χαρακτηρίζονται από έναν εύκολο διαχωρισμό των εργασιών τους χωρίς την ανάγκη ενδιάμεσης επικοινωνίας μεταξύ τους καθιστώντας δυνατή την ανεξάρτητη και παράλληλη επεξεργασία των εργασιών. Ένας αγωγός εργασιών υλοποιείται χρησιμοποιώντας το μοντέλο προγραμματισμού Apache Beam. Αυτός ο αγωγός περιλαμβάνει ξεχωριστές και διακριτές εργασίες που υλοποιούνται ως μέθοδοι μετασχηματισμού, απεικονίζονται στο Σχήμα 4.6. Αυτός ο αγωγός εκτείνεται σε εννέα στάδια με το στάδιο περισσότερων απαιτήσεων υπολογιστικής ισχύς που είναι η σύγκριση γραφήματος να επεκτείνεται σε k παράλληλους κόμβους όπου k είναι ο αριθμός των προκαθορισμένων κατηγοριών (k είναι 5 στο σχήμα 4.6). Στα δύο τελευταία στάδια πραγματοποιείται η διαδικασία της κατεύθυνσης των κατηγοριοποιημένων κειμένων στις αντίστοιχες εξόδους ροής δεδομένων.

Το μοντέλο προγραμματισμού Beam χρησιμοποιεί αγωγούς για να κατευθύνει και να επεξεργάζεται δεδομένα. Αυτοί οι αγωγοί οργανώνονται ως κατευθυνόμενοι ακυκλικοί γράφοι και διευκολύνουν την κατανεμημένη επεξεργασία δεδομένων σε μεγάλη κλίμακα, με τρόπο που ενοποιεί την επεξεργασία συνόλου και ροής δεδομένων. Τα κατανεμημένα δεδομένα, για να είναι συμβατά με την ορολογία του Beam, καλούνται PCollections και μπορούν να δημιουργηθούν από μια εξωτερική πηγή δεδομένων ή από δεδομένα μνήμης. Τα PCollections αποτελούνται από στοιχεία οποιουδήποτε τύπου με τον περιορισμό ότι όλα τα στοιχεία που ανήκουν σε PCollection πρέπει να είναι του ίδιου τύπου και η τυχαία πρόσβαση δεν υποστηρίζεται. Τα στοιχεία συνδέονται επίσης με ένα χρονικό αποτύπωμα που αποδίδεται στις περισσότερες περιπτώσεις όταν το στοιχείο διαβάζεται ή προστίθεται.

Τα PCollections επεξεργάζονται και διαβιβάζονται στα συνεχόμενα διακριτά στάδια του αγωγού. Η επεξεργασία που λαμβάνει χώρα σε ένα στάδιο του αγωγού ονομάζεται μετασχηματισμός και μπορεί να αφορά την αλλαγή, τη επιλογή, την ομαδοποίηση, την κατάτμηση, την εξαγωγή τμημάτων της εισόδου και την ανάλυση των στοιχείων PCollection. Οι μετασχηματισμοί εξετάζουν κάθε στοιχείο στο PCollection μεμονωμένα και πρέπει να ικανοποιούν τα κριτήρια της σειριοποίησης, της συμβατότητας με νήματα εκτέλεσης και του να είναι ταυτοδύναμα (idempotent). Άλλα τρία χαρακτηριστικά του μοντέλου προγραμματισμού Beam είναι τα παράθυρα που υποδιαιρούν τα στοιχεία ενός PCollection με βάση τα υδατογραφήματα timestamps τους, η διαχείριση των στοιχείων που έφτασαν αργά και οι ενεργοποιητές (triggers) για να προσδιορίσουν πότε πρέπει να πραγματοποιηθούν μετασχηματισμοί συνάθροισης. Τέλος, η χρήση του αγωγού Beam που παρέχεται από το Dataflow λαμβάνει όλα τα πλεονεκτήματα του Google cloud όσον αφορά την διαθεσιμότητα, την αυτοματοποιημένη δέσμευση υπολογιστικών πόρων, την έγκαιρη απόκριση, την αξιοπιστία, με οικονομικά αποδοτικό τρόπο.

Ο αγωγός ταξινόμησης ΓΝΓ ξεκινά με την συνδρομή (subscription) σε μια ροή κειμένων και μετατρέπει αυτή τη ροή σε ένα απεριόριστο PCollection όπου τα στοιχεία του αντιπροσωπεύουν μεμονωμένα μη κατηγοριοποιημένα κείμενα. Ο δεύτερος κόμβος ορίζει το στάδιο προεπεξεργασίας που βελτιώνει κάθε κείμενο με αφαίρεση των κοινών λέξεων και λημματοποίηση. Ο τρίτος κόμβος αντιστοιχεί στη μετατροπή των κειμενικών σε ΓΝΓ. Το τέταρτο επίπεδο κόμβων είναι το πιο κρίσιμο στάδιο του αγωγού γιατί εδώ γίνεται η σύγκριση γράφων κειμένου με κάθε ΓΝΓ που αντιπροσωπεύει μια θεματική κατηγορία και παράγει την τιμή της ομοιότητάς τους. Επειδή αυτές οι συγκρίσεις είναι δεδομένα ανεξάρτητα το ένα από το άλλο και συνεπάγονται ένα βαρύτερο φόρτο εργασίας επεξεργασίας σχεδιάζονται για να αντιστοιχίζονται σε διαφορετικούς κόμβους. Το έκτο επίπεδο περιλαμβάνει έναν μετασχηματισμό παραθύρου που είναι

απαραίτητος για να γίνει στο έβδομο επίπεδο γη συγκέντρωση τις ομοιότητες τιμών και να σχηματίσει ένα διανυσματικό χώρο διαστάσεων k που αντιπροσωπεύει κάθε κείμενο. Το έκτο επίπεδο περιλαμβάνει ένα σύνολο απαραίτητων μετατροπών των δεδομένων που είναι απαραίτητο ούτως ώστε το έβδομο επίπεδο να μπορεί να ομαδοποιήσει τους βαθμούς ομοιότητας σε ένα k-διαστάσεων διάνυσμα που αναπαριστά κάθε κείμενο. Το έβδομο επίπεδο εκτελεί την κατηγοριοποίηση και βάζει μια ετικέτα σε κάθε κείμενο. Το όγδοο επίπεδο δέχεται τα κείμενα με τις ετικέτες τους από το έβδομο επίπεδο και κατευθύνει κάθε επισημασμένο κείμενο στην κατάλληλη ροή εξόδου. Τέλος, το όγδοο επίπεδο αποτελείται από k κόμβους (publisher) που δημοσιεύουν τα κείμενα στις αντίστοιχες ροές εξόδου.



Σχήμα 4.6 Ταξινόμηση κειμένων μέσω ΓΝΓ και του Beam pipeline

4.5.Πειραματική Αξιολόγηση του Μοντέλου Κατηγοριοποίησης Κειμένων με ΓΝΓ

Το μοντέλο κατηγοριοποίησης κειμένων με χρήση ΓΝΓ έχει αξιολογηθεί πειραματικά με δύο διαφορετικά σύνολα δεδομένων από όπου φαίνεται η δυνατότητα εφαρμογής του και η αποτελεσματικότητά του σε μια σειρά ροής κειμένων. Η κατηγοριοποίηση σε πραγματικό χρόνο είναι μια κρίσιμη εφαρμογή που περιλαμβάνει δύο βασικούς παράγοντες, την ακρίβεια των προβλέψεων κατηγοριοποίησης και την εφαρμοσιμότητα ενός αγωγού Beam για την πρόβλεψη κατηγοριών σε κείμενα κλιμακούμενης ή υψηλής ροής. Το μοντέλο ταξινόμησης γράφων N-γραμμάτων αναπτύχθηκε με τη γλώσσα προγραμματισμού Java και τη προγραμματιστική βιβλιοθήκη Weka που περιλαμβάνει τους ταξινομητές SVM και Bayes. Ο αγωγός Beam αναπτύχθηκε και εκτελέστηκε στον δρομολογητή (runner) Dataflow στο Google Cloud. Η αξιολόγηση της κατηγοριοποίησης έγινε με τη χρήση της μεθόδου αξιολόγησης δέκα διασταυρούμενων αναδιπλώσεων (10-fold cross validation) και χρήση μίκρο και μάκρο μετρικών αξιολόγησης [82].

Το μοντέλο ταξινόμησης που κάνει χρήση των ΓΝΓ εφαρμόστηκε στο σύνολο δεδομένων 20newsgroup (έκδοση matlab/octave) με κείμενα που χωρίζονται εξίσου σε είκοσι διαφορετικές κατηγορίες. Κάθε μία από αυτές τις είκοσι κατηγορίες αντιστοιχεί σε μια ομάδα θεματικής συζήτησης και όλες περιλαμβάνουν κείμενα για ένα συγκεκριμένο θέμα, ενδεικτικά κάποιες κατηγορίες διαπραγματεύονται θέματα όπως γραφικά υπολογιστών, λειτουργικό σύστημα υπολογιστών Windows, αυτοκίνητα, μοτοσικλέτες, διαστημικό χώρο, θρησκεία κλπ. Όλες αυτές οι κατηγορίες είναι διακριτές και διαφορετικές και όλα τα κείμενα ανήκουν μόνο σε μια κατηγορία, αν και ορισμένες κατηγορίες είναι πολύ στενά συνδεδεμένες μεταξύ τους, π.χ. pc hardware to mac hardware και θρησκεία στον αθεϊσμό. Όλο το σύνολο κειμένων χρησιμοποιείται στα πειράματα και στην αξιολόγηση εκτός από λίγα κείμενα που απορρίφθηκαν γιατί περιλαμβάναν λιγότερες από δύο λέξεις.

Το Reuters - 21578 είναι μια συλλογή από άρθρα που συλλέχθηκαν και ταξινομήθηκαν σε κατηγορίες από το προσωπικό του Reuters Ltd. Τα κείμενα είναι άνισα κατανεμημένα σε θεματικές κατηγορίες. Στα πειράματά μας χρησιμοποιήσαμε τις δέκα κατηγορίες που περιείχαν τα περισσότερα κείμενα (earn, acq, money-fx, grain, crude, trade, interest, ship, wheat και corn) όπως περιγράφονται στα πειράματα του [83]. οι δέκα κατηγορίες είναι εξαιρετικά ανόμοιες όσον αφορά το πλήθος κειμένων που περιλαμβάνει η κάθε μια. Ενδεικτικά η μεγαλύτερη κατηγορία περιλαμβάνει 32 φορές περισσότερα κείμενα απ' ότι η μικρότερη. Το χαρακτηριστικό της ανόμοιας κατανομής κειμένων καθιστά την κατηγοριοποίηση ένα πιο δύσκολο έργο επειδή οι περισσότερες από τις μεθόδους ταξινόμησης τείνουν να ταξινομούν όλα τα κείμενα στις μεγάλες κατηγορίες αφήνοντας κενές τις μικρότερες.

Ο σκοπός των πειραμάτων είναι διπλός. Πρώτον, να πραγματοποιήσουμε μια έρευνα σε όλες τις παραμέτρους που επηρεάζουν την ακρίβεια και την χρονική ανταπόκριση του προτεινόμενου μοντέλου και δεύτερον, να πραγματοποιηθεί μια σύγκριση όσων αφορά την ακρίβεια προβλέψεων με άλλες διαθέσιμες μεθόδους κατηγοριοποίησης κειμένων. Οι εξεταζόμενες παράμετροι είναι ο βαθμός Ν των Ν-γραμμάτων, η λημματοποίηση, η αφαίρεση των κοινών λέξεων, η χρήση σταθμισμένων ή μη σταθμισμένων γράφων, οι μετρήσεις ομοιότητας γράφων και οι τεχνικές ταξινόμησης διανυσμάτων.

4.6.Πειραματικά Αποτελέσματα

Τα πειράματα κατηγοριοποίησης έγιναν στα σύνολα δεδομένων του Reuters - 21578 και του 20 Newsgroup ως ένα Java project στο Eclipse εγκατεστημένο σε έναν υπολογιστή βασικών προδιαγραφών με επεξεργαστή Pentium 5 2.9GHz, 16GB ram το οποίο έπειτα έτρεξε ως υπηρεσία στο Google Cloud μέσω του δρομολογητή Cloud Dataflow. Για τις ανάγκες της αξιολόγησης της κατηγοριοποίησης, διαιρέσαμε το πλήθος των κειμένων σε δύο τμήματα, της εκπαίδευσης και των δοκιμών και διεξήγαμε τα πειράματα σύμφωνα με την δεκαπλή-αναδίπλωση διασταυρούμενης αξιολόγησης. Από την πλευρά ενός κατανεμημένου μοντέλου που επεξεργάζεται δεδομένα ροής σε πραγματικό χρόνο, αναπτύξαμε το μοντέλο κατηγοριοποίησης ΓΝΓ ως υπηρεσία στο Dataflow του Google Cloud και εξετάσαμε την εφαρμογή του, την χρονική του απόκριση και την αξιοπιστία του μοντέλου.

Τα αρχεία κειμένου μετατρέπονται από έναν κάδο (bucket) που βρίσκεται στην πλατφόρμα αποθήκευσης του Google Cloud σε ροή κειμένου σε πραγματικό χρόνο με την υπηρεσία μηνυμάτων Pub / Sub. Κάθε ξεχωριστό κείμενο του συνόλου δεδομένων δημοσιεύεται σε ένα θέμα που παρακολουθείται από το μοντέλο ταξινόμησης κειμένων. Χρησιμοποιώντας την ορολογία του Google Cloud, τα θέματα (topics) είναι οι πόροι (resources) στους οποίους ένας ή περισσότεροι εκδότες (publisher) μπορούν να δημοσιεύουν μηνύματα τους και αντίστοιχα ένας ή περισσότεροι συνδρομητές μπορούν να τα καταναλώσουν. Το Pub / Sub λειτουργεί ως ενδιάμεσο λογισμικό με χαρακτηριστικό ότι τα μηνύματα κειμένου είναι ασύγχρονα και οι αποστολείς είναι αποσυνδεδεμένοι από τις υπηρεσίες των δεκτών. Με παρόμοιο τρόπο, όταν τα κείμενα εισερχόμενων ροών που έχουν επισημανθεί από τη διαδικασία ταξινόμησης κατευθύνονται στον αντίστοιχο εκδότη θα δημοσιεύονται στην κατάλληλη θεματική κατηγορία. Το Pub / Sub έχει την δυνατότητα αυτόματης κλιμάκωσης είναι αξιόπιστο και μεταβιβάζει μηνύματα σε πραγματικό χρόνο.

Οι ακόλουθες παράγραφοι συνοψίζουν τα πειραματικά αποτελέσματα του προτεινόμενου μοντέλου ταξινόμησης ΓΝΓ από την πλευρά της ακρίβειας πρόβλεψης. Η επόμενη υποενότητα περιέχει μια συζήτηση σχετικά με τα αποτελέσματα πρόβλεψης, την απόκριση χρόνου, την εφαρμογή και την ανάπτυξη του αγωγού Beam στο Google Cloud. Οι όροι που χρησιμοποιούνται στους παρακάτω πίνακες παρατίθενται παρακάτω:

CS: Ομοιότητα περιεχομένου για μη σταθμισμένους γράφους

VS: Ομοιότητα τιμής για σταθμισμένους γράφους

NVS: Ομοιότητα κανονικοποιημένης τιμής για σταθμισμένους γράφους

MCSNS: Μέγιστη ομοιότητα κοινού υπογράφου που βασίζεται στους κοινούς κόμβους

MCSUES: Μέγιστη ομοιότητα κοινού υπογράφου που βασίζεται στα κοινές μη κατευθυνόμενες ακμές

MCSDES: Μέγιστη ομοιότητα κοινού υπογράφου που βασίζεται στις κοινές κατευθυνόμενες ακμές

Stm: Λημματοποίηση των κειμένων με χρήση του αλγορίθμου Porter [84]

SWR: Αφαίρεση των κοινών λέξεων (Stop Words) από το σύνολο των κειμένων.

SWR & Stm: Ένας συνδυασμός αφαίρεσης των κοινών λέξεων και λημματοποίησης.

SVM: Κατηγοριοποίηση με τον αλγόριθμο Support Vector Machines

BayPr: Κατηγοριοποίηση με χρήση της Μπεϋζιανής στατιστικής

20Newsgroup

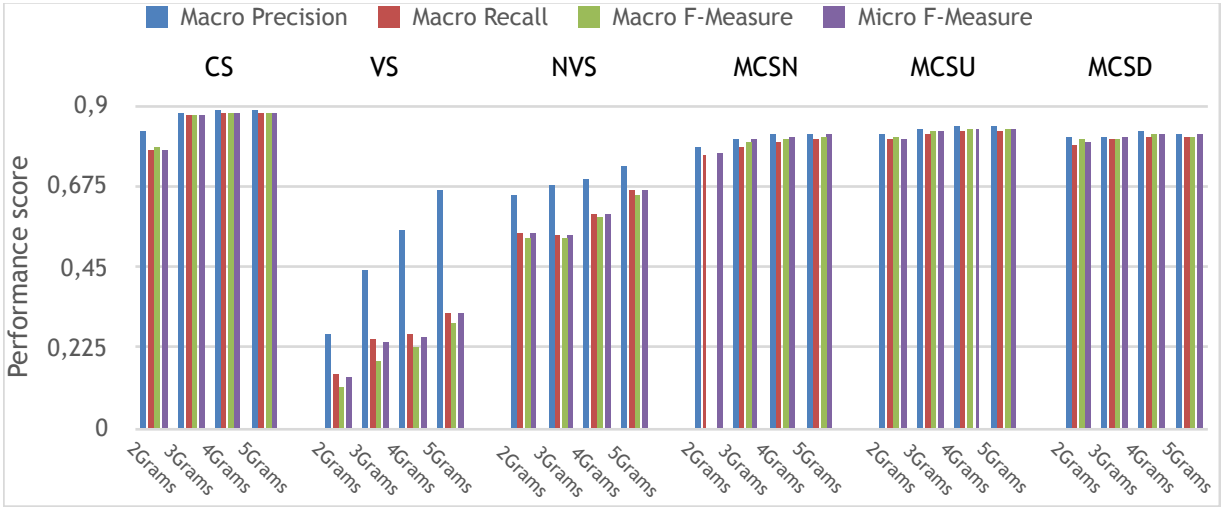
Και οι δύο εκδοχές των γράφων, σταθμισμένες και μη σταθμισμένες, δοκιμάζονται πειραματικά για πολλές τάξεις N-γραμμάτων χρησιμοποιώντας την Μπεϋζιανή κατηγοριοποίηση. Για μη σταθμισμένους γράφους χρησιμοποιήσαμε της μετρικές, ομοιότητα περιεχομένου (CS), ομοιότητα μέγιστου κοινού υπογράφου με άθροισμα κοινών κόμβων (MCSNS), μέγιστου κοινού υπογράφου με βάση το σύνολο μη κατευθυνόμενων ακμών (MCSUES), και μέγιστου κοινού υπογράφου με βάση το σύνολο κατευθυνόμενων ακμών (MCSDES). ενώ για σταθμισμένους γράφους χρησιμοποιήθηκαν δύο μετρικές ομοιότητας, η ομοιότητα τιμής (VS) και η κανονικοποιημένη ομοιότητα τιμής (NVS), οι πίνακας 4.1 συνοψίζει τα πειραματικά αποτελέσματα.

Από το πρώτο σύνολο πειραμάτων βλέπουμε ότι η μέθοδος κατηγοριοποίησης ΓΝΓ μπορεί να έχει καλύτερα αποτελέσματα από άλλες μεθόδους. Συγκεκριμένα, η ακρίβεια του μοντέλου με μη σταθμισμένους ΓΝΓ και N = 5 είναι 88,16%. Η αντίστοιχη πρόβλεψη της μελέτης using Error-Correcting Output Coding and Sub-class Partitions(Li and Vogel, 2010) είναι ίση με 81,84%. Επιπλέον, μπορούμε να δούμε ότι οι μη σταθμισμένοι γράφοι με ομοιότητα περιεχομένου υπερβαίνουν τους σταθμισμένους γράφους και η χρήση 5-γραμμάτων δίνει τα καλύτερα αποτελέσματα.

Πίνακας 4.1 Αξιολόγηση μετρήσεων ομοιότητας γράφων με το σύνολο δεδομένων 20Newsgroup

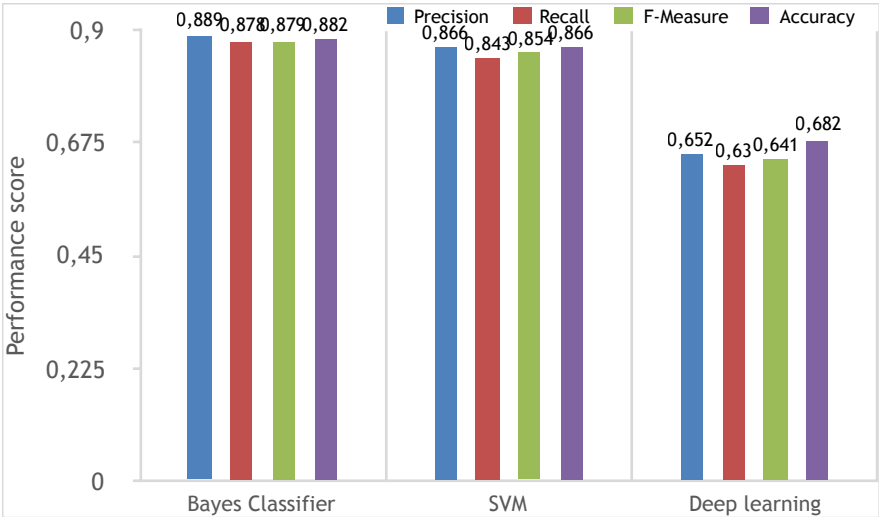
	Macro Precision	Macro Recall	Macro F-Measure	Micro F-Measure
Containment Similarity				
2Grams	0.8303	0.7711	0.7818	0.7728
3Grams	0.8792	0.8695	0.8688	0.8725

4Grams	0.8879	0.8768	0.8767	0.8801
5Grams	0.8894	0.8781	0.8786	0.8816
Value Similarity				
2Grams	0.2652	0.149	0.1130	0.1413
3Grams	0.4425	0.2497	0.1844	0.2414
4Grams	0.5490	0.2614	0.2263	0.257
5Grams	0.6597	0.3205	0.2935	0.3186
Normalized Value Similarity				
2Grams	0.6492	0.5432	0.5331	0.5465
3Grams	0.6806	0.5381	0.5274	0.54
4Grams	0.6938	0.5971	0.5862	0.5964
5Grams	0.7316	0.6609	0.6516	0.6611
Maximum Common subgraph with Nodes Similarity				
2Grams	0.7817	0.7598	0,7705	0.7692
3Grams	0.8046	0.7842	0.7942	0.8081
4Grams	0.8218	0.7939	0.8076	0.8125
5Grams	0.8227	0.8073	0.8149	0.8193
Maximum Common subgraph with Undirected Edges Similarity				
2Grams	0.8197	0.8076	0.8136	0.8064
3Grams	0.8356	0.8196	0.8275	0.8289
4Grams	0.8384	0.824	0.8311	0.834
5Grams	0.8419	0.8268	0.8342	0.8378
Maximum Common subgraph with Directed Edges Similarity				
2Grams	0.8110	0.7926	0.8016	0.7991
3Grams	0.8129	0.8034	0.8081	0.8103
4Grams	0.8258	0.8127	0.8191	0.8197
5Grams	0.8172	0.8114	0.8142	0.8214



Σχήμα 4.7 Αξιολόγηση μετρήσεων ομοιότητας γράφων με το σύνολο δεδομένων 20Newsgroup

Συγκρίναμε τον ταξινομητή που εφαρμόζει μπεϋζιανή στατιστική με τον SVM και τον ταξινομητή που εφαρμόζει βαθιά μάθηση, όπως φαίνεται στο Σχήμα 4.8, ο ταξινομητής μπεϋζιανή στατιστικής είχε καλύτερα αποτελέσματα πρόβλεψης. Ο λόγος για τον οποίο ο SVM εμφανίζει χαμηλότερη αποτελέσματα είναι ότι χρησιμοποιεί ένα χώρο χαμηλών διαστάσεων. Από την άλλη πλευρά η βαθιά εκμάθηση απαιτεί μεγάλη ποσότητα δεδομένων για να δείξει καλά αποτελέσματα πρόβλεψης. Σε περίπτωση μόνο χιλιάδων παρατηρήσεων, είναι δύσκολο να ξεπεράσει το μπεϋζιανό ταξινομητή. Περαιτέρω πειραματικά αποτελέσματα με το σταθμισμένο ΓΝΓ δεν θα παρουσιαστούν επειδή η ακρίβειά τους είναι πολύ χαμηλότερη από τον μη σταθμισμένο ΓΝΓ. Στα ακόλουθα πειράματα διατηρήσαμε την ομοιότητα περιεχομένου σε συνδυασμό με τον μπεϋζιανό ταξινομητή και εξετάσαμε τις τεχνικές προεπεξεργασίας χρησιμοποιώντας τον αλγόριθμο Porter και τη διαγραφή των κοινών λέξεων όπως βλέπουμε στον πίνακα 4.2



Σχήμα 4.8 Αξιολόγηση της απόδοσης ταξινομητών με το σύνολο δεδομένων 20Newsgroup

Η προεπεξεργασία της λημματοποίησης έχει μικρή βελτίωση μικρότερη από 0,10% τόσο στα Μικρο όσο και στις Μακρό μετρικές, επομένως δεν πραγματοποιήσαμε περαιτέρω πειράματα με υψηλότερη βαθμίδα N. Μετά την αφαίρεση των κοινών λέξεων, υπάρχει βελτίωση των μετρικών που υπερβαίνει το 1,20% για N = 6. Η σταδιακή βελτίωση των αποτελεσμάτων της αξιολόγησης με τις υψηλότερες τάξεις του N μας κάνει να συνεχίσουμε τα πειράματά μας με υψηλότερο βαθμό N. Στόχος μας είναι να βρούμε την καλύτερη ακρίβεια προβλέψεων συναρτήσει του N.

Το τελευταίο σύνολο πειραμάτων που διεξήγαμε είναι ο συνδυασμός της αφαίρεσης των κοινών λέξεων με την λημματοποίηση. Σε αυτήν την περίπτωση μπορούμε να δούμε ακόμα καλύτερα αποτελέσματα

προβλέψεων για όλες τις μετρήσεις. Στο στάδιο προεπεξεργασίας πραγματοποιήσαμε πρώτα την αφαίρεση των κοινών λέξεων και στη συνέχεια τη λημματοποίηση. Τα πειράματα στην περίπτωση όπου πρώτα πραγματοποιήθηκε η λημματοποίηση και μετά η αφαίρεση των κοινών λέξεων έχουν ελαφρώς χειρότερα αποτελέσματα αξιολόγησης.

Πίνακας 4.2 Αξιολόγηση τεχνικών προεπεξεργασίας με ομάδα 20Newsgroup

	Macro Precision	Macro Recall	Macro F-Measure	Micro F-Measure
Stemmed CS				
2Grams	0.8285	0.7686	0.7791	0.7698
3Grams	0.8783	0.8672	0.8665	0.8701
4Grams	0.8870	0.8753	0.8753	0.8788
5Grams	0.8900	0.8790	0.8795	0.8826
Stop Words Removal CS				
2Grams	0.8296	0.7733	0.7828	0.7749
3Grams	0.8774	0.8684	0.8677	0.8715
4Grams	0.8892	0.8808	0.8807	0.8840
5Grams	0.8947	0.8879	0.8879	0.8910
6 Grams	0.8965	0.8911	0.8913	0.8938
7Grams	0.8959	0.8911	0.8914	0.8935
8Grams	0.8928	0.8886	0.8888	0.8905
9Grams	0.8877	0.8836	0.8839	0.8852
10Grams	0.8808	0.8765	0.8770	0.8777
Stop Words Removal Stemmed CS				
2Grams	0.8290	0.7693	0.7792	0.7713
3Grams	0.8791	0.8697	0.8692	0.8727
4Grams	0.8922	0.8843	0.8844	0.8874
5Grams	0.8977	0.8913	0.8916	0.8943
6 Grams	0.9010	0.8961	0.8964	0.8987
7Grams	0.8980	0.8936	0.8940	0.8957
8Grams	0.8940	0.8901	0.8905	0.8920
9Grams	0.8877	0.8838	0.8842	0.8852
10Grams	0.8815	0.8772	0.8779	0.8784

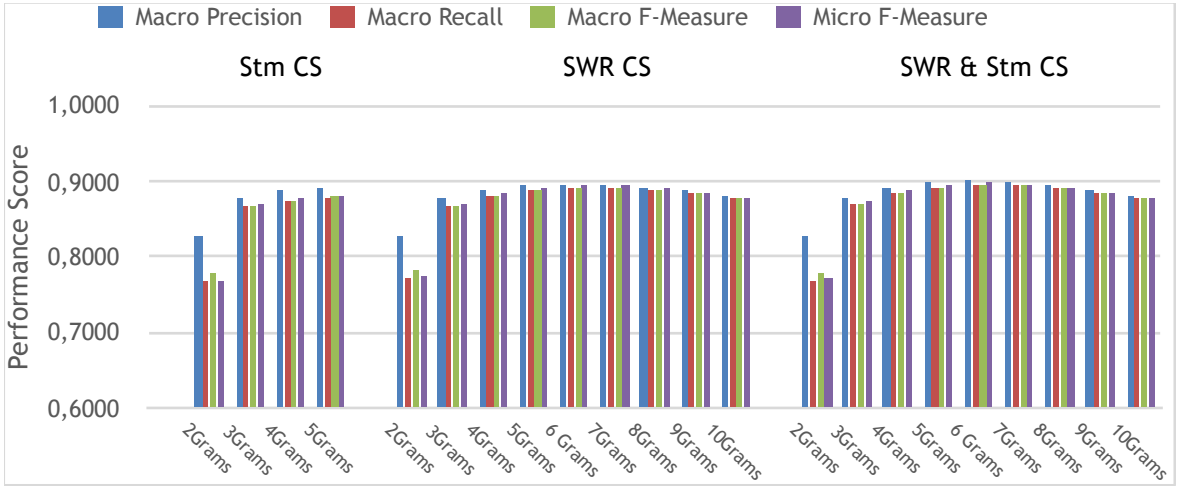


Figure 4.9 Αξιολόγηση τεχνικών προεπεξεργασίας με το σύνολο δεδομένων 20Newsgroup

Τα πειραματικά αποτελέσματα δείχνουν ότι οι μη σταθμισμένοι γράφοι 6-γραμμάτων όπου οι κοινές λέξεις έχουν αφαιρεθεί έχουν τα καλύτερα αποτελέσματα προβλέψεων με την Μίκρο Φ-μετρική να φτάνει 89,87% ξεπερνώντας ακόμα περισσότερο το ποσοστό του 81,84% της έρευνας των [85]. Η προεπεξεργασία της αφαίρεσης των κοινών λέξεων και της λημματοποίησης έχει ένα ακόμη σημαντικό όφελος. Μειώνει την είσοδο δεδομένων, το σύνολο χαρακτήρων των κειμένων κατά ποσοστό 44,5% καθιστώντας τη μέθοδο ταξινόμησης να απαιτεί λιγότερους υπολογιστικούς πόρους.

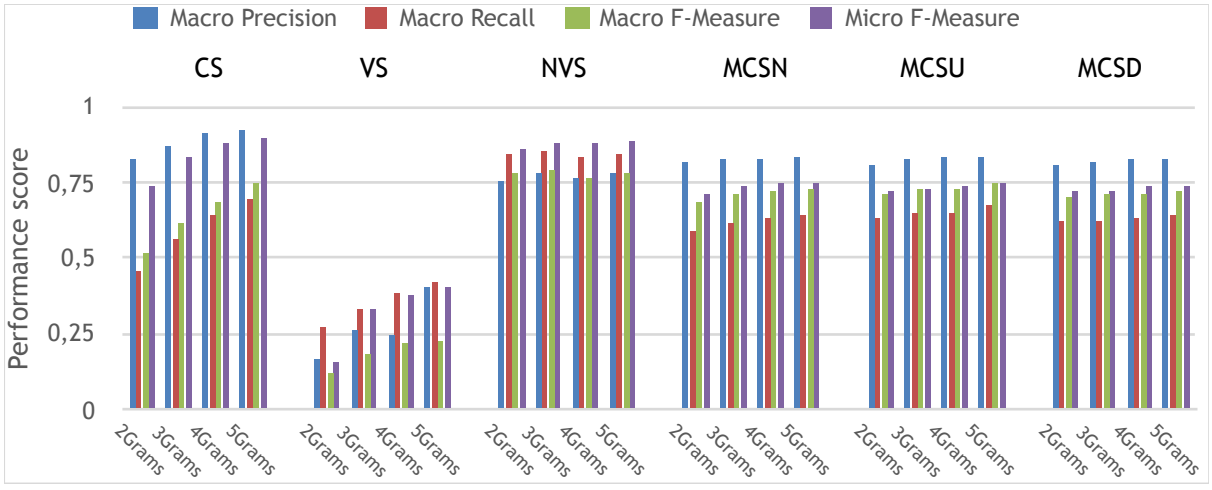
Reuters-21578

Τα κείμενα στο σύνολο δεδομένων του Reuters-21578 είναι με ανόμοιο πλήθος κατανεμημένα στις κατηγορίες. Αυτό το χαρακτηριστικό δυσχεραίνει την διαδικασία κατηγοριοποίησης δημιουργώντας την τάση τα κείμενα που ανήκουν σε μικρές κατηγορίες να ταξινομούνται εσφαλμένα στις μεγάλες κατηγορίες. Πραγματοποιήσαμε ξανά πειράματα με τις δύο εκδοχές γράφων, σταθμισμένες και μη σταθμισμένες, πολλούς βαθμούς N-γραμμάτων και μετρικές ομοιότητας γράφων χρησιμοποιώντας τον μπεϋζιανό ταξινομητή. Όπως μπορούμε να δούμε στον Πίνακα 4.3, ο κανονικοποιημένος παράγοντας για χαμηλής τάξης N έχει θετική επίδραση στις Μάκρο μετρικές αξιολόγησης Recall και F-Measure. Τα καλύτερα αποτελέσματα Μίκρο Φ-μέτρων προκύπτουν για μη σταθμισμένους γράφους με χρήση ομοιότητας περιεχομένου με N = 5. Το ίδιο σύνολο παραμέτρων ήταν η καλύτερο και στο σύνολο δεδομένων 20Newsgroup.

Πίνακας 4.3 Αξιολόγηση μετρήσεων ομοιότητας γράφων με το Reuters-21578

	Macro Precision	Macro Recall	Macro F-Measure	Micro F-Measure
Containment Similarity				
2Grams	0.822	0.4537	0.5169	0.7361
3Grams	0.8705	0.5650	0.6159	0.8317
4Grams	0.9148	0.6364	0.6866	0.8747
5Grams	0.9185	0.6964	0.7495	0.8966

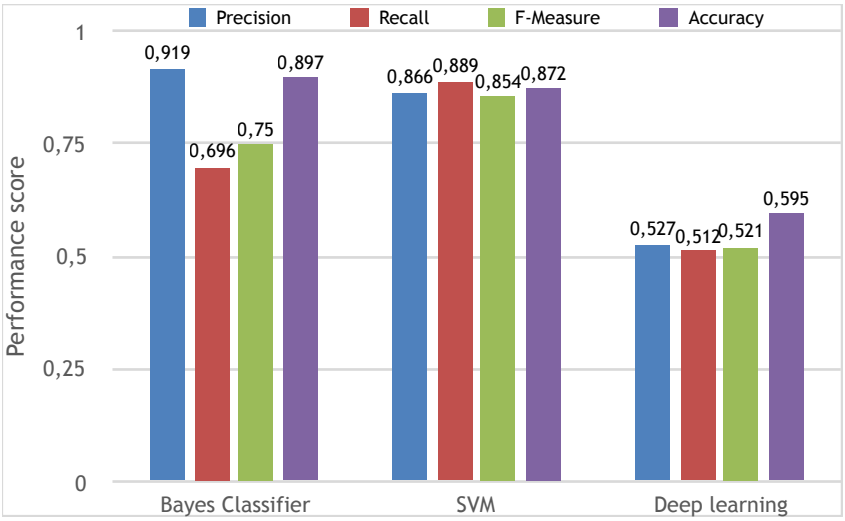
Value Similarity				
2Grams	0.1628	0.2674	0.1221	0.1572
3Grams	0.2625	0.3337	0.1850	0.3341
4Grams	0.2432	0.3879	0.2153	0.3741
5Grams	0.4048	0.4242	0.2272	0.4028
Normalized Value Similarity				
2Grams	0.7511	0.8435	0.7784	0.8631
3Grams	0.7804	0.8498	0.7880	0.8807
4Grams	0.7673	0.8384	0.7681	0.8796
5Grams	0.7785	0.8388	0.7774	0.8905
Maximum Common Subgraph with Nodes Similarity				
2Grams	0.8127	0.5916	0.6847	0.7142
3Grams	0.8261	0.6177	0.7068	0.7357
4Grams	0.8284	0.6314	0.7166	0.7421
5Grams	0.8347	0.6421	0.7258	0.7467
Maximum Common Subgraph with Undirected Edges Similarity				
2Grams	0.8103	0.6291	0.7082	0.7216
3Grams	0.8224	0.6473	0.7244	0.7281
4Grams	0.8298	0.6519	0.7301	0.7367
5Grams	0.8341	0.6722	0.7444	0.7438
Maximum Common Subgraph with Directed Edges Similarity				
2Grams	0.8116	0.6197	0.7027	0.7198
3Grams	0.8165	0.6247	0.7078	0.7219
4Grams	0.8232	0.6315	0.7147	0.7341
5Grams	0.8276	0.6381	0.7205	0.7376



Σχήμα 4.10 Αξιολόγηση μετρήσεων ομοιότητας γράφων με το σύνολο δεδομένων Reuters-21578

Οι Μίκρο μετρικές αξιολόγησης δηλώνουν την ποσότητα κειμένων που έχουν ταξινομηθεί καλά. Τα κείμενα που ανήκουν σε μεγάλες κατηγορίες είναι περισσότερα από τα κείμενα που ανήκουν σε μικρές κατηγορίες, οπότε ο κανονικοποιημένος παράγοντας δεν επηρεάζει αυτό το αποτέλεσμα. Από την άλλη πλευρά, τα αποτελέσματα των μετρήσεων Φ-Μετρικών είναι καλύτερα χρησιμοποιώντας σταθμισμένους γράφους και την κανονικοποιημένη ομοιότητα τιμών. Ο παράγοντας κανονικοποίησης βελτίωσε σημαντικά τον βαθμό όπου τα κείμενα κατηγοριοποιούνται σε μια κλάση (Macro Recall), ενώ είχε ως αποτέλεσμα τη μείωση του βαθμού των επιλεγμένων στοιχείων που έχουν σχέση με μια κατηγορία (Macro Precision). Οι Μάκρο Φ-μετρικές είναι επίσης καλύτερες χρησιμοποιώντας σταθμισμένους γράφους με κανονικοποιημένη ομοιότητα τιμών.

Συγκρίναμε τον μπεϋζιανό ταξινομητή με τον SVM και τον ταξινομητή που χρησιμοποιεί βαθιά εκμάθηση, όπως κάναμε στο σύνολο δεδομένων 20Newsgroup. Ο μπεϋζιανός ταξινομητής είχε και πάλι τα καλύτερα αποτελέσματα πρόβλεψης για αυτό τον διατηρήσαμε για τα επόμενα πειράματα που κάναμε όσων αφορά την προεπεξεργασία κειμένων.



Σχήμα 4.11 Αξιολόγηση της απόδοσης των ταξινομητών με το Reuters-21578

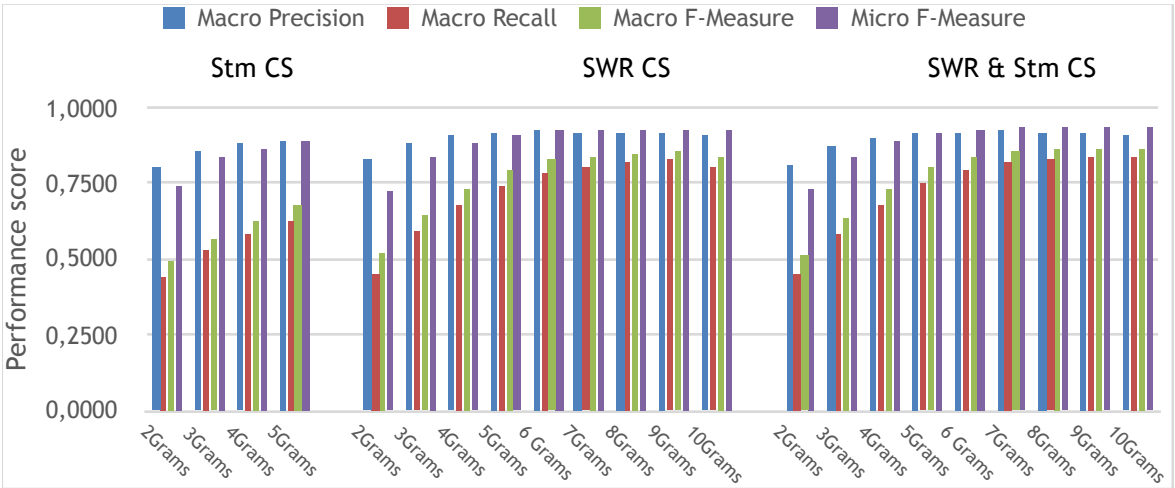
Οι δύο μέθοδοι προεπεξεργασίας που εφαρμόστηκαν στο μη σταθμισμένο και στο σταθμισμένο μοντέλο ΓΝΓ και τα αποτελέσματα παρουσιάζονται στον Πίνακα 4.4 και τον Πίνακα 4.5

Η λημματοποίηση μείωσε τις μετρήσεις αξιολόγησης τόσο στους σταθμισμένους όσο και στους μη σταθμισμένους ΓΝΓ, αλλά η αφαίρεση των κοινών λέξεων είχε αξιοσημείωτη θετική επίδραση στις Μίκρο και Μάκρο Φ-Μετρικές . Επιπλέον, η κατάργηση των κοινών λέξεων είχε πιο σημαντική επίδραση στο μοντέλο μη σταθμισμένων ΓΝΓ από ότι στο σταθμισμένο.

Πίνακας 4.4 Αξιολόγηση τεχνικών προεπεξεργασίας με το Reuters-21578 και μετρική περιεχομένου

	Macro Precision	Macro Recall	Macro F-Measure	Micro F-Measure
Stemmed Containment Similarity				
2Grams	0.7989	0.4423	0.4927	0.7388
3Grams	0.8523	0.5290	0.5691	0.8332
4Grams	0.8826	0.5791	0.6228	0.8666
5Grams	0.8910	0.6284	0.6756	0.8886
Stop Words Removal Containment Similarity				
2Grams	0.8250	0.4496	0.5167	0.7234

3Grams	0.8827	0.5904	0.6450	0.8355
4Grams	0.9040	0.6783	0.7317	0.8824
5Grams	0.9174	0.7410	0.7909	0.9067
6 Grams	0.9210	0.7817	0.8252	0.9214
7Grams	0.9184	0.8032	0.8413	0.9261
8Grams	0.9157	0.8156	0.8500	0.9267
9Grams	0.9135	0.8253	0.8563	0.9280
10Grams	0.9120	0.8011	0.8384	0.9271
Stop Words Removal & Stemmed Containment Similarity				
2Grams	0.8114	0.4482	0.5157	0.7361
3Grams	0.8756	0.5813	0.6394	0.8391
4Grams	0.9033	0.6753	0.7296	0.8871
5Grams	0.9181	0.7506	0.7992	0.9134
6 Grams	0.9208	0.7967	0.8382	0.9287
7Grams	0.9226	0.8167	0.8531	0.9353
8Grams	0.9197	0.8287	0.8607	0.9380
9Grams	0.9168	0.8365	0.8652	0.9384
10Grams	0.9110	0.8362	0.8625	0.9363

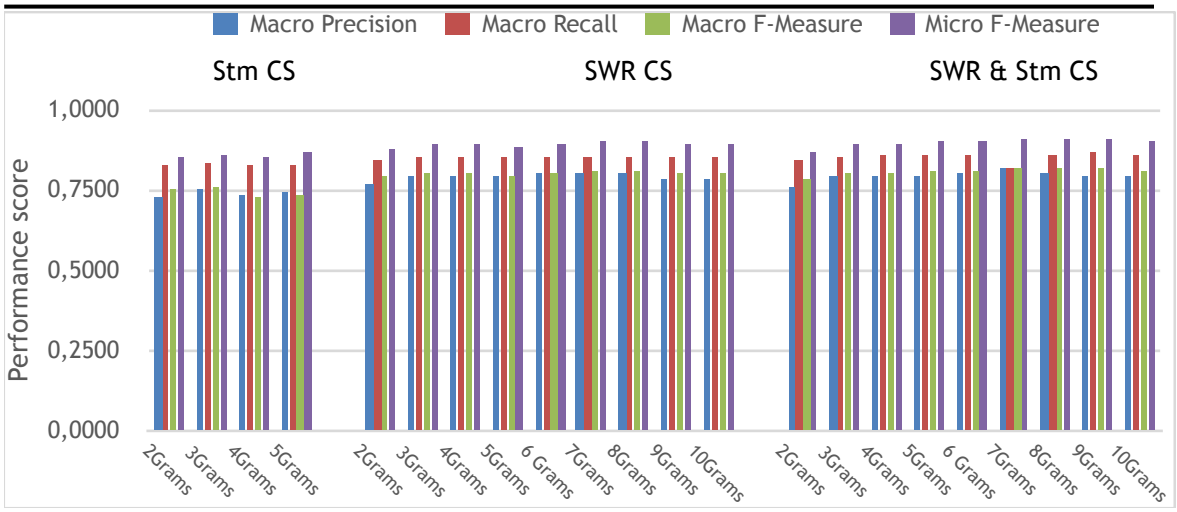


Σχήμα 4.12 Αξιολόγηση των τεχνικών προεπεξεργασίας με το Reuters-21578 και την ομοιότητα περιεχομένου

Η αφαίρεση των κοινών λέξεων αύξησε τις μετρήσεις αξιολόγησης περισσότερο στους μη σταθμισμένους γράφους από ότι στους σταθμισμένους. Το μέγιστο Μίκρο και Μάκρο μέτρο F είναι για N = 9 σε μη σταθμισμένους ΓΝΓ. Ακόμα κι αν τα αποτελέσματα της αξιολόγησης είναι πολύ καλά, αυτό είναι ένα μεγάλο N που καθιστά τη μέθοδο πιο απαιτητική σε υπολογιστικούς πόρους. Για να μειώσουμε περαιτέρω την ποσότητα δεδομένων, φιλτράραμε τα δεδομένα εισόδου χρησιμοποιώντας τον συνδυασμό κοινών λέξεων και λημματοποίησης όπως κάναμε και στο σύνολο δεδομένων 20Newsgroup. Τα πειράματα έδειξαν βελτίωση στην ακρίβεια που φτάνει το 0,89% για την Μάκρο F μετρική και 1,04% για την Μικρό Φ-Μετρική.

Πίνακας 4.5 Αξιολόγηση τεχνικών προεπεξεργασίας με το Reuters-21578 και την κανονικοποιημένη ομοιότητα αξίας

	Macro Precision	Macro Recall	Macro F-Measure	Micro F-Measure
Stemmed Normalized Value Similarity				
2Grams	0.7297	0.8230	0.7526	0.8494
3Grams	0.7506	0.8390	0.7568	0.8635
4Grams	0.7321	0.8244	0.7261	0.8520
5Grams	0.7409	0.8300	0.7391	0.8667
Stop Words Removal Normalized Value Similarity				
2Grams	0.7695	0.8475	0.7919	0.8758
3Grams	0.7945	0.8493	0.8009	0.8910
4Grams	0.7938	0.8530	0.7994	0.8911
5Grams	0.7894	0.8494	0.7965	0.8888
6 Grams	0.8030	0.8481	0.8059	0.8973
7Grams	0.8037	0.8505	0.8094	0.8987
8Grams	0.7994	0.8506	0.8087	0.8987
9Grams	0.7882	0.8489	0.8024	0.8966
10Grams	0.7874	0.8478	0.8024	0.8972
Stop Words Removal & Stemmed Normalized Value Similarity				
2Grams	0.7597	0.8411	0.7813	0.8693
3Grams	0.7929	0.8514	0.8002	0.8920
4Grams	0.7929	0.8561	0.8032	0.8975
5Grams	0.7950	0.8578	0.8079	0.9015
6 Grams	0.8001	0.8560	0.8128	0.9050
7Grams	0.8162	0.8162	0.8162	0.9066
8Grams	0.7987	0.8638	0.8189	0.9069
9Grams	0.7968	0.8654	0.8191	0.9060
10Grams	0.7927	0.8615	0.8142	0.9028



Σχήμα 4.13 Αξιολόγηση τεχνικών προεπεξεργασίας με την ομοιότητα του Reuters-21578 και την ομοιότητα αξίας

4.7.Συζήτηση στα Πειραματικά Αποτελέσματα

Τα πειραματικά αποτελέσματα επιβεβαιώνουν ότι το μοντέλο ταξινόμησης με ΓΝΓ μπορεί να παράγει ακριβείς προβλέψεις ταξινόμησης με χαμηλές υπολογιστικές απαιτήσεις. Μη σταθμισμένοι γράφοι με ομοιότητα περιεχομένου βαθμού Ν-γραμμάτων ίσος με 6 έχει τα καλύτερα Μάκρο αποτελέσματα και στα δύο σύνολα δεδομένων. Η τάξη Ν = 6 έχει τα καλύτερα αποτελέσματα αξιολόγησης και στις μετρήσεις για το σύνολο δεδομένων 20Newsgroup που είναι ισόποσα κατανεμημένο στις θεματικές κατηγορίες. Στην περίπτωση του ανόμοια κατανεμημένου συνόλου δεδομένων Reuters με ομοιότητα περιεχομένου σε μη ζυγισμένους γράφους, η τάξη Ν στα Ν-γράμματα ίση με 9 μετριάζει το φαινόμενο τα κείμενα που ανήκουν στις μικρές κατηγορίες να τείνουν να κατηγοριοποιούνται στις μεγάλες κατηγορίες.

Η χρήση μη σταθμισμένων γράφων Ν-γραμμάτων λειτουργεί καλύτερα από τη χρήση σταθμισμένων γράφων Ν-γραμμάτων σε όλες τις περιπτώσεις όπου τα κείμενα είναι με όμοιο τρόπο κατανεμημένα στις θεματικές κατηγορίες. Εκτενής έρευνα έχει γίνει για να δικαιολογήσει τους λόγους αυτών των συμπερασμάτων. Ένας γράφος κειμένου αποτελείται από ακμές που αντικατοπτρίζουν την έννοια του κειμένου αλλά και του θορύβου που μπορεί να υπάρχει σε οποιοδήποτε τυχαίο κείμενο όπως αναφέρεται στην ενότητα που διαπραγματευθήκαμε την προεπεξεργασία των κειμένων. Οι ακμές που συμβάλουν στην κατηγοριοποίηση των κειμένων έχουν μικρότερα βάρη από τις ακμές που αποτελούν θόρυβο, επειδή οι ακμές θορύβου επαναλαμβάνονται συχνότερα.

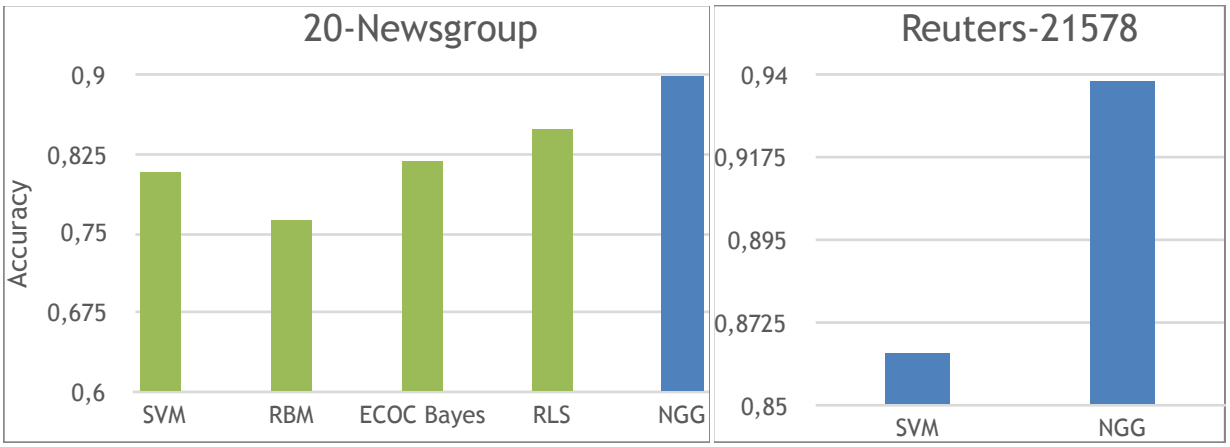
Επιπλέον, η απλή ομοιότητα αξίας αποτυγχάνει εντελώς να συγκρίνει επαρκώς δύο γράφους επειδή δεν λαμβάνει υπόψη το μέγεθος των γράφων. Η κανονικοποίηση της ομοιότητας γράφων βελτίωσε το ποσοστό ορθής πρόβλεψης αλλά δεν ξεπέρασε την ομοιότητα περιεχομένου των μη σταθμισμένων γράφων.

Η χρήση σταθμισμένων γράφων δεν δείχνει τόσο καλά αποτελέσματα σε σχέση με τους μη σταθμισμένους γράφους. Ωστόσο, μπορούν να ανιχνευθούν οι παράγοντες που συμβάλλουν στη μη ορθή προβλεψιμότητα. Αυτό σημαίνει ότι με τη χρήση κριτηρίων επιλογής χαρακτηριστικών, οι ακμές που συστηματικά μειώνουν την ακρίβεια της μεθόδου μπορούν να παραλειφθούν.

Στην περίπτωση των άνισα κατανεμημένων κειμένων μπορούμε να δούμε ότι οι σταθμισμένοι γράφοι με την κανονικοποιημένη ομοιότητα τιμών μπορούν να συνεισφέρουν σε καλύτερες προβλέψεις, ειδικά στη Μάκρο μετρική Recall και χρησιμοποιώντας κανονικοποίηση στις μετρικές μπορεί να συμβάλει στη σωστή κατηγοριοποίηση κειμένων που ανήκουν σε μικρές κατηγορίες.

Η χρήση ενός συνδυασμού παράληψης των κοινών λέξεων με λημματοποίηση στο στάδιο της προεπεξεργασίας μπορεί να αποδώσει ακόμα καλύτερα αποτελέσματα προβλέψεων. Επιπλέον, ο όγκος των δεδομένων μειώνεται σημαντικά και μειώνει τις αντίστοιχες απαιτήσεις σε υπολογιστικούς πόρους.

Η ταξινόμηση με τη χρήση μη σταθμισμένων γράφων 6-γραμμάτων με αφαίρεση των κοινών λέξεων και λημματοποίηση έχει ως αποτέλεσμα η Μάκρο Φ-Μετρική να φτάνει το 0.8964 και η Μικρό Φ-Μετρική το 0.8987 στο σύνολο δεδομένων του 20NewsGroup. Αυτά τα αποτελέσματα είναι καλύτερα από τα αποτελέσματα άλλων ερευνών της βιβλιογραφίας όπως Χρησιμοποιώντας το SVM [86] όπου το Μίκρο Φ-Μέτρο είναι 0.808. Ο hybrid discriminative RBM [87] όπου έχει Μίκρο Φ-Μέτρο 0.762, ο ECIC Naive Bayes [85] με 0.818 και ο regularized least squares classifier [88] με 0.8486.



Εικόνα 4.14 Σύγκριση μεθόδων ταξινόμησης κειμένων

Το σύνολο δεδομένων του Reuters έχει χρησιμοποιηθεί με πολλούς τρόπους στην διεθνή επιστημονική βιβλιογραφία. Μια από τις καλύτερες πειραματικές έρευνες που χρησιμοποιεί τις ίδιες πρώτες δέκα κατηγορίες είναι η κατηγοριοποίηση κειμένου με SVM [83], όπου το Μίκρο Φ-Μέτρο είναι 0,864. Χρησιμοποιώντας το μοντέλο ΓΝΓ, το Μίκρο Φ-μέτρο έφθασε στο 0,9384 το οποίο είναι μια μεγάλη βελτίωση.

Το Google Cloud μπορεί να μας απελευθερώσει από τον φόρτο διαχείρισης των υπολογιστικών υποδομών, την διαθεσιμότητα και την επεκτασιμότητα μιας υπηρεσίας. Συγκεκριμένα για διαδικασίες που μπορούν να χωριστούν σε μια σειρά από ανεξάρτητες υπο-εργασίες, παρέχει το προγραμματιστικό μοντέλο Beam που με τη μορφή ενός αγωγού τα δεδομένα επεξεργάζονται σε μια ακολουθία σταδίων. Η αυτόματη κατηγοριοποίηση ροών κειμένου υψηλής συχνότητας πραγματοποιείται σε πραγματικό χρόνο χωρίς να χρειάζονται να απαιτείται κατά την εκτέλεση κάποια ενέργεια όσων αφορά τους πόρους υλικού που πρέπει να διατεθούν. Μετρήσαμε την απόκριση χρόνου του μοντέλου που το τροφοδοτούσε από 1 κείμενο ανά δευτερόλεπτο έως και 10.000 κείμενα ανά δευτερόλεπτο και σε κάθε περίπτωση τα αρχεία καταγραφών (Log Files) είχαν καταγράψει ένα σταθερό χρόνο επεξεργασίας κοντά στα 24 msec για κάθε κείμενο στη ροή.

4.8.Συμπεράσματα

Σε αυτή την έρευνα ακολουθούμε μια διαφορετική πορεία για την κατηγοριοποίηση κειμένων από αυτές που υπάρχουν στην βιβλιογραφία. Οι γράφοι Ν-γραμμάτων είναι ένα μοντέλο αναπαράστασης που έχει χρησιμοποιηθεί σε άλλες τεχνικές μηχανικής μάθησης και ήταν μια πρόκληση να εφαρμοστεί για την ροή κειμένων που παράγονται με υψηλή ταχύτητα και ταξινομούνται σε πραγματικό χρόνο. Το μοντέλο κατηγοριοποίησης ΓΝΓ έχει την εγγενή ιδιότητα να μπορεί να εκτελεστεί σε μια σειρά ξεχωριστών εργασιών. Σχεδιάστηκε με το μοντέλο προγραμματισμού Beam και αναπτύχθηκε στο Google Cloud που διανέμεται μεταξύ πολλών εξυπηρετητών για να παρέχει υψηλή επεκτασιμότητα, ανθεκτικότητα, ανοχή σε σφάλματα και απόκριση σε πραγματικό χρόνο.

Πολλές παραλλαγές της μεθόδου κατηγοριοποίησης ΓΝΓ προτείνονται, υλοποιούνται και δοκιμάζονται πειραματικά για να καταλήξουμε σε μια βέλτιστη ρύθμιση. Οι σταθμισμένοι γράφοι, οι μη σταθμισμένα γράφοι, οι μέθοδοι κατηγοριοποίησης και οι πολλές μετρήσεις ομοιότητας γράφων χρησιμοποιούνται σε συνδυασμό με τεχνικές προεπεξεργασίας κειμένων. Το προτεινόμενο μοντέλο έχει χαρακτηριστικά που εκμεταλλεύονται τα προτερήματα των αναπαραστάσεων ΓΝΓ και μπορούν να δώσουν λύση σε δύσκολα θέματα ταξινόμησης κειμένων, όπως η ακολουθία λέξεων, και τα ορθογραφικά λάθη. Τα πειράματα διεξήχθησαν με τη αξιολόγησης δέκα διασταυρούμενων αναδιπλώσεων χρησιμοποιώντας τα σύνολα δεδομένων 20-Newsgroup και Reuters και προσεκτικά συγκρίθηκαν με άλλες μεθόδους που αναφέρονται στην διεθνή επιστημονική βιβλιογραφία που ακολουθούν παρόμοιες μετρήσεις αξιολόγησης και χρήσης των συνόλων δεδομένων.

Το μέγεθος του πλαισίου που χρησιμοποιήθηκε κατά τη διάρκεια των δοκιμών αξιολόγησης ήταν ίσο με το μέγεθος του βαθμού Ν αλλά αυτό δεν αποτελεί περιορισμό. Οι τεχνικές επιλογής χαρακτηριστικών μπορούν να χρησιμοποιηθούν για να εκτιμηθεί η σημασία των ακμών σύμφωνα με το γεγονός ότι μερικές ακμές συμβάλλουν περισσότερο στη σωστή ταξινόμηση από άλλες. Επιπλέον, το κριτήριο ομοιότητας θα πρέπει να επιλέγεται λαμβάνοντας υπόψη τη σημασία κατηγοριοποίησης των ακμών και οι ακμές θορύβου θα πρέπει να φιλτραριστούν από τους γράφους. Το μοντέλο αναπαράστασης ΓΝΓ με μετρήσεις ομοιότητας μπορεί να προσαρμοστεί και να συνδυαστεί με διάφορους αλγορίθμους ταξινόμησης προκειμένου να επιτευχθούν καλύτερα αποτελέσματα αξιολόγησης και να επιλυθούν πιο πολύπλοκα προβλήματα όπως η ιεραρχική ταξινόμηση πολλαπλών ετικετών.

5. Το Μοντέλο Συσταδοποίησης Κειμένων με Γράφους 3-γραμμάτων

Σε αυτή την ενότητα παρουσιάζουμε μια καινοτόμο μέθοδο συσταδοποίησης κειμένων χρησιμοποιώντας το μοντέλο αναπαράστασης γράφων 3-γραμμάτων και αναγάγουμε το πρόβλημα της ομαδοποίησης εγγράφων στο πρόβλημα του διαμερισμού γράφων. Για τον διαμερισμό των γράφων χρησιμοποιούμε τον αλγόριθμο k-μέσων (kernel k-means). Αξιολογούμε την προτεινόμενη μέθοδο χρησιμοποιώντας την συλλογή δεδομένων του Reuters-21578 και συγκρίνουμε την ακρίβεια προβλέψεων με αυτή που επιτυγχάνουμε με τον Αλγόριθμο Latent Dirichlet Allocation (LDA). Τα αποτελέσματα είναι ενθαρρυντικά φανερώνοντας ότι η μέθοδος γράφων 3-γραμμάτων παρουσιάζει πολύ καλά αποτελέσματα ανάκλησης και Φ-μέτρου αλλά ελαφρώς χειρότερη ακρίβεια.

5.1.Εισαγωγή στην Συσταδοποίηση Κειμένων με Γράφους 3-γραμμάτων

Ο όγκος κειμένων που είναι διαθέσιμοι στο διαδίκτυο καθιστά αναγκαία την ανάπτυξη μηχανισμών που διεξάγουν μια αυτόματη θεματική κατηγοριοποίηση. Οι εφαρμογές που μπορούν να την εφαρμόσουν είναι πολυάριθμες, ωστόσο το ίδιο το έργο της αυτόματης συσταδοποίησης είναι εξαιρετικά δύσκολο λόγω της ανάγκης να χρησιμοποιηθεί περιεχόμενο για την εξαγωγή - γνωστών έως τώρα κατηγοριών που συχνά καθορίζονται με βάση τα συμφραζόμενα. Αυτή η εργασία είναι σε μεγάλο βαθμό γνωστή ως "ομαδοποίηση κειμένων" και υπάρχουν αρκετές προσεγγίσεις για την αντιμετώπισή της. Σε αυτή την έρευνα, εστιάζουμε σε μια νέα προσέγγιση που αποφέρει ενθαρρυντικά αποτελέσματα και βασίζεται στον συνδυασμό ενός μοντέλου αναπαράστασης κειμένων και ενός αλγόριθμου ομαδοποίησης.

Το προτεινόμενο μοντέλο αναπαράστασης κειμένων είναι μια μέθοδος που χρησιμοποιείται στην Φυσική Επεξεργασία Γλώσσας (Natural Language Processing NLP) που ονομάζεται γράφοι Ν-γραμμάτων που εισήχθη αρχικά στην έρευνα [44]. Ένα γράφημα Ν-γραμμάτων μοντελοποιεί τη συχνότητα που δύο Ν-γράμματα (διαδοχικές ακολουθίες χαρακτήρων) εμφανίζονται σε γειτονικές θέσεις σε ένα κείμενο. Τα Ν-γράμματα υποδηλώνονται ως κόμβοι, ενώ η γειτνίαση υποδηλώνεται με ακμές, το βάρος των ακμών αντιστοιχεί στη συχνότητα που συναντάμε τις αντίστοιχες ακολουθίες γραμμάτων στα αρχικά κείμενα. Ο μετασχηματισμός ενός κειμένου σε γράφο είναι σε θέση να αποδώσει τις μορφές χρήσης της γλώσσας, οι οποίες συχνά περιλαμβάνουν πολύτιμες πληροφορίες συμφραζομένων. Ταυτόχρονα, η χρήση Ν-γραμμάτων αντί λέξεων ως βασικά στοιχεία του μοντέλου, επιτρέπει μια καλύτερη ανθεκτικότητα στα γραμματικά και συντακτικά λάθη, στους νεολογισμούς, στις συντομογραφίες, στην ταυτόχρονη χρήση περισσότερων από μια γλώσσας και σε άλλα παρόμοια φαινόμενα τα οποία συνηθίζονται σε κείμενα στον παγκόσμιο ιστό.

Ο προσδιορισμός των κοινών στοιχείων στα έγγραφα μπορεί τώρα να μετατοπιστεί από την σύγκριση του περιεχομένου των κείμενων στην σύγκριση γράφων με έναν τρόπο που αναδεικνύονται τα χαρακτηριστικά των κειμένων μέσω της αναπαράστασης με γράφους. Έτσι, το πρόβλημα της ομαδοποίησης κειμένων αναγάγετε στο πρόβλημα του διαμερισμού γράφων. Για το σκοπό αυτό χρησιμοποιούμε τον αλγόριθμο k-μέσων [89], ο οποίος διαμερίζει τον γράφο ενός κειμένου ή ενός συνόλου κειμένων σε k υπογράφους με στόχο να μεγιστοποιήσει το άθροισμα των ακμών βάρους μέσα στις συστάδες και να ελαχιστοποιήσει το άθροισμα των ακμών βάρους μεταξύ διαφορετικών συστάδων. Συγκρίνοντας το μοντέλο γράφων κάθε μεμονωμένου κειμένου με τους k-υπογράφους και υπολογίζοντας μια μετρική ομοιότητας, μπορούμε να κατηγοριοποιήσουμε τα κείμενα στην ομάδα που παρουσιάζεται η πιο υψηλή ομοιότητα.

Για να υποστηρίξουμε περαιτέρω τον ισχυρισμό ότι ο συνδυασμός γράφων Ν-γραμμάτων και των k-μέσων επιτυγχάνει μια αποτελεσματική συσταδοποίηση και συγκριτικά καλύτερα αποτελέσματα από τις τρέχουσες τεχνικές συσταδοποίησης κειμένων, εκτελούμε ένα σύνολο πειραμάτων με το σύνολο δεδομένων του

Reuters-21578, και παρουσιάζουμε τον τρόπο με τον οποίο ο προτεινόμενος μηχανισμός ξεπερνάει άλλες παρόμοιες επιστημονικές προσεγγίσεις.

Στη συνέχεια, το έγγραφο είναι δομημένο ως εξής: Οι τεχνολογίες πάνω στις οποίες στηρίχθηκε αυτή η έρευνα παρουσιάζονται στο Τμήμα 5.2. Στην Ενότητα 5.3 παρέχονται λεπτομέρειες για το μοντέλο αναπαράστασης και τον αλγόριθμο ομαδοποίησης. Η ενότητα 5.4 παρουσιάζει τα πειραματικά αποτελέσματα και τη σύγκριση με άλλες μεθόδους ομαδοποίησης εγγράφων. Τέλος, στο τμήμα 5.5, παραθέτουμε τα κύρια συμπεράσματα σχετικά με αυτή την έρευνα.

5.2. Διαμερισμός Γράφων και Συσταδοποίηση Κειμένων

Η μέθοδος μας βασίζεται στη χρήση γράφων N-γραμμάτων τα οποία μπορούν να αντιπροσωπεύουν οποιοδήποτε είδος κειμένου ως γράφο [90]. Διαφορετικές εκδοχές στις παραμέτρους που συνιστούν τους γράφους έχουν ως αποτέλεσμα διαφορετικά μοντέλα απεικόνισης γράφων. Μια διαφοροποίηση προκύπτει από τη χρήση μη σταθμισμένων γράφων στους οποία οι ακμές απεικονίζουν την γειτνίαση μεταξύ δύο N-γραμμάτων όπως εμφανίζεται στα κείμενα. Άλλες παραλλαγές μπορεί να υπάρχουν με τη χρήση κατευθυνόμενων γράφων, αν και - σε αυτή την έρευνα χρησιμοποιούμε μη-κατευθυνόμενους γράφους. Διαφορετικές παραλλαγές δημιουργούνται εάν ληφθεί υπόψη η παραγωγή N-γραμμάτων, χρησιμοποιώντας διαφορετικά παράθυρα ολίσθησης. Δεδομένου ότι ενδιαφερόμαστε για όλους τους πιθανούς συνδυασμούς N-γραμμάτων σε ένα κείμενο, ακολουθούμε την προσέγγιση, για κάθε κόμβο να δημιουργούμε N ακμές που τον ενώνουν με τους N κόμβους που ακολουθούν. Για παράδειγμα, εάν χρησιμοποιούμε γράφους με 3-γράμματα για κάθε κόμβο, δημιουργούμε τρεις ακμές που ενώνουν τον τρέχοντα κόμβο με τους επόμενους τρεις κόμβους. Επαναλαμβάνουμε αυτή τη διαδικασία από τους πρώτους χαρακτήρες κειμένου έως το τέλος.

Αφού έχουμε κατασκευάσει τους ξεχωριστούς ΓΝΓ για κάθε κείμενο και έναν κύριο γράφο για όλα τα κείμενα, χρησιμοποιούμε έναν αλγόριθμο διαμέρισης γράφων για τον προσδιορισμό των γράφων που αντιπροσωπεύουν τις συστάδες. Συνήθως τα προβλήματα διαχωρισμού γράφων είναι NP-δύσκολα. Γι 'αυτό χρησιμοποιούμε ευριστικούς αλγορίθμους. Είναι σημαντικό να δηλώνουμε ότι παρόλο που υπάρχουν πολλές επιλογές για ευριστικούς αλγορίθμους, όλοι δεν λύνουν αποτελεσματικά το πρόβλημα.

Ένας πολύ γνωστός αλγόριθμος διαμερισμού γράφων είναι ο αλγόριθμος Kernighan-Lin [56] ο οποίος ανταλλάσσει κόμβους μεταξύ των διαμερισμάτων χρησιμοποιώντας μια μετρική εσωτερικού και εξωτερικού κόστους. Οι Newman και Girman [53] υποδεικνύουν τη χρήση της betweeness μετρικής για την κατασκευή των διαμερισμάτων. Αυτός ο αλγόριθμος έχει ικανοποιητική απόδοση για σχετικά περιορισμένο αριθμό κόμβων, λιγότερων από 10000.

Ο αλγόριθμος K-μέσων (K-Means) [21] είναι ένας δημοφιλής αλγόριθμος που χωρίζει τις παρατηρήσεις σε k συστάδες. Σε αυτή τη μέθοδο, κάθε παρατήρηση ανήκει στη συστάδα με τον πλησιέστερο κέντρο. Οι παρατηρήσεις είναι διανύσματα d-διαστάσεων. Ο αλγόριθμος K-Μέσων έχει έναν περιορισμό: δεν μπορεί να χωρίσει αποτελεσματικά μη γραμμικές παρατηρήσεις. Συνεπώς, σε πολλές περιπτώσεις δεν είναι ο κατάλληλος για πραγματικά προβλήματα.

Μια επέκταση του αλγορίθμου K-Μέσων, ο οποίος αντιμετωπίζει αποτελεσματικά τον προηγούμενο περιορισμό γίνεται με την σταθμισμένη συσταδοποίηση που κάνει χρήση μιας απεικόνισης σε έναν διαφορετικό διανυσματικό χώρο (Kernel K-Means). Το πρόβλημα της συσταδοποίησης δεδομένων είναι μαθηματικά ισοδύναμο με το πρόβλημα διαχωρισμού γράφων [89]. Αρκετοί αλγόριθμοι για τον διαχωρισμό γράφων που χρησιμοποιούν μιαν αλλαγή στο διανυσματικό χώρο έχουν προταθεί. Επιλέγουμε να χρησιμοποιήσουμε έναν αλγόριθμο που κάνει χρήση της αλλαγής διανυσματικού χώρου ως την καταλληλότερη λύση για την κλίμακα και τις απαιτήσεις του προβλήματος που εξετάζουμε.

Για να εκτιμηθεί η ομοιότητα μεταξύ δύο γράφων N-γραμμάτων χρησιμοποιήσαμε το κριτήριο της ομοιότητας περιεχομένου [45] το οποίο εκφράζει την αναλογία των κοινών ακμών μεταξύ των γράφων που συγκρίνονται. Επιλέξαμε αυτή τη μετρική επειδή στο στάδιο της σύγκρισης γράφων χρησιμοποιούνται μη

σταθμισμένοι γράφοι και περιλαμβάνει λιγότερους υπολογισμούς από τις άλλες μετρικές. Ο αριθμός των συγκρίσεων είναι μεγάλος στις προτεινόμενες μεθόδους και θέτουν ζήτημα αποδοτικότητας.

Πολλές προσεγγίσεις έχουν προταθεί από τομείς όπως η Ανάκτηση Πληροφοριών και η Εξόρυξη Δεδομένων για την αυτόματη κατηγοριοποίηση κειμένων, με τις πιο γνωστές και αποδοτικές να είναι η PLSA [24] και η LDA [28]. Η PLSA βασίζεται στην συνύπαρξη λέξεων στο ίδιο έγγραφο και με τη χρήση στατιστικών τεχνικών δημιουργούν ορισμένες λανθάνουσες (latent) μεταβλητές που αντιπροσωπεύουν τις παρατηρούμενες λέξεις. Η LDA είναι μια πιο ακριβής μέθοδος από την PLSA. Η LDA χρησιμοποιεί τη κατανομή Dirichlet και παράγει την πιθανότητα κάθε κείμενο να ανήκει σε κάθε θεματική κατηγορία. Στο τέλος της αξιολόγησής μας χρησιμοποιήσαμε επίσης τη μέθοδο ομαδοποίησης LDA με το ίδιο σύνολο δεδομένων για να συγκρίνουμε τα αποτελέσματα της προτεινόμενης μεθόδους.

Υπάρχουν συγκεκριμένες μετρήσεις αξιολόγησης της συσταδοποίησης [14], όπως οι Precision BCubed, Recall BCubed και F-μετρική που χρησιμοποιούνται για την εκτίμηση της ποιότητας των αλγορίθμων ομαδοποίησης κειμένου. Αυτές οι μέθοδοι καθιστούν δυνατή την σύγκριση μεταξύ διαφορετικών μεθόδων και αναδεικνύουν τα χαρακτηριστικά τους.

5.3.Προτεινόμενο Μοντέλο

5.3.1.Μοντέλο αναπαράστασης κειμένων

Ένα N-γραμμα είναι μια συνεχής ακολουθία N στοιχείων όπως βρίσκονται σε ένα κείμενο. Τα στοιχεία μπορούν να είναι φωνήματα, συλλαβές, γράμματα ή λέξεις και κάθε φορά επιλέγονται με βάση την εφαρμογή. Ένα N-γραμμα μεγέθους 1 αναφέρεται ως μονόγραμμα (unigram), μεγέθους 2 δίγραμμα (bigram), μεγέθους 3 τρίγραμμα (threegram) κ.ο.κ. Σε αυτό το μοντέλο χρησιμοποιήσαμε τις ακολουθίες τριών γραμμάτων που μπορούν να εξαχθούν από οποιοδήποτε σύνολο κειμένων που θέλουμε να συσταδοποιήσουμε. Χρησιμοποιούμε τα τριγράμματα ως τα βασικά συστατικά του μοντέλου για να φτιάξουμε τους γράφους τριγραμμάτων.

Ένας γράφος N-γραμμάτων G = $\{V^G, E^G\}$ έχει N-γράμματα ως κόμβους $v^G \in V^G$ και ακμές $e^G \in E^G$ που συνδέουν τα αντίστοιχα N-γράμματα και δηλώνουν την γειτνίαση τους όπως υπάρχει στα αντίστοιχα κείμενα. Η συχνότητα που συνδέονται τα N-γράμματα μέσω μια ακμής και βρίσκονται κοντά στο αρχικό κείμενο υποδηλώνεται από μια τιμή που είναι το βάρος των ακμών. Ο επίσημος ορισμός είναι ο ακόλουθος.

Ορισμός 5.1. Αν S = $\{S_1, S_2, \ldots\}$, $S_k \neq S_l$, για k ≠ l , k, l ∈ ℕ είναι το σύνολο των διακριτών N-γραμμάτων που εξάγονται από ένα κείμενο T^l , και S_l είναι το ιοστό N-γραμμα, τότε ο G είναι ένας γράφος όπου υπάρχει μια αμφιμονοσήμαντη σχέση f: S → V.

Μετά τον σχηματισμό των κόμβων N-γράμματων, σχηματίζονται οι ακμές σύμφωνα με ένα κριτήριο γειτνίασης. Κάθε κόμβος N-γράμματος συνδέεται με τους ακόλουθους κόμβους N-γραμμάτων σύμφωνα με ένα παράθυρο D_{win}.

Στις ακμές E αποδίδονται βάρη $c_{i,j}$ που υποδηλώνουν την συχνότητα που ένα ζεύγος S_i, S_j N-γραμμάτων είναι γειτονικά σε κάποια απόσταση D_{win} (η γειτονικότητα υπολογίζεται με βάση έναν αριθμό χαρακτήρων). Δεδομένου ότι πιθανώς δεν έχουν σημασία αποστάσεις μεγαλύτερες από έναν συγκεκριμένο μέγεθος, λαμβάνουμε υπόψη ένα παράθυρο γύρω από το S_i στο αρχικό κείμενο και καθορίζουμε ποια S_j θεωρούνται γειτονικά. Οι κορυφές v_i, v_j που αντιστοιχούν στα N-γράμματα S_i, S_j που βρίσκονται μέσα στην απόσταση D_{win} συνδέονται την αντίστοιχη ακμή e ≡ $\{v_i , v_j\}$.

Στην έρευνα μας με την μέθοδο αναπαράστασης ΓΝΓ, χρησιμοποιήσαμε ως μήκος N των N-γραμμάτων και μήκος παραθύρου D_{win} ίσο με 3.

5.3.2.Αλγόριθμος διαμέρισης

Η διαδικασία της διαμέρισης ξεκινά με την μετατροπή ενός συνόλου κειμένων στους αντίστοιχους γράφους 3-γραμμάτων. Στη συνέχεια δημιουργείται ένα νέος γράφος με τη συγχώνευση του αρχικού συνόλου των γράφων των επιμέρους κειμένων. Π.χ. για ένα σύνολο 100 εγγράφων, θα δημιουργηθούν σύνολο 101 ΓΝΓ: Οι 100 ΓΝΓ θα αντιστοιχούν σε κάθε ένα από τα κείμενα και ένας ακόμη γράφος που αντιστοιχεί στο άθροισμα των προηγούμενων εκατό γράφων κειμένων.

Κάθε κόμβος αντιπροσωπεύει ένα 3-γραμμα όπως βρίσκεται στο αρχικό κείμενο και κάθε κατευθυνόμενη ακμή αντιπροσωπεύει την διάταξη μεταξύ των κόμβων και την γειτονικότητα τους σε ένα πλαίσιο τριών χαρακτήρων. Σε αυτό το στάδιο της μεθόδου ο γράφος κάθε εγγράφου είναι κατευθυνόμενος και μη σταθμισμένος. Δηλαδή, δεν έχει σημασία πόσες φορές δύο 3-γράμματα βρίσκονται κοντά σε ένα κείμενο. Ο λόγος αυτής της επιλογής είναι ότι όλα τα κείμενα συμβάλλουν εξίσου στην διαδικασία της συσταδοποίησης ανεξάρτητα από το μέγεθος τους. Αν χρησιμοποιούσαμε σταθμισμένους γράφους το αποτέλεσμα θα επηρεαζόταν από το μήκος του κειμένου (όπου τα μεγαλύτερα κείμενα πιθανότατα θα έχουν μεγαλύτερα βάρη) και θα οδηγούμασταν σε μια άνιση επιρροή κάθε κειμένου που εξαρτάται από το μέγεθος του.

Από την άλλη πλευρά, ο γράφος 3-γραμμάτων που συνδυάζει όλα τα κείμενα είναι σταθμισμένος, επειδή θα χρειαστούν τα βάρη των ακμών για να τον διαμερισμό του όπως εξηγείται στις επόμενες παραγράφους.

Έχοντας ένα γράφημα 3-γραμμάτων που αντιπροσωπεύει το σύνολο όλων των κειμένων, ο αλγόριθμος k-μέσων το χωρίζει σε k διαμερίσματα. Κάθε διαμέρισμα έχει ορισμένους κύριους κόμβους, καθώς και όλες τις κατευθυνόμενες γειτονικές ακμές των κύριων κόμβων και τους τελικούς κόμβους αυτών των ακμών. Εάν ένα διαμέρισμα έχει τους κύριους κόμβους {A, B, C} θα περιλαμβάνει επίσης όλες τις ακμές που έχουν ως κόμβο έναν από τους κόμβους A, B και C οπότε θα περιλαμβάνονται και οι ακμές {(A,B), (A,D), (A,E), (A,C)} καθώς και οι κόμβοι D και E.

Στόχος μας είναι να συσταδοποιήσουμε μεγάλους όγκους κειμένων με βάση τα θέματά τους. Ένας αλγόριθμος διαμέρισης γράφων χρησιμοποιείται ως συστατικό της μεθόδου που προτείνουμε. Χρησιμοποιούμε τον αλγόριθμο διαμέρισης γράφων για να κατασκευάσουμε το γράφημα Ν-γραμμάτων το οποίο αντιστοιχεί και αντιπροσωπεύει κάθε συστάδα. Αυτά οι γράφοι συστάδων Ν-γραμμάτων συγκρίνονται με τους γράφους Ν-γραμμάτων των κειμένων προκειμένου να κατηγοριοποιηθούν τα έγγραφα στην σωστή θεματική συστάδα.

Για να χωρίσετε τον γράφο 3-γραμμάτων που αντιπροσωπεύει το σύνολο όλων των κειμένων χρησιμοποιήσαμε τον αλγόριθμο βάσης πυρήνα k-Μέσων όπως δίνεται στον αλγόριθμο 5.1

Αλγόριθμος 5.1 Συσταδοποίηση γράφων με χρήση πυρήνα

Weighted Kernel k-means(K, k, w, t_{max}, $\{\pi_c^{(0)}\}_{c=1}^{k}$, $\{\pi_c^{(0)}\}_{c=1}^{k}$)

Είσοδος: K: kernel matrix, k: ο αριθμός των συστάδων, w: τα βάρη για κάθε σημείο, t_{max}: ο μέγιστος αριθμός επαναλήψεων, $\{\pi_c^{(0)}\}_{c=1}^{k}$: η αρχική συσταδοποίηση

Έξοδος: $\{\pi_c\}_{c=1}^{k}$: Η τελική συσταδοποίηση των σημείων

1. Αν δεν υπάρχει αρχική συσταδοποίηση, αρχικοποιούμε τις k συστάδες $\pi_1^{(0)}, \ldots, \pi_k^{(0)}$ τυχαία. Θέτουμε t = 0.

2. Για κάθε i γραμμή του K και κάθε συστάδα c, υπολογίζουμε την απόσταση

$$d(i, m_c) = K_{ii} - \frac{2 \cdot \sum_{j \in \pi_c^{(t)}} w_j \cdot K_{ij}}{\sum_{j \in \pi_c^{(t)}} w_j} + \frac{\sum_{j \in \pi_c^{(t)}} w_j \cdot w_l \cdot K_{ij}}{(\sum_{j \in \pi_c^{(t)}} w_j)^2}$$

3. Βρίσκουμε τα νέα κέντρα των συστάδων $c^*(i) = argmin_c \, d(i, m_c)$. Υπολογίζουμε τις νέες συστάδες

$$\pi_c^{t+1} = \{i : c^*(i) = c\}$$

4. Αν δεν υπάρχει σύγκληση ή t_{max}, > t, θέτουμε t = t + 1 και πηγαίνουμε στο βήμα 3; Διαφορετικά, σταματάμε και εξάγουμε τις τελικές συστάδες $\{\pi_c^{(t+1)}\}_{c=1}^{k}$

Ο πίνακας πυρήνα (kernel matrix) υπολογίζεται χρησιμοποιώντας ένα από τα ακόλουθα κριτήρια του πίνακα 1.

Κριτήρια	Βάρη των Ακμών	Πίνακας πυρήνα
Ratio Association	1 ∀ ακμή	K = σI + A
Ratio Cut	1 ∀ ακμή	K = σI -D + A
Normalized Cut	βαθμός κάθε ακμής	K=σD^{-1} +$D^{-1}A\,D^{-1}$ A

Πίνακας 5.1: Κριτήρια για διαμέριση γράφων

Ο πίνακας συγγένειας (Affinity matrix) A είναι ένας πίνακας διαστάσεων | V | × | V | του οποίου τα κελιά αντιπροσωπεύουν τα βάρη των ακμών ή 0 αν δεν υπάρχει ακμή μεταξύ των αντίστοιχων κορυφών.

Ο αλγόριθμος 5.1 υπολογίζει για κάθε κόμβο i και κάθε κέντρο συστάδας m_c το κόστος $d(i, m_c)$ και συσταδοποιεί τον κόμβο i στην συστάδα που παράγει το ελάχιστο κόστος. Τα βήματα 1 έως 4 του αλγορίθμου επαναλαμβάνονται έως ότου κανένας κόμβος να μην αλλάξει συστάδα ή για ένα συγκεκριμένο αριθμό επαναλήψεων t_{max}.

Στο τέλος έχουμε k ΓΝΓ που αντιπροσωπεύουν τις k συστάδες. Από τώρα και στο εξής, τα βάρη των γράφων δεν θα χρησιμοποιηθούν και για κάθε συστάδα θα υπάρχει ένας μη σταθμισμένος γράφος 3-γραμμάτων που θα συγκρίνεται με το μη σταθμισμένο γράφο 3-γραμμάτων κάθε κειμένου.

Για τη σύγκριση γράφων χρησιμοποιήσαμε την ομοιότητα περιεχομένου (5.1), η οποία εκφράζει την αναλογία των ακμών του γράφου G_{ti} που είναι κοινοί με τον γράφο G_{Tp}. Υποθέτοντας ότι G είναι ένας γράφος N-γραμμάτων, το e είναι μια ακμή γράφου και ότι για τη συνάρτηση μ (e, G) δηλώνει ότι μ (e, G) = 1, αν και μόνο αν e ∈ G και 0 διαφορετικά, τότε:

$$CS(Gti, GTp) = \sum_{e \in G_{t_i}} \frac{\mu(e, GTp)}{\min(|G_{t_i}|, |G_{t_p}|)} \quad (5.1)$$

Όπου | G | υποδηλώνει το μέγεθος του γράφου N-γραμμάτων G και στο μοντέλο μας είναι ο αριθμός των ακμών που περιέχει ο γράφος. Το G_{t_i} δηλώνει τον γράφο που αντιπροσωπεύει κάθε συστάδα και το G_{t_p} αντιπροσωπεύει τον γράφο κάθε κειμένου.

Η ομοιότητα περιεχομένου εκφράζει το ποσοστό που ένα κείμενο ανήκει σε κάθε συστάδα και μπορούμε να εκφράσουμε μια ποσοστιαία σχέση κάθε έγγραφου με κάθε συστάδα. Αν θέλουμε ένα κείμενο να ανήκει σε μία μόνο θεματική κατηγορία τότε επιλέγουμε την συστάδα με την οποία έχει την υψηλότερη ομοιότητα περιεχομένου. Σε περίπτωση που θέλουμε ένα κείμενο να ανήκει σε m αριθμό θεματικές κατηγορίες, επιλέγουμε τις m συστάδες που έχουν την υψηλότερη ομοιότητα περιεχομένου.

5.4. Πειραματική Αξιολόγηση

Χρησιμοποιήσαμε τη Συλλογή κειμένων Reuters-21578 ως το σύνολο δεδομένων αξιολόγησης για να εξετάσουμε την απόδοση του μοντέλου μας. Τα κείμενα συγκεντρώθηκαν και ταξινομήθηκαν με κατηγορίες από προσωπικό από το Reuters Ltd. Η συλλογή αυτή περιλαμβάνει πολλά έγγραφα που ανήκουν από μηδέν έως είκοσι εννέα κατηγορίες. Η εκκαθάριση του συνόλου δεδομένων είχε ως αποτέλεσμα την απόρριψη κειμένων που δεν ανήκουν σε καμία κατηγορία καθώς και πολλά έγγραφα τα οποία η ερμηνεία τους ήταν ιδιαίτερα δύσκολη. Συνολικά χρησιμοποιήθηκαν 18457 έγγραφα που ανήκουν σε 428 κατηγορίες.

Ορισμένες κατηγορίες ήταν κενές, έτσι δεν χρησιμοποιήθηκαν. Το μέγεθος των υπόλοιπων κατηγοριών ήταν αρκετά κυμαινόμενο με κάποιες κατηγορίες να περιέχουν λίγα και κάποιες αρκετά κείμενα. Το σετ κειμένων είχε μια ποικιλία μικρών, μεσαίων και μεγάλων εγγράφων και όλα τα κείμενα είναι γραμμένα στην αγγλική γλώσσα.

Η πλήρης μέθοδος υλοποιήθηκε χρησιμοποιώντας την Java SE. Το πρόγραμμα έλαβε το σύνολο δεδομένων Reuters ως είσοδο και σχημάτισε τον γράφο 3-γραμμάτων όλων των κειμένων. Στη συνέχεια διαμερίσαμε τον γράφο σε 428 διαμερίσεις. Κάθε μια από αυτές είχε μερικούς κόμβους και όλες τις γειτονικές ακμές τους. Πολύ λίγες διαμερίσεις ήταν κενές και δεν είχαν γράφο 3-γραμμάτων για να τις αντιπροσωπεύει. Αυτό το ζήτημα επηρέασε ελαφρώς τα αποτελέσματα ειδικά τις κατηγορίες που έχουν πολύ λίγα κείμενα. Στη συνέχεια δημιουργήθηκαν οι γράφοι 3-γραμμάτων για κάθε κείμενο και κάθε ένα συγκρίθηκε με τους 428 γράφους 3-γραμμάτων που αντιπροσωπεύουν τις διαμερίσεις. Μια μερική ιδιότητα μέλους (soft membership) θεωρήθηκε για κάθε κείμενο σε σχέση με όλες τις ομάδες, δηλαδή εκτιμήθηκε η πιθανότητα για κάθε κείμενο να ανήκει σε κάθε μία από της θεματικές κατηγορίες. Ένα κείμενο συμπεριλήφθηκε σε μια κατηγορία εάν ο γράφος του παρουσιάζει μεγάλη ομοιότητα με τον γράφο ενός διαμερίσματος / συστάδας.

Στο ίδιο σύνολο δεδομένων διεξήχθησαν πειράματα, αυτή τη φορά χρησιμοποιώντας την μέθοδο συσταδοποίησης LDA με σκοπό να συγκρίνουμε τα αποτελέσματα και να έχουμε ένα μέτρο σύγκρισης και αξιολόγησης του προτεινόμενου αλγορίθμου μας. Το LDA είναι μία από τις πιο δημοφιλείς μεθόδους

κατηγοριοποίησης κειμένων. Για την εφαρμογή του LDA χρησιμοποιήσαμε το εργαλείο Machine Learning for Language Toolkit (Mallet) το οποίο είναι ένα πακέτο γραμμένο σε Java και προσφέρει την LDA μέθοδο.

Όπως έχουμε ήδη αναφέρει σε προηγούμενη ενότητα, για τη σύγκριση των δύο προσεγγίσεων συσταδοποίησης χρησιμοποιήσαμε τρεις μετρήσεις που χρησιμοποιούνται για να εκτιμηθεί η ακρίβεια μιας μεθόδου συσταδοποίησης, η Precision, Recall και F-Measure. Στην αξιολόγηση της μεθόδου χρησιμοποιήσαμε μέτρα ανάκτησης / κατηγοριοποίησης επειδή στόχος μας είναι να αξιολογήσουμε την συσταδοποίηση κειμένων σε σχέση με τις ομάδες θεμάτων. Οι μετρήσεις ακρίβειας δίνονται ξανά συνοπτικά παρακάτω όπως τις εφαρμόσαμε για τις ανάγκες της έρευνας μας.

$$Avg_e[Avg_{e'.C(e)\cap C(e')\neq 0}[Multiplicity\ precision(e,e')]] \quad (5.2)$$

$$\text{Recall} = Avg_e[Avg_{e'.L(e)\cap L(e')\neq 0}[Multiplicity\ recall(e,e')]] \quad (5.3)$$

$$\text{Multiplicity Precision}(e,e') = \frac{Min(|C(e)\cap C(e')|, |L(e)\cap L(e')|)}{|C(e)\cap C(e')|} \quad (5.4)$$

$$\text{Multiplicity Recall}(e,e') = \frac{Min(|C(e)\cap C(e')|, |L(e)\cap L(e')|)}{|L(e)\cap L(e')|} \quad (5.5)$$

$$\text{F-measure} = \frac{2\cdot Recall(L,C)\cdot Precision(L,C)}{Recall(L,C)+Precision(L,C)} \quad (5.6)$$

όπου e και e΄ είναι δύο έγγραφα, L (e) δηλώνει το σύνολο κατηγοριών (που προέρχονται από το σύνολο δεδομένων) και C(e) το σύνολο συστάδων που σχετίζονται με e. Η ακρίβεια (Precision) του στοιχείου αντιπροσωπεύει πόσα αντικείμενα στην ίδια συστάδα ανήκουν στην κατηγορία του. Η ανάκληση (Recall) για ένα στοιχείο αντιπροσωπεύει πόσα αντικείμενα από την κατηγορία του εμφανίζονται στην συστάδα του. Το μέτρο F είναι μια μέτρηση που συνδυάζει την ακρίβεια και την ανάκληση.

5.5.Αποτελέσματα

Τα αποτελέσματα έδειξαν ότι η μέθοδος γράφων 3-γραμμάτων για την εφαρμογή της συσταδοποίησης παρουσιάζει καλύτερα αποτελέσματα στην μετρική Recall, αλλά χειρότερα στην μετρική Precision. Ο λόγος είναι ότι οι γράφοι 3-γραμμάτων αναγνωρίζουν τις συστάδες που αποτελούνται από μεγάλο πλήθος κειμένων και συσταδοποιούνται σωστά πολλά από τα κείμενα σε αυτές. Από την άλλη πλευρά, η LDA δημιουργεί πολλές μικρές συστάδες που αποτελούνται από μικρό πλήθος κειμένων και δεν δίνει την απαραίτητη σημασία στις μεγάλες συστάδες. Ως παράδειγμα, μια συστάδα που δημιουργήθηκε από την μέθοδο LDA περιέχει τέσσερα κείμενα, δύο από αυτά είναι σωστά συσταδοποιημένα μαζί έτσι ώστε η ακρίβεια να είναι 2/4 = 50%, αλλά η αντίστοιχη συστάδα όπως υπάρχει στο σύνολο δεδομένων αποτελείτο από 100 κείμενα οπότε η ανάκληση είναι 2/100 = 2% Ο Πίνακας 5.2 συνοψίζει τα αποτελέσματα της σύγκρισης μεταξύ της μεθόδου συσταδοποίησης που χρησιμοποιεί τους γράφους 3-γραμμάτων και της μεθόδου LDA χρησιμοποιώντας τις εξισώσεις (5.1) έως (5.6)

	Precision	Recall	F-measure
3-Gram graph	0.2870524	0.2045877	0.2419
LDA	0.5757706	0.025551	0.0498

Πίνακας 5.2: Αξιολόγηση πειραματικών αποτελεσμάτων

Στις περισσότερες περιπτώσεις όπου τα κείμενα συσταδοποιήθικαν εσφαλμένα με την μέθοδο που προτείνουμε παρατηρήσαμε ότι η σωστή συστάδα βρισκόταν στην δεύτερη ή στην τρίτη επιλογή. Από αυτό συμπεράνουμε ότι η μέθοδος που εφαρμόζει τον γράφο 3-γραμμάτων μπορεί να αναγνωρίσει τις περισσότερες φορές την σωστή συστάδα που ανήκει ένα κείμενο μέσα στις τρεις πρώτες επιλογές. Από την άλλη πλευρά η μέθοδος LDA δημιούργησε ένα πλήθος από μικρές συστάδες οι οποίες είχαν μια μορφή που έμοιαζε σαν σπασμένα τμήματα των πραγματικών συστάδων.

Σε σχέση με την εφαρμογή μπορεί να είναι σημαντικότερη η ακρίβεια ή η ανάκληση. Μια υψηλή ακρίβεια μπορεί να μας προσφέρει καθαρές συστάδες που έχουν μόνο τα σωστά συγκεντρωμένα κείμενα. Είναι πολύ θετικό για την προτεινόμενη μέθοδο ότι η ακρίβεια και η ανάκληση είναι κοντά και ότι η F-μέτρηση της είναι πολύ μεγαλύτερη από αυτήν της LDA. Η μέθοδος LDA ήταν σε θέση να εντοπίσει τα κείμενα που είχαν πολλές κοινές λέξεις, έτσι ώστε να τα συσταδοποιήσει μαζί, αλλά όπως διαπιστώσαμε πειραματικά αυτό δεν ήταν αρκετό για να έχουμε υψηλό βαθμό ανάκλησης. Η συσταδοποιήση με την χρήση γράφων 3-γραμμάτων χρησιμοποιεί ένα κριτήριο ομοιότητας γράφων για κάθε συστάδα και κάθε κείμενο και κατηγορία να έχουν ένα αντιπροσωπευτικό γράφο. Αν οι γράφοι θεματικών συστάδων και εγγράφων παρουσιάζουν υψηλή ομοιότητα τότε αναμένουμε ότι τα κείμενα που συσταδοποιούνται μαζί θα έχουν υψηλή θεματική ομοιότητα. Τα πειραματικά αποτελέσματα επιβεβαίωσαν την εφαρμοσιμότητα του κριτηρίου ομοιότητας των γράφων 3-γραμμάτων.

Εκτελέσαμε το πρόγραμμα java σε έναν υπολογιστή γενικής χρήσης. Πραγματοποιήσαμε 400 επαναλήψεις του αλγόριθμου ομαδοποίησης K-Means (3.1) και παρατηρήσαμε ότι σύγκλινε με γρήγορο ρυθμό. Υπάρχουν πολλές εκδοχές σχετικά με τον αρχικό διαμερισμό των γράφων 3-γραμμάτων οι οποίοι μπορούν να επιταχύνουν τη σύγκλιση. Εμείς χρησιμοποιήσαμε την coarsening method of A Fast Kernel based Multilevel Algorithm for graph Clustering [21].

5.6.Συμπεράσματα

Η συσταδοποίηση κειμένων χρησιμοποιώντας το μοντέλο αναπαράστασης γράφων 3-γραμμάτων είναι μια καινοτόμος μέθοδος που μπορεί να συσταδοποιήσει αποτελεσματικά χιλιάδες κείμενα και να εκτιμήσει τον βαθμό στον οποίο κάθε κείμενο ανήκει σε κάθε συστάδα. Η μέθοδος είναι ανεξάρτητη από τη γλώσσα και μπορεί να εφαρμοστεί σε προβλήματα ανάκτησης πληροφορίας και εξόρυξης κειμένων. Επίσης παρουσιάζει καλύτερα αποτελέσματα από άλλες δημοφιλείς μεθόδους συσταδοποίησης κειμένων σύμφωνα με τα πειράματά που διεξήγαμε.

6. Εφαρμογή στην Αναγνώριση Κοινοτήτων σε Μέσα Κοινωνικών Δικτύων

Τα μέσα Κοινωνικών Δικτύων είναι ένας συνεχώς αναπτυσσόμενος ερευνητικός τομέας. Έχει παρατηρηθεί ότι οι χρήστες των κοινωνικών δικτύων συνιστούν ομάδες οι οποίες μπορούν να ανιχνευθούν από την τοπολογική τους σχέση μέσα στον κοινωνικό γράφο και από το περιεχόμενο των δεδομένων, συχνά κειμένων, με τα οποία σχετίζονται.

Η ανίχνευση θεματικών κοινοτήτων (Topic communities Detection TCD) μπορεί να πραγματοποιηθεί με βάση τις αναρτήσεις των χρηστών των κοινωνικών δικτύων (Social Network SN). Ένας αλγόριθμος κατηγοριοποίησης κειμένων μπορεί να αναλύσει και να παράγει αναπαραστάσεις κειμένων που γράφτηκαν από τους χρήστες, και να τις συγκρίνει με τις αναπαραστάσεις των κοινοτήτων. Συγκρίνοντας τις αναπαραστάσεις κειμένων των χρηστών με αυτές των κοινοτήτων μπορούμε να έχουμε την πρόβλεψη της κοινότητας στην οποία ανήκει κάθε χρήστης.

Μια διαφορετική προσέγγιση του TCD εκμεταλλεύεται την κοινωνική αλληλεπίδραση των χρηστών. Η κοινωνική δράση της σχολιασμού, της προσθήκης ετικετών και της αρεσκείας μεταξύ δύο χρηστών δηλώνει ότι οι δύο χρήστες αλληλεπιδρούν και σχετίζονται. Το σύνολο των χρηστών που έχουν μια πυκνή αλληλεπίδραση διαμορφώνει τις κοινότητες. Ένας αλγόριθμος συσχετισμού μπορεί να χρησιμοποιηθεί για να εκτιμηθούν οι ομοιότητες αλληλεπιδράσεων μεταξύ των χρηστών και μετά η κοινότητα θα διαμορφωθεί χρησιμοποιώντας έναν αλγόριθμο ο οποίος μεγιστοποιεί την ομοιότητα μεταξύ των συστάδων των χρηστών.

Παραθέτουμε μια σειρά αλγορίθμων που ανιχνεύουν τις κοινότητες αυτές και με τους δύο τρόπους αντίστοιχα. Επίσης οι κοινότητες που συνιστούν οι χρήστες εξελίσσονται με την πάροδο του χρόνου είτε με το να διαχωρίζονται σε επιμέρους υπό κοινότητες είτε στο να ενώνονται σχηματίζοντας μεγαλύτερες. Περιγράφουμε μια μοντελοποίηση της χρονικής τους αυτής εξέλιξης. Μελετάμε με συστηματικό τρόπο ποιοι χρήστες είναι αυτοί που επηρεάζουν περισσότερο την κοινότητα μέσα στην οποία βρίσκονται. Τέλος παρουσιάζουμε την εφαρμογή και αξιολόγηση του μοντέλου κατηγοριοποίησης κειμένων που κάνει χρήση της αναπαράστασης ΓΝΓ.

6.1.Εισαγωγή στην Αναγνώριση Κοινοτήτων σε Κοινωνικά Δίκτυα

Η μοντελοποίηση δικτύων είναι ένα θέμα που έχει βρει εφαρμογές σε πολλά επιστημονικά και τεχνολογικά πεδία όπως αυτά των εφαρμοσμένων μαθηματικών, της στατιστικής φυσικής (statistical physics) [56], της επιστήμης των υπολογιστών, της αναγνώριση εικόνων [91], τις κοινωνικές επιστήμες [92], της βιολογίας [93] και των κοινωνικών δικτύων [94]. Στην παρούσα ενότητα θα τη μελετήσουμε από την πλευρά των κοινωνικών δικτύων.

Η πιο συνηθισμένη αναπαράσταση των κοινωνικών δικτύων γίνεται με την βοήθεια κοινωνικών γράφων (social graph) όπου κάθε κόμβος μπορεί να αντιστοιχεί συνήθως σε έναν χρήστη ή ένα αντικείμενο. Πιθανά αντικείμενα μπορεί να είναι μια φωτογραφία, ένας σύνδεσμος (url), ένα κείμενο, μια διαφήμιση για ένα καταναλωτικό προϊόν. Συνήθως συναντάμε δίκτυα όπου οι κόμβοι τους είναι μόνο χρήστες ή μόνο αντικείμενα και σπάνια συνδυασμούς τους. Στην μελέτη μας αυτή, οι κόμβοι των κοινωνικών δικτύων θα αναπαριστούν χρήστες αλλιώς θα αναφέρεται.

Κάθε ακμή του δικτύου υποδηλώνει μια συσχέτιση μεταξύ δύο κόμβων. Αν έχουμε ως κόμβους χρήστες οι ακμές μπορούν να υποδηλώνουν φιλίες μεταξύ χρηστών ή ότι εκδηλώνουν συχνά ενδιαφέρον για αντικείμενα που βρίσκονται στην ίδια κατηγορία. Αυτό σημαίνει ότι αν δύο χρήστες προσπελαύνουν συχνά τα ίδια αντικείμενα μια ακμή μπορεί να σύνδεση τους αντίστοιχους κόμβους τους και αν έχουμε γράφο με

βάρη στις ακμές τους τότε η ακμή θα έχει ένα βάρος που θα αντιστοιχεί στο πόσο συχνά αυτοί οι χρήστες ενδιαφέρονται για τα ίδια αντικείμενα.

Μετά την αναπαράσταση του κοινωνικού δικτύου με την βοήθεια ενός γράφου ένα πρόβλημα που προκύπτει είναι να βρεθούν ομαδοποιήσεις κόμβων που έχουν κοινά χαρακτηριστικά. Αυτό το ερώτημα μπορούμε να το διατυπώσουμε με το να κάνουμε μια διαμέριση του γράφου ή αλλιώς μια συσταδοποίηση των κόμβων (χρηστών ή αντικειμένων) με τέτοιο τρόπο, ώστε κόμβοι που συσχετίζονται σε μεγαλύτερο βαθμό αναμεταξύ τους να βρίσκονται στην ίδια συστάδα και κάθε συστάδα να συνδέεται, μέσω ακμών, όσο το δυνατό λιγότερο με τις υπόλοιπες συστάδες του δικτύου. Επίσης αναμένουμε οι ακμές που συνδέουν εσωτερικά τους κόμβους μιας συστάδας να είναι κατά πολύ πυκνότερες από τις υπόλοιπες.

Πιο συγκεκριμένα αρχικά παρουσιάζουμε δημοφιλείς τεχνικές ανεύρεσης κοινοτήτων χρηστών που τους απασχολούν κοινά θέματα μέσω των γράφων κοινωνικών δικτύων και το σημασιολογικό περιεχόμενο των δεδομένων που γράφουν ή διαβάζουν. Έπειτα παρουσιάζουμε πως το μοντέλο κατηγοριοποίησης κειμένων με ΓΝΓ έχει εφαρμοστεί και αξιολογηθεί για την αναγνώριση κοινοτήτων με βάση τα κείμενα που αναρτούν.

Εν συντομία τα περιεχόμενα αυτής της ενότητας είναι ως εξής. Στο τμήμα 6.2 αναπτύσσουμε αλγοριθμικές μεθόδους που βρίσκουν της κοινότητες με βάση τα τοπολογικά χαρακτηριστικά του ενός κοινωνικού γράφου. Στο τμήμα 6.3 Η ανεύρεση των κοινοτήτων γίνεται με βάση τη σημασιολογική πληροφορία των δεδομένων με τα οποία σχετίζεται κάθε χρήστης. Στο τμήμα 6.4 αναγνωρίζουμε τις κοινότητες όχι ως στατικές αλλά ως δυναμικά εξελισσόμενες όπως είναι στον πραγματικό κόσμο και με συστηματικό τρόπο ανιχνεύουμε την χρονική τους εξέλιξη. Στο τμήμα 6.5 μας απασχολεί το θέμα της εύρεσης των χρηστών που ασκούν περισσότερη επιρροή στην κοινότητα που βρίσκονται. Στο τμήμα 6.6 παρουσιάζουμε την ανεύρεση θεματικών κοινοτήτων με βάση το προτεινόμενο μοντέλο κατηγοριοποίησης κειμένων που κάνει χρήση των ΓΝΓ. Τέλος στο τμήμα 6.7 συγκρίνουμε τις προηγούμενες μεθόδους σε ποιες περιπτώσεις ταιριάζει καλύτερα να χρησιμοποιήσουμε την κάθε μια και εξάγουμε τα τελικά συμπεράσματα μας.

6.2.Αναγνώριση Κοινοτήτων με Τοπολογικά Χαρακτηριστικά του Κοινωνικού Γράφου

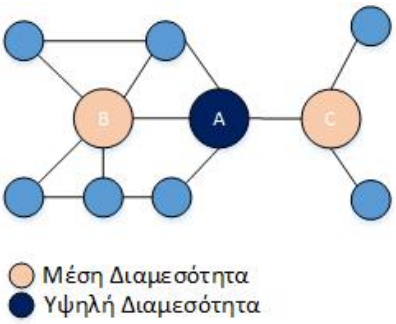
Με δεδομένο τον γράφο που αναπαριστά ένα κοινωνικό δίκτυο υπάρχουν δύο βασικές προσεγγίσεις για την αναγνώριση των κοινοτήτων που περιλαμβάνει. Η πρώτη έγκειται σε μια επαναληπτική διαδικασία όπου αφαιρούνται ακμές από τον γράφο με κάποιο κριτήριο. Μετά από ένα αριθμό επαναλήψεων και έπειτα ο αρχικός μας γράφος θα αρχίσει να χωρίζεται σε ένα πλήθος υπογράφων όπου κάθε υπογράφος θα υποδηλώνει μια κοινότητα[53].

Η δεύτερη προσέγγιση λειτουργεί αντίστροφα. Ξεκινά μόνο με τους κόμβους όλου του γράφου και σε μια σειρά διαδοχικών βημάτων προστίθενται με κάποιο κριτήριο ακμές [95]. Αρχικά θα θεωρούμε ότι κάθε κόμβος υποδηλώνει μια κοινότητα. Καθώς ακμές θα προστίθενται μικρές κοινότητες θα ενώνονται σχηματίζοντας μεγαλύτερες. Σε κάθε βήμα της επαναληπτικής διαδικασίας οι υπογράφοι που σχηματίζονται θα δηλώνουν της κοινότητες που υπάρχουν σε αυτό το στάδιο.

Υπάρχουν και άλλες τοπολογικές μέθοδοι που έχουν προταθεί για τον διαχωρισμό ενός γράφου σε επιμέρους συστάδες.

6.2.1.Βρίσκοντας κοινότητες σε έναν γράφο κοινωνικού δικτύου αφαιρώντας ακμές

Η πρώτη μέθοδος που θα περιγράψουμε βασίζεται στην διαδικασία αφαίρεσης ακμών από έναν γράφο έτσι όπως προτάθηκε από τους M.E.J.Newman και M.Girvan[53].



Σχήμα 6.1 Διαμεσότητα μεταξύ χρηστών

Έχουμε ως δεδομένο έναν γράφο ενός τύπου κορυφών οι οποίες αντιστοιχούν σε χρήστες και ακμές μη κατευθυνόμενες χωρίς βάρη που υποδηλώνουν τις σχέσεις μεταξύ δύο χρηστών. Προσπαθούμε να βρούμε της ακμές με το υψηλότερο betweenness (διαμεσότητα).

Αρχικά το betweenness ήταν ένας όρος που χαρακτήριζε κόμβους και όχι ακμές Ο ορισμός του betweenness για κόμβους είναι ο αριθμός που δηλώνει πόσες φορές οποιοσδήποτε κόμβος του γράφου χρειάζεται ένα συγκεκριμένο κόμβο για να φτάσει οποιοδήποτε άλλον κόμβο μέσω ενός συντομότερου μονοπατιού.

Στο σχήμα 6.1 Παρατηρούμε ότι πάρα πολύ κόμβοι για να επικοινωνήσουν αναμεταξύ τους διέρχονται μέσα από τον κόμβο Α για αυτό ο κόμβος αυτός έχει υψηλό betweenness. Το ίδιο ισχύει αλλά σε μικρότερο βαθμό για τους κόμβους Β και C.

Οι M.E.J.Newman και M.Girvan επέκτειναν τον όρο του betweenness για να προσδιορίζει και ακμές. Το betweenness ακμής το όρισαν ως τον αριθμό τον συντομότερων μονοπατιών που περνάν μέσα από αυτή την ακμή για να ενώσουν δύο κόμβους.

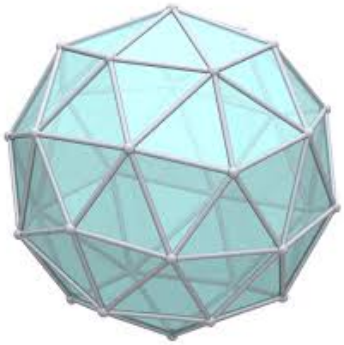
Διαισθητικά θα μπορούσαμε να πούμε πως το betweenness μιας ακμής, ενός κόμβου ακόμη και ενός υπογράφου μέσα σ' έναν γράφο δηλώνει την ικανότητα που έχει να συνδέει επιμέρους κομμάτια του γράφου. Καταλαβαίνουμε ότι όποια ακμή έχει υψηλό betweenness κατέχει μια σημαντική θέση μέσα στο δίκτυο και στην μετάδοση πληροφοριών των συστάδων του.

Ο αλγόριθμος μας βασίζεται στην ιδέα ότι οι ακμές με το μεγαλύτερο betweenness αναμένεται ότι θα βρίσκονται μεταξύ των διάφορων κοινοτήτων. Δηλαδή ακμές με υψηλό betweenness θα συνδέουν τις διάφορες συστάδες κόμβων αναμεταξύ τους και δεν θα βρίσκονται μέσα σε μια κοινότητα για να συνδέουν δύο κόμβους που ανήκουν στην ίδια συστάδα.

Δύο διαφορετικές κοινότητες αναμένουμε να συνδέονται με λίγες ακμές, οι οποίες θα έχουν υψηλό βαθμό betweenness. Επίσης όλα τα μονοπάτια του δικτύου που περνάνε από τις κορυφές μιας κοινότητας στις κορυφές μια άλλης κοινότητας πρέπει να διέρχονται μέσα από αυτές τις ακμές με το υψηλό betweenness.

6.2.2.Υπολογισμός διαμεσότητας κάθε ακμής

Για να μετρήσουμε το μέγεθος betweenness κάθε ακμής, η πιο αποδοτική μέθοδος είναι να βρούμε τα συντομότερα geodesic paths (γεωδαιτικά μονοπάτια) [96] μεταξύ όλων των ακμών και να υπολογίσουμε πόσα από αυτά τα μονοπάτια περνάνε από κάθε ακμή [97]. Το πόσα μονοπάτια περνάνε από κάθε ακμή θα ισούται με το betweenness της ακμής.



Σχήμα 6.2 γεωδαιτικό μονοπάτι

Σχήμα 2.2 Ο όρος geodesic path προέρχεται από την επιστήμη της γεωδαισίας (τομέας της γεωγραφίας) και είχε αρχικά το νόημα να δηλώνει το συντομότερο μονοπάτι μεταξύ δύο σημείων πάνω στην γη. Ο όρος γενικεύτηκε για να χρησιμοποιείται και σε μετρήσεις πιο γενικές. Στην θεωρία γράφων δηλώνει το μονοπάτι με το λιγότερο πλήθος ακμών που συνδέει δύο κόμβους.

Έχουν προταθεί και άλλες μέθοδοι για την μέτρηση του betweenness. Μία παραλλαγή της προηγούμενης είναι να θεωρήσουμε ότι κατά μήκος όλων των geodesic network path (γεωδαιτικών διαδρομών δικτύου) ταξιδεύουν σήματα από την κορυφή πηγής στην κορυφή προορισμού. Ο ρυθμός με τον οποίο περνάνε τα σήματα μέσα από τις ακμές ισούται με το betweenness της ακμής. Παραλλαγή και γενίκευση της μεθόδου αυτής είναι τα σήματα να ταξιδεύουν σε τυχαία μονοπάτια μέσα στο δίκτυο και εμείς απλώς να υπολογίζουμε τον αριθμό του πλήθος σημάτων που πέρασε μεταξύ δύο κορυφών.

Μια ακόμη μέθοδος μέτρησης του betweenness έρχεται από τον χώρο των ηλεκτρικών κυκλωμάτων. Θεωρούμε τον γράφο ως ένα ηλεκτρικό κύκλωμα όπου για κάθε ζευγάρι κόμβων θεωρείται ο ένας ως γεννήτρια και ο άλλος ως καταβόθρα φορτίου (sink). Επίσης κάθε ακμή ως μια μοναδιαία αντίσταση. Τέλος με εφαρμογές κανόνων Κίρχοφ βρίσκεται το ρεύμα που περνάει από κάθε ακμή. Γίνεται αντιληπτό ότι όπου υπάρχει μεγαλύτερη ροή ρεύματος εκεί θα υπάρχει ο μέγιστος βαθμός betweenness.

6.2.3.Αλγόριθμος εύρεσης κοινοτήτων που αφαιρεί ακμές με μέγιστη τιμή διαμεσότητας

Η εύρεση του betweenness κάθε ακμής είναι η καρδιά του αλγορίθμου μας και θα γίνεται σε κάθε βήμα όπως θα δούμε. Ο αλγόριθμος ξεκινά με τον αρχικό γράφο και τον υπολογισμό του betweenness για κάθε ακμής. Έπειτα αφαιρούμε από τον γράφο την ακμή με το υψηλότερο betweenness και επαναϋπολογίζουμε το betweenness.

Η αφαίρεση της ακμής με την υψηλότερη τιμή betweenness και ο επανυπολογισμός των betweenness τιμών για τον καινούριο γράφο που προκύπτει κάθε φορά είναι μια επαναληπτική διαδικασία η οποία τελειώνει όταν ο γράφος θα έχει διαιρεθεί σε τόσα partition όσες είναι οι κοινότητες που επιθυμούμε να έχουμε.

Ο λόγος που επαναλαμβάνουμε σε κάθε βήμα τον υπολογισμό του betweenness είναι ότι μετά την αφαίρεση κάθε ακμής ο γράφος αλλάζει. Οι προηγούμενες τιμές betweenness για τις ακμές όχι απλώς δεν αντιπροσωπεύουν το καινούριο μας δίκτυο αλλά αν βασιστούμε σε αυτές μπορεί να οδηγηθούμε σε εντελώς λανθασμένα αποτελέσματα. Αυτό θα γίνει κατανοητό με το παράδειγμα που ακολουθεί.

Έστω ότι υπάρχουν δύο συστάδες του γράφου που συνδέονται μόνο με δύο ακμές και η μια από τις δύο έχει υψηλό betweenness ενώ η άλλη όχι. Μετά την αφαίρεση της ακμής που έχει υψηλό betweenness η δεύτερη ακμή που θα συνδέει τις δύο συστάδες θα αποχτήσει στον καινούριο γράφο και αυτή υψηλό betweenness (Αφού τώρα πια θα είναι η μόνη ακμή που μπορεί να συνδέσει της δυο συστάδες με τους κόμβους τους) οπότε αναμένουμε ότι θα είναι αυτή η επόμενη που θα πρέπει να αφαιρεθεί. Αν όμως δεν κάνουμε τον επανυπολογισμό του betweenness μετά από την αφαίρεση κάθε ακμής και χρησιμοποιούμε την παλιά τιμή της δεν θα αφαιρεθεί και θα μας οδηγήσει προς λάθος αποτελέσματα. Οπότε αξίζει να αυξήσουμε

την πολυπλοκότητα του αλγορίθμου με το να κάνουμε αυτόν τον επανυπολογισμό κάθε φορά προς χάριν της ακρίβειας των αποτελεσμάτων.

Αφού εισήγαμε τις βασικές έννοιες της μεθόδου ακολουθεί ο αλγόριθμος εύρεσης κοινοτήτων που αφαιρεί ακμές με μέγιστη τιμή betweenness.

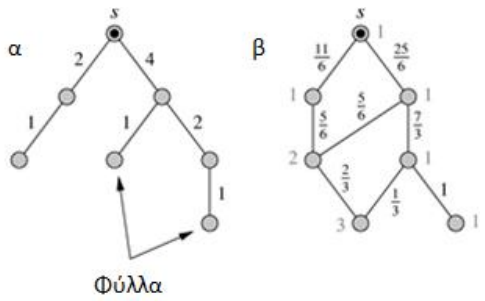
	Αλγόριθμος 6.1
1:	Υπολογίζουμε το betweenness βαθμό για κάθε ακμή στο κοινωνικό δίκτυο.
2:	Βρίσκουμε την ακμή με το υψηλότερο βαθμό και την διαγράφουμε (αν δύο ή περισσότερες ακμές έχουν το ίδιο μέγιστο βαθμό επιλέγουμε κάποια με τυχαία κριτήρια).
3:	Υπολογίζουμε ξανά το betweenness για όλες τις ακμές που έμειναν.
4:	Επιστρέφουμε στο βήμα 2 μέχρι να διαμεριστεί ο γράφος στο επιθυμητό πλήθος συστάδων.

Μετά από μια σειρά εκτελέσεων του αλγόριθμου θα αρχίσουν να σχηματίζονται οι πρώτες συστάδες που θα αντιστοιχούν στις κοινότητες. Το πότε θα σταματήσουμε εξαρτάται από το πλήθος των κοινοτήτων που θέλουμε να σχηματιστεί. Συνήθως ο αριθμός αυτός θα μας δίνεται Το βήμα 2 της εύρεσης της ακμής με το υψηλότερο βαθμό betweenness και αφαίρεσης της είναι τετριμμένο. Το σημαντικό βήμα είναι το πως υπολογίζουμε το betweenness βαθμό για κάθε ακμή του γράφου για αυτό θα ασχοληθούμε μαζί του αναλυτικότερα.

6.2.4.Αλγόριθμοι υπολογισμού διαμεσότητας κάθε ακμής

Με τον αλγόριθμο Breadth-first search μπορούμε να βρούμε τα συντομότερα μονοπάτια από μια κορυφή προς οποιαδήποτε άλλη. Ο λόγος που επιλέγουμε τον αλγόριθμο Breadth-first search είναι ότι έχουμε ακμές που δεν έχουν βάρη. Αν είχαμε ακμές με βάρη θα έπρεπε να γενικευθεί η μέθοδος αυτή και να χρησιμοποιηθούν αλγόριθμοι εύρεσης συντομότερων μονοπατιών όπως αυτοί των Dijkstra, Floyd–Warshall, Johnson, Viterbi και ο A*

Στην απλή περίπτωση που έχουμε μόνο ένα μονοπάτι προς κάθε άλλη κορυφή το σύνολο όλων των μονοπατιών από αυτή την κορυφή προς τις άλλες κορυφές θα σχηματίζει ένα δέντρο στην Περίπτωση Α αλλιώς θα σχηματίζει ένα γράφο στην Περίπτωση Β.



Σχήμα 6.3 Αναπαράσταση συντομότερων μονοπατιών

Στο σχήμα 6.3 απεικονίζεται η αναπαράσταση συντομότερων μονοπατιών από τον κόμβο s. Στο 6.3a αν έχουμε ένα μοναδικό συντομότερο μονοπάτι και στο 6.3b αν έχουμε πολλά συντομότερα μονοπάτια.

Περίπτωση Α: που υπάρχει μόνο ένα μονοπάτι προς κάθε άλλο κόμβο

Αν θεωρήσουμε την απλή περίπτωση που υπάρχει μόνο ένα geodesic path που συνδέει κάθε κόμβο με οποιονδήποτε άλλο τότε παραθέτοντας για κάθε κόμβο ξεχωριστά τα μονοπάτια που τον ενώνουν με τους άλλους κόμβους σχηματίζεται ένα πλήθος δέντρων. Το κάθε δέντρο θα ξεκινά από το κόμβο s ρίζα και καταλήγει προς οποιονδήποτε άλλο κόμβο. Το betweenness κάθε ακμής υπολογίζεται σε αυτή την περίπτωση ως εξής.

Ξεκινάμε για κάθε δέντρο ξεχωριστά από τα φύλλα του δέντρου, που αντιστοιχούν στους κόμβους που θέλουν να καταλήξουν τα μονοπάτια, και αναθέτουμε σε κάθε ακμή που είναι προσκείμενη σε αυτά τον βαθμό 1 όπως φαίνεται στο σχήμα 6.3a. Έπειτα διατρέχουμε το δέντρο προς τα πάνω δηλαδή με κατεύθυνση από τα φύλλα προς την κορυφή και για κάθε ακμή βάζουμε ένα βάρος που ισούται με το άθροισμα των βαθμών που έχουν οι από κάτω ακμές συν ένα. Όταν θα έχουμε φτάσει στην κορυφή του δέντρου θα έχουμε ένα βαθμό για όλες τις ακμές.

Αυτή την διαδικασία την κάνουμε για κάθε κορυφή του γράφου. Δηλαδή κάθε κορυφή θεωρείται ως ρίζα ενός δέντρου που σχηματίστηκε από την παράθεση όλων των μονοπατιών που ενώνουν αυτόν τον κόμβο της ρίζας με κάθε άλλο κόμβο.

Τέλος αν αθροίσουμε το βάρος που έχει μια ακμή σε κάθε ένα από αυτά τα δέντρα θα πάρουμε το betweenness της ακμής αυτής.

Περίπτωση Β: που υπάρχουν πολλά μονοπάτια προς κάθε άλλο κόμβο

Στην περίπτωση που έχουμε από δύο και πάνω ίσου βάρους συντομότερα μονοπάτια που ενώνουν δύο κορυφές σχηματίζεται αντί για δέντρο ένας γράφος που έχει κύκλους 6.3b. Να διευκρινίσουμε πως για κάθε κόμβο σχηματίζεται ένας ξεχωριστός γράφος που είναι το άθροισμα όλων των μονοπατιών κάθε άλλου κόμβου προς τον κόμβο που εξετάζουμε.

Για την εύρεση του betweenness κάθε ακμής εργαζόμαστε για κάθε γράφο ξεχωριστά. Εδώ η διαδικασία είναι πιο περίπλοκη. Σύμφωνα με τον παραδοσιακό ορισμό του betweenness κόμβων [98] όταν υπάρχουν πολλά συντομότερα μονοπάτια που ενώνουν δύο κόμβους σε κάθε ένα από αυτά αναθέτουμε ένα κλασματικό βαθμό που το άθροισμα τους δίνει ένα και είναι ίσο αναμεταξύ τους. Δηλαδή αν έχουμε ένα τελικό κόμβο που μπορούμε να καταλήξουμε σ' αυτόν μέσω τεσσάρων μονοπατιών κάθε ένα από αυτά τα μονοπάτια θα πάρει ως βάρος τον αριθμό ¼. Να σημειώσουμε ότι τα μονοπάτια αυτά δίνουν στις ακμές από τις οποίες αποτελούνται κάποιο βάρος. Για να υπολογίσουμε σωστά το κλασματικό μέρος που παίρνει κάθε ακμή χρησιμοποιούμε τον αλγόριθμο 2.2 που είναι μια γενίκευση του bfs.

	Αλγόριθμος 6.2
1:	Ο αρχικός κόμβος s έχει απόσταση d_s=0 και βάρος w_s=1
2:	Κάθε γειτονικός κόμβος i στον s αποχτά απόσταση d_i=d_s +1 και βάρος w_i= w_s =1.

3: Για κάθε κόμβο j προσκείμενο σ' έναν από τους κόμβους i (που αναφέραμε) κάνουμε ένα από τα εξής τρία βήματα ανάλογα την περίπτωση.
(a) Αν στον j κόμβο δεν έχει αποδοθεί απόσταση τότε βάζουμε απόσταση $d_j = d_i+1$ και βάρος $w_j = w_i$.
(b) Αν στον j κόμβο έχει ήδη αποδοθεί απόσταση και αυτή είναι $d_j = d_i+1$ τότε το βάρος του γίνεται $w_j \leftarrow w_i + w_i$.
(c) Αν στον κόμβο j έχει ήδη ανατεθεί μια απόσταση και αυτή είναι $d_j < d_i + 1$ δεν κάνουμε τίποτα.

4: Επαναλαμβάνουμε το βήμα 3 έως ότου να μην μείνει κανένας κόμβος που να του έχουν ανατεθεί αποστάσεις αλλά στους γείτονές του να μην έχουν ανατεθεί αποστάσεις.

Όπως γίνεται αντιληπτό το βάρος ενός κόμβου i δηλώνει τον αριθμό των ξεχωριστών μονοπατιών που διέρχονται από την ρίζα s προς τον κόμβο i.

Αυτά τα βάρη είναι ακριβώς ότι χρειαζόμαστε για να υπολογίσουμε το betweenness των ακμών διότι αν δύο κόμβοι i και j είναι συνδεδεμένοι με τον j πιο μακριά από τον i σε σχέση με την πηγή s τότε το κλάσμα του geodesic path από το j μέσω του του i προς το s είναι ίσο με w_i/w_j. Έτσι για να υπολογίσουμε (την συμβολή) του betweenness κάθε ακμής από όλα τα συντομότερα μονοπάτια ξεκινώντας από το s κάνουμε τα ακόλουθα τέσσερα βήματα:

Αλγόριθμος 6.3

1: Βρίσκουμε κάθε φύλο – κόμβο t δηλαδή ένα κόμβο που δεν περνάνε άλλα μονοπάτια από μέσα του.

2: Για κάθε κόμβο i που συνορεύει προς το t αποδίδουμε έναν βαθμό στην ακμή t-i w_i/w_j.

3: Ξεκινώντας από τις ακμές που είναι πιο μακριά από την πηγή s σύμφωνα με το σχήμα 4β από τους πιο χαμηλούς κόμβους ανεβαίνοντας προς τα πάνω προς την πηγή s. Για κάθε ακμή από τον κόμβο i προς τον κόμβο j, ο κόμβος j είναι πιο μακριά από την πηγή από τον i, απονέμουμε έναν βαθμό στην ακμή i j που είναι ίσος με 1 συν το άθροισμα των βαθμών των γειτονικών ακμών που είναι ακριβώς από κάτω του πολλαπλασιασμένο με w_i/w_j.

4: Επαναλαμβάνουμε το βήμα 3 μέχρι να φτάσουμε στην κορυφή s.

Η εκτέλεση του προηγούμενου αλγόριθμου θα πρέπει να γίνει για κάθε έναν κόμβο ξεχωριστά και να αθροίσουμε τα betweenness όλων των ακμών όπως κάναμε και στην προηγούμενη ενότητα που είχαμε μόνο ένα μονοπάτι προς κάθε άλλο κόμβο.

Αξιολόγηση και πολυπλοκότητα της μεθόδου

Οι κοινωνικοί γράφοι μπορεί να είναι κατευθυνόμενοι ή μη κατευθυνόμενοι. Η μέθοδος που περιγράψαμε είναι για μη κατευθυνόμενα δίκτυα. Πολύ εύκολα μπορούμε να την γενικεύσουμε και για κατευθυνόμενους

γράφους απλά με το να υπολογίζουμε στους προηγούμενους αλγόριθμους για κάθε δύο κόμβους μόνο τα μονοπάτια που όλες οι ακμές τους προσανατολίζονται προς μία κατεύθυνση. Δηλαδή να κάνουμε χρήση των κατευθυνόμενων μονοπατιών.

Ο M. E. J. Newman και ο M. Girvan που δημοσίευσαν την μέθοδο αυτή [53] ισχυρίζονται ότι μπορούμε σε πολλές περιπτώσεις έναν κατευθυνόμενο γράφο να τον μετατρέψουμε σε μη κατευθυνόμενο και να εφαρμόσουμε τους αλγόριθμους χωρίς να οδηγηθούμε σε λάθος συμπεράσματα. Αυτό το στηρίζουν στο ότι οι ακμές μεταξύ κόμβων δηλώνουν απλά μια σύνδεση μεταξύ χρηστών και δεν προσφέρει περισσότερη πληροφορία ποια θα είναι η κατεύθυνση τους.

Ο αλγόριθμος εύρεσης του betweenness των ακμών του Newman έχει πολυπλοκότητα *O(mn)* και λόγω του ότι θα τον εφαρμόσουμε για m το πλήθος ακμές που θα αφαιρούνται θα έχουμε συνολική πολυπλοκότητα της τάξεως του $O(m^2n)$ ή για έναν αραιό πίνακα της τάξεως του $O(n^3)$.

Η πολυπλοκότητα της εύρεσης του betweenness των ακμών στην περίπτωση που υπάρχει μόνο ένα μονοπάτι προς κάθε άλλο κόμβο, υπολογίζεται από το ότι ο bfs αλγόριθμος που βρίσκει τα μονοπάτια έχει πολυπλοκότητα O(m) και εφόσον έχουμε n κόμβους για τους όποιους θα χρειαστεί να βρούμε τα geodesic paths το γινόμενο τους θα είναι η συνολική πολυπλοκότητα *O(mn)*. Αντίστοιχα και η πολυπλοκότητα που προκύπτει από τους αλγορίθμους 6.2 και 6.3 είναι *O(mn)*.

Σε πειράματα που έκανε ο Newman με 10.000 κόμβους ένας απλός επιτραπέζιος υπολογιστής μπορούσε να αναγνωρίσει τις κοινότητες σε αποδοτικό χρονικό διάστημα και μάλιστα σε δίκτυα με σαφή κοινωνική δομή ο σχηματισμός των κοινοτήτων προέκυπτε από τις πρώτες κιόλας αφαιρέσεις ακμών.

6.2.5.Μια ευριστική διαδικασία για διαμερισμό γράφων

Μια ευριστική επίλυση του προβλήματος της διαμέρισης του γράφου ενός κοινωνικού δικτύου σε κοινότητες προκύπτει από τον χώρο της ηλεκτρονικής [56]. Η μέθοδος αυτή είχε προταθεί για τον βέλτιστο διαμερισμό ηλεκτρονικών στοιχείων ανάμεσα σ' ένα πλήθος ηλεκτρονικών πλακετών ούτως ώστε οι συνδέσεις που ενώνουν διαφορετικές πλακέτες αναμεταξύ τους να είναι όσο το δυνατόν λιγότερες.

Οι συνδέσεις μεταξύ στοιχείων στην ίδια πλακέτα είναι κάτι επιθυμητό, ενώ οι συνδέσεις στοιχείων μεταξύ διαφορετικών πλακετών θα πρέπει να αποφεύγονται. Καταλαβαίνουμε πως αυτό το πρόβλημα αν αναγάγουμε όπου χρήστης ένα ηλεκτρονικό στοιχείο, όπου σύνδεση μια ακμή και όπου πλακέτα μια συστάδα έχει την ίδια βάση με το πρόβλημα της αναγνώρισης κοινοτήτων από τον κοινωνικό γράφο.

Αρχικά θα διατυπώσουμε την επίλυση του προβλήματος για δύο συστάδες με ίσο πλήθος κόμβων η κάθε μια και έπειτα θα γενικεύσουμε το πρόβλημα σε k το πλήθος διαφορετικές συστάδες με ανόμοιο πλήθος κόμβων η κάθε μία.

Α: Διαμέριση του γράφου σε δύο ίσα σύνολα

Έστω S είναι το σύνολο 2n κόμβων και έστω ο πίνακας $C=(c_{ij})$ που το c_{ij} δηλώνει το κόστος κάθε ακμής και εξαρτάται από το αν συνδέονται οι κόμβους i και j. Ο πίνακας C είναι συμμετρικός και κάθε στοιχείο της διάμεσου του είναι μηδέν $c_{ij}=0$ λόγο του ότι ο γράφος είναι μη κατευθυνόμενος. Η διαμέριση που θα παραχθεί έχει ως στόχο να ελαχιστοποιήσει το εξωτερικό κόστος

$$T= \sum_{AxB} c_{ab} \qquad (6.1)$$

Δηλαδή σκοπός μας είναι το κόστος που προκαλείται από τις ακμές που συνδέουν την μια συστάδα με την άλλη να είναι ελάχιστο. Εδώ να αναφέρουμε ότι υπάρχουν τέσσερα προβλήματα που είναι ισοδύναμα και η μέθοδος που θα περιγράψουμε επιλύει και τα τέσσερα αφού ουσιαστικά είναι διαφορετικές διατυπώσεις του ίδιου προβλήματος.

(α) Η εύρεση μια διαμέρισης που ελαχιστοποιεί το εξωτερικό κόστος

(β) Η εύρεση μια διαμέρισης που μεγιστοποιεί το εσωτερικό κόστος

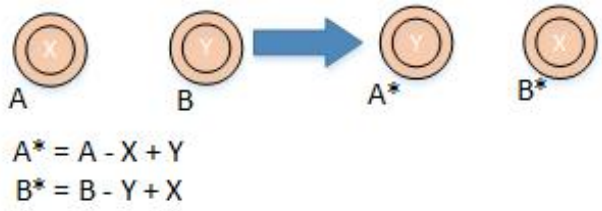
Επίσης αν αλλάξουμε τις τιμές του c_{ij}

(γ) Η εύρεση μια διαμέρισης που μεγιστοποιεί το εξωτερικό κόστος

(δ) Η εύρεση μια διαμέρισης που ελαχιστοποιεί το εσωτερικό κόστος

Η βασική ιδέα της μεθόδου είναι η εξής: Ξεκινώντας με μια αυθαίρετη διαμέριση του γράφου S στους υπογράφους A και B ανταλλάσσοντας κόμβους από τον A στο B με σκοπό να ελαχιστοποιήσουμε το εξωτερικό κόστος.

Έχοντας τον γράφο S και τον πίνακα με τα κόστη ακμών c_{ij} , θεωρούμε A*, B* την διαμέριση με το ελάχιστο εξωτερικό κόστος και A, B οποιαδήποτε αυθαίρετη διαμέριση. Σε αυτή την περίπτωση υπάρχουν X κόμβοι που ανήκουν στο A και Y κόμβοι που ανήκουν στο B τέτοιοι που αν τους ανταλλάξουμε από το A στο B θα προκύψει το A* και B* όπως δείχνει το σχήμα 6.4



Σχήμα 6.4 Αντιμετάθεση κόμβων

Στο σχήμα 6.4 το σύνολο A έχει X κόμβους που θα έπρεπε να ανήκουν στο B και το σύνολο B έχει Y κόμβους που θα έπρεπε να ανήκουν στο A. Το βέλτιστο σύνολο A* προκύπτει αν του αφαιρέσουμε τους X κόμβους και του προσθέσουμε τους Y. Αντίστοιχα το βέλτιστο σύνολο B* προκύπτει αν του αφαιρέσουμε τους Y κόμβους και προσθέσουμε τους X.

Σκοπός της μεθόδου μας είναι να βρούμε τα υποσύνολα X και Y από τα A και B που θα πρέπει να αφαιρεθούν και να τα ανταλλάξουμε αναμεταξύ τους.

Ορίζουμε το εξωτερικό κόστος Ea ως

$$\text{Ea} = \sum_{y \epsilon B} c_{ay} \quad (6.2)$$

Όπου α οποιοσδήποτε κόμβος ανήκει στο σύνολο Α. Η εξίσωση 6.2 σημαίνει ότι το εξωτερικό κόστος Ea είναι ίσο με το πόσοι κόμβοι υπάρχουν μέσα στο σύνολο Β που ενώνονται με μία ακμή με το σύνολο Α. Το εσωτερικό κόστος Ia

$$Ia= \sum_{x \varepsilon A} c_{ax} \quad (6.3)$$

Δηλώνει πόσοι κόμβοι υπάρχουν μέσα στο σύνολο Α που ενώνονται μέσω μια ακμής με κάποιον άλλο κόμβο του συνόλου Α. Αντίστοιχα ορίζονται και τα Eb Ib για κάθε b που ανήκει στο Β. Καταλαβαίνουμε ότι επιθυμούμε να έχουμε μεγάλο εσωτερικό κόστος Ia και μικρό εξωτερικό κόστος Ea από την διαμέριση που θα κάνουμε.

Η διαφορά εξωτερικού με εσωτερικού κόστους για το z ε S δίνεται από τον τύπο:

$$Dz = Ez - Iz \quad (6.4)$$

Ισχύει το επόμενο πολύ σημαντικό λήμμα:

Αν ο κόμβος α ανήκει στο σύνολο Α, ο κόμβος b ανήκει στο σύνολο Β και ανταλλάξουμε τον a με τον b το κέρδος που θα αποκτήσουμε θα είναι

$$g = D_a + D_b - 2C_{ab} \quad (6.5)$$

Ο Αλγόριθμος 6.4 κάνει χρήση του παραπάνω λήμματος και διαμερίζει τον γράφο σε δύο σύνολα συστηματοποιώντας όλες τις παραπάνω ιδέες.

	Αλγόριθμος 6.4
1:	Υπολογίζουμε τις D τιμές για όλα τα στοιχεία του γράφου S
2:	Διαλέγουμε τα a_i ε A και b_j ε B τέτοια ώστε να είναι μέγιστη η ποσότητα $g_1=D_{ai}+D_{bj}-2C_{aibj}$. Τα a_i και b_j αντιστοιχούν στο μεγαλύτερο κέρδος που μπορούμε να έχουμε από μια απλή εναλλαγή.
3:	Επαναϋπολογίζουμε τις Τιμές D για όλα τα στοιχεία του A - {a_i} και B - {b_j} με $D'_x= D_x + 2c_{xai} - 2c_{xbj}$ x ε A - {a_i}, $D'_y= D_y + 2c_{ybj} - 2c_{yai}$ y ε B - {b_j}.
4:	Επαναλαμβάνουμε το (ii) επιλέγοντας ένα ζευγάρι a'_2, b'_2 από το A - {a'_1} και B - {b'_1} έτσι ώστε να είναι μέγιστη η ποσότητα $g_2=D_{a2'} + D_{b2'} - 2\ C_{a2'b2'}$
5:	Συνεχίζουμε έως ότου όλοι οι κόμβοι (a'_i,b'_i) με τα αντίστοιχα κέρδη τους g_n να εξεταστούν.
6:	Στο τέλος επιλέγουμε τους k $g_1 + g_2 + \ldots + g_k$ που μεγιστοποιούν το μερικό άθροισμα G= $\sum_{i=1}^{k} g_i$

7: Αν το G>0 τότε μια ακόμη μείωση στο κόστος μπορεί να γίνει με το να αλλάξουν θέση οι κόμβοι από το σύνολο X στο σύνολο Y. Μετά την καινούρια αλλαγή την διαμέριση μπορούμε να την διαχειριστούμε ως να είναι η αρχική και να συνεχίσει να επαναλαμβάνεται η διαδικασία αυτή.

8: Αν το G=0 βρισκόμαστε στην τοπικά βέλτιστη λύση της διαμέρισης

Η τοπικά βέλτιστη λύση είναι μια λύση αλλά λόγω του ότι ο αλγόριθμός μας είναι ευριστικός μπορεί να υπάρχει κάποια άλλη τοπικά βέλτιστη λύση που να διαμερίζει καλύτερα τους κόμβους. Φυσικά έχουμε την επιλογή να επαναλάβουμε τον αλγόριθμο με μια άλλη αρχική διαμέριση με την ελπίδα να προκύψει μια καλύτερη διαμέριση.

Β: Διαμέριση σε άνισα σύνολα

Ας υποθέσουμε την πιο πιθανή περίπτωση να θέλουμε να διαμοιράσουμε τους κόμβους ενός γράφου όχι σε δύο ίσα σύνολα αλλά σε δύο άνισα. Το ένα σύνολο θα έχει n1 κόμβους, το άλλο n2 και n1 + n2 =n. Έστω ότι n1 < n2 Ο μόνος περιορισμός που έχουμε είναι ότι σε μια επανάληψη του αλγορίθμου που περιγράψαμε δεν μπορούμε να μεταθέσουμε περισσότερους από n1 κόμβους.

Η διαδικασία μπορεί εύκολα να επιτευχθεί με τον αλγόριθμο 6.4 αν προσθέσουμε κάποιους ψεύτικους (dummy) κόμβους οι οποίοι δεν θα συνδέονται με κανέναν άλλον κόμβο έτσι θα έχουν μηδενικές τιμές στα κόστη του πίνακα C. Το πλήθος των κόμβων που θα προσθέσουμε θα είναι 2n2 - n. Αφού εκτελέσουμε τον αλγόριθμο αφαιρούμε τους dummy κόμβους και έχουμε την διαμέριση που θέλουμε.

Γ: Διαμέριση σε περισσότερα από δύο σύνολα

Έστω ότι έχουμε kn κόμβους και θέλουμε να τους διαμοιράσουμε σε k σύνολα αρχίζουμε με το να διαμερίσουμε τους κόμβους στα k σύνολα. Έπειτα εφαρμόζουμε τον αλγόριθμο της διαμέρισης σε δύο σύνολα επαναλαμβάνοντας τον για όλα τα ζευγάρια συνόλων που έχουμε τόσες φορές όσες να κάνουμε την διαμέριση πιο βέλτιστη. Τα πειραματικά δεδομένα έχουν δείξει πως ο αριθμός των πόσων φορών θα επαναλάβουμε την διαδικασία είναι σχετικά μικρός και η όλη διαδικασία συγκλίνει αρκετά γρήγορα.

Δ: Αρχική διαμέριση

Σε όλες τις προηγούμενες παραγράφους αναφέραμε ότι ξεκινάμε την εκτέλεση του αλγόριθμου μας από μια αρχική διαμέριση. Καταλαβαίνουμε πως το να πάρουμε μια τυχαία διαμέριση για αρχική δεν είναι καλή επιλογή. Παρακάτω παραθέτουμε δύο μεθόδους για να έχουμε καλύτερες αρχικές διαμερίσεις.

Αν πρέπει να κάνουμε μια διαμέριση σε k σύνολα ξεκινάμε με μια διαμέριση σε r σύνολα έπειτα το κάθε ένα από αυτά το διαμερίζουμε ώστε τελικά να έχουμε s σύνολα και συνεχίζουμε να διαμερίζουμε το κάθε υποσύνολο έως ότου παραχθούν k στο πλήθος σύνολα. Για παράδειγμα αν το k είναι δύναμη του δύο θα μπορούσαμε αρχικά να κάνουμε μια διαμέριση σε δύο σύνολα έπειτα το κάθε ένα σύνολο από αυτά σε δύο και αναδρομικά θα συνεχίζαμε να διαμερίζουμε κάθε υποσύνολο έως ότου φτάναμε σ’ ένα πλήθος k συνόλων που επιθυμούμε.

Η δεύτερη μέθοδος διαμέρισης βασίζεται στην ιδέα διαμέρισης μεταξύ άνισου μεγέθους συνόλων με χρήση dummy κόμβων όπως συζητήσαμε πιο πριν. Αν έχουμε kn το πλήθος κόμβους θα τους διαμερίσουμε σε k σύνολα με την εξής μέθοδο. Στην αρχή διαμερίζουμε τους κόμβους σε δύο σύνολα το ένα με n κόμβους και το άλλο με (k - 1)n κόμβους έπειτα συνεχίζουμε με το να διαμερίσουμε τους (k - 1)n κόμβους σε n και (k - 2)n κόμβους και συνεχίζουμε επαναληπτικά έως ότου φτάσουμε να διαμερίσουμε όλος τους κόμβους σε k σύνολα.

Πολυπλοκότητα της μεθόδου

Η αλγοριθμική μέθοδος αυτή θα χρειαστεί να κάνει μια σειρά περασμάτων για να επιλέξει τα (a'1 , b'1) ,... , (a'n , b'n) καθώς και τα σύνολα X και Y που θα ανταλλάξουμε. Κάθε πέρασμα θα αποτελείται από τον υπολογισμό των D τιμών. Μια διαδικασία που έχει πολυπλοκότητα O(n^2) λόγω του ότι για κάθε στοιχείο του S πρέπει να υπολογίσουμε την σχέση του με όλα τα υπόλοιπα και από μια διαδικασία με πολυπλοκότητα O(n^2logn) που θα είναι η ταξινόμηση αυτών των D τιμών. Η ολική πολυπλοκότητα για όλα τα περάσματα θα είναι O (n^2logn.)

6.2.6.Αναγνώριση κοινοτήτων μέσω σημασιολογικών και στατιστικών δεδομένων

Παραπάνω περιγράψαμε διαδικασίες ανεύρεσης κοινοτήτων που χρησιμοποιούν την τοπολογία του κοινωνικού γράφου. Γενικώς οι πρώτες μέθοδοι που εφαρμόστηκαν βασίζονταν στον διαμερισμό του γράφου (graph partitioning) [56] και της ιεραρχικής συσταδοποίησης (hierarchical clustering) [99][100]. Σε αυτές τις μεθόδους υπάρχουν εγγενώς τα προβλήματα ότι πρέπει να γνωρίζουμε από πριν το πλήθος των κοινοτήτων, ποια θα είναι τα κριτήρια με τα οποία θα θεωρούμε κάποιους κόμβους ότι ανήκουν σε μια συστάδα και βέβαια πάντα υπάρχει η πρόκληση του να βρεθούν αλγόριθμοι πιο αποδοτικοί και με καλύτερη πολυπλοκότητα [101]. Μεταγενέστεροι αλγόριθμοι[102] [103] [104] προτάθηκαν που φέρανε μεγάλες βελτιώσεις στην αποδοτικότητα και στην μικρότερη πολυπλοκότητα.

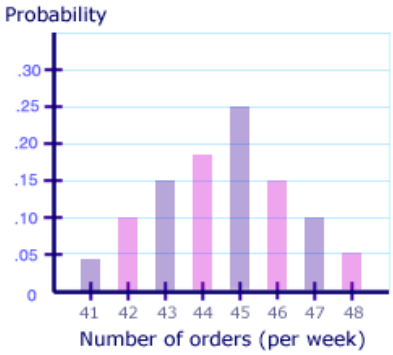
Η μεγάλη αλλαγή στην προσέγγιση και στον τρόπο αντιμετώπισης του προβλήματος της ανεύρεσης κοινοτήτων που σχηματίζονται με βάση τα θέματα που διαπραγματεύονται βασίζεται στην ιδέα πως όλα τα μέλη ενός κοινωνικού δικτύου (social actors) γράφουν ή διαβάζουν κείμενα όπως προσωπικά μηνύματα, συνομιλίες, σχόλια σε φωτογραφίες και άλλα. Σε αυτό το σημείο συνειδητοποιούμε πως η συσχέτιση και συσταδοποίηση των χρηστών μπορεί να γίνει βάση σημασιολογικών κριτηρίων και όχι απλώς από έναν γράφο που δηλώνει σχέσεις μεταξύ χρηστών.

Έξυπνοι αλγόριθμοι διαβάζουν κείμενα που έχουν προσπελάσει χρήστες, έπειτα τα κείμενα αυτά τα συσχετίζουν με θέματα (topics) και τα θέματα με κοινότητες (communities) μέσω πιθανοτικών κατανομών (Probability distribution). Αυτοί οι αλγόριθμοι είναι σε θέση από την σημασιολογία των κειμένων που έχουν διαβάσει και γράψει οι χρήστες, να παράγουν πιθανότητες για κάθε κοινότητα που να δηλώνουν, σε τι βαθμό ο χρήστης ανήκει σε αυτήν.

6.2.7.Πιθανοτικές κατανομές

Στην υποενότητα αυτή διαπραγματευόμαστε την έννοια της πιθανοτικής κατανομής στην οποία στηρίζονται οι παρακάτω μέθοδοι που αναπτύσσουμε. Ο όρος της πιθανοτικής κατανομής προκύπτει από την θεωρία πιθανοτήτων και την επιστήμη της στατιστικής και διαισθητικά δηλώνει έναν τρόπο που οργανώνεται και παρουσιάζετε μια λίστα πιθανοτήτων. Μια κατανομή πιθανότητας είναι μια συνάρτηση ή ένας νόμος σύμφωνα με τον οποίο κατανέμουμε μια πιθανότητα για κάθε τιμή μιας συγκεκριμένης μεταβλητής.

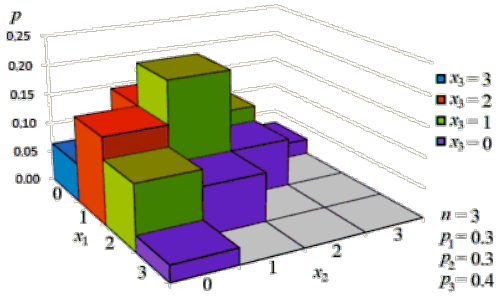
Η κατανομή μια μεταβλητής [105] συνήθως αναπαρίσταται γραφικά ως ένα ιστόγραμμα διαφορετικά ως μια λίστα ή ένας δισδιάστατος πίνακας που για κάθε μεταβλητή αντιστοιχούμε μια τιμή. Αν επιλέξουμε την αναπαράσταση με ιστόγραμμα και έχουμε μια μεταβλητή ο άξονας χ θα έχει το σύνολο των τιμών που μπορεί να πάρει η μεταβλητή της πιθανοτικής κατανομής και ο άξονας ψ την τιμή της πιθανότητας που έχει η μεταβλητή χ για κάθε τιμή της.



Σχήμα 6.5 Κατανομή πιθανότητας μίας μεταβλητής

Στο σχήμα 6.5 απεικονίζεται ένα παράδειγμα κατανομής πιθανότητας μίας μεταβλητής. Ο άξονας χ δηλώνει την μεταβλητή του πόσες παραγγελίες γίνονται σε μια βδομάδα και ο άξονας ψ με τη πιθανότητα γίνεται η διακριτή μεταβλητή χ να πάρει την κάθε τιμή.

Όταν έχουμε μια κατανομή που δέχεται πολλές μεταβλητές [106] η αναπαράσταση και η κατανόηση της γίνεται πιο περίπλοκη. Αν έχουμε δύο μεταβλητές μπορούμε να την αναπαραστήσουμε στον τρισδιάστατο χώρο όπως δείχνει το σχήμα 6.6. Οι τιμές της μπορούν να είναι αποθηκευμένες σε μια δομή δεδομένων ή να υπολογίζονται από μια συνάρτηση.



Σχήμα 6.6 Παράδειγμα Κατανομή Πιθανότητας πολλών μεταβλητών

Στο σχήμα 6.6 απεικονίζεται ένα Παράδειγμα Κατανομής Πιθανότητας πολλών μεταβλητών. Ο άξονας χ1 δηλώνει το πλήθος διακριτών τιμών που μπορεί να πάρει η πρώτη μεταβλητή. Ο άξονας χ2 δηλώνει το πλήθος διακριτών τιμών που μπορεί να πάρει η δεύτερη μεταβλητή και ο κάθετος άξονας την πιθανότητα που έχει κάθε συνδυασμός P(χ1,χ2)

6.2.8.Τα μοντέλα που σχηματίζονται από τα θέματα, λέξεις, και τους συγγραφείς

Αρχικά θα μπορούσαμε να διαχωρίσουμε τις μεθόδους σημασιολογικών μοντέλων ανεύρεσης κοινοτήτων που σχηματίζονται με βάση τα θέματα που διαπραγματεύονται σε τρεις κατηγορίες Topic - Word model (μοντέλο Θέματος - Λέξης), Author - Word model (μοντέλο Συγγραφέα - Λέξης) και Author - Topic model (μοντέλο Συγγραφέα - Θέματος)

Αρχικά έχουμε ένα σύνολο από κείμενα D όπου το κάθε ένα αποτελείται από μια ακολουθία λέξεων wd {Nd}. Η παραγωγή κάθε λέξης wdi wd για κάθε κείμενο d μπορεί να μοντελοποιηθεί είτε από την πλευρά του συγγραφέα A, είτε από την πλευρά του θέματος T είτε από έναν συνδυασμό αυτών των δύο. Οπότε θεωρούμε κάθε θέμα ως μια πιθανοτική πολυωνυμική κατανομή σε λέξεις.

Παρακάτω ακολουθούμε τους ακόλουθους συμβολισμούς.

φ για την κατανομή κάθε θέματος σε λέξεις. Δηλαδή κάθε θέμα σε τι ποσοστό περιέχει κάθε λέξη.

θ για την κατανομή κάθε κειμένου σε θέματα. Δηλαδή κάθε κείμενο σε τι ποσοστό εμπίπτει σε κάθε θέμα.

Θα περιγράψουμε τα τρία μοντέλα με συντομία:

Μοντέλο Θέματος – Λέξης

Σε αυτό το μοντέλο ένα κείμενο d θεωρείται ως μια μίξη θεμάτων. Κάθε θέμα αντιστοιχεί σε μια πολυωνυμική κατανομή σε ένα λεξικό, δηλαδή σε μια σειρά λέξεων.

Η ύπαρξη της λέξης w σ' ένα κείμενο d προκύπτει από την κατανομή φz πάνω σε ένα συγκεκριμένο θέμα z. Αντίστοιχα το θέμα z προέρχεται από την κατανομή του συγκεκριμένου κειμένου σε θέματα θd. Αυτό συνήθως το αναπαριστούμε με έναν πίνακα όπου κάθε γραμμή δηλώνει την κατανομή θ.

Αν θέλαμε να το διατυπώσουμε αντίστροφα θα λέγαμε η κατανομή θεμάτων για το συγκεκριμένο κείμενο d παράγει το θέμα z. Το θέμα z μέσα στην κατανομή θέματος σε λέξεις Φz παράγει τις λέξεις w.

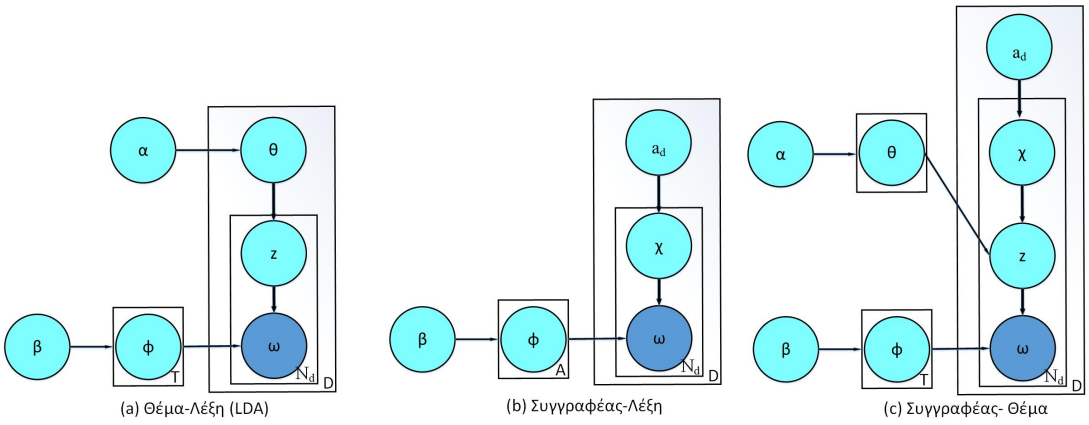
Μοντέλο Συγγραφέα – Λέξης

Παρόμοια με το μοντέλο Θέματος - Λέξης. Βαθμολογούμε τα ενδιαφέροντα ενός συγγραφέα A με βάση τις λέξεις που έχει χρησιμοποιήσει. Με ad συμβολίζουμε το σύνολο κειμένων που έχει γράψει ένας συγγραφέας.

Κάθε λέξη w στο κείμενο d επιλέγεται από μια κατανομή ενός συγκεκριμένου συγγραφέα σε λέξεις. Σε αυτό το μοντέλο Συγγραφέα - Λέξης ο συγγραφέας είναι υπεύθυνος και συσχετίζεται με κάθε συγκεκριμένη λέξη που έχει επιλεχθεί από το σύνολο κειμένων ad που έχει γράψει.

Μοντέλο Συγγραφέα – Θέματος

Για κάθε λέξη w σε ένα κείμενο d ένας συγγραφέας x επιλέγεται από το σύνολο συγγραφέων ad. Έπειτα από την κατανομή θεμάτων με δεδομένα το x, θα παράγεται ένα θέμα z. Το θέμα z με την σειρά του παράγει την λέξη w όπως την είδαμε στο κείμενο d. Το μοντέλο αυτό θα το παρουσιάσουμε αναλυτικότερα.



Σχήμα 6.7 Μοντέλο Θέματος – Λέξης, Συγγραφέα – Λέξης και Συγγραφέα – Θέματος

Στο σχήμα 6.7 βλέπουμε τα τρία σχήματα που αναπαριστούν τα τρία μοντέλα. Με σκιασμένο κύκλο αναπαριστούμε τα παρατηρούμενα δεδομένα ενώ με φωτεινά τα έμμεσα συνεπαγόμενα. Κάθε ορθογώνιο παραλληλόγραμμο αναπαριστά ένα σύνολο μεταβλητών. Στο (a) φαίνεται πως η κάθε λέξη σ' ένα κείμενο σχετίζεται από την κατανομή του θέματος σε λέξεις φ και από ένα θέμα z από την κατανομή του κειμένου σε θέματα θ. Στο (b) η λέξη σ' ένα κείμενο σχετίζεται από την κατανομή του θέματος σε λέξεις φ και έναν συγγραφέα x από ένα σύνολο συγγραφέων. Στο (c) έχουμε έναν συνδυασμό των προηγούμενων δύο μεθόδων. Η κάθε λέξη σ' ένα κείμενο σχετίζεται από την κατανομή του θέματος σε λέξεις φ αλλά κάθε θέμα z επιλέγεται ως ένας συνδυασμός της κατανομής του κειμένου σε θέματα θd και ενός συγγραφέα x από το σύνολο συγγραφέων.

Το μοντέλο Συγγραφέα – Θέματος αναλυτικότερα

Το μοντέλο Συγγραφέα - Θέματος περιέχει δύο σημαντικούς παράγοντες τον συγγραφέα και το θέμα. Μοντελοποιώντας και τους δύο παράγοντες ως μεταβλητές σε ένα Bayesian network (Μπεϋζιανό δίκτυο) προσφέρει στο μοντέλο την ικανότητα να ομαδοποιεί τις λέξεις σε σημασιολογικά θέματα και ταυτόχρονα να δείχνει για κάθε συγγραφέα τις λέξεις που συνηθίζει να χρησιμοποιεί και τα θέματα με τα οποία ασχολείται.

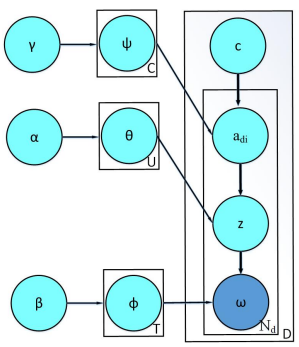
Στα προηγούμενα τρία συστήματα ο υπολογισμός του Bayesian network βασιζόταν στις παρατηρούμενες (observed) παραμέτρους του συγγραφέα, των λέξεων και σε κάποιες άλλες κριμένες (hidden) παραμέτρους όπως τα θέματα τα οποία υπολογίζει η ίδια η μέθοδος. Τέλος κάθε λέξη την επεξεργαζόμαστε ως μια οντότητα που προήλθε ως αποτέλεσμα πιθανοτήτων του μοντέλου.

Παρακάτω θα περιγράψουμε πως μπορούμε να μοντελοποιήσουμε μια κοινότητα να αποτελείται από χρήστες καθώς και πως μπορούμε να μοντελοποιήσουμε την κάθε κοινότητα να σχετίζεται με θέματα.

Μοντελοποιώντας μια Κοινότητα σε Χρήστες

Από τη στιγμή που λαμβάνουμε υπόψη την επίδραση της κοινότητας στην δημιουργία των κειμένων επικοινωνίας το πρώτο πράγμα που θα μας απασχολήσει είναι η αλληλεπίδραση μεταξύ των κρυμμένων (latent) μεταβλητών.

Ενώ σε προηγούμενες μεθόδους η σύνδεση μεταξύ δύο χρηστών αντιστοιχούσε στην συχνότητα με την οποία οι δύο τους επικοινωνούν και από αυτές τις συνδέσεις προέκυπταν οι κοινότητες. Στην μοντελοποίηση μιας κοινότητας με χρήστες θεωρούμε κάθε κοινότητα ως μια πολυωνυμική κατανομή σε χρήστες. Όπου κάθε χρήστης σχετίζεται με μια πιθανότητα P(u|c) η οποία δηλώνει τον βαθμό που ο χρήστης u ανήκει στην κοινότητα c. Σκοπός μας είναι να βρούμε την υπό συνθήκη πιθανότητα, δοσμένου μιας κοινότητας που να προσδιορίζει ποια η πιθανότητα να ανήκει εκεί ο χρήστης u. Έπειτα οι χρήστες θα συσχετιστούν με ένα σύνολο από θέματα όπου κάθε ένα έχει μια κατανομή σε λέξεις.



Σχήμα 6.8 Μοντελοποίηση κοινότητας με χρήστες

Στο σχήμα 6.8 βλέπουμε την μοντελοποίηση κοινότητας με χρήστες. Το C δηλώνει το community (κοινότητα), u τον χρήστη (user), T το θέμα (topic). Το α και β είναι οι πρότεροι (prior) Dirichlet παράμετροι για τα θέματα και τις λέξεις. Το ψ δηλώνει την πολυωνυμική κατανομή των χρηστών για κάθε κοινότητα το οποίο παραμετροποιείται από το γ.

Σε αυτό το μοντέλο ένα κείμενο παράγεται από τέσσερα στάδια

(i) Μια κοινότητα c γράφει ή αναρτά ένα κείμενο d.

(ii) Ένας χρήστης u επιλέγεται από την κοινότητα c ως αναγνώστης του κειμένου d.

(iii) Ο χρήστης u συσχετίζεται με το θέμα z.

(iv) Από το θέμα z μια λέξη γράφεται στο κείμενο d.

Επαναλαμβάνοντας αυτή την διαδικασία ένα κείμενο γράφεται λέξη προς λέξη.

Η εκ των υστέρων πιθανότητα (posterior probability) P(c,u,z|w) δεσμευμένης της λέξης w σε μια συγκεκριμένη κοινότητα c, χρήστη u και θέμα z υπολογίζεται από τον τύπο

$$P(c, u, z, w) = P(w|z)\, P(c, u, z) = P(w|z)\, P(z|u)\, P(c,u) = P(w|z)\, P(z|u)\, P(u|c) P(c) \quad (6.6)$$

Θεωρητικά, η υπό συνθήκη πιθανότητα P(c, u, z| w) μπορεί να υπολογιστεί από την κατανομή πιθανότητας P(c, u, z, w).

Ένα πιθανό μειονέκτημα της μεθόδου που μπορεί να προκύψει είναι ότι η μέθοδος μοντελοποίησης μια κοινότητας με χρήστες θεωρεί μια κοινότητα απλώς ως μια πολυωνυμική κατανομή σε χρήστες και αυτό μειώνει την επιρροή της κοινότητας στα θέματα.

Εγγενώς μια κοινότητα σχηματίζεται διότι οι χρήστες της επικοινωνούν συχνά και μοιράζονται κοινά θέματα στις επικοινωνίες τους. Στην μέθοδο αυτή οι κοινότητες παράγουν χρήστες και τα θέματα φτιάχνονται από τους χρήστες με αποτέλεσμα να διαδίδεται μια μείωση της επιρροής μεταξύ κοινότητας και θεμάτων.

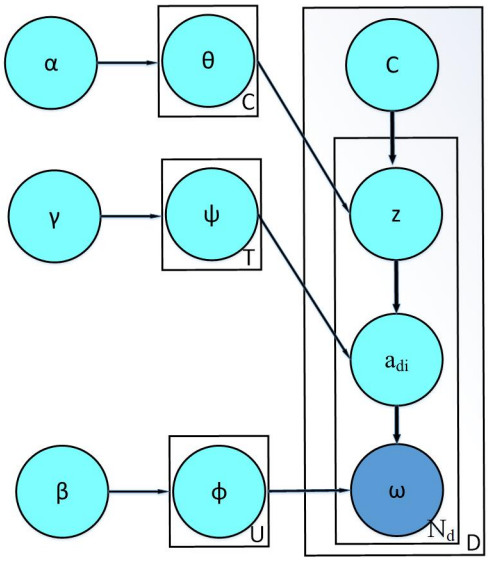
Μοντελοποιώντας μια κοινότητα με θέματα

Αυτό το μοντέλο θεωρεί ότι μια κοινότητα ενός κοινωνικού δικτύου αποτελείται από ένα σύνολο θεμάτων που αντιστοιχούν σε ομάδες χρηστών.

Κάθε λέξη w σ' ένα κείμενα d επιλέγεται από μία πολυωνυμική κατανομή σ' ένα χρήστη adi η οποία προέρχεται από την λίστα αυτών που θα διαβάσουν το κείμενο d. Πριν από αυτό ο χρήστης adi έχει προκύψει από ένα θέμα z και το z από την κατανομή της κοινότητας σε θέματα.

Η μέθοδος αυτή παράγει ένα σύνολο από υπό συνθήκη πιθανότητες P(z|c) οι οποίες καθορίζουν ποια από τα θέματα είναι πιο πιθανό να συζητηθούν στην κοινότητα c. Έχοντας μια ομάδα θεμάτων όπου η κοινότητα c συσχετίζεται με κάθε θέμα z, οι χρήστες που αναφέρονται στο θέμα z μπορούν να βρεθούν υπολογίζοντας την πιθανότητα P(u|z)

Η μοντελοποίηση κοινότητας με θέματα διαφέρει από την μοντελοποίηση κοινότητας με χρήστες στο ότι κάνει πιο δυνατές τις σχέσεις μεταξύ κοινότητας και θέματος διότι η σημασιολογία έχει έναν πιο σημαντικό ρόλο στην ανακάλυψη κοινοτήτων. Όμως μειονεκτεί ως το ότι κάνει πιο αδύναμες τις σχέσεις μεταξύ κοινότητας και χρηστών.



Σχήμα 6.9 Μοντελοποίηση κοινότητας με θέματα

Στο σχήμα 6.9 βλέπουμε την μοντελοποίηση κοινότητας με θέματα. Παρατηρούμε πως η κατανομή του κειμένου σε θέματα παράγει ένα συγκεκριμένο θέμα z το οποίο σε συνδυασμό με την πολυωνυμική κατανομή των χρηστών για κάθε κοινότητα ψ παράγει τον χρήστη a_{di}

Η πιθανότητα της κατανομής είναι

$$P(c, u, z, w) = P(w|u)P(u|z)P(z|c)P(c) \quad (6.7)$$

Στην υπό συνθήκη πιθανότητα P(c, u, z|w) μια λέξη w συσχετίζεται με τρεις μεταβλητές την κοινότητα c, τον χρήστη u και το θέμα z και δίνεται από τον τύπο

$$P(c, u, z|w) = \frac{P(c,\ u,\ z, w)}{\sum_{c,u,z} P(c, u, z, w)} \quad (6.8)$$

Ο οποίος είναι δύσκολο και μη αποδοτικό να υπολογιστεί για αυτό καταφεύγουμε στην δειγματοληψία Gibbs.

Δειγματοληψία Gibbs

Η δειγματοληψία Gibbs είναι ένας αλγόριθμος που προσεγγίζει την κατανομή πολλών μεταβλητών από μια σειρά δειγμάτων. Η δειγματοληψία Gibbs είναι ένας Μαρκοβιανής αλυσίδας Monte Carlo αλγόριθμος που χρησιμοποιείται όταν η υπό συνθήκη κατανομή πιθανότητας μπορεί να υπολογιστεί.

Ο αλγόριθμος διατρέχει όλα τα κείμενα λέξη προς λέξη και για κάθε λέξη wi, το θέμα zi και ο συγγραφέας xi συσχετίζονται με αυτή την λέξη με βάση την εκ τον υστέρων υπό συνθήκη πιθανότητα P(zi, xi | wi, z-i, x-i, w-i, ad). Τα zi και xi δηλώνουν το θέμα και τον συγγραφέα που εκχωρεί την λέξη wi. Ενώ τα z-i και x-i δηλώνουν όλες τις άλλες καταχωρήσεις των θεμάτων και των συγγραφέων εκτός από το τρέχον στιγμιότυπο. Το w-i δηλώνει όλες τις άλλες λέξεις που παρατηρήθηκαν στο σύνολο κειμένων και το ad είναι το σύνολο συγγραφέων για αυτό το κείμενο.

Η διατίμηση της υπό συνθήκης εκ των υστέρων πιθανότητας στο μοντέλο Συγγραφέα - Θέματος με δοσμένο τα θέματα T και W τις λέξεις υπολογίζεται με τον ακόλουθο τύπο:

$$P(zi = j, xi = k \mid wi = m, z\text{-}i, x\text{-}i, w\text{-}i, ad) \approx P(wi = m|xi = k)\, P(xi = k|zi = j) \approx \frac{C_{mj}^{WT} + \beta}{\sum_{m'} C_{m'j}^{WT} + V\beta} \frac{C_{kj}^{AT} + \alpha}{\sum_{m'} C_{kj'}^{AT} + Ta} \qquad (6.9)$$

Όπου m΄≠ m και j΄≠j, α και β είναι οι πρότερες παράμετροι για τις λέξεις και τα θέματα. Το C_{mj}^{WT} αντιπροσωπεύει τον αριθμό των φορών που η λέξη $w_i = m$ συσχετίζεται με το θέμα $z_i = j$ και το C_{kj}^{AT} αντιπροσωπεύει τον αριθμό των φορών που ο συγγραφέας $x_i = k$ συσχετίζεται με το θέμα j.

Η μετάβαση μεταξύ των εξισώσεων έγινε με το να μην λάβουμε υπόψη μας τις μεταβλητές z_{-i}, x_{-i}, w_{-i}, a_d διότι κάθε εμφάνιση της λέξης wi θεωρείται ανεξάρτητη από τις άλλες λέξεις στο κείμενο.

Με την εφαρμογή της δειγματοληψίας του Gibbs είμαστε σε θέση να ανακαλύψουμε τις σημασιολογικές κοινότητες εφαρμόζοντας τα μοντέλα που αναπτύξαμε πιο πάνω, κοινότητα με θέματα και κοινότητα με χρήστες. Με την πιθανότητα P(c, u, z| w) όπου οι τρεις μεταβλητές κοινότητα c, χρήστης u, θέμα z σχετίζονται με την λέξη w μπορούμε να χαρακτηρίσουμε μια κοινότητα με τα θέματα και τους σχετικούς χρήστες της. Οπότε το πρόβλημα της ανακάλυψης της σημασιολογικής κοινότητας περιορίζεται στον υπολογισμό της πιθανότητας P(c, u, z| w)

Παρακάτω παραθέτουμε τον αλγόριθμο δειγματοληψίας Gibbs

	Αλγόριθμος 6.5
1:	Για κάθε κείμενο d.
2:	Για κάθε λέξη w_i στο κείμενο d.
3:	Τοποθετούμε την λέξη w_i σε μια τυχαία κοινότητα, θέμα και χρήστη.
4:	i ← 0
5:	I ← Πλήθος επιθυμητών επαναλήψεων
6:	Ενώ I < I
7:	Για κάθε κείμενο d.
8:	Για κάθε λέξη $w_i \in$ d.

9: Υπολογίζουμε το $P(c_i, u_i, z_i| w_i)$, $u \in a_d$.

10: $(p,q,r) \leftarrow \text{argmax}(P(c_p, u_p, z_r| w_i))$

/*Εναποθέτουμε στην κοινότητα p, τον χρήστη q και το θέμα r την λέξη w_i*/

11: Καταγράφουμε το $(c_p, u_p, z_r| w_i)$

12: i++

Παρατηρούμε ότι έχοντας ως είσοδο το σύνολο των χρηστών U, το σύνολο των κειμένων D, το σύνολο των επιθυμητών θεμάτων |T|, τον αριθμό των επιθυμητών κοινοτήτων |C| ο αλγόριθμος ξεκινά με μια τυχαία καταχώρηση λέξεων σε μια κοινότητα, χρήστη, θέμα. Μια Μαρκοβιανή αλυσίδα κατασκευάζεται για να συγκλίνει στην επιθυμητή κατανομή. Σε κάθε δοκιμή αυτής της Monte Carlo προσομοίωσης ένα μπλοκ από (κοινότητα, χρήστη, θέμα) αποδίδεται στην λέξη w_i. Έπειτα από μια σειρά στάδια στην αλυσίδα η αρχική κατανομή P(c, u, z| w) προσεγγίζει την επιθυμητή κατανομή.

Για να προσαρμόσουμε την δειγματοληψία του Gibbs στα δύο προηγούμενα μοντέλα το σημαντικό είναι να υπολογίσουμε την πιθανότητα $P(c_i, u_i, z_i| w_i)$. Θα περιγράψουμε την μέθοδο που χρησιμοποιούμε και για τα δύο μοντέλα.

Έστω $P(c_i=p, u_i=q, z_i=r \mid w_i=m, z_{-i}, x_{-i}, w_{-i})$ είναι η πιθανότητα ότι η λέξη w_i να παράγεται από την κοινότητα p, τον χρήστη q και το θέμα r η οποία εξαρτάται από όλες τις καταχωρήσεις λέξεων, εκτός από την τωρινή παρατήρηση των w_i, z_{-i}, x_{-i}, και w_{-i} τα οποία αναπαριστούν όλα τα θέματα, χρήστες και λέξεις που δεν περικλείουν την τρέχουσα καταχώρηση της λέξης w_i.

Στο μοντέλο της κοινότητας με χρήστες συνδυάζουμε την εξίσωση 6.8 με την εξίσωση 6.9 και παίρνουμε την εξίσωση 6.10

$$P(ci = p,ui = q,zi = r|wi = m,z\text{-}i,x\text{-}i,w\text{-}i) \approx P(wi = m|zi = r)\, P(zi = r|ui = q)\, P(ui = q|ci = p) \approx$$

$$\frac{C_{mr}^{WT} + \beta}{\sum_{m'} C_{m'r}^{WT} + V\beta} \frac{C_{rq}^{TU} + \alpha}{\sum_{r'} C_{rq}^{TU} + Ta} \frac{C_{qp}^{UC} + \gamma}{\sum_{q'} C_{q'p}^{UC} + U\gamma} \quad (6.10)$$

Από όπου τελικά προκύπτει.

$$P(wi = m|zi = r) \approx \frac{C_{mr}^{WT} + \beta}{\sum_{m'} C_{m'r}^{WT} + V\beta} \quad (6.11)$$

$$P(zi = r|ui = q) \approx \frac{C_{rq}^{TU} + \alpha}{\sum_{r'} C_{rq}^{TU} + Ta} \quad (6.12)$$

$$P(u_i = q|c_i = p) \approx \frac{C_{qp}^{UC} + \gamma}{\sum_{q'} C_{q'p}^{UC} + U\gamma} \quad (6.13)$$

Στις εξισώσεις από πάνω το C_{mr}^{WT} είναι ο αριθμός των φορών που η λέξη $w_i = m$ συσχετίστηκε με το θέμα $z_i = r$ χωρίς να υπολογίζει την τρέχουσα φορά. Το C_{rq}^{TU} είναι ο αριθμός των φορών που ένα θέμα z=r συσχετίζεται με τον χρήστη u = q και C_{qp}^{UC} είναι ο αριθμός των φορών που ο χρήστης u = q ανήκει στην κοινότητα c = p και τα δύο χωρίς να περιλαμβάνουν το τρέχον στιγμιότυπο. C είναι ο αριθμός των κοινοτήτων στο κοινωνικό δίκτυο και είναι μια παράμετρος που την δίνουμε εμείς.

Στην μέθοδο αυτή έχουμε τρεις πίνακες τους C_{mr}^{WT}, C_{rq}^{TU}, C_{qp}^{UC}. Ο πίνακας C_{mr}^{WT}έχει μέγεθος W x T και για κάθε κελί C_{ij}^{WT} έχουμε καταχωρήσει τον αριθμό των φορών που η λέξη i έχει συσχετιστεί με το θέμα j . Παρόμοια έχουμε και για τους υπόλοιπους δύο πίνακες.

Στην μοντελοποίηση κοινοτήτων με θέματα η πιθανότητα $P(c_i=p, u_i=q, z_i=r \mid w_i=m, z_{-i}, x_{-i}, w_{-i})$ υπολογίζεται αντίστοιχα ως

$$P(c = p, u = q, z = r \mid w_i = m, z_{-i}, x_{-i}, w_{-i}) \approx \frac{C_{mq}^{WU} + \beta}{\sum_{m'} C_{m'r}^{WT} + V\beta} \frac{C_{qr}^{UT} + \gamma}{\sum_{q'} C_{q'r}^{UT} + U\gamma} \frac{C_{rp}^{TC} + \alpha}{\sum_{r'} C_{r'p}^{TC} + T\alpha} \quad (6.14)$$

όπου πάλι έχουμε να υπολογίσουμε τρεις δισδιάστατους πίνακες C^{WU}, C^{UT}, C^{TC} από τα συνεχόμενα στάδια της Μαρκοβιανής αλυσίδας και οι σημασιολογικές κοινότητες μπορούν να προκύψουν από ετικέτες που αντιστοιχούν στα θέματα.

Κλείνοντας αυτή την ενότητα να διευκρινίσουμε ότι οι χρήστες που ανήκουν σε κάθε κοινότητα c είναι αυτοί που η συνάρτηση πιθανότητας P(u|c) είναι μέγιστη σύμφωνα με τον πίνακα C^{UC}. Έπειτα τα θέματα με τα οποία ασχολούνται αυτοί οι χρήστες τα βρίσκουμε αντίστοιχα από τον πίνακα C^{TU} και μια εξήγηση για το κάθε θέμα μπορεί να προκύψει από τον πίνακα C^{WT} όπου κάθε θέμα δείχνει με ποιες λέξεις συσχετίζεται.

Πολυπλοκότητα της μεθόδου και περαιτέρω βελτιώσεις

Η δειγματοληψία Gibbs έχει δυστυχώς μεγάλη πολυπλοκότητα της τάξεως του $O(I \cdot N \cdot (U \cdot C \cdot T))$ όπου N είναι το πλήθος των λέξεων, U το πλήθος των χρηστών, C το πλήθος των κοινοτήτων, T το πλήθος των θεμάτων και I το πλήθος των επαναλήψεων που θα χρειαστεί να κάνει ο αλγόριθμος.

Αν θεωρήσουμε ένα πιθανό σενάριο όπου έχουμε 106 λέξεις, 150 χρήστες, 10 κοινότητες, 20 θέματα και θα εκτελέσουμε 1000 επαναλήψεις συνολικά θα εκτελεστούν $3 \cdot 1013$ υπολογισμοί.

Μια βασική βελτίωση της μεθόδου που πάντοτε ακολουθείται όταν επεξεργαζόμαστε σημασιολογικά διάφορα κείμενα είναι η αφαίρεση των stop words δηλαδή λέξεων όπως “και”, “όταν” “ίσως” “θα” που δεν προσφέρουν σημασιολογική πληροφορία για το τι διαπραγματεύεται ο χρήστης. Επίσης υπάρχει μια βελτιωμένη έκδοση του αλγορίθμου, ο αλγόριθμος της EnF-Gibbs δειγματοληψίας [107] που μειώνει περαιτέρω την αποδοτικότητα αφαιρώντας πολλές λέξεις που δεν προσφέρουν πληροφορία μέσω κάποιων μέτρων εντροπίας.

6.3.Αναγνώριση Κοινοτήτων σε Δυναμικά-Συνεχώς Εξελισσόμενα Κοινωνικά Δίκτυα

Στις προηγούμενες παραγράφους αντιλαμβανόμασταν και διαχειριζόμασταν τις κοινότητες ενός κοινωνικού δικτύου ως κάτι στατικό που από την στιγμή που δημιουργήθηκαν τις θεωρούμε σταθερές καθ' όλη την διάρκεια που έχουμε το κοινωνικό δίκτυο. Οι κοινότητες παράγονταν είτε από το πλήθος των δεδομένων που έχουν μαζευτεί για το κοινωνικό δίκτυο μέχρι εκείνη την στιγμή, είτε από ένα στιγμιότυπο που δείχνει πως είναι το κοινωνικό δίκτυο την στιγμή που αποφασίσαμε να βρούμε τις κοινότητες που το συνιστούν.

Από ότι καταλαβαίνουμε με τους παραπάνω τρόπους δεν έχουμε την καλύτερη εικόνα για τις κοινότητες που συνιστούν ένα κοινωνικό δίκτυο. Οι κοινότητες, οι ομάδες που συνιστούν τα άτομα είναι κάτι δυναμικό που με την πάροδο του χρόνου εξελίσσεται και ολοένα καινούριες δημιουργούνται από τις ήδη υπάρχουσες ή οι παλιές σπάνε σε δύο ή περισσότερες.

Παρακάτω θα περιγράψουμε μεθόδους που έχουν την ικανότητα να αντιλαμβάνεται τον σχηματισμό και την εξέλιξη των κοινοτήτων με την πάροδο του χρόνου. Οι πρώτες μελέτες [108] [109] [110] [111] [112] [113] [114] [115] που λάμβαναν υπόψη τους την χρονική εξέλιξη των κοινωνικών δικτύων βασίζονται στην ιδέα να ανιχνεύονται οι κοινότητες για κάθε μονάδα χρόνου. Έπειτα να συγκρίνονται οι κοινότητες που δημιουργήθηκαν για κάθε χρονική μονάδα ούτως ώστε να βρούμε τις διαφορές και την εξέλιξη τους. Δηλαδή είχαν το χαρακτηριστικό να παίρνουν υπόψη τους την χρονική εξέλιξη. Ο σχηματισμός των κοινοτήτων και η εξέλιξη τους όμως ήταν κάτι που το μελετούσαν ξεχωριστά.

Σε αυτούς τους αλγόριθμους η δομή των κοινοτήτων εξάγεται ανεξάρτητα σε συνεχόμενα χρονικά βήματα και έπειτα αναδρομικά εξετάζονται τα στοιχεία που φανερώνουν την εξέλιξη της κάθε κοινότητας με σκοπό να εντοπίσουμε τις διαφορές στη δομή των κοινοτήτων με την πάροδο του χρόνου. Αυτή η προσέγγιση του θέματος σε δύο στάδια μπορεί να εφαρμοστεί μόνο όταν η δομή των κοινοτήτων είναι σαφής και δεν υπάρχει "θόρυβος" στα δεδομένα. Διαφορετικά θα προκύψουν μεγάλες χρονικές μεταβολές.

Ο αλγόριθμος FacetNet [116] που προτάθηκε από τους Yu-Ru Lin, Yun Chi, Shenghuo Zhu, Hari Sundaram και Belle L. Tseng αναλύει τις κοινότητες και την εξέλιξη τους σ' ένα ενιαίο πλαίσιο όπου η τωρινή δομή των κοινοτήτων είναι η βάση για την μελλοντική εξέλιξη τους. Οπότε το πώς θα είναι η κοινότητα σε κάθε χρονικό στάδιο βασίζεται στο πώς ήταν η μορφή των κοινοτήτων στο προηγούμενο.

Σε αυτήν την προσέγγιση αν για ένα μικρό χρονικό διάστημα υπάρξει μια μεγάλη αλλαγή στις κοινότητες αυτή δεν θα επηρεάσει σε μεγάλο βαθμό την διαμέριση και μορφοποίηση που έχουμε κάνει. Μόνο αν αυτή η αλλαγή συνεχίσει να υπάρχει και σε επόμενα στάδια θα διαμορφώσει διαφορετικά την δομή των κοινοτήτων.

Μία επίσης σημαντική βελτίωση και διαφορά που έχει ο αλγόριθμος FacetNet είναι ότι κάθε χρήστης μπορεί να ανήκει σε περισσότερες από μια κοινότητες την ίδια στιγμή. Σε αντίθεση με τις περισσότερες άλλες μεθόδους που απλά κατέτασσαν κάθε χρήστη σε μια κοινότητα. Αυτό είναι πολύ λογικό αν αναλογιστούμε ότι ένας χρήστης μπορεί να ενδιαφέρεται για τον ραδιοερασιτεχνισμό αλλά εξίσου και για την γλυπτική οπότε θα έπρεπε να συσχετίζεται και με τις δύο κοινότητες ή ακόμη και με περισσότερες. Σε αυτή την μέθοδο ένας αριθμός θα προσδιορίζει σε τι βαθμό σχετίζεται κάθε χρήστης με την κάθε κοινότητα. Για αυτό εισαγάγουμε τον όρο soft community membership (μαλακή συμμετοχή σε κοινότητα) που προσδιορίζει αυτό τον αριθμό.

Συμβολισμοί για τον αλγόριθμο FacetNet

Στην συνέχεια αυτής της ενότητας χρησιμοποιούμε μικρά γράμματα x για να αναπαραστήσουμε βαθμωτά μεγέθη, διανυσματικής μορφής γράμματα $\vec{u}$ για να αναπαραστήσουμε διανύσματα, κεφαλαία γράμματα W για πίνακες, κεφαλαία άλλα και μικρά γράμματα με δείκτες w_{ij}, W_{ij} για στοιχεία πίνακα που βρίσκονται στην

i γραμμή και στην j στήλη, vec(W) για την διανυσματοποίηση του πίνακα W. Ο δείχτης t στης μεταβλητές Wt και wt;ij δηλώνει την τιμή της μεταβλητής την χρονική στιγμή t. Τέλος όλες οι ακμές του γράφου συνοδεύονται με τον δείκτη i για να προσδιορίσουν την χρονική στιγμή που αντιστοιχούν.

Ένα στιγμιότυπο του γράφου την χρονική στιγμή t το απεικονίζουμε ως $G_t(V_t,E_t)$. Κάθε ακμή $e_{ij} \in E_t$ δηλώνει την αλληλεπίδραση του κόμβου $v_i \in V_t$ με τον κόμβο v_j. Αν υποθέσουμε ότι ο γράφος Gt έχει n κόμβους ορίζουμε τον πίνακα $W \in R_+^{nxn}$ (ο οποίος είναι της μορφής W_t) για να αναπαραστήσουμε την ομοιότητα μεταξύ δύο κόμβων του γράφου G_t. Κάθε κελί του w_{ij} παίρνει μια τιμή που υποδηλώνει την σχέση μεταξύ του κόμβου vi (από τον πρώτο γράφο) με τον κόμβο v_j (από τον δεύτερο γράφο). Αν υπάρχει ομοιότητα μεταξύ των δύο κόμβων $w_{ij} > 0$ αλλιώς $w_{ij} = 0$ και ισχύει ότι $\sum_{i,j} w_{ij} = 1$. Με την πάροδο του χρόνου η ιστορία της αλληλεπίδρασης των κοινοτήτων στον γράφο φαίνεται από μια σειρά στιγμιότυπων που η κάθε μια αντιστοιχεί σ' ένα γράφο $<G_1,... ,G_t, ..>$

Συνάρτηση Κόστους μεταξύ διαδοχικών γράφων ενός εξελισσόμενου κοινωνικού δικτύου

Το ότι μπορούμε να αναλύσουμε τις κοινότητες και την εξέλιξη τους σε μία ενιαία διαδικασία βασίζεται στην παρακάτω σχέση κόστους που μετρά την ποιότητα της δομής των κοινοτήτων την χρονική στιγμή t. Η σχέση 6.14 παίρνει υπόψη της δύο παράγοντες ένα κόστος στιγμιότυπου και ένα χρονικό κόστος. Με αυτό τον τρόπο σκοπεύουμε να χρησιμοποιήσουμε την δομή της κοινότητας την χρονική στιγμή t - 1 για να κανονικοποιήσουμε την μορφή της κοινότητας την χρονική στιγμή t. Η συνάρτηση κόστους είναι η εξής:

$$cost = a \cdot CS + (1 - a) \cdot CT \quad (6.15)$$

Αυτή την συνάρτηση την εισήγαγε και χρησιμοποίησε ο Chakrabarti στην μελέτη του πάνω στην εξελισσόμενη συσταδοποίηση δεδομένων [117] και αποτελείται βασικά από δύο όρους. Το snapshot cost CS (κόστος στιγμιότυπου) που δηλώνει πόσο καλά η δομή των κοινοτήτων ταιριάζει με τις αλληλεπιδράσεις τον κόμβων που έχουμε την χρονική στιγμή t. Αυτό φαίνεται από τον πίνακα W. Και από το temporal cost CT (χρονικό κόστος) που δηλώνει πόσο συνεπής είναι η τρέχουσα δομή κοινοτήτων σε σχέση με τις προηγούμενες δομές κοινοτήτων. Η παράμετρος $a \in (0, 1)$ ορίζεται από αυτόν που θα εκτελέσει την μέθοδο αυτή για να επιλέξει σε ποιον από τους δύο παράγοντες θα δώσει πιο πολύ έμφαση.

Ένα μεγάλο a σημαίνει πως η μέθοδος μας βασίζεται περισσότερο στο πώς είναι ο κοινωνικός γράφος στο στιγμιότυπο t για να σχηματίσουμε τις κοινότητες (στο παρόν). Ένα μικρό a σημαίνει πώς η μέθοδος μας βασίζεται περισσότερο στο πώς ήταν σχηματισμένες οι κοινότητες στα προηγούμενα χρονικά στιγμιότυπα (στο παρελθόν). Αλλά ας δούμε κάθε ένας από αυτούς τους δύο όρους CS και CT πώς προκύπτει.

Η δομή των κοινοτήτων την χρονική στιγμή t θα πρέπει να ταιριάζει με τα δεδομένα του πίνακα W που δηλώνει την αλληλεπίδραση - ομοιότητα των κόμβων την χρονική στιγμή t. Αρχικά θα δούμε πώς μοντελοποιούμε την δομή των κοινοτήτων και έπειτα πώς ορίζουμε το κόστος στιγμιότυπου.

Έστω ότι υπάρχουν m κοινότητες την χρονική στιγμή t, Το στοιχείο w_{ij} του πίνακα W δηλώνει την ομοιότητα και αλληλεπίδραση μεταξύ των κόμβων u_i και u_j και μια συνδυασμένη δράση σε σχέση με όλες τις m κοινότητες. Ο υπολογισμός της τιμής w_{ij} προκύπτει από την ακόλουθη σχέση.

$$\text{wij} \approx \sum_{k=1}^{m} p_k \cdot p_{k\to i} \cdot p_{k\to j} \quad (6.16)$$

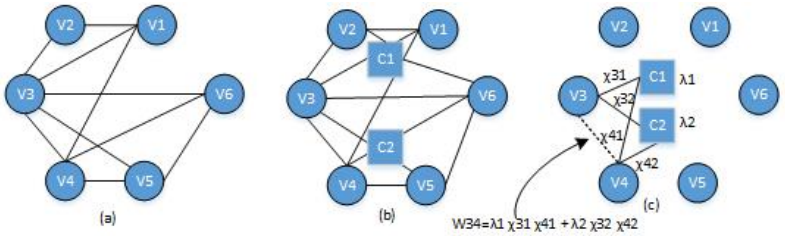
Όπου pk είναι η πρότερη πιθανότητα ότι η αλληλεπίδραση w_{ij} προκύπτει από την k κοινότητα και οι $p_{k\to i}$ και $p_{k\to j}$ είναι οι πιθανότητες ότι μια αλληλεπίδραση στην κοινότητα k περιλαμβάνει τους κόμβους u_i και u_j αντίστοιχα.

Συνδυάζοντας τα w_{ij} έχουμε τον πίνακα W όπου

$$W \approx X\Lambda X^T \quad (6.17)$$

Όπου $X \in R_+^{nxm}$ είναι ένας μη αρνητικός πίνακας όπου $x_{ik} = p_{k\to i}$ και $\sum_i x_{ik} = 1$. Ο πίνακας Λ είναι ένας m x m μη αρνητικός διαγώνιος πίνακας με $\lambda_{kk} = \lambda_k = p_k$.

Οι πίνακες X και Λ και αντίστοιχα το γινόμενο τους ΧΛ χαρακτηρίζουν πλήρως την δομή της κοινότητας. Αυτό το μοντέλο χρησιμοποιήθηκε πρώτη φορά στην δημοσίευση Soft Clustering on Graphs[118]



Σχήμα 6.10 Σχηματική αναπαράσταση της "μαλακής" συμμετοχή σε κοινότητες

(a) Ο αυθεντικός γράφος, (b) Ο διμερής γράφος με δύο κοινότητες (c) Υπολογισμός της ακμής w_{34}

Στο σχήμα 6.10 (b) δείχνουμε πως από τον αρχικό γράφο των χρηστών (a) προσθέτουμε δύο ακόμη κόμβους c_1 και c_2 που αντιστοιχούν στις κοινότητες και έχουμε ένα διμερή γράφο (διότι ακμές μπορούν να υπάρξουν μεταξύ ενός χρήστη u και μιας κοινότητας c). Αυτός ο διμερής γράφος (b) αναπαριστά κάθε κόμβος χρήστη σε τι ποσοστό ανήκει σε κάθε κόμβο κοινότητας. Στο (c) δείχνουμε πως μια ακμή w_{34} παράγεται ως το άθροισμα δύο γινομένων.

Με αυτό τον τρόπο μπορούμε να προσεγγίσουμε τον πίνακα W από το γινόμενο $X\Lambda X^T$ και με βάση αυτό το μοντέλο ορίζουμε το κόστος στιγμιότυπου CS ως το λάθος που εισαγάγετε από αυτή την προσέγγιση.

$$CS = D(W||X\Lambda X^T) \quad (6.18)$$

Όπου $D(A||B) = \sum_{i,j} (a_{ij} log \frac{a_{ij}}{b_{ij}} - a_{ij} + b_{ij})$ είναι η απόκλιση (divergence)Kullback – Leibler [119] μεταξύ του πίνακα A και B. Έτσι το κόστος στιγμιότυπου CS είναι υψηλό όταν η προσέγγιση της δομής των κοινοτήτων $X\Lambda X^T$ δεν ταιριάζει καλά με τα δεδομένα που παρατηρούμε από τον πίνακα W.

Χρονικό κόστος.

Στην εξίσωση (6.15) το χρονικό κόστος CT χρησιμοποιείται για να κανονικοποιήσει την δομή των κοινοτήτων έτσι ώστε αν προκύψουν παράλογα δεδομένα, για παράδειγμα για μικρό χρονικό διάστημα μεγάλες αλλαγές στην σύσταση των κοινοτήτων, το μοντέλο μας να μην είναι επιρρεπές στο να αλλάξει πολύ τη δομή των κοινοτήτων. Θα μπορούσαμε να ορίσουμε το χρονικό κόστος CT ως την διαφορά μεταξύ της δομής των κοινοτήτων από την χρονική στιγμή t - 1 στην t. Θα ξαναχρησιμοποιήσουμε το ότι η δομή της κοινότητας μπορεί να αναπαρασταθεί από το γινόμενο ΧΛ και το ότι $Y = X_{t-1}\Lambda_{t-1}$ Το χρονικό κόστος ορίζεται ως

$$CT = D(Y|| X\Lambda) \quad (6.19)$$

Όπου D είναι η απόκλιση (divergence) Kullback – Leibler [119] όπως και πριν. Οπότε το χρονικό κόστος CT είναι υψηλό όταν υπάρχουν μεγάλες αλλαγές στην δομή των κοινοτήτων από την χρονική στιγμή t - 1 στην t.

Ενώνοντας το κόστος στιγμιότυπου και το χρονικό κόστος

Έχοντας εξηγήσει το νόημα του κόστους στιγμιότυπου και του χρονικού κόστους την συνάρτηση κόστους (6.15) μπορούμε να την ξαναδιατυπώσουμε στην πιο αναλυτική της μορφή

$$cost = a \cdot D(W|| X\Lambda XT) + (1 - a) \cdot D(Y ||X\Lambda) \quad (6.20)$$

Όπου κάθε όρος της έχει επεξηγηθεί λεπτομερώς στις προηγούμενες παραγράφους. Σκοπός μας είναι να βρούμε για κάθε χρονική στιγμή t την δομή κοινοτήτων που ελαχιστοποιεί το παραπάνω κόστος. Αυτή η εξίσωση είναι ο πυρήνας του προβλήματος βελτιστοποίησης για τον FacetNet αλγόριθμο.

Ερμηνεία της συνάρτησης κόστους

Η συνάρτηση κόστους (6.15) μπορεί να ερμηνευθεί με όρους της θεωρίας πληροφορίας (information theory). Στην θεωρία πληροφοριών η απόκλιση KL - divergence D(P||Q) [119] είναι επίσης γνωστή ως σχετική εντροπία (relative entropy) και αναπαριστά το κέρδος πληροφορίας αν χρησιμοποιήσουμε την ακριβή κατανομή P αντί για το προσεγγιστικό μοντέλο Q. Όπου το Q ουσιαστικά προσπαθεί να μοντελοποιήσει το P. Στην δική μας δομή κοινοτήτων $X\Lambda X^T$ είναι η οριακή κατανομή που εξάγεται από το διμερές μοντέλο και προσπαθεί να προσεγγίσει το W. Ως αποτέλεσμα το $D(W|| X\Lambda X^T)$ μας δίνει το κέρδος πληροφορίας ή το λάθος που εισαγάγετε από την δομή των κοινοτήτων $X\Lambda X^T$ που φτιάξαμε εμείς σε σχέση με την πραγματική κατανομή W. Ένα υψηλότερο κέρδος πληροφορίας σημαίνει ότι εισάγεται ένα μεγαλύτερο λάθος από το $X\Lambda X^T$ οπότε και ένα υψηλότερο κόστος στιγμιότυπου.

Παρόμοια στο D(Y||XΛ), το Y αναπαριστά την δομή των κοινοτήτων την χρονική στιγμή t - 1. Αν η πληροφορία που λαμβάνουμε από το ΧΛ είναι μεγαλύτερη από ότι από αυτή του $Y = X_{t-1} \cdot \Lambda_{t-1}$ τ τότε η αλλαγή της δομής των κοινοτήτων από την χρονική στιγμή t-1 στην t θα γίνει μεγαλύτερη και άρα το χρονικό κόστος CT θα μεγαλώσει.

Επαναληπτικός Αλγόριθμος

Ο αλγόριθμος μας βασίζεται σε επαναλήψεις που ενημερώνουν τους πίνακες Χ και Λ με τέτοιο τρόπο που η συνάρτηση κόστους (6.15) να μειώνεται μονότονα σε κάθε επανάληψη.

Οι ακόλουθες δύο εξισώσεις είναι αυτές που ενημερώνουν τα κελιά των πινάκων x_{ik} και λ_k, μειώνουν το αποτέλεσμα της συνάρτησης κόστους και μετά από μια σειρά επαναλήψεων συγκλίνουν σε μια βέλτιστη λύση.

$$\text{xik} \leftarrow \text{xik} \cdot 2a \cdot \sum_j \frac{w_{ij} \cdot \lambda_k \cdot x_{jk}}{(X \Lambda X^T)_{ij}} + (1-a) \cdot y_{ik} \quad (6.21)$$

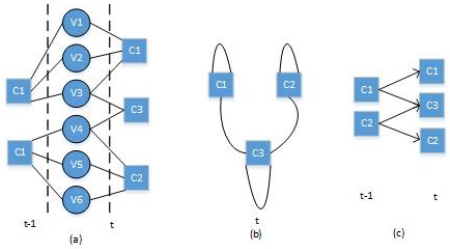
Έπειτα κανονικοποιούμε τους παράγοντες x_{ik} έτσι ώστε $\sum_i x_{ik} = 1, \forall\, k$

$$\lambda_k \leftarrow \lambda_k \cdot a \cdot \sum_{ij} \frac{w_{ij} \cdot x_{ik} \cdot x_{jk}}{(X \Lambda X^T)_{ij}} + (1-a) \cdot \sum_i y_{ik} \quad (6.22)$$

Έπειτα κανονικοποιούμε τους παράγοντες λ_k έτσι ώστε $\sum_k \lambda_k = 1$

Μέλος κοινότητας

Αρχικά υπολογίζουμε τους πίνακες X_{t-1}, Λ_{t-1} την χρονική στιγμή t - 1 και τους πίνακες X_t, Λ_t την χρονική στιγμή t. Έπειτα ορίζουμε τον διαγώνιο πίνακα Dt του οποίου τα διαγώνια στοιχεία είναι τα αθροίσματα της γραμμής $X_t\Lambda_t$ έτσι ώστε $d_{t;ii} = \sum_j (X_t \Lambda_t)_{ij}$.



Σχήμα 6.11 Αναπαράσταση των κοινοτήτων και της εξέλιξης τους

Η μαλακή συμμετοχή του κόμβου ui σε κάθε κοινότητα προσδιορίζεται από την i-οστή γραμμή του πίνακα $D_t^{-1}X_t\Lambda_t$. Δηλαδή κάθε κελί της γραμμής πίνακα αντιστοιχεί σε μια κοινότητα και η τιμή που έχει δείχνει σε τι ποσοστό ο κάθε χρήστης ανήκει σε αυτή την κοινότητα την χρονική στιγμή t.

Στο σχήμα 6.11 απεικονίζεται η αναπαράσταση των κοινοτήτων και της εξέλιξης τους: (a) Ο διμερής γράφος την χρονική στιγμή t-1 και t, (b) Το δίκτυο κοινοτήτων την χρονική στιγμή t όπως προκύπτει από τον διμερή γράφο, (c) Το δίκτυο εξέλιξης από την χρονική στιγμή t -1 στην t.

1. Δίκτυο κοινοτήτων

Η δομή των κοινοτήτων εκφράζεται από το γινόμενο των πινάκων $\Lambda_t X_t^T D_t^{-1} X_t \Lambda_t$ το οποίο δίνει την οριακή κατανομή των υπογράφων των κόμβων και είναι η δομή των κοινοτήτων που ψάχναμε. Οι επαγόμενοι υπογράφοι των κόμβων των κοινοτήτων ονομάζεται δίχτυ κοινοτήτων. Να σημειώσουμε ότι για να εξάγουμε δίχτυ κοινοτήτων κάθε κόμβος u_i συμμετάσχει σε όλες τις κοινότητες σε διαφορετικό βαθμό

Δίχτυ εξέλιξης

Το δίχτυ εξέλιξης δείχνει πως εξελίσσονται οι κοινότητες από το χρονικό στάδιο t - 1 στο t είτε με το να διαχωρίζονται είτε με το να ενώνονται. Η εξέλιξη των κοινοτήτων μπορεί να οριστεί ως η περίπτωση να ξεκινήσουμε από την κοινότητα $c_{t-1;i}$ την χρονική στιγμή t - 1, να περάσουμε μέσα από τον διμερή γράφο και να φτάσουμε την χρονική στιγμή t στην κοινότητα $c_{t;j}$. Αυτή η διαδικασία που περνάμε από κοινότητα σε κοινότητα είναι το δίχτυ εξέλιξης που παρουσιάζει την εξέλιξη των κοινοτήτων όπως φαίνεται στο σχήμα 4.2 (c). Η πιθανότητα που ενώνει τις κοινότητες $c_{t-1;i}$ και $c_{t;j}$ είναι ίση με $P(c_{t-1;i},c_{t;j}) = (\Lambda_{t-1}X_{t-1}^T D_t^{-1} X_t \Lambda_t)_{ij}$ και $P(c_{t;j}|c_{t-1;i}) = (X_{t-1}^T D_t^{-1} X_t \Lambda_t)_{ij}$. Ξανά κάθε κόμβος και κάθε ακμή συμβάλουν στην εξέλιξη από το το $c_{t-1;i}$ στο $c_{t;j}$. Αυτό σημαίνει πως όλοι οι χρήστες (κόμβοι) καθώς και όλες οι αλληλεπιδράσεις (ακμές) τους υπολογίζονται στην εξέλιξη των κοινοτήτων αλλά σε διαφορετικό βαθμό.

Πολυπλοκότητα της μεθόδου

Η αλγοριθμική μας διαδικασία βασίζεται στις εξισώσεις 6.20 και 6.21 που έχουν σκοπό να μειώσουν το κόστος της συνάρτησης 6.19. Αυτό το οποίο έχει την μεγαλύτερη χρονική πολυπλοκότητα είναι ο υπολογισμός του $(X\Lambda T^T)_{ij}$ για όλα τα $i,j \in \{1, \ldots, n\}$. Ευτυχώς δεν χρειάζεται να υπολογίσουμε το γινόμενο των πινάκων για κάθε ζευγάρι (i,j) λόγω του ότι ο πίνακας W είναι αρκετά αραιός. Το πλήθος των μη μηδενικών στοιχείων είναι ο αριθμός των ακμών του γράφου στιγμιότυπου και ορίζεται ως l. Οπότε για κάθε μη μηδενικό στοιχείο υπολογίζουμε το $(X\Lambda T^T)_{ij}$ το οποίο έχει πολυπλοκότητα O(lm). Με το m να είναι το πλήθος των κοινοτήτων.

Στην ειδική περίπτωση που το πλήθος των κοινοτήτων είναι σταθερό και ο αριθμός των κόμβων είναι μικρότερος από έναν μέγιστο αριθμό η πολυπλοκότητα μπορεί να θεωρηθεί γραμμική ως προς το πλήθος των κόμβων O(n).

6.4.Αναγνώριση των Χρηστών που Επηρεάζουν τις Κοινότητες

Κάποιοι χρήστες - κόμβοι του κοινωνικού μας γράφου παρατηρείται ότι έχουν την τάση να επηρεάζουν πιο πολύ τους υπόλοιπους χρήστες από ότι άλλοι. Σκοπός μας σε αυτή την ενότητα είναι να βρούμε ποιοι είναι αυτοί οι χρήστες.

΄Ενα κλασικό παράδειγμα που επεξηγεί την φύση του προβλήματος που θα επιλύσουμε είναι το εξής. Αν έχουμε μια εταιρία η οποία θέλει να διαφημίσει ένα προϊόν και σκοπεύει να διαθέσει α το πλήθος κομμάτια από αυτό το προϊόν σε καταναλωτές για να το διαφημίσει. Ποιοι είναι αυτοί οι α καταναλωτές που αφού δοκιμάσουν το προϊόν θα επηρεάσουν και θα ενημερώσουν μεγαλύτερο πλήθος φίλων, συγγενών και γνωστών τους για την ύπαρξη του.

Σε αυτό το σημείο θα χρειαστεί να εισάγουμε την έννοια του information cascade model (μοντέλο διάδοσης πληροφορίας καταρράκτη) που δηλώνει των τρόπο που μοντελοποιούμε την διάδοση πληροφοριών στον γράφο ενός κοινωνικού δικτύου.

Όταν ο κόμβος v ενεργοποιείται δηλαδή υιοθετεί μια συμπεριφορά που μπορεί να επηρεάσει την κοινότητα έχει μια και μοναδική πιθανότητα να επηρεάσει δηλαδή να ενεργοποιήσει κάθε έναν από τους γειτονικούς του ανενεργούς κόμβους w. Η προσπάθεια που καταβάλει κάθε κόμβος για να ενεργοποιήσει έναν γειτονικό του κόμβου συμβολίζεται με pp(v, w). Οπότε έχουμε μια σειρά σταδίων όπου στο κάθε ένα διαδοχικά κάποιοι ενεργοποιημένοι κόμβοι προσπαθούν μέσω μιας πιθανότητας να ενεργοποιήσουν τους γειτονικούς τους κόμβους. Όσοι από αυτούς ενεργοποιηθούν ξανά συνεχίζουν την ίδια διαδικασία διάδοσης της ενεργοποίησης έως ότου δεν θα υπάρχουν άλλοι κόμβοι να ενεργοποιηθούν

Διατυπώνοντας το πρόβλημα μας με μια πιο αυστηρή μορφή θα λέγαμε. Έχοντας τον γράφο G του κοινωνικού δικτύου, ένα information cascade model και έναν μικρό αριθμό k σκοπός μας είναι να βρούμε τους k το πλήθος αρχικούς κόμβους που στο εξής θα τους ονομάζουμε seeds (σπόρους) με τους όποιους θα ξεκινήσει η διάδοση πληροφορίας μέσο του cascade model έτσι ώστε στο τέλος να έχει ενεργοποιηθεί το μέγιστο πλήθος κόμβων.

Εύρεση κόμβων μέγιστης επιρροής μέσω άπληστου αλγόριθμου.

Οι Kempe, Kleinberg, και Tardos απέδειξαν ότι η βέλτιστη λύση του προβλήματος είναι NP-hard και πρότειναν έναν greedy (άπληστο) αλγόριθμο για την επίλυσή του. Η βασική ιδέα του αλγόριθμου αυτού είναι να ξεκινήσει με ένα κενό σύνολο S να κάνει k επαναλήψεις όπου σε κάθε επανάληψη να προσθέτει έναν κόμβο v ο οποίος θα μεγιστοποιεί την τιμή f(S +v) - f(S).

	Αλγόριθμος (k,f) 6.6
1:	Αρχικοποιώ S = Ø
2:	για i=1 έως k
3:	Επιλέγω u= arg $max_{w \in V \setminus S}(f(S \cup \{w\}) - f(S))$
4:	S = S ∪ {u}
5:	τέλος βρόχου για
6:	Αποτέλεσμα S

Η συνάρτηση f(S) δηλώνει τον αναμενόμενο αριθμό των κόμβων που θα έχουν ενεργοποιηθεί στο τέλος της διαδικασίας αν το αρχικό σύνολο κόμβων που είναι ενεργοποιημένο είναι το S. Η συνάρτηση f(S) είναι μη αρνητική, μονότονα αύξουσα και ικανοποιεί την ανισότητα 6.22 για όλα τα u ∈ V και για όλα τα ζευγάρια υποσυνόλων S και T με S ⊆ T ⊆ V

$$f(S \cup \{u\}) - f(S) \geq f(T \cup \{u\}) - f(T). \quad (6.23)$$

Που σημαίνει ότι το οριακό της κέρδος της f(S) αν προσθέσουμε τον κόμβο u στο σύνολο S να είναι μεγαλύτερο ή ίσο με το οριακό κέρδος της f(T) που έχουμε όταν προσθέτουμε τον κόμβο u στο Σύνολο T.

Δυστυχώς ακόμη και ο αλγόριθμος αυτός δεν είναι αποδοτικός για μεγάλα δίκτυα. Ευτυχώς προτάθηκε από τους Wei Chen, Chi Wang, Yajun Wang ένας ευρετικός αλγόριθμος [12] ο οποίος μπορεί να διαχειριστεί δίκτυα εκατομμυρίων χριστών και να βρει συνεπή αποτελέσματα. Τον αλγόριθμο αυτόν που θεωρείται ο πιο διαδεδομένος στην επίλυση του προβλήματος της αναγνώρισης των χρηστών που επηρεάζουν περισσότερο την κοινότητα θα παρουσιάσουμε πιο αναλυτικά.

Διάδοση συμπεριφοράς μέσω δέντρων επιρροής

Η βασική ιδέα της μεθόδου αυτής είναι να χρησιμοποιήσει ένα κατευθυνόμενο δέντρο (Arborescence [121]) για κάθε κόμβο του γράφου. Αυτή η δομή δεδομένων θα αναπαριστά τον τρόπο που επηρεάζονται οι κόμβοι δηλαδή πώς διαδίδεται η απόχτηση μια νέας συμπεριφοράς από τους χρήστες.

Το πρώτο πράγμα που πρέπει να γίνει για να σχηματιστούν αυτά τα δέντρα είναι να υπολογίσουμε τα maximum influence paths (μονοπάτια μέγιστης επιρροής) μεταξύ κάθε ζευγαριού κόμβων μέσω του αλγόριθμου του συντομότερου μονοπατιού του Dijkstra. Από όλα τα μονοπάτια που υπολογίσαμε θα αγνοήσουμε αυτά που η πιθανότητα τους είναι μικρότερη από ένα κατώφλι θ. Έπειτα ενώνουμε όλα αυτά τα μονοπάτια. Η ένωση θα γίνει με τέτοιο τρόπο που να ενώνουμε αναμεταξύ τους μόνο αρχικούς ή τελικούς κόμβους των μονοπατιών για να σχηματιστεί το δέντρο.

Τα μονοπάτια μέγιστης επιρροής Maximum Influence Paths στην συνέχεια θα τα αναφέρουμε ως MIPs και το μοντέλο που χρησιμοποιεί τα δέντρα μέγιστης επιρροής που προέκυψαν μετά την ένωση των μονοπατιών ως MIA και αντίστοιχα το όλο μοντέλο που περιγράφουμε ως MIA μοντέλο.

Ο αλγόριθμος του μοντέλου MIA είναι αποδοτικός γιατί βασίζεται στα εξής δύο βήματα

(i) Ο υπολογισμός της οριακής διάδοσης επιρροής υπολογίζεται με αναδρομική κλήση συναρτήσεων

(ii) Αφού επιλέξουμε έναν κόμβο με μέγιστη διάδοση επιρροής για να τον προσθέσουμε στο S το μόνο που χρειάζεται είναι να ενημερώσουμε τα δέντρα όπου σχετίζονται με τον κόμβο αυτόν που προσθέσαμε ούτως ώστε να γίνουν οι επόμενες επιλογές κόμβων.

Πιθανότητα διάδοσης και δέντρο μέγιστης διάδοσης επηρεασμού

Ας δούμε πιο αναλυτικά πώς λειτουργεί η μέθοδος αυτή. Αρχικά έχουμε όπως έχουμε πει έναν κατευθυνόμενο γράφο G=(V,E) και για κάθε ακμή $(u,v) \in E$ μια τιμή pp(u,v) που δηλώνει την πιθανότητα να διαδοθεί ο επηρεασμός από τον κόμβο u στον κόμβο v. Δηλαδή το pp(u,v) δηλώνει ποια είναι η πιθανότητα εφόσον ο κόμβος u είναι ενεργοποιημένος να επηρεάσει και να ενεργοποιήσει τον v στο επόμενο στάδιο.

Έχοντας ένα υποσύνολο των κόμβων $S \subset V$ το μοντέλο μορφής καταρράκτη λειτουργεί ως εξής. $St \subset V$ είναι το σύνολο κόμβων που είναι ενεργοποιημένοι στο στάδιο t. Σε κάθε στάδιο t+1 κάθε κόμβος $u \in St$ μπορεί να ενεργοποιήσει κάποιους από τους γειτονικούς του $u \in V \backslash U_{0 \leq i \leq t} S_i$ με μια πιθανότητα pp(u,v). Η όλη διαδικασία συνεχίζεται και κόμβοι σε κάθε χρονικό στάδιο ενεργοποιούν άλλους κόμβους που οι

τελευταίοι με την σειρά τους στο επόμενο χρονικό στάδιο θα ενεργοποιήσουν άλλους. Η όλη διαδικασία τελειώνει στο στάδιο t όταν St = ∅.

Να σημειώσουμε ότι σε κάθε κόμβο που ενεργοποιείται έχει μόνο σ' ένα στάδιο την ευκαιρία να ενεργοποιήσει τους γειτονικούς του και όχι σε επόμενα. Επίσης αφού ενεργοποιηθεί ένας κόμβος παραμένει ενεργοποιημένος καθ' όλη την διάρκεια της διαδικασίας.

Στην συνέχεια θα συμβολίσουμε με $\sigma_I(S)$ την διάδοση επηρεασμού που αναμένουμε να προκύψει αφού έχουμε ενεργοποιήσει τους αρχικούς κόμβους S. Δηλαδή το πλήθος των κόμβων που αναμένουμε να ενεργοποιηθούν

Έχοντας ως είσοδο τον αριθμό k σκοπός μας είναι να βρούμε το υποσύνολο $S^* \subset V$ τέτοιο ώστε $|S^*|=k$ και $\sigma_I(S^*)= \max\{\sigma_I(S) \mid |S| = k, S \subseteq V\}$

Για ένα μονοπάτι $P = <u = p1, p2, \dots , p_m = v>$ ορίζουμε την πιθανότητα διάδοσης όπως φαίνεται από την εξίσωση 6.24 που εκφράζει την πιθανότητα ο κόμβος u να ενεργοποιήσει τον κόμβο v μέσα από το μονοπάτι P. Η φύση του γινομένου γίνεται κατανοητή από το ότι χρειάζεται όλοι οι κόμβοι κατά μήκος του μονοπατιού να ενεργοποιηθούν.

$$pp(P) = \prod_{i=1}^{m-1} pp(p_i, p_{i+1}) \quad (6.24)$$

Ορίζουμε ως P(G,u,v) το σύνολο όλων των μονοπατιών στον γράφο G που ενώνει τον κόμβο u με τον κόμβο v και από όλα τα μονοπάτια που ενώνουν τον κόμβο u με τον v ορίζουμε ως MIP αυτό με την μεγαλύτερη τιμή pp(P)

$$MIPG(u,v)=\arg\max_{P}\{pp(P) \mid P \in P(G,u,v)\} \quad (6.25)$$

Όπως έχουμε πει τα μονοπάτια μέγιστης επιρροής μπορούν να υπολογιστούν από τον αλγόριθμο του Dijstra [122]. Έπειτα για κάθε κόμβο u ενώνουμε όλα τα μονοπάτια που ξεκινούν ή τελειώνουν από αυτό τον κόμβο και η πιθανότητα διάδοσης P(u,v) τους ξεπερνάει ένα κατώφλι θ. Το κατώφλι θ το εφαρμόζουμε για να περιορίσουμε τα μονοπάτια MIPs που έχουν μικρή πιθανότητα διάδοσης. Με αυτό τον τρόπο σχηματίζουμε για κάθε κόμβο το δέντρο μέγιστης διάδοσης επηρεασμού (MIA) που ξεκινά από τον κόμβο που εξετάζουμε προς οποιονδήποτε άλλο κόμβο.

$$MIA(u,\theta) = \bigcup_{u \in V,\, pp(MIP_G(u,v)) \geq \theta} MIP_G(u,v) \quad (6.26)$$

Διαισθητικά η συνάρτηση MIA(u,θ) δίνει τους κόμβους που επηρεάζονται από τον κόμβο u με μια πιθανότητα που να ξεπερνά το κατώφλι θ. Αναλόγως με την τιμή του θ θα έχουμε διαφορετικού μεγέθους δέντρο και πλήθος χρηστών που επηρεάζονται.

Πιθανότητα ενεργοποίησης ενός κόμβου δεδομένου του δέντρου MIA(u,θ)

Ορίζουμε ως ap(u,S,MIA(u,θ)) την πιθανότητα ο κόμβος u να ενεργοποιηθεί δεδομένου του δέντρου MIA(u,θ) που προκύπτει από ένα αρχικό σύνολο κόμβων S. Στον αλγόριθμο 6.2 δείχνουμε τον τρόπο που το ap(u,S,MIA(u,θ)) μπορεί να υπολογιστεί. ToN^{in} (u, MIA(v,θ)) είναι το σύνολο των κόμβων που είναι γείτονες του κόμβου u στο MIA(u,θ) και οι ακμές τους έχουν κατεύθυνση προς το u.

	Αλγόριθμος 6.7 ap(u,S,MIA(u,θ))
1:	Αν u ∈ S τότε
2:	ap(u) = 1
3:	αλλιώς αν $N^{in}(u) = \varnothing$ τότε
4:	ap(u) = 1
5:	αλλιώς
6:	$ap(u) = 1 - \prod_{w \in N^{in}} (1 - ap(w) \cdot pp(w,u))$
7:	τέλος αν

Διάδοση επιρροής από αρχικό πλήθος κόμβων και δέντρου MIA(u,θ)

Ορίζουμε ως σ_M την τελική διάδοση επιρροής δηλαδή το πλήθος των κόμβων που θα έχουν ενεργοποιηθεί στο τέλος με βάση τα δέντρο MIA(u,θ) και αρχικό πλήθος κόμβων S

$$\sigma_M(S) = \sum_{u \in V} ap(u, S, MIA(u, \theta)) \quad (6.27)$$

Σκοπός μας είναι να βρούμε ένα σύνολο μεγέθους k από τους κόμβους S που θα μεγιστοποιείται η συνάρτηση $\sigma_M(S)$. Δυστυχώς ο υπολογισμός k κόμβων που μεγιστοποιούν την συνάρτηση $\sigma_M(S)$ ανήκει στην κλάση πολυπλοκότητας NP-hard

Παρόλα αυτά μπορούμε να χρησιμοποιήσουμε τον greedy αλγόριθμο 6.6(k,f) που αναφέραμε στην αρχή με παράμετρο το σM αντί του φ δηλαδή αλγόριθμος 6.6(k, σM). Περαιτέρω αλγόριθμοι έχουν προταθεί που χρησιμοποιούν το λήμμα της γραμμικής επιρροής [120].

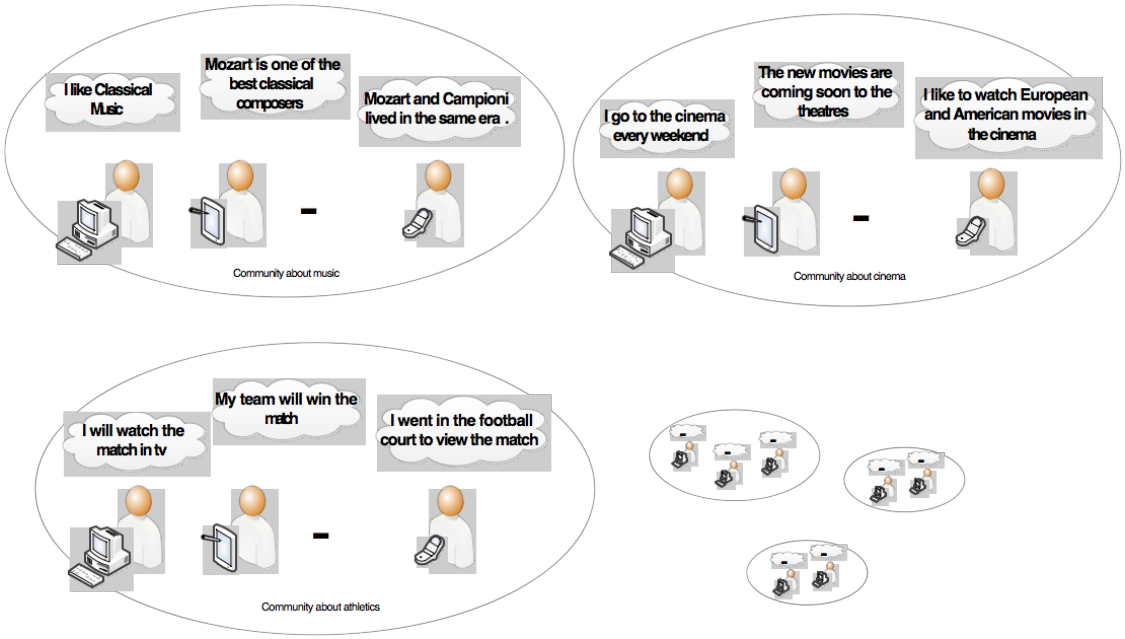
Πολυπλοκότητα της μεθόδου

Ο αναδρομικός υπολογισμός του ap(u) από τον αλγόριθμο 6.7 και την εξίσωση 6.27 έχει πολυωνυμική πολυπλοκότητα. Αν χρησιμοποιήσουμε και τον greedy αλγόριθμο 6.6 πάλι η πολυπλοκότητα της μεθόδου θα παραμείνει πολυωνυμική. Αυτό αιτιολογείται από το ότι η αναδρομή του αλγορίθμου 6.6 μπορεί να μετατραπεί σε μια επαναληπτική διαδικασία όπου οι κόμβοι u στο MIA(u,θ) μπορούν να υπολογισθούν με ένα πέρασμα του δέντρου από τα φύλλα προς την ρίζα.

6.5.Αναγνώριση Κοινοτήτων με την Κατηγοριοποίηση Κειμένων με Γράφους Ν-γραμμάτων

Στα κοινωνικά δίκτυα παρουσιάζονται ομάδες χρηστών που έχουν κοινά ενδιαφέροντα, δραστηριότητες, επικοινωνούν και δημιουργούν κοινότητες. Η κοινότητα στην οποία μπορεί να ανήκει ένας χρήστης κοινωνικού δικτύου εξαρτάται από το πεδίο εφαρμογής και τις ανάγκες κάθε εφαρμογής. Οι εφαρμογές που διαχειρίζονται κείμενα που γράφουν οι χρήστες μπορούν να χρησιμοποιήσουν τεχνικές κατηγοριοποίησης κειμένων για την αναγνώριση και σχηματισμό συγκεκριμένων θεματικών κοινοτήτων. Σε αυτό το κεφάλαιο εξετάζουμε πως η μέθοδος κατηγοριοποίησης κειμένων που κάνει χρήση του μοντέλου αναπαράστασης ΓΝΓ μπορεί να χρησιμοποιηθεί για την αναγνώριση κοινοτήτων.

Οι χρήστες κοινωνικών δικτύων σχηματίζουν θεματικές κοινότητες με βάση τα θέματα που διαπραγματεύονται στις αναρτήσεις τους, την θέση στην οποία βρίσκονται και τα ενδιαφέροντα τους. Το πρόβλημα της αναγνώρισης κοινοτήτων στα κοινωνικά δίκτυα μπορεί να αναχθεί σε ένα πρόβλημα κατηγοριοποίησης των κειμένων που γράφουν οι χρήστες όπου κάθε θεματική κοινότητα αναπαριστάται από έναν θεματικό ΓΝΓ και κάθε χρήστης από έναν ΓΝΓ που σχηματίστηκε σύμφωνα με τα κείμενα που δημοσίευσε. Οι γράφοι σχηματίζονται, συγκρίνονται και οι χρήστες κατηγοριοποιούνται ακολουθώντας το μοντέλο επιβλεπόμενης μηχανικής μάθησης όπως το είδαμε στην ενότητα 4.



Σχήμα 6.12 Αναγνώριση θεματικών κοινοτήτων με βάση τα κείμενα που γράφουν οι χρήστες

Για την εφαρμογή και αξιολόγηση της αναγνώρισης θεματικών κοινοτήτων μέσω της κατηγοριοποίησης κειμένων με το μοντέλο αναπαράστασης ΓΝΓ χρησιμοποιήσαμε το σύνολο δεδομένων Benevento που περιγράφεται στην επόμενη ενότητα. Επίσης αξίζει να αναφερθεί ότι το προτεινόμενο μοντέλο αναγνώρισης κοινοτήτων και το σύνολο δεδομένων Benevento χρησιμοποιήθηκαν στο ερευνητικό έργο Super.

6.5.1.Το σύνολο δεδομένων Benevento

Το σύνολο δεδομένων Benevento έχει φτιαχτεί για την εκπαίδευση και αξιολόγηση του μοντέλου αναγνώρισης κοινοτήτων που σχηματίζονται πριν και κατά τη διάρκεια μιας φυσικής καταστροφής ή έκτακτης ανάγκης. Συγκεκριμένα στις 15 Οκτωβρίου 2015 στην περιοχή Benevento της Ιταλίας υπήρξε μια

βροχόπτωση που πλημμύρισε πολλούς δρόμους και σπίτια και προκάλεσε πολλές καταστροφές σε περιουσίες πολιτών. Κατά τη διάρκεια και μετά από αυτό το γεγονός ανακτήσαμε tweets χρησιμοποιώντας τον crawler SocIos τα οποία κατηγοριοποιήθηκαν από το εξειδικευμένο προσωπικό της Υπηρεσίας Πολιτικής Προστασίας της Περιφέρειας της Καμπανίας (Civil Protection Service of Campania Region). Αυτό το σύνολο δεδομένων χρησιμοποιήθηκε για την εκπαίδευση και την πειραματική δοκιμή της μεθόδου μας.

Αυτό το σύνολο δεδομένων έχει μια ακόμη σημασία επειδή είναι η πρώτη φορά που αναλυτές κοινωνικών δικτύων έχουν διερευνήσει ποιες κοινότητες σχηματίζονται και αναρτούν κείμενα πριν, μετά και κατά τη διάρκεια μιας φυσικής καταστροφής. Στις παρακάτω παραγράφους δίνουμε κάποια στοιχεία για το περιστατικό που έλαβε μέρος στην περιοχή Benevento της Ιταλίας και για την κατασκευή του αντίστοιχου συνόλου δεδομένων.

Benevento 15 Οκτωβρίου

Πολλές περιοχές στην κεντρική Ιταλία υποβλήθηκαν σε ισχυρές καταιγίδες και έντονες βροχές στις 15 Οκτωβρίου 2015 και τις επόμενες ημέρες [123] [124]. Περισσότερο από 150 χιλιοστά βροχοπτώσεων έπεσαν μόνο σε λίγες ώρες στην περιοχή του Benevento. Οι συνέπειες των ακραίων καιρικών συνθηκών ήταν η υπερχείλιση του τοπικού ποταμού, σπίτια και δρόμοι να πλημμυρίσουν, κατολισθήσεις, δέντρα ξεριζωμένα, πέντε αναφορές θανάτου και περισσότεροι από 1.500 εργαζόμενοι στο Benevento δεν μπόρεσαν να εργαστούν αυτές τις μέρες.

Η απάντηση των αρχών ήταν άμεση. Η αστυνομία, ο στρατός, οι εργαζόμενοι πολιτικής προστασίας, οι πυροσβέστες και δεκάδες εθελοντές εργάστηκαν για να μετριάσουν τη φυσική καταστροφή. Επιπλέον, η τράπεζα της Νάπολης προσέφερε πέντε εκατομμύρια ευρώ σε οικογένειες και επιχειρήσεις που επλήγησαν από τις πλημμύρες. Οι κάτοικοι αντιμετώπισαν ακόμη μεγαλύτερα προβλήματα από τις έντονες βροχοπτώσεις, όπως για παράδειγμα διακοπές ρεύματος που τους άφησαν παγιδευμένους σε ανελκυστήρες και οδηγούς αυτοκινήτων που εγκλωβίστηκαν στα αυτοκίνητά τους από την λάσπη και την πλημμύρα. Όλα αυτά τα περιστατικά υποδεικνύουν τη μεγάλη ανάγκη άμεσης και καλής επικοινωνίας μεταξύ των μέσων ενημέρωσης, των θυμάτων και των αρχών.

Διαδικασία ανάκτησης δεδομένων

Κατά την διάρκεια των ημερών που συνέβησαν τα ακραία καιρικά φαινόμενα καθώς και τις επόμενες, δημοσιεύτηκαν πολλά tweets που αναφέρονταν στη φυσική καταστροφή και τις συνέπειές της. Το SocIoS api χρησιμοποιήθηκε για την ανάκτηση χιλιάδων από αυτά τα tweets από την 15η Οκτωβρίου 2015 έως την 5η Νοεμβρίου 2015. Η ανάκτηση των tweets έγινε με βάση τις ημέρες και τη θέση του γεγονότος, τον περιορισμό ότι το κείμενο τους να είναι στην ιταλική γλώσσα και να περιλαμβάνουν κάποιες συγκεκριμένες λέξεις-κλειδιά. Οι λέξεις-κλειδιά ορίστηκαν από την Υπηρεσία Πολιτικής Προστασίας της Περιφέρειας της Καμπανίας και αναφέρονται στον παρακάτω πίνακα.

Λέξη-κλειδί	Μετάφραση-Επεξήγηση
allertameteocam	ειδοποίηση καιρού
alluvione	πλημμύρα
benevento	μπενεβέντο
beneventorialzati	σήκω μπενεβεντο
rummo	ένα εργοστάσιο ζυμαρικών που έχει καταστραφεί από την πλημμύρα
saverummo	σώστε το rummo
savesannio	το Sannio είναι ένα όνομα που χρησιμοποιείται συχνά για την επαρχία Benevento

Πίνακας 6.1 λέξεις-κλειδιά που χρησιμοποιήθηκαν για το σύνολο δεδομένων Benevento

Από το API SocIos παρήχθησαν 121 JSON αρχεία. Κάθε ένα από αυτά τα αρχεία αφορούσε μια συγκεκριμένη ημέρα και λέξη-κλειδί. Τα αρχεία αυτά δόθηκαν στην Υπηρεσία Πολιτικής Προστασίας της Περιφέρειας της Καμπανίας για να τα κατηγοριοποιήσουν. Οι υπάλληλοι της Πολιτικής Προστασίας διάβασαν τα αρχεία, αφαίρεσαν τα μη σχετικά tweets, ανίχνευσαν τις θεματικές κοινότητες χρηστών που σχετίστηκαν με την φυσική καταστροφή και τα ταξινόμησαν. Από την μελέτη και την επεξεργασία του συνόλου δεδομένων Benevento εμφανίστηκαν οι ακόλουθες πέντε θεματικές κοινότητες.

Detected Communities	Explanation
Γενικοί χρήστες	SN χρήστες που συμμετείχαν στην εκδήλωση φυσικών καταστροφών.
Μέσα ενημέρωσης	Νέα εφημερίδων, τηλεοράσεων, καναλιών και άλλων μέσων
Μέσα μεταφοράς	Ανακοινώσεις μέσων μεταφοράς
Εθελοντές	Χρήστες SN που επιθυμούν να παρέχουν τις υπηρεσίες τους
Μη εφαρμόσιμο	Χρήστες SN που δεν ανήκουν σε κάποια συγκεκριμένη κοινότητα.

Πίνακας 6.2 Κοινότητες που ανιχνεύθηκαν

Μορφή του συνόλου δεδομένων Benevento

Το σύνολο δεδομένων Benevento περιέχει ένα σύνολο κειμένων από 2097 tweets τα οποία διανέμονται στις πέντε προαναφερόμενες κατηγορίες. Το σύνολο δεδομένων είναι ανόμοια κατανεμημένο με ορισμένες κατηγορίες να περιέχουν πολύ περισσότερα tweets από άλλες όπως φαίνεται στον πίνακα 6.3. Το χαρακτηριστικό αυτό του συνόλου δεδομένων καθιστά την διαδικασία ταξινόμησης μια ακόμη πιο δύσκολη διαδικασία επειδή οι χρήστες που δημοσιεύουν tweets και ανήκουν σε μικρές κατηγορίες τείνουν να ταξινομούνται στις μεγάλες κατηγορίες που περιέχουν πολλά tweets. Για να αντιμετωπίσουμε αυτό το πρόβλημα χρησιμοποιήσαμε κανονικοποιημένες μετρήσεις για την σύγκριση των γράφων και την ταξινόμηση των χρηστών.

	Μέσα ενημέρωσης	Μέσα μεταφοράς	Γενικοί χρήστες	Εθελοντές	Μη εφαρμόσιμο
Αριθμός Tweets	186	25	1012	297	577

Πίνακας 6.3 Διανομή tweets σε θεματικές κοινότητες

Χρήση του συνόλου δεδομένων Benevento

Το σύνολο δεδομένων Benevento χρησιμοποιήθηκε για δύο διαφορετικούς σκοπούς. Ο πρώτος σκοπός ήταν να κάνουμε πειράματα και να αξιολογήσουμε και να συγκρίνουμε το προτεινόμενο μοντέλο αναγνώρισης θεματικών κοινοτήτων χρησιμοποιώντας ένα πραγματικό σύνολο δεδομένων. Ο δεύτερος σκοπός ήταν να δημιουργηθεί ένα σύνολο δεδομένων που θα μπορεί να χρησιμοποιηθεί για την ανάλυση και την εκπαίδευση μοντέλων αναγνώρισης κοινοτήτων που αφορά περιστατικά έκτακτης ανάγκης και φυσικών καταστροφών.

6.5.2.Μοντέλο κατηγοριοποίησης

Στο μοντέλο που προτείνουμε, τα κείμενα που αναρτούν οι χρήστες κοινωνικών δικτύων και οι θεματικές κατηγορίες μπορούν να αναπαρίστανται ως ΓΝΓ. Συγκρίνοντας τον ΓΝΓ του χρήστη με τον ΓΝΓ μιας θεματικής κατηγορίας, μπορούμε να έχουμε μια καλή εκτίμηση αν ο χρήστης διαπραγματεύεται ένα θέμα που χαρακτηρίζει μια ομάδα ανθρώπων.

Ένας ΓΝΓ είναι σε θέση να εκφράσει συνοπτικά το θέμα ενός κειμένου ή ενός συνόλου κειμένων. Οι ΓΝΓ προσφέρουν μια συνοπτική αναπαράσταση κειμένων και μπορούν να χρησιμοποιηθούν σε εφαρμογές κατηγοριοποίησης.

Το σύνολο των κειμένων που δημοσιεύει και ανταλλάσσει ένας χρήστης αναπαρίσταται ως ΓΝΓ και μπορεί να συγκριθεί με όλους τους ΓΝΓ των θεματικών κοινοτήτων. Το υψηλότερο αποτέλεσμα σύγκρισης μπορεί να υποδεικνύει την πιο σχετική κοινότητα που ενδέχεται να ανήκει ένας χρήστης. Έτσι, το πρόβλημα αναγνώρισης θεματικών κοινοτήτων ανάγεται σε ένα πρόβλημα σύγκρισης γράφων.

Η διαδικασία κατηγοριοποίησης των κειμένων που δημοσίευσαν οι χρήστες πραγματοποιήθηκε όπως περιγράφεται στην ενότητα 4 μοντέλο κατηγοριοποίησης κειμένων με Γράφους Ν-γραμμάτων. όπου προστέθηκε ένα επίπεδο παραπάνω με τους χρήστες να αναπαρίστανται από το σύνολο των κειμένων που έχουν γράψει. Η κατασκευή των γράφων, οι μετρικές σύγκρισης γράφων, οι διανυσματικοί ταξινομητές και οι μέθοδοι αξιολόγησης χαρακτηριστικών που χρησιμοποιήθηκαν περιγράφονται στην ενότητα 4. Στο επόμενο κεφάλαιο περιγράφουμε τα πειραματικά αποτελέσματα με το σύνολο δεδομένων Benevento.

6.5.3.Πειραματικά αποτελέσματα

Τα πειράματα πραγματοποιήθηκαν σε επιτραπέζιο υπολογιστή βασικών προϊόντων. Η γλώσσα προγραμματισμού Java χρησιμοποιείται για την υλοποίηση και εφαρμογή της κατασκευής γράφων, των μετρικών ομοιότητας, της επιλογής χαρακτηριστικών και της ταξινόμησης. Η αξιολόγηση πραγματοποιήθηκε χρησιμοποιώντας τη βιβλιοθήκη scikit-learn της python και εφαρμόστηκε η δεκαπλή-αναδιπλώσεων διασταυρούμενη αξιολόγηση.

Εφαρμογή του Μπεϋζιανού ταξινομητή

Το πρώτο σύνολο πειραμάτων πραγματοποιήθηκε χρησιμοποιώντας μη σταθμισμένους γράφους, με ομοιότητα περιεχομένου και την απλή μέθοδο ταξινόμησης Bayes για Ν-γραμμάρια με Ν από δύο έως δέκα. Παρατηρούμε ότι τα 3Grams και τα 4Grams έχουν τις υψηλότερες Μίκρο και Μάκρο μετρήσεις αξιολόγησης.

	Macro Precision	Macro Recall	Macro F-Measure	Micro F-Measure
2Grams	0,7027	0,7238	0,6992	0,7371
3Grams	0,7507	0,7496	0,7392	0,7852
4Grams	0,7429	0,7563	0,7377	0,7762
5Grams	0,7237	0,7486	0,7229	0,7562
6Grams	0,7102	0,7520	0,7145	0,7476
7Grams	0,6927	0,7559	0,7011	0,7338
8Grams	0,6872	0,7618	0,6974	0,7319
9Grams	0,6764	0,7566	0,6867	0,7214

10Grams	0,6717	0,7532	0,6812	0,7167

Πίνακας 6.4 Πειράματα με τον ταξινομητή Bayes

	Macro Precision	Macro Recall	Macro F-Measure	Micro F-Measure
2Grams	0,6945	0,7062	0,6870	0,7218
3Grams	0,7445	0,7399	0,7308	0,7718
4Grams	0,7348	0,7493	0,7264	0,7632
5Grams	0,7178	0,7478	0,7140	0,7480
6Grams	0,7051	0,7495	0,7041	0,7327
7Grams	0,6806	0,7467	0,6828	0,7108
8Grams	0,6652	0,7462	0,6685	0,6980
9Grams	0,6648	0,7553	0,6699	0,7126
10Grams	0,6609	0,7002	0,6419	0,6667

Πίνακας 6.5 Πειράματα με τον ταξινομητή Bayes και κριτήριο αμοιβαίας πληροφορίας

Στη συνέχεια πραγματοποιήθηκαν πειράματα με βάση το κριτήριο επιλογής χαρακτηριστικών της Αμοιβαίας Πληροφορίας (Mutual Information). Το κριτήριο αυτό συνέβαλε στο να μειωθούν οι ανάγκες των πόρων μνήμης και του χρόνου εκτέλεσης, αλλά η ακρίβεια της μεθόδου μειώθηκε επίσης.

Πειράματα με τον ταξινόμητη SVM

Η ακρίβεια της μεθόδου βελτιώνεται με τον ταξινομητή SVM. Αλλά όπως φαίνεται στον παρακάτω πίνακα δεν είναι σαφές ποιος είναι ο καλύτερος βαθμός N για τα N-γράμματα. Το υψηλότερο αποτέλεσμα Μάκρο Φ-μέτρου επιτυγχάνεται με N = 2 και Μίκρο Φ-μέτρου με N=4.

Χρησιμοποιώντας το κριτήριο της αμοιβαίας πληροφορίας παρατηρούμε ξανά μείωση σε όλες τις μετρήσεις αξιολόγησης αλλά και ταυτόχρονα υπήρξε μείωση των υπολογιστικών πόρων και του χρόνου απόκρισης.

	Macro Precision	Macro Recall	Macro F-Measure	Micro F-Measure
2Grams	0,7520	0,7864	0,7467	0,7271
3Grams	0,7054	0,8088	0,6948	0,7757
4Grams	0,7249	0,8073	0,7317	0,7929
5Grams	0,7177	0,8068	0,7275	0,7838
6Grams	0,6935	0,8098	0,7083	0,7543
7Grams	0,6737	0,8038	0,6880	0,7281
8Grams	0,6648	0,7974	0,6776	0,7152

9Grams	0,6555	0,7983	0,6672	0,7038
10Grams	0,6484	0,7981	0,6594	0,6962

Πίνακας 6.6 Πειράματα με πυρήνα ταξινόμησης SVM

	Macro Precision	Macro Recall	Macro F-Measure	Micro F-Measure
2Grams	0,8622	0,9811	0,8747	0,6958
3Grams	0,8303	0,9122	0,8126	0,7663
4Grams	0,7130	0,7946	0,7125	0,7897
5Grams	0,7117	0,7915	0,7181	0,7830
6Grams	0,6937	0,7966	0,7042	0,7625
7Grams	0,6794	0,7889	0,6914	0,7439
8Grams	0,6631	0,7989	0,6775	0,7253
9Grams	0,6475	0,7984	0,6616	0,7058
10Grams	0,6413	0,7978	0,6549	0,6981

Πίνακας 6.7 Πειράματα με τον ταξινομητή SVM και κριτήριο Αμοιβαίας Πληροφορίας

6.6.Συμπεράσματα

Η αναγνώριση κοινοτήτων σ' ένα κοινωνικό δίκτυο είναι ένα πρόβλημα που μπορεί να επιλυθεί από τρεις διαφορετικές προσεγγίσεις, τοπολογικά, σημασιολογικά και εξελικτικά. Η επιλογή της μεθόδου εξαρτάται από τις πληροφορίες που μας παρέχονται από το κοινωνικό δίκτυο και από το πώς μας ζητούν να θεωρήσουμε και να σχηματίσουμε τις κοινότητες.

Οι πληροφορίες που μας παρέχονται μπορεί να είναι απλώς σχέσεις μεταξύ χρηστών. Δηλαδή χρήστες οι οποίοι είναι φίλοι, επικοινωνούν αναμεταξύ τους, προσπελαύνουν κοινά δεδομένα χωρίς να έχουμε πρόσβαση στα δεδομένα αυτά ή σημασιολογικά δεδομένα δηλαδή κείμενα τα οποία θα διαβαστούν από ένα αυτοματοποιημένο πρόγραμμα και θα κατανοήσουν με τι θέματα ασχολείται ο συγγραφέας τους ή ο αναγνώστης του.

Το πώς θα σχηματίσουμε τις κοινότητες εξαρτάται από το αν θεωρούμε τις κοινότητες ως κάτι στατικό ή δυναμικό. Οι στατικές κοινότητες δημιουργήθηκαν κάποια στιγμή και το μόνο που μας απασχολεί είναι ποιοι χρήστες θα είναι μέλη τους. Οι δυναμικές κοινότητες εξελίσσονται με την πάροδο του χρόνου. Κοινότητες θα δημιουργούνται με την πάροδο του χρόνου που θα απεικονίζουν την εξέλιξη των ενδιαφερόντων που έχουν οι χρήστες.

Ερωτήματα που επίσης μας απασχολούν είναι αν είναι γνωστός από πριν ο αριθμός και τα θέματα των κοινοτήτων που θα έχουμε, αν θέλουμε οι χρήστες να ανήκουν σε περισσότερες από μια κοινότητες με διαφορετικό βαθμό, καθώς βέβαια και το μέγεθος των δεδομένων και η πολυπλοκότητα επίλυσης του προβλήματος.

Αναλόγως με τις απαντήσεις στα παραπάνω ερωτήματα θα πρέπει να ακολουθήσουμε μια από τις μεθόδους που αναφέραμε στις προηγούμενες ενότητες. Ευτυχώς οι αλγόριθμοι είναι αρκετά γενικοί και ευέλικτοι οπότε οι μέθοδοι μπορούν να προσαρμοστούν για να δώσουν απαντήσεις σε οποιεσδήποτε απαιτήσεις και κοινωνικό δίκτυο εφαρμοστούν.

Ένα σημαντικό πλεονέκτημα της μεθόδου ταξινόμησης με το μοντέλο αναπαράστασης ΓΝΓ για την ανίχνευση θεματικών κοινοτήτων είναι ότι δεν χρησιμοποιεί το μοντέλο του σάκου λέξεων. Οι περισσότερες από τις μεθόδους ταξινόμησης αντιπροσωπεύουν τα κείμενα ως ένα σύνολο όρων ή αναγάγουν ένα κείμενο σε ένα σύνολο από λανθάνουσες μεταβλητές. Αυτές οι μέθοδοι μπορούν να καταγράψουν ποιοι όροι περιλαμβάνονται σε κάθε κατηγορία αλλά δεν μπορούν να αποτυπώσουν την ακολουθία αυτών των όρων. Η ακολουθία των όρων στο αρχικό κείμενο είναι πληροφορία που είναι σημαντική στα κριτήρια ταξινόμησης. Το μοντέλο ταξινόμησης με ΓΝΓ χρησιμοποιεί και αποδίδει την διάταξη και την αλληλουχία των λέξεων.

Το μοντέλο ταξινόμησης με ΓΝΓ μπορεί να έχει γενική χρήση. Δεν είναι εξειδικευμένο μόνο για μηνύματα στο Twitter ή αναρτήσεις στο Facebook, αλλά μπορεί να έχει γενική χρήση σε οποιοδήποτε είδος κειμένων ή και ακόμη και μεταδεδομένα μπορούν να αναπαρασταθούν ως ΓΝΓ με την προϋπόθεση ότι τα αντικείμενα ταξινόμησης συνοδεύονται από δεδομένα κειμένου που είναι μεγαλύτερα από τρεις ή τέσσερις λέξεις.

Η υπολογιστική πολυπλοκότητα της μεθόδου είναι χαμηλή. Η πολυπλοκότητα της κατασκευής γράφων είναι O (l) όπου l είναι ο συνολικός αριθμός των γραμμάτων και η πολυπλοκότητα της μεθόδου ομοιότητας γραφημάτων είναι O ($e_1 \cdot e_2$) όπου e_1 και e_2 είναι ο αριθμός των ακμών σε δύο γράφους που συγκρίνονται. Η πολυπλοκότητα μειώνεται σε περίπτωση που αφαιρεθούν οι ακμές που δεν συμβάλλουν στην διαδικασία ταξινόμησης.

Η μερική αντιστοίχιση λέξεων δεν υποστηρίζεται στις περισσότερες μεθόδους ταξινόμησης. Εάν μια λέξη διαφέρει από μιαν άλλη μόνο σε ένα γράμμα, θα θεωρηθεί ως μια διαφορετική λέξη. Μια λέξη με ορθογραφικά λάθη μπορεί να υπάρχει κατά λάθος ή από πρόθεση (π.χ. νεολογισμός). Το. μοντέλο ΓΝΓ μπορεί να ανιχνεύσει αυτές τις μικρές διαφοροποιήσεις και να διαχειριστεί τις λέξεις με μια μερική ομοιότητα.

Η αναπαράστση με ΓΝΓ μπορεί να συμπυκνώσει το ύφος ενός συγγραφέα ή μιας ομάδας συγγραφέων. Οι λέξεις που είναι γραμμένες και η ακολουθία τους διατηρούνται στη δομή των γράφων και υποδηλώνουν ένα προσωπικό στιλ που μπορεί να συμβάλει θετικά στη διαδικασία ταξινόμησης.

Η παραλληλοποίηση του μοντέλου είναι εύκολο να επιτευχθεί όπως δείξαμε στην ενότητα 4. Κάθε αναπαράσταση ΓΝΓ μπορεί να συγκριθεί ταυτόχρονα με πολλές αναπαραστάσεις ΓΝΓ θεματικών κατηγοριών. Δεν υπάρχουν εξαρτήσεις μεταξύ των διαδικασιών σύγκρισης γράφων. Μετά από όλες τις παράλληλες συγκρίσεις, το κείμενο θα ανήκει στην κατηγορία με την υψηλότερη βαθμολογία ομοιότητας.

Η προσαρμοστικότητα της κατηγοριοποίησης κειμένων με ΓΝΓ είναι ένα σημαντικό πλεονέκτημα. Στα περισσότερα από τα προβλήματα ανίχνευσης θεματικών κοινοτήτων, οι κοινότητες εξελίσσονται συνεχώς. Νέες δημοσιεύσεις γράφονται οι οποίες είναι χρήσιμο να ενσωματωθούν στο σύνολο αναπαράστασης δεδομένων των κατηγοριών. Επίσης το προτεινόμενο μοντέλο έχει το πλεονέκτημα να ανακατασκευάσει εύκολα τους γράφους αναπαράστασης κάθε κατηγορίας προσθέτοντας μόνο μερικούς κόμβους και ακμές ή αλλάζοντας κάποια βάρη ακμών.

7. Εφαρμογή στην Ανάλυση Συναισθήματος σε Μέσα Κοινωνικών Δικτύων

Η αυτόματη και γρήγορη αξιολόγηση της συναισθηματικής θέσης έχει γίνει ένα απαραίτητο εργαλείο στον χώρο της ανάλυσης των μέσων κοινωνικών δικτύων καθώς και σε άλλους τομείς όπως το μάρκετινγκ. Η ανάλυση συναισθήματος περιλαμβάνει την αναγνώριση της συναισθηματικής θέσης που εκφράζει ο συγγραφέας ενός κειμένου. Συνήθως οι συναισθηματικές θέσεις είναι τρεις, θετική, ουδέτερη και αρνητική. Η ανάλυση συναισθήματος σε κείμενα που προέρχονται από κοινωνικά δίκτυα είναι ένα πρόβλημα που μπορεί να αναχθεί σε μια διαδικασία αυτόματης κατηγοριοποίησης κειμένων όπου η κατηγοριοποίηση των κειμένων γίνεται στις τρεις προαναφερθείσες συναισθηματικές κατηγορίες.

Το μοντέλο αναπαράστασης ΓΝΓ με χαρακτήρες έχεις εφαρμοστεί για της ανάγκες της συναισθηματικής ανάλυσης και παρουσίασε πολύ καλά αποτελέσματα. Στο παρόν κεφάλαιο θα εφαρμόσουμε και θα αξιολογήσουμε την τεχνική κατηγοριοποίησης κειμένων που κάνει χρήση μιας παραλλαγής των ΓΝΓ όπου αντί για χαρακτήρες έχουμε λέξεις.

Με παρόμοιο τρόπο βλέπουμε ότι η ακολουθία των λέξεων μπορεί να αναπαρασταθεί χρησιμοποιώντας γράφους στους οποίους εφαρμόζονται μετρήσεις ομοιότητας γράφων και αλγόριθμοι ταξινόμησης. Τα πειράματα που πραγματοποιήθηκαν με αυτήν τη μέθοδο σε ένα σύνολο δεδομένων Twitter για την πρόβλεψη συναισθημάτων επικυρώνουν το προτεινόμενο μοντέλο και μας επιτρέπουν να κατανοήσουμε περαιτέρω τις μετρήσεις και τα χαρακτηριστικά της μεθόδου.

7.1. Εισαγωγή στην Ανάλυση Συναισθήματος σε Μέσα Κοινωνικών Δικτύων

Το πρόβλημα της ανάλυσης συναισθήματος από αναρτήσεις στα μέσα κοινωνικών δικτύων περιλαμβάνει δύο σημαντικές προκλήσεις, τον περιορισμό χαρακτήρων που τίθεται σε κοινωνικά δίκτυα όπως το Twitter και η χρήση της γλώσσας που δεν τηρεί πλήρως τους γραμματικούς και συντακτικούς κανόνες. Ως αντιμετώπιση αυτών έρχεται η προτεινόμενη προσέγγιση η οποία χρησιμοποιεί μια μοντελοποίηση και συναισθηματική κατηγοριοποίηση κειμένων, η οποία ονομάζεται μέθοδος ανάλυσης συναισθημάτων με γράφους λέξεων.

Αυτή η προσέγγιση συνδυάζει τη δομή των γράφων, τις μετρήσεις ομοιότητας γράφων και μεθόδους ταξινόμησης προκειμένου να αναγνωριστεί το συναίσθημα που εκφράζει ο συντάκτης του εγγράφου προς μια συναισθηματική κατηγορία. Στο μοντέλο γράφων λέξεων οι λέξεις που περιέχονται σε ένα σύντομο κείμενο αντιπροσωπεύονται ως κόμβοι και η εγγύτητα μεταξύ των λέξεων ως ακμές.

Το μοντέλο γράφων λέξεων χρησιμοποιείται για να αντιπροσωπεύει τις συναισθηματικές κατηγορίες με βάση ένα σύνολο δεδομένων εκπαίδευσης. Στη συνέχεια εφαρμόζεται μια τεχνική ομοιότητας γράφων για την σύγκριση κάθε κειμένου με τις τρεις συναισθηματικές κατηγορίες. Τα αποτελέσματα της σύγκρισης δημιουργούν μια νέα αναπαράσταση κειμένων ως διανύσματα τριών διαστάσεων και η πρόβλεψη συναισθήματος πραγματοποιείται με τη χρήση μεθόδων ταξινόμησης διανυσμάτων.

Για την εύρεση των παραμέτρων και των τεχνικών που παράγουν τις πιο ακριβείς προβλέψεις εξετάζονται πολλές εναλλακτικές προσεγγίσεις σε κάθε στάδιο. Η καταλληλότητα και η δυνατότητα εφαρμογής της μεθόδου αποδεικνύονται με μετρήσεις που ποσοτικοποιούν την ομοιότητα των συναισθημάτων που εκφράζονται με βάση την γειτνίαση των λέξεων.

Πειράματα διεξάγονται για κάθε μία από τις παραμέτρους και τις τεχνικές που περιλαμβάνονται. Επιπλέον, για την αξιολόγηση και σύγκριση της προτεινόμενης μεθόδου πραγματοποιούνται δοκιμές με άλλες

μεθόδους συναισθηματικής ανάλυσης στο ίδιο σύνολο δεδομένων. Η σύγκριση των αποτελεσμάτων επαληθεύει το προτεινόμενο μοντέλο και την υπόθεση ότι η συναισθηματική θέση σε μεγάλο βαθμό βρίσκεται όχι μόνο στις λέξεις που χρησιμοποιούνται αλλά και στη θέση τους στο κείμενο και στην γειτνίαση τους.

7.2.Τεχνικές Συναισθηματικής ανάλυσης

Η ανάλυση συναισθήματος μπορεί να αναχθεί σε μια διαδικασία κατηγοριοποίησης κειμένων σε δύο ή τρεις κατηγορίες, ανάλογα με το αν γίνεται διάκριση μεταξύ μόνο θετικών και αρνητικών κατηγοριών ή περιλαμβάνεται και ως τρίτη η ουδέτερη. Συχνά το πρόβλημα αντιμετωπίζεται ως πρόβλημα ταξινόμησης πολλαπλών ετικετών [125] [126] [127] [77] [128], στο οποίο στα κείμενα αποδίδονται βάρη που δείχνουν το βαθμό συσχέτισης τους με την αντίστοιχη συναισθηματική κατηγορία.

Η βασική πρόκληση σε αυτά τα είδη προβλημάτων είναι η αναγνώριση των στοιχείων που χαρακτηρίζουν ένα κείμενο και τα μεταδεδομένα του. Αυτά τα δύο θα παράγουν την αναπαράσταση του κειμένου για τους σκοπούς της ταξινόμησης. Η αναπαράσταση αυτή συνήθως περνά ως είσοδος σε μια διαδικασία ταξινόμησης που απεικονίζει την αναπαράσταση σε μια συναισθηματική αποτίμηση. Η λειτουργία του ταξινομητή διαφέρει με βάση το μοντέλο με το οποίο έχει υλοποιηθεί. Τα περισσότερα μοντέλα ταξινόμησης βασίζονται σε τεχνικές μηχανικής μάθησης όπως οι Naïve Bayes, C4.5 και SVM και οι οποίες προσδιορίζουν τα όρια μεταξύ των συναισθηματικών κατηγοριών.

Η αναπαράσταση των κειμένων βασίζεται στην εξαγωγή χαρακτηριστικών που συνήθως λαμβάνουν χώρα με μια τεχνική επεξεργασίας φυσικής γλώσσας (NLP) όπως ή πιο απλά με την χρήση ενός λεξικού ([129, 130]) όπως το SentiWordNet [131] που σχεδιάστηκε ρητά για την εφαρμογή συναισθηματικής ανάλυσης και εξόρυξης γνώμης. Μια άλλη προσέγγιση που χρησιμοποιείται ευρέως στις εργασίες ανάλυσης του συναισθήματος είναι ο «σάκος λέξεων» και ή και ακόμη σάκος N-γραμμάτων [132] όπου αναπαριστούν κάθε κείμενο από ένα σύνολο λέξεων ή N-γραμμάτων αντίστοιχα.

Από την άλλη έχουν προταθεί μοντέλα που δεν είναι NLP, όπως οι προσεγγίσεις που βασίζονται σε οντολογίες [133]. Η Οντολογία χρησιμοποιείται ως ένα σύνολο αντιπροσωπευτικών οντοτήτων που μπορεί να μοντελοποιήσει ένα πεδίο γνώσης. Στη περίπτωση αυτή, η συναισθηματική ανάλυση πραγματοποιείται με βάση τις έννοιες και τις ιδιότητες που περιέχονται στην οντολογία [134] ή με τη χρήση μιας αλληλουχίας ερωταπαντήσεων που αναλύουν διαφορετικές εξαγόμενες λέξεις [135]. Αυτές οι λέξεις χαρτογραφούνται σε έννοιες, ιδιότητες και οντολογικές παρουσίες και χρησιμοποιούνται για μια σύσταση μιας οντολογίας που συνδέει τη συναισθηματική απόσταση μεταξύ ενός ερωτήματος και ενός κειμένου.

7.3.Προτεινόμενο Μοντέλο

Η μέθοδος συναισθηματικής ανάλυσης με γράφους λέξεων [136] έχει πολλά κοινά με την μέθοδο κατηγοριοποίησης κειμένων με το μοντέλο αναπαράστασης ΓΝΓ, αλλά διαφοροποιείται ως προς το ότι εδώ γίνεται χρήση λέξεων ως κόμβοι στον γράφο αναπαράστασης και ότι υπάρχουν συγκεκριμένα τρεις συναισθηματικές κατηγορίες, θετική, αρνητική και ουδέτερη και όχι ένα πλήθος θεματικών κατηγοριών.

Στο μοντέλο που προτείνουμε κάθε κείμενο αναπαρίσταται ως μη σταθμισμένος κατευθυνόμενος γράφος βάσει των λέξεων που υπάρχουν στο αρχικό κείμενο καθώς και της εγγύτητάς τους. Οι κατηγορίες συναισθημάτων (θετική, αρνητική και ουδέτερη) μπορούν επίσης να εκπροσωπούνται ως κατευθυνόμενοι μη σταθμισμένοι γράφοι που έχουν σχηματιστεί στην φάση της εκπαίδευσης. Η σύγκριση του γράφου κειμένου με τους τρεις γράφους συναισθημάτων αποδίδει έναν τρισδιάστατο διάνυσμα. Τέλος, στο αρχικό

κείμενο αποδίδεται μια συναισθηματική πολικότητα χρησιμοποιώντας έναν ταξινομητή διανυσμάτων. Κάθε ένα από αυτά τα βήματα θα περιγραφεί στις ακόλουθες παραγράφους.

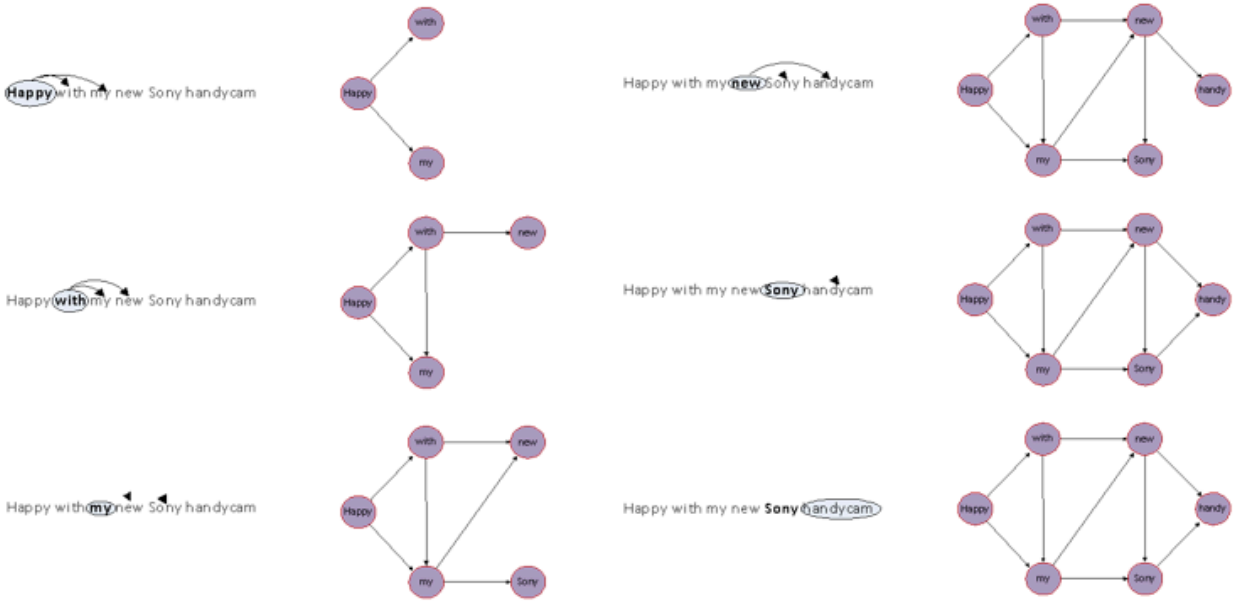
Κάθε κείμενο που θέλουμε να του αποδώσουμε την συναισθηματική πολικότητα του μετατρέπεται σε γράφο και στη συνέχεια μετράμε την "ομοιότητά" του με κάθε γράφο από τις συναισθηματικές κατηγορίες. Η ομοιότητα μεταξύ των γράφων όπως συζητήσαμε στην ενότητα 4 μπορεί να μετρηθεί με πολλές μετρικές όπως η ομοιότητα περιεχομένου, ομοιότητα αξίας, η κανονικοποιημένη ομοιότητα αξίας και τρεις διαφορετικές εκδοχές της ομοιότητας με βάση τον μέγιστο κοινό υπογράφο.

7.4.Κατασκευή Γράφων

Η κατασκευή των γράφων-λέξεων βασίζεται στις λέξεις που περιέχονται σε ένα κείμενο και στην εγγύτητά τους. Η διαδικασία είναι παρόμοια για τα κείμενα και για τις κατηγορίες των συναισθημάτων. Κάθε λέξη που περιέχεται στο αρχικό κείμενο αντιπροσωπεύεται από έναν κόμβο που έχει ως ετικέτα την λέξη αυτή. Δύο κόμβοι ενώνονται με μια ακμή αν οι αντίστοιχες λέξεις τους είναι κοντά στο αρχικό κείμενο. Οι ακμές είναι κατευθυνόμενες για να αποδίδουν την ακολουθία των λέξεων όπως υπάρχουν στα αρχικά κείμενα.

Η εγγύτητα μεταξύ των λέξεων αναπαρίσταται από τις ακμές που συνδέουν τους κόμβους και ορίζεται από έναν συγκεκριμένο αριθμό λέξεων που ακολουθούν μια λέξη. Ένα πλαίσιο λέξεων ολισθαίνει στο κείμενο και ορίζει τους κόμβους και τις ακμές του γράφου όπως απεικονίζεται στο σχήμα 7.1. Το μέγεθος του πλαισίου είναι μια παράμετρος που επηρεάζει σε μεγάλο βαθμό την ακρίβεια της μεθόδου. Όπως θα φανεί στο τμήμα αξιολόγησης, έχουμε πραγματοποιήσει πειράματα με μέγεθος πλαισίου που κυμαίνεται από 2 έως 10. Ο λόγος που το πλαίσιο δεν ξεπερνάει το όριο των 10 είναι ότι κείμενα όπως tweets περιέχουν συνήθως μια μικρή ποσότητα λέξεων είτε επειδή υπάρχει περιορισμός στον αριθμό των χαρακτήρων είτε γιατί οι χρήστες προτιμούν να εκφράζονται με συνοπτικό τρόπο.

Αν και χρησιμοποιούμε τον όρο γράφοι λέξεων οι κόμβοι του γράφου μπορεί να είναι οποιαδήποτε ακολουθία χαρακτήρων που διαχωρίζεται από δύο κενά διαστήματα. Αυτές οι ακολουθίες χαρακτήρων ακόμη και αν δεν αντιστοιχούν σε λέξεις μπορεί να εκφράζουν ένα νόημα για τους συγγραφείς και τους αναγνώστες τους.



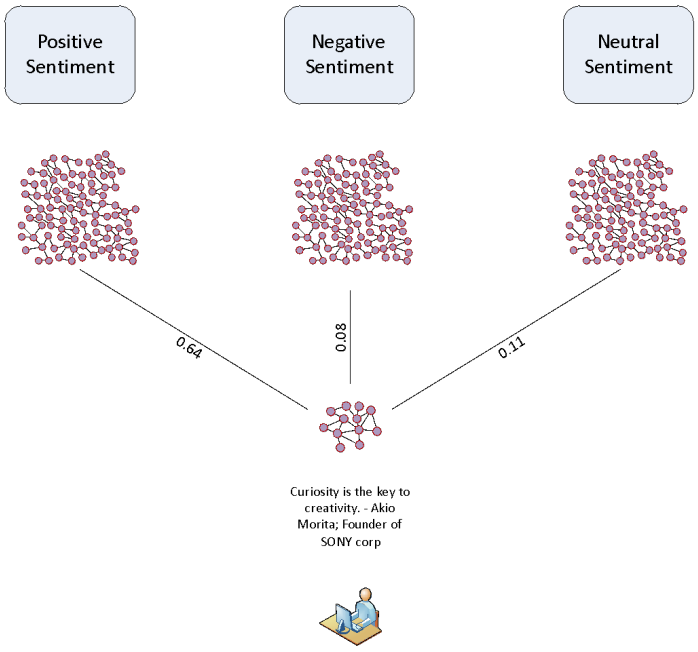
Σχήμα 7.1. Κατασκευή γράφων λέξεων

Όσον αφορά τους γράφους των τριών συναισθηματικών κατηγοριών, κάθε μία από αυτές κατασκευάζεται από το άθροισμα όλων των γράφων που αναπαριστούν τα κείμενα που ανήκουν στην αντίστοιχη κατηγορία. Δηλαδή, όλοι οι γράφοι κειμένων που εμπίπτουν σε μια ενιαία κατηγορία συγχωνεύονται για να αποτυπώσουν έναν αντιπροσωπευτικό γράφο συναισθηματικής κατηγορίας. Η συγχώνευση πραγματοποιείται με τις ακμές και τους κόμβους να διατηρούνται και να συναθροίζονται στον νέο γράφο.

Στο σχήμα 7.1 απεικονίζει τον τρόπο ολίσθησης του πλαισίου από την πρώτη λέξη προς την τελευταία του κειμένου. Σε αυτό το παράδειγμα το πλαίσιο είναι τάξης δύο όπου για κάθε λέξη προστίθενται δύο ακμές και κόμβοι στον γράφο εφόσον δεν υπάρχουν ήδη.

7.5.Ομοιότητα Γράφων

Η ομοιότητα γράφων μεταξύ του γράφου ενός κειμένου και του γράφου μιας συναισθηματικής κατηγορίας υποδεικνύει το βαθμό στο οποίο ένα κείμενο εκφράζει το αντίστοιχο συναίσθημα. Υπάρχει μια πληθώρα από μεθόδους για την εκτίμηση της ομοιότητας γράφων. Για τους σκοπούς της έρευνας μας χρησιμοποιούμε μεθόδους που έχουν ως παραμέτρους τις ετικέτες των κόμβων, την κατεύθυνση των ακμών και τον αριθμό των κοινών ακμών.



Σχήμα 7.2. Σύγκριση των γράφων λέξης ενός κειμένου με τις συναισθηματικές κατηγορίες

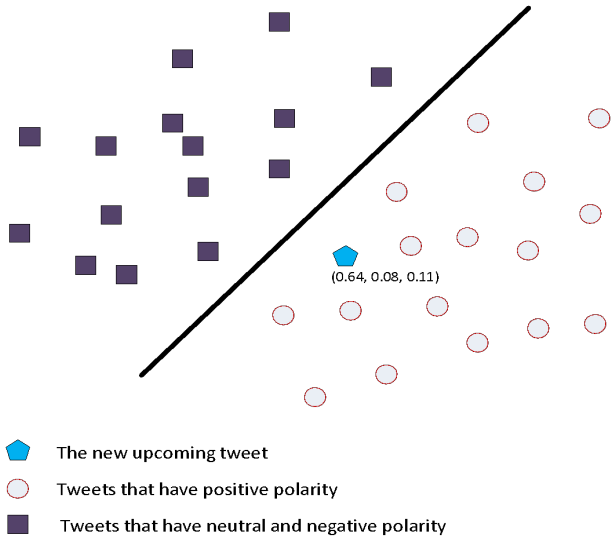
Στο σχήμα 7.2 φαίνεται ότι ένας γράφος κειμένου συγκρίνεται με τον γράφο κάθε συναισθηματικής κατηγορίας προκειμένου να παραχθούν τρεις αριθμοί ομοιότητας. Κάθε αριθμός ομοιότητας ποσοτικοποιεί την σχέση του κειμένου με τις τρεις κατηγορίες συναισθημάτων. Αυτοί οι τρεις αριθμοί σχηματίζουν το διάνυσμα που χρησιμοποιείται από ένα ταξινομητή για να κατηγοριοποιηθεί σε μια συναισθηματική διάθεση. Ως μετρικές ομοιότητας γράφων χρησιμοποιούμε της μετρικές ομοιότητας περιεχομένου,

ομοιότητας αξίας, κανονικοποιημένης ομοιότητας αξίας και τις τρεις παραλλαγές της ομοιότητας μέγιστου κοινού υπογράφου όπως περιεγράφησαν στην ενότητα 4.

7.6.Ταξινόμηση

Η προτεινόμενη μέθοδος ανάλυσης των συναισθημάτων διαχωρίζει τα δεδομένα εκπαίδευσης των κειμένων σε δύο μέρη. Το πρώτο μέρος χρησιμοποιείται για την αναπαράσταση του γράφου λέξεων για κάθε συναισθηματική κατηγορία και το δεύτερο μέρος των κειμένων χρησιμοποιείται για την εκπαίδευση ενός ταξινομητή όπως οι SVM και Bayes.

Κάθε κείμενο του δεύτερου μέρους του συνόλου δεδομένων κατάρτισης παρουσιάζεται με έναν γράφο λέξεων και συγκρίνονται με τους γράφους των τριών συναισθηματικών κατηγοριών χρησιμοποιώντας ένα από τα μέτρα ομοιότητας γράφων που περιγράφονται στην Ενότητα 4. Τα αποτελέσματα της σύγκρισης σχηματίζουν έναν τρισδιάστατο διάνυσμα. Αυτό το διάνυσμα αντιπροσωπεύει την συναισθηματική πολικότητα του κειμένου. Τα διανύσματα του δεύτερου μέρους του συνόλου δεδομένων εκπαίδευσης χρησιμοποιούνται για την εκπαίδευση του ταξινομητή.
Ένα νέο κείμενο αντιπροσωπεύεται και πάλι ως γράφος λέξεων και συγκρίνεται με τους γράφους λέξεων των τριών συναισθηματικών κατηγοριών με αντίστοιχο τρόπο όπως τα κείμενα του δεύτερου μέρους του συνόλου δεδομένων εκπαίδευσης. Με αυτό τον τρόπο δημιουργείται ένα διάνυσμα το οποίο κατηγοριοποιείται από τον εκπαιδευμένο ταξινομητή σε μία από τις τρεις κατηγορίες που αντιπροσωπεύουν τις τάξεις συναισθημάτων.



Σχήμα 7.3.Καθορισμός ενός νέου κειμένου στο μοντέλο κατηγοριοποίησης SVM

Χρησιμοποιήθηκαν δύο διαφορετικές μέθοδοι ταξινόμησης οι Μηχανές Διανυσμάτων Υποστήριξης Support Vector Machines (SVM) και η ταξινόμηση Bayes για την κατηγοριοποίηση των διανυσμάτων. Το μοντέλο SVM αντιπροσωπεύει τις παρατηρήσεις ως σημεία σε έναν χώρο N-διαστάσεων. Στη συνέχεια θα πρέπει να υπολογιστεί ένα επίπεδο για να διαχωρίσει τις παρατηρήσεις που ανήκουν σε διαφορετικές κατηγορίες όπως

απεικονίζεται στο Σχήμα 7.3. Η απόσταση μεταξύ των περιπτώσεων που ανήκουν σε διαφορετικές κατηγορίες θα πρέπει να είναι όσο το δυνατόν μεγαλύτερη. Ο SVM αναπτύχθηκε αρχικά για δυαδικά προβλήματα κατηγοριοποίησης. Το πρόβλημα της ανάλυσης του συναισθήματος απαιτεί τρεις κατηγορίες, έτσι χρησιμοποιήθηκε μια παραλλαγή του SVM για ταξινόμηση πολλαπλών κατηγοριών Συγκεκριμένα επιλέχθηκε ένας γραμμικός SVM με κατάταξη ενός προς ενός [137]. Στην ταξινόμηση κατά ενός προς ενός διαχωρισμένου δυαδικού ταξινομητή η εκπαίδευση γίνεται για κάθε ζεύγος κατηγοριών. Όταν ένα νέο κείμενο εμφανίζεται, εφαρμόζεται σε όλους τους δυαδικούς ταξινομητές και η κατηγορία που αποδίδει τον υψηλότερο αριθμό προβλέψεων θα είναι αυτή όπου θα καταχωρηθεί.

Ο ταξινομητής Gaussian Bayes [138] χρησιμοποιεί ένα μοντέλο πιθανότητας υπό όρους στο οποίο ένα κείμενο μπορεί να αναπαρασταθεί ως διάνυσμα όπως περιγράφεται στις προηγούμενες παραγράφους. Οι τρεις τιμές του διανύσματος Si (s +, s-, s =) είναι οι τρεις ανεξάρτητες μεταβλητές και η θέση συναισθημάτων είναι η εξαρτημένη μεταβλητή y. Η πιθανότητα ενός εγγράφου να εκφράσει ένα συναίσθημα θεωρείται ότι είναι όπως δίνεται στην εξίσωση 4.8 της ενότητας 4.

7.7.Φιλτράρισμα Γράφων

Πολλές τεχνικές μηχανικής εκμάθησης χρησιμοποιούν κριτήρια επιλογής χαρακτηριστικών (Feature Selection) για να μειώσουν το μέγεθος των διαστάσεων και να βελτιώσουν την ακρίβεια τους [139]. Τα κριτήρια επιλογής χαρακτηριστικών φιλτράρουν όρους όπως λέξεις, N-γράμματα ή αριθμούς που δεν συνεισφέρουν θετικά στην διαδικασία της ταξινόμησης. Στη μελέτη μας θα δοκιμάσουμε ένα γνωστό κριτήριο χαρακτηριστικών την Αμοιβαία Πληροφορίας (Mutual Information) [140] χρησιμοποιώντας ως όρους τις ακμές των γράφων.

Το κριτήριο της αμοιβαίας πληροφορίας θα εφαρμοστεί μετά την δημιουργία του γράφου λέξεων των κειμένων εκπαίδευσης και των συναισθηματικών κατηγοριών και αποσκοπεί στο να απορριφθεί ένα ποσοστό των ακμών από τους γράφους των συναισθηματικών κατηγοριών και των κειμένων που θα κατηγοριοποιηθούν.

Τα οφέλη από τη χρήση της αμοιβαίας πληροφορίας είναι ότι οι γράφοι γίνονται μικρότεροι, με αποτέλεσμα η μέθοδος συναισθηματικής ανάλυσης να έχει λιγότερες απαιτήσεις σε υπολογιστικούς πόρους. Επιπλέον, υπάρχουν περιπτώσεις στις οποίες οι μέθοδοι μηχανικής μάθησης παρουσίαζαν βελτιωμένη ακρίβεια. Δυστυχώς, όπως θα δούμε και θα εξηγήσουμε στο πειραματικό τμήμα, αυτό δεν συμβαίνει με το μοντέλο που εξετάζουμε.

Τα κριτήρια όπου φιλτράρονται οι ακμές χρησιμοποιούνται μόνο με τη μέτρηση ομοιότητας περιεχομένου και όχι με τις μετρήσεις που βασίζονται στον μέγιστο κοινό υπογράφο. Αυτό οφείλεται στο ότι διαπιστώσαμε ότι οι πιθανές απορριπτόμενες ακμές περιέχονται συχνά στον μέγιστο κοινό υπογράφο και η απόρριψη τους επηρέαζε αρνητικά την ακρίβεια της μεθόδου. Η αμοιβαία πληροφορία μεταξύ μιας ακμής του γράφου μιας κατηγορίας αισθήματος ορίζεται από την εξίσωση 7.1

$$I(e, G_S) = log \frac{A \cdot N}{(A + C)(A + B)} (7.1)$$

όπου:

Α είναι ο αριθμός των περιπτώσεων που η ακμή e υπάρχει στους γράφους κειμένων του δεύτερου μέρους των δεδομένων εκπαίδευσης και στον γράφο της κατηγορίας συναισθήματος.

B είναι ο αριθμός των περιπτώσεων που η ακμή e υπάρχει στους γράφους των κειμένων του δεύτερο μέρους των δεδομένων εκπαίδευσης, αλλά δεν υπάρχει στον γράφο της κατηγορίας συναισθήματος.

C είναι ο αριθμός των περιπτώσεων που το δεύτερο μέρος των δεδομένων εκπαίδευσης εκφράζει το συναίσθημα του G_S αλλά δεν περιέχει την άκρη e.

N είναι ο συνολικός αριθμός κειμένων στο δεύτερο μέρος των δεδομένων εκπαίδευσης.

Για να μετρήσουμε τη συμβολή μιας ακμής στη διαδικασία πρόβλεψης μιας κατηγορίας, υπολογίζουμε τον συντελεστή $I(e, G_S)$ για τις θετικές S+, αρνητικές S- και ουδέτερες S= κατηγορίες συναισθημάτων. Τέλος, η εξίσωση 7.2 χρησιμοποιείται για τον υπολογισμό της συνολικής αμοιβαίας πληροφορίας για κάθε ακμή.

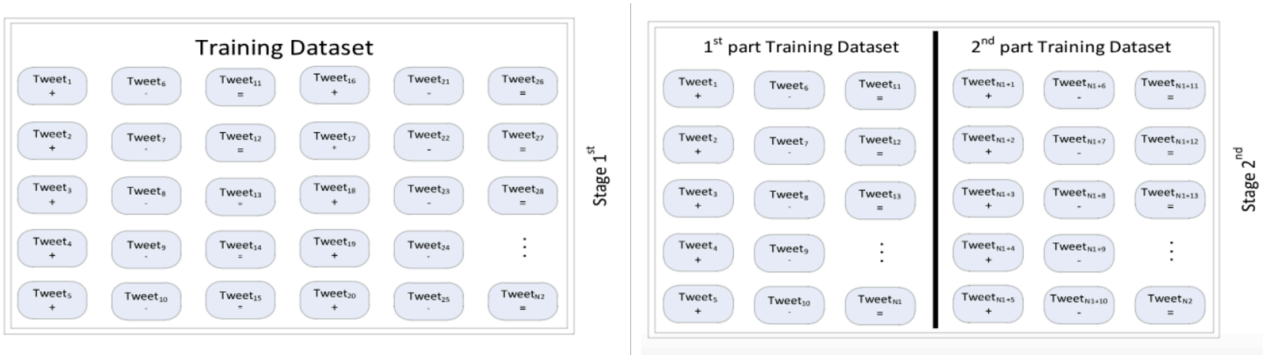
$$I_{avg}(e) = \sum_{i\epsilon(+,-,=)} P_r(S^i) \cdot I(e, G_i) \quad (7.2)$$

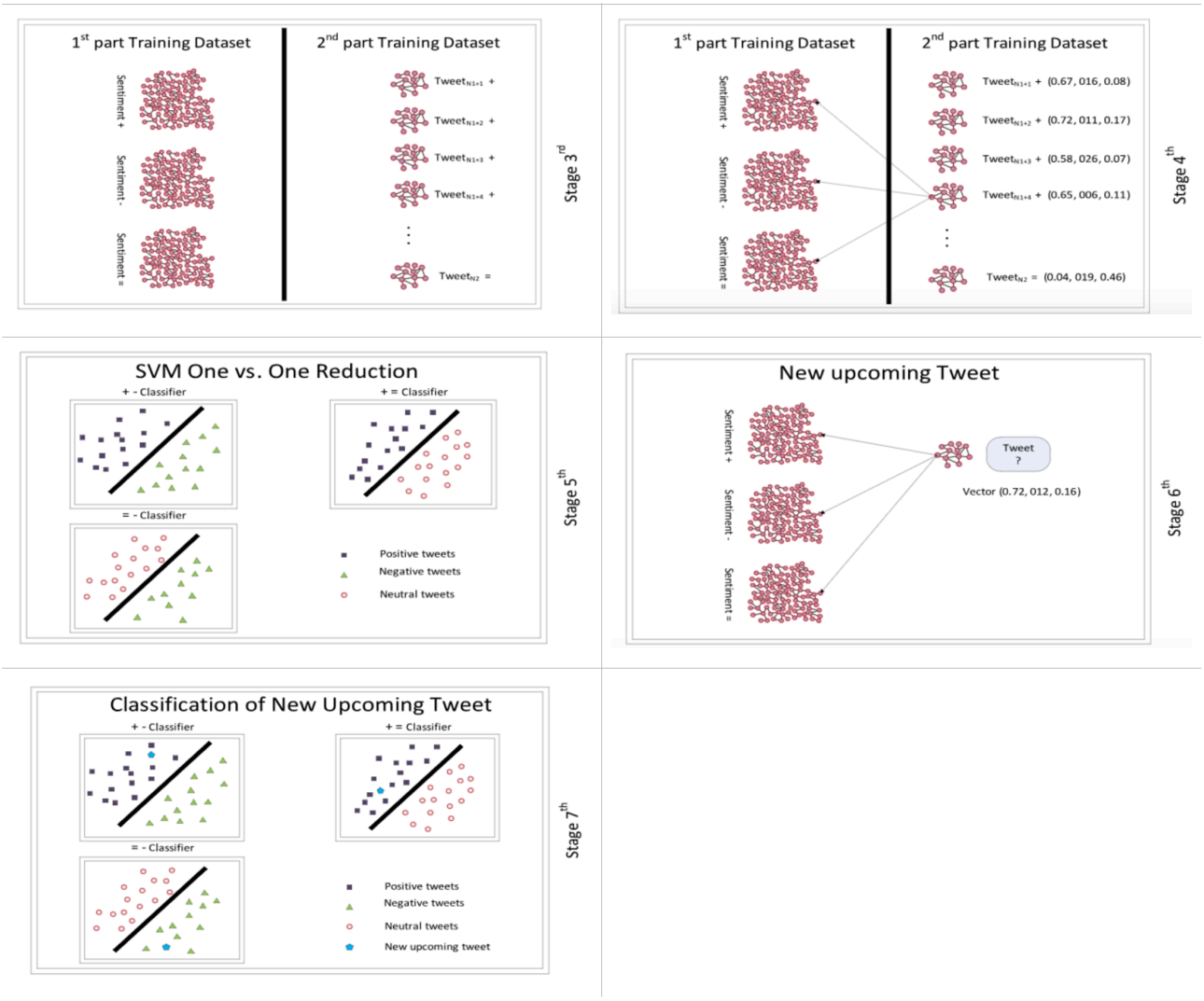
όπου $P_r(S^i)$ είναι το ποσοστό των κειμένων από το δεύτερο μέρος του συνόλου δεδομένων εκπαίδευσης που εκφράζουν το συναίσθημα S^i.

Το $I_{avg}(e)$ εκφράζει την καταλληλότητα και τη συμβολή της κάθε ακμής για την πρόβλεψη της συναισθηματικής κατηγορίας. Μπορούμε να ορίσουμε ένα όριο για το $I_{avg}(e)$ και να απορρίψουμε όλες τις ακμές που είναι μικρότερες από αυτό.

7.8.Επισκόπηση της Μεθόδου Συναισθηματικής Ανάλυσης με Γράφους Λέξεων

Στις προηγούμενες ενότητες περιγράψαμε την κατασκευή των γράφων-λέξεων για τα κείμενα και τις συναισθηματικές κατηγορίες. Επίσης αναφέραμε τις μετρήσεις ομοιότητας γράφων, την διανυσματική αναπαράσταση και κατηγοριοποίηση των κειμένων μετά την σύγκριση γράφων. Αυτή η ενότητα συνοψίζει όλα τα στάδια με ένα παράδειγμα. Εξηγεί την μέθοδο και περιγράφει πώς μπορεί να εντοπιστεί το συναίσθημα ενός νέου κειμένου. Το σχήμα 7.4 απεικονίζει όλα τα στάδια της προτεινόμενης μεθόδου.





Σχήμα 7.4 Τα στάδια του μοντέλου συναισθηματικής ανάλυσης με γράφους λέξεων

Στάδιο 1

Για την εκπαίδευση του μοντέλου συναισθηματικής ανάλυσης με γράφους λέξεων χρειάζεται ένα σύνολο δεδομένων εκπαίδευσης που περιλαμβάνει ένα πλήθος κειμένων σε συνδυασμό με την αντίστοιχη συναισθηματική κατηγορία που ανήκουν.

Στάδιο 2

Το σύνολο δεδομένων χωρίζεται σε δύο ίσα μέρη τα οποία χρησιμοποιούνται με διαφορετικό τρόπο. Στο 1ο μέρος, τα κείμενα που εκφράζουν το ίδιο συναίσθημα ομαδοποιούνται.

Στάδιο 3

Οι ομάδες των κειμένων που εκφράζουν το ίδιο συναίσθημα χρησιμοποιούνται για την κατασκευή των γράφων λέξεων για τις τρεις κατηγορίες συναισθημάτων. Κάθε γράφος κατασκευάζεται από τα κείμενα που εκφράζουν την συναισθηματική πολικότητα. Κάθε κείμενο του 2ου μέρους παρουσιάζεται επίσης ως γράφος λέξεων.

Στάδιο 4

Οι γράφοι κειμένων του 2ου μέρους συγκρίνονται μέσω της μετρική ομοιότητας γράφων με τους τρεις γράφους λέξεων. Η σύγκριση αυτή έχει ως αποτέλεσμα τρεις αριθμούς. Κάθε νούμερο εκφράζει την ομοιότητα του γράφου του κειμένου με το γράφο της συναισθηματικής κατηγορίας. Από εδώ και πέρα, κάθε κείμενο του 2ου μέρους μπορεί να εκπροσωπείται ως διάνυσμα αυτών των τριών αριθμών.

Στάδιο 5

Τα διανύσματα που προέκυψαν από τα κείμενα του 1ου μέρους του συνόλου δεδομένων εκπαίδευσης συνοδεύονται από το συναίσθημα που εκφράζουν και χρησιμοποιούνται για την εκπαίδευση του ταξινομητή SVM. Στην πραγματικότητα, τρεις ταξινομητές SVM εκπαιδεύονται λόγω της ανάγκης για ταξινόμηση πολλαπλών κατηγοριών. Κάθε κατηγορία αντιπροσωπεύει μια πολικότητα συναισθημάτων.

Στάδιο 6

Ένα νέο κείμενο αναπαρίσταται επίσης ως γράφος λέξεων με παρόμοιο τρόπο όπως τα κείμενα του 2ου μέρους στο στάδιο 3. Στη συνέχεια, ο γράφος του νέου κείμενου συγκρίνεται με τους γράφους των συναισθηματικών κατηγοριών και αντιπροσωπεύεται ως ένα διάνυσμα όπως κάναμε στο στάδιο 4.

Στάδιο 7

Το διάνυσμα του νέου κειμένου στόχου αποδίδεται ως ένα σημείο στον διανυσματικό χώρο των ταξινομητών SVM που κατασκευάσαμε στο 5ο στάδιο από όπου προκύπτει και η αντίστοιχη πρόβλεψη του συναισθήματος.

Η διαδικασία φιλτραρίσματος των γράφων δεν απεικονίζεται στο σχήμα 4 αν και το τμήμα αφαίρεσης ακμών λαμβάνει χώρα στα στάδια 3 και 6. Η επισκόπηση του προτεινόμενου μοντέλου περιγράφεται με την ταξινόμηση SVM. Η διαδικασία είναι παρόμοια στην περίπτωση ταξινόμησης Bayes. Για τη μέτρηση ομοιότητας στο στάδιο 4 και 6, μπορούμε να χρησιμοποιήσουμε οποιαδήποτε μέτρηση από αυτές που περιγράφονται στην Ενότητα 4 με τον περιορισμό ότι θα χρησιμοποιηθεί η ίδια μετρική και στα δύο στάδια.

Η πολυπλοκότητα της μεθόδου είναι χαμηλή. Η διαδικασία της δημιουργίας γράφων είναι γραμμική. Όλες οι λέξεις ενός κειμένου επαναλαμβάνονται από τον πρώτο μέχρι τον τελευταίο και για κάθε λέξη δημιουργείται ο αντίστοιχος κόμβος και κατευθυνόμενες ακμές που συνδέουν τους κόμβους με ένα κριτήριο εγγύτητας. Η σύγκριση μεταξύ δύο γράφων στο προετινόμενο μοντέλο έχει πολυπλοκότητα $O(V^2 + V \cdot E)$ όπου Ε είναι ο αριθμός των ακμών και V ο αριθμός των κόμβων. Η πολυπλοκότητα του κριτηρίου ομοιότητας περιεχομένου είναι $O(E_{G1} \cdot E_{G2})$. Όπου E_{G1} και E_{G2} είναι οι αριθμοί των ακμών του γραφήματος G1 και G2 αντίστοιχα.

Ένα ζήτημα που εμφανίζεται πάντοτε στις τεχνικές εποπτευόμενης μηχανικής μάθησης είναι η ανάγκη ενός συνόλου δεδομένων εκπαίδευσης. Το προτεινόμενο μοντέλο ανάλυσης συναισθημάτων για την πραγματοποίηση προβλέψεων χρειάζεται ένα αντίστοιχο σύνολο δεδομένων. Η διαδικασία σχηματισμού ενός συνόλου δεδομένων εκπαίδευσης είναι μια κουραστική εργασία επειδή τα κείμενα πρέπει να διαβάζονται από άτομα που τα αντιλαμβάνονται με υποκειμενικά κριτήρια.

7.9.Πειραματική Αξιολόγηση

Για τη διεξαγωγή των πειραμάτων χρησιμοποιήσαμε το δημόσια διαθέσιμο σύνολο δεδομένων από την δημοσίευση των Sascha, Hufenhaus και Albayrak Language-Independent Twitter Sentiment Analysis [141]. Αυτό το σύνολο δεδομένων αποτελείται από tweets που επιλέχθηκαν τυχαία και tweets που περιέχουν εμπορικές λέξεις κλειδιά όπως "sony", "audi", κλπ.

Το σύνολο δεδομένων έχει κατηγοριοποιηθεί από τους εργαζόμενους στο Mehanical Turk του Amazon. Περιέχει 10594 tweets. 1486 είναι αρνητικά, 2334 είναι θετικά και 6774 είναι ουδέτερα. Ο μέσος όρος μήκους των tweets είναι 14,2 λέξεις

Δοκιμάσαμε το προτεινόμενο μοντέλο, όπως περιγράφεται στις προηγούμενες παραγράφους, χρησιμοποιώντας την δεκαπλή-αναδιπλώσεων διασταυρούμενη αξιολόγηση. Σε κάθε αναδίπλωση χρησιμοποιήσαμε το 90% των tweets για την εκπαίδευση του αλγορίθμου συναισθηματικής ανάλυσης και το 10% για τις δοκιμές. Δεν χρησιμοποιήθηκαν εξωτερικά κείμενα για την εκπαίδευση του μοντέλου όπως η έρευνα των Sascha, Hufenhaus και Albayrak που χρησιμοποιεί ένα σύνολο δεδομένων 800 εκατομμυρίων tweets, κάνοντας τις δύο προσεγγίσεις ασύγκριτες.

Τα πειράματα πραγματοποιήθηκαν σε έναν υπολογιστή βασικών προδιαγραφών όπως και στις προηγούμενες ενότητες. Η γλώσσα προγραμματισμού Java χρησιμοποιείται για την υλοποίηση και την εφαρμογή του λογισμικού που πραγματοποιεί την κατασκευή των γράφων, τη σύγκριση τους, την επιλογή χαρακτηριστικών και την κατασκευή της διανυσματικής αναπαράστασης των tweets. Η ταξινόμηση των διανυσμάτων που αντιπροσωπεύουν τα tweets και η αξιολόγηση της μεθόδου πραγματοποιήθηκαν χρησιμοποιώντας τη βιβλιοθήκη scikit στη γλώσσα προγραμματισμού Python. Για να μετρήσουμε το χρόνο εκτέλεσης, χρειάστηκε να μετρήσουμε τον χρόνο για την εκτέλεση της δεκαπλής-αναδιπλώσεων διασταυρούμενη αξιολόγηση και περιελάμβανε την κατασκευή των γράφων των τριών συναισθηματικών κατηγοριών, την κατασκευή των διανυσμάτων και την εφαρμογή του ταξινομητή. Στις περισσότερες περιπτώσεις η αλληλουχία αυτών των βημάτων διήρκεσε λιγότερο από 30 δευτερόλεπτα.

Η προσέγγισή μας συγκρίθηκε με άλλες προσεγγίσεις που χρησιμοποιούν το ίδιο σύνολο δεδομένων. Ειδικότερα, συγκρίναμε ένα σύνολο NLP μεθόδων και αλγορίθμων μηχανικής μάθησης, τα αποτελέσματα ακρίβειας παρουσιάζονται στον Πίνακα 7.1.

Method	4Grams	4Gram Graphs
Bayesian Network	0.6788	0.6791
C4.5	0.6828	0.6896
Support Vector Machines	0.6777	0.6847
Logistic Regression	0.6822	0.7115
Simple Logistic Regression	0.6816	0.7109
Multi-Layer Perceptron	0.6788	0.7069
Best-First Tree	0.6790	0.6840
Functional Tree	0.6822	0.7079

Πίνακας 7.1. Πειραματικά αποτελέσματα χρησιμοποιώντας άλλες μεθόδους συναισθηματικής ανάλυσης

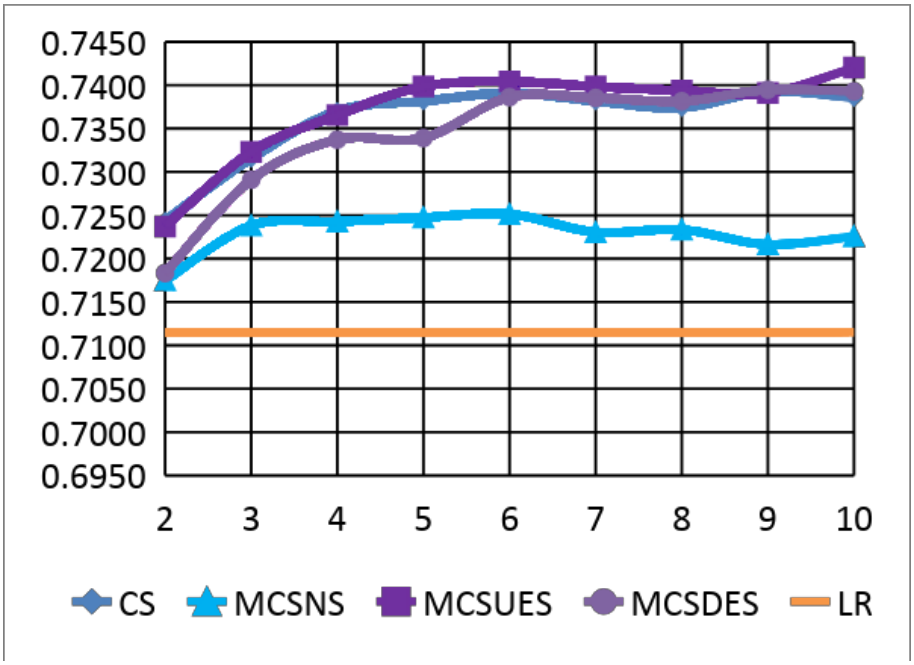
Η υψηλότερη ακρίβεια επιτυγχάνεται με την μέθοδο Logistic Regression (LR) χρησιμοποιώντας γράφους 4-γραμμάτων 0,7115. Αυτό το αποτέλεσμα χρησιμοποιείται στη συνέχεια ως τιμή σύγκρισης. Η Ανάλυση συναισθήματος με γράφους λέξεων ξεπερνάει αυτή την τιμή στις περισσότερες περιπτώσεις.

Η μέθοδος ανάλυσης αισθήματος με γράφους λέξεων έχει διερευνηθεί με πολλές παραλλαγές προκειμένου να κατανοηθούν οι δυνατότητες της μεθόδου στην ανάλυση συναισθημάτων και να καταλήξουμε σε μια σειρά παραμέτρων και τεχνικών που αυξάνουν την ακρίβεια της μεθόδου. Τα πειράματα διεξήχθησαν χρησιμοποιώντας ένα πλαίσιο λέξεων που εκτείνεται από δύο έως δέκα.

Όπως αναφέρθηκε στις προηγούμενες παραγράφους, θεωρούμε την επιλογή των μετρικών γράφων ομοιότητας ως ένα πολύ σημαντικό βήμα. Το Σχήμα 7.5 συνοψίζει τα αποτελέσματα με τις μετρικές ομοιότητας, την Ομοιότητα περιεχομένου CS, και τις τρεις ομοιότητες που βασίζονται στον μέγιστο κοινό υπογράφο, MCSNS, MCUES, MCDES όπως τις περιγράψαμε στην ενότητα 4.

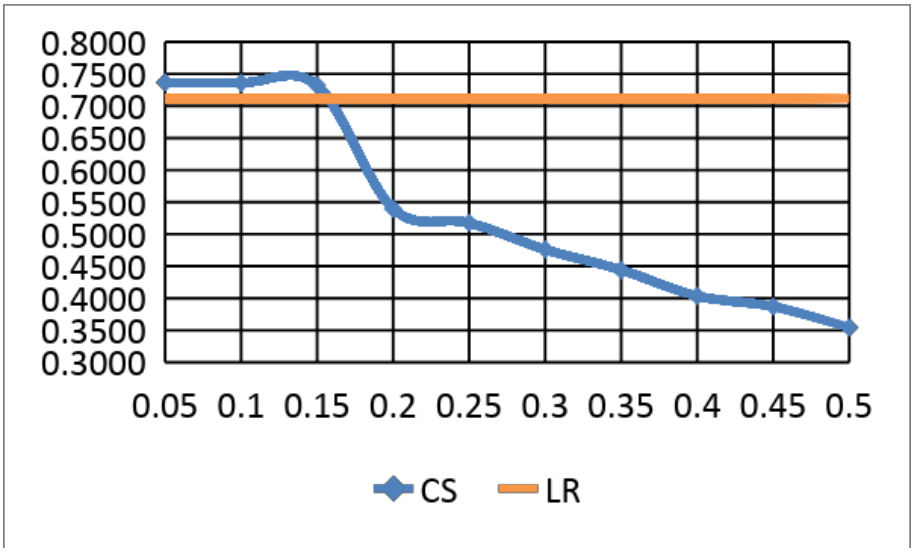
Η CS βασίζεται στην ποσότητα των κοινών ακμών, δηλαδή στο πλήθος των περιπτώσεων που δύο λέξεις είναι γειτονικές. Οι μέθοδοι που χρησιμοποιούν τον μέγιστο κοινό υπογράφο (MCS) αποδίδουν καλύτερα τη δομή των υπογράφων. Η MCSNS εκφράζει την ποσότητα των κόμβων στο MCS. Από την άλλη, οι MCSUES και MCSDES βασίζονται στο πλήθος των ακμών στο MCS. Τα πειραματικά αποτελέσματα επαληθεύουν ότι η εγγύτητα μεταξύ των λέξεων εκφράζει τη συναισθηματική θέση ενός κειμένου με έναν καλύτερο τρόπο από την απλή ύπαρξη των λέξεων. Τα κριτήρια που χρησιμοποιούν τον MCS δεν εκφράζουν μόνο πόσα ζευγάρια λέξεων είναι κοντά το ένα με το άλλο αλλά εκφράζουν επίσης ότι ένα σύνολο λέξεων σχετίζεται επειδή είναι κοντά στα αρχικά κείμενα.

Οι MCSUES και MCSDES έχουν καλύτερη ακρίβεια από την MCSNS. Οι MCSUES και MCSDES μετράνε τις ακμές που υπάρχουν στον MCS ανάμεσα στον γράφο του κειμένου και τον γράφο λέξεων που αντιπροσωπεύει μια συναισθηματική κατηγορία. Ενώ οι MCSNS καταμετρούν τους κόμβους που υπάρχουν μεταξύ τους. Εξετάζοντας αυτό το αποτέλεσμα συμπεραίνουμε ότι είναι πιο σημαντικός ο βαθμός που οι λέξεις είναι γειτονικές στον MCS από το πλήθος των κοινών λέξεων.



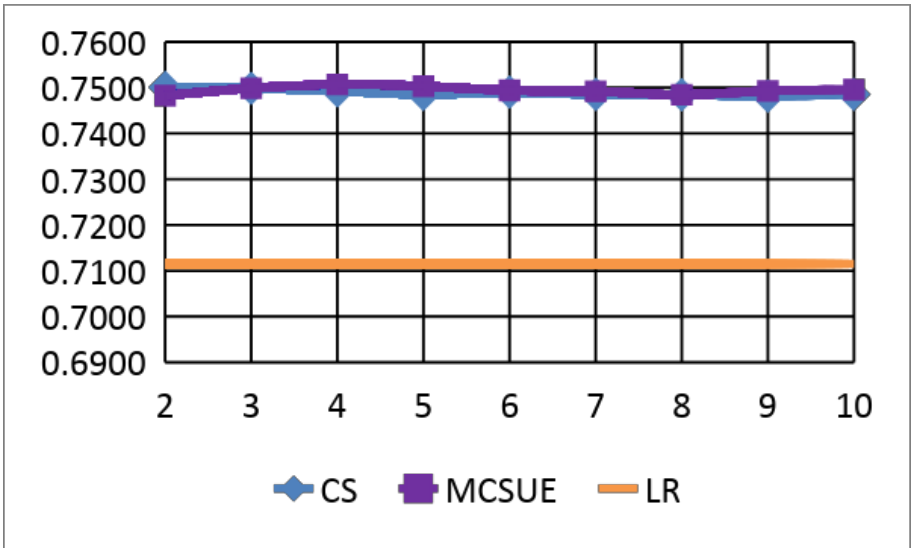
Σχήμα 7.5 Εφαρμογή μετρικών σύγκρισης γράφων στην συναισθηματική ανάλυση

Τα επόμενα πειράματα που πραγματοποιήθηκαν ήταν με το κριτήριο επιλογής χαρακτηριστικών αμοιβαίας πληροφορίας (MI), προκειμένου να αφαιρεθεί ένα ποσοστό ακμών. Στα πειράματα που παρουσιάζουμε το MI δεν εφαρμόζεται σε συνδυασμό με τις MCSN, MCSUES και MCDES επειδή παρατηρήθηκε ότι χρησιμοποιώντας το, απορρίπτονται αρκετές ακμές που υπάρχουν στον MCS με αποτέλεσμα ο MCS που παραμένει μετά να είναι αρκετά μικρός και να μην αρκεί για να εκφράσει την ομοιότητα των δύο γράφων. Σε πολλές περιπτώσεις αποτελείται μόνο από έναν ή δύο κόμβους.



Εικόνα 7.6 Εφαρμογή της αμοιβαίας πληροφορίας

Η μετρική CS μπορεί να χρησιμοποιηθεί σε συνδυασμό με την MI, επειδή η CS βασίζεται στις κοινές ακμές που υπάρχουν μεταξύ δύο γράφων. Τα πειράματα που απεικονίζονται στο Σχήμα 7.6 πραγματοποιήθηκαν χρησιμοποιώντας CS με βαθμό πλαισίου 4. Όπως φαίνεται, όσο περισσότερες ακμές φιλτράρονται, τόσο περισσότερο μειώνεται η ακρίβεια της μεθόδου. Η διαδικασία φιλτραρίσματος των ακμών οδηγεί σε μικρότερους γράφους που απαιτούν λιγότερους υπολογιστικούς πόρους, αλλά η ακρίβεια μειώνεται δραματικά. Ο άξονας x του Σχήματος 7.6 αντιπροσωπεύει την MI όπως περιγράφεται στην εξίσωση 7.2. Είναι φανερό ότι η ακρίβεια μειώνεται ακόμη και αν απορρίψουμε τις ακμές που έχουν πολύ μικρό MI.



Σχήμα 7.7 Αξιολόγηση κατάταξης Gaussian Naeve Bayes

Τα πειράματα στο Σχήμα 7.5 διεξήχθησαν χρησιμοποιώντας τη μέθοδο ταξινόμησης SVM. Στις περισσότερες περιπτώσεις, ο SVM για εφαρμογές ταξινόμησης υπερβαίνει τις άλλες μεθόδους ταξινόμησης

[142]. Οι μέθοδοι ανάλυσης συναρτήσεων γράφων λέξεων συνδυάστηκαν επίσης με άλλες μεθόδους ταξινόμησης και διαπιστώθηκε ότι χρησιμοποιώντας έναν ταξινομητή Gaussian Bayes μπορούμε να έχουμε ακόμα καλύτερα πειραματικά αποτελέσματα. Τα πειράματα πραγματοποιήθηκαν χρησιμοποιώντας τις μετρήσεις ομοιότητας γράφων MCSUE και CS. Η ακρίβεια της μεθόδου χρησιμοποιώντας γράφους τεσσάρων λέξεων, MCSUE και Bayes Gaussian Classifier φτάνει το 75,07%. Ο λόγος για τον οποίο ο SVM εμφανίζει χαμηλότερη ακρίβεια ταξινόμησης από ότι ο Μπεϋζιανό ταξινομητής είναι ότι χρησιμοποιεί ένα χώρο χαμηλών διαστάσεων.

7.10.Συμπεράσματα

Εφαρμόσαμε την μέθοδο κατηγοριοποίησης κειμένων με μια παραλλαγή των ΓΝΓ για την αντιμετώπιση του προβλήματος της συναισθηματικής ανάλυσης. Συγκεκριμένα διαφοροποιήθηκε η μέθοδος που προτείνουμε μόνο στο ότι αναπαραστήσαμε τα κείμενα ως γράφους λέξεων, ενώ κρατάμε το μοντέλο σύγκρισης και ταξινόμησης ίδιο.

Η συναισθηματική ανάλυση με γράφους λέξεων συνδυάζει την καλά καθορισμένη δομή των γράφων με αλγορίθμους ταξινόμησης υψηλής ακρίβειας. Οι γράφοι λέξεων μπορούν να καταγράψουν την ακολουθία των λέξεων που περιέχονται σε ένα κείμενο. Χρησιμοποιήθηκαν αρκετές τεχνικές ομοιότητας γράφων για να εκτιμηθεί η ομοιότητα μεταξύ των γράφων κειμένου με τους γράφους που αντιπροσωπεύουν τις κατηγορίες των συναισθημάτων. Το αποτέλεσμα της σύγκρισης είναι ένα διάνυσμα 3 χαρακτηριστικών το οποίο αποδίδει την συναισθηματική θέση του κειμένου. Η πρόβλεψη του συναισθήματος γίνεται με το να εισαγάγουμε αυτό το διάνυσμα σε έναν ταξινομητή. Σε όλα τα στάδια της μεθόδου εφαρμόσαμε ένα σύνολο από μετρήσεις και τεχνικές που είναι εγγενείς στις ανάγκες της έρευνάς μας.

Η εγγύτητα και η διάταξη των λέξεων αποδεικνύεται ότι αποτελούν αξιόπιστο κριτήριο για την αποτίμηση του συναισθήματος ενός κειμένου και συγκεκριμένα ενός tweet. Συμπεραίνουμε ότι πιο εξελιγμένες μετρήσεις ομοιότητας γράφων που μετράνε τον βαθμό σχέσης μεταξύ των κοινών λέξεων (MCSUES) μπορούν να έχουν καλύτερα αποτελέσματα από τις μετρήσεις ομοιότητας γράφων που βασίζονται στον αριθμό των κοινών λέξεων που είναι γειτονικές (MCSNS) και το πλήθος των ζευγαριών λέξεων που είναι κοντά (CS).

Το μέγεθος του πλαισίου λέξεων που χρησιμοποιείται για την κατασκευή των γράφων λέξεων επηρεάζει την ακρίβεια της μεθόδου σε μεγάλο βαθμό. Παρατηρήσαμε ότι καθώς χρησιμοποιούσαμε μεγαλύτερο πλαίσιο αυξανόταν η ακρίβεια σε συνδυασμό με τις μετρικές CS, MCSUES και MCSDES. Από την άλλη πλευρά, παραμένει σταθερή ή μειώνεται ελαφρώς χρησιμοποιώντας ένα πλαίσιο μεγαλύτερο από 3 με το κριτήριο MCSNS. Ο λόγος είναι ότι τα κριτήρια CS, MCSUES και MCSDES βασίζονται στο πλήθος των κοινών ακμών που υπάρχουν μεταξύ των δύο γράφων. Ένα μεγάλο πλαίσιο λέξεων θα αποδώσει τη σχέση λέξεων καλύτερα από ένα μικρό. Η μετρική MCSNS βασίζεται στους κοινούς κόμβους που υπάρχουν μεταξύ των δύο γράφων, οπότε δεν επηρεάζεται πολύ από το μέγεθος του πλαισίου.

Δύο ακόμη θέματα που μελετώνται στην έρευνα μας είναι η ταξινόμηση των διανυσμάτων που αντιπροσωπεύουν τα tweets και ο συνδυασμός της μεθόδου με ένα κριτήριο επιλογής χαρακτηριστικών. Η μέθοδος αναπαράστασης γράφων λέξεων θα πρέπει να συνδυαστεί με μια μέθοδο ταξινόμησης που κάνει καλές προβλέψεις σε ένα χώρο χαμηλών διαστάσεων. Ο ταξινομητής Gaussian Bayes μπορεί να εκπαιδευτεί ευκολότερα και να παράγει ακριβέστερες προβλέψεις από άλλους ταξινομητές.

Μια μέθοδος επιλογής χαρακτηριστικών μπορεί να χρησιμοποιηθεί για τη μείωση της ποσότητας των δεδομένων που πρέπει να επεξεργαστούν. Πραγματοποιήσαμε πειράματα χρησιμοποιώντας τη μέτρηση της Αμοιβαίας Πληροφορίας, αλλά διαπιστώσαμε ότι η ακρίβεια της μεθόδου μειώθηκε δραματικά καθώς αυξήθηκε η ποσότητα των ακμών που αφαιρέθηκαν.

Τα πειραματικά αποτελέσματα δείχνουν ότι το προτεινόμενο μοντέλο είναι πρακτικό, αποτελεσματικό και στις περισσότερες περιπτώσεις υπερβαίνει τις άλλες μεθόδους ανάλυσης συναισθηματικής ανάλυσης. Η δομή των γράφων λέξεων για την αντιπροσώπευση των μηνυμάτων tweets και των συναισθηματικών κατηγοριών είναι μια δομή δεδομένων που αποτυπώνει την συναισθηματική θέση που εκφράζεται και χρησιμοποιώντας μια μέτρηση ομοιότητας γράφων και μια μέθοδο ταξινόμησης μπορεί να προβλεφθεί με μεγάλη ακρίβεια.

8. Εφαρμογή στην Αναγνώριση Γεγονότων σε Μέσα Κοινωνικών Δικτύων

Η αναγνώριση γεγονότων στα μέσα κοινωνικών δικτύων [143] έχει σκοπό την ανίχνευση περιστατικών όπως συμβαίνουν στον πραγματικό κόσμο εστιάζοντας στην έγκαιρη αναγνώριση τους καθώς εξελίσσονται από τα πρώτα λεπτά που χρήστες μιλάν για αυτά. Κοινωνικά δίκτυα όπως το Twitter αποτελούν πολύτιμη πηγή πληροφορίας όπου δημοσιεύονται κείμενα για ένα γεγονός κατά την διάρκεια που συμβαίνει, αφού συμβεί και σε κάποιες περιπτώσεις πριν συμβεί ως μια φάση προετοιμασίας.

Οι τεχνικές που αναγνωρίζουν γεγονότα είναι τεχνικές που έρχονται από επιστημονικά πεδία όπως η μηχανική μάθηση, η φυσική επεξεργασία γλώσσας, η εξόρυξη δεδομένων, η ανάκτηση πληροφορίας και η εξόρυξη κειμένων με κυρίαρχες προσεγγίσεις την κατηγοριοποίηση, την συσταδοποίηση και την ανίχνευση ανωμαλιών. Ένας άλλος διαχωρισμός γίνεται μεταξύ αναδρομικής αναγνώρισης γεγονότων και νέας αναγνώρισης γεγονότων.

Η αναδρομική αναγνώριση γεγονότων περιλαμβάνει την επαναληπτική επεξεργασία των κειμένων έως ότου να οργανωθούν σε θεματικές κατηγορίες. Βασική προσέγγιση στην αναδρομική αναγνώριση κειμένων είναι η από κάτω προς τα πάνω αθροιστική συσταδοποίηση όπου αρχικά αντιλαμβάνεται κάθε κείμενο ως υποψήφιο γεγονός και με μια συνάρτηση ομοιότητας τα αθροίζει μέχρις ότου κάποια κριτήρια τερματισμού ικανοποιηθούν. Η αναγνώριση γεγονότων παράγει προβλέψεις για την άμεση εμφάνιση καινούριων γεγονότων, καθώς τα κείμενα παράγονται χρησιμοποιώντας άπληστους αλγορίθμους που επεξεργάζονται τις δομές δεδομένων ακουλουθιακά και κάθε νέο κείμενο το συσχετίζουν με ένα ήδη υπάρχον γεγονός ή δημιουργούν ένα καινούριο.

	Micro F-Measure	Macro Precision	Macro Recall	Macro F-Measure
JCS 2Grams	0,7962	0,7175	0,8097	0,7282
JCS 3Grams	0,8161	0,7281	0,8494	0,7499
JCS 4Grams	0,8171	0,7172	0,8560	0,7423
JCS 5Grams	0,8074	0,7015	0,8525	0,7232
JCS 6Grams	0,7758	0,6748	0,8319	0,6829
JCS 7Grams	0,7500	0,6538	0,8140	0,6582
JCS 8Grams	0,7538	0,6447	0,8129	0,6569
JCS 9Grams	0,7669	0,6515	0,8191	0,6728
JCS 10Grams	0,7695	0,6701	0,8106	0,6832

Παρακάτω θα δούμε τα βασικά χαρακτηριστικά που διέπουν τις τεχνικές αναγνώρισης γεγονότων, θα παρουσιάσουμε πως η κατηγοριοποίηση κειμένων που χρησιμοποιεί ΓΝΓ εφαρμόζεται για τις ανάγκες αναγνώρισης κειμένων και θα την αξιολογήσουμε σε τρία σύνολα δεδομένων.

8.1.Εισαγωγή στις Τεχνικές Αναγνώρισης Γεγονότων

Τεχνικές επεξεργασίας κειμένων

Οι τεχνικές που διαχειρίζονται τις δημοσιεύσεις χρηστών ως κείμενα αποτελούνται από τρεις βασικές εργασίες, τον διαχωρισμό, την αναγνώριση και την παρακολούθηση. Το σύνολο των διαθέσιμων κειμένων

διαχωρίζεται σε συνεκτικά θέματα και προσπαθούμε με μια μετρική ομοιότητας να βρούμε τα πρωτοεμφανιζόμενα θέματα σε σχέση με αυτά που προϋπήρχαν και εξελίχθηκαν με την πάροδο του χρόνου. Οι τεχνικές επεξεργασίας κειμένων περιλαμβάνουν τα τρία στάδια: προεπεξεργασία κειμένου, αναπαράσταση κειμένου και οργάνωση ή συσταδοποίηση των κειμένων όπως συζητήσαμε στην ενότητα 2.

Τα κείμενα αναπαριστώνται ως διανύσματα ή με το μοντέλο σάκου λέξεων σύμφωνα με τι λέξεις που περιέχουν ή με το διάνυσμα οντότητας το οποίο περιέχει πληροφορία που απαντά στα ερωτήματα Ποιος, Που, Πότε και Γιατί. Η ομοιότητα μεταξύ των γεγονότων υπολογίζεται από μετρικές ομοιότητας όπως η συσχέτιση Pearson ή η Hellinger [144] απόσταση.

Τεχνικές επεξεργασίας χαρακτηριστικών

Η αναγνώριση γεγονότων μπορεί να πραγματοποιηθεί με την παρακολούθηση της συχνότητας χρήσης συγκεκριμένων όρων από τους χρήστες κοινωνικών δικτύων. Αυτοί οι όροι μπορεί πριν να μην χρησιμοποιούνταν και μετά από μια χρονική στιγμή να ξεκίνησαν να χρησιμοποιούνται ή να παρατηρήθηκε μια απότομη αύξηση της χρήσης τους από ένα σύνολο χρηστών. Ένα γεγονός αναπαρίσταται και στις δύο περιπτώσεις ως ένα σύνολο όρων οπότε η αλλαγή συχνότητας των όρων πυροδοτεί την αναγνώριση του αντίστοιχου γεγονότος.

Η αναγνώριση γεγονότων μπορεί να περιλαμβάνει καθορισμένα ή μη καθορισμένα γεγονότα. Στην αναγνώριση καθορισμένων γεγονότων έχουμε μια περιγραφή των γεγονότων που αναμένουμε να αναγνωρίσουμε είτε ως ένα σύνολο όρων είτε ως ένα σύνολο κειμένων. Η χρήση προδιαγεγραμμένων όρων στα κοινωνικά δίκτυα σηματοδοτεί την πιθανή εμφάνιση των αντίστοιχων γεγονότων. Από την άλλη πλευρά στην αναγνώριση μη καθορισμένων γεγονότων δεν υπάρχει διαθέσιμη γνώση που να σχετίζεται με συγκεκριμένα γεγονότα και η αναγνώριση γίνεται με χωρο-χρονικά κριτήρια, το απότομο ενδιαφέρον μεγάλου πλήθους χρηστών και τον συσχετισμό των χαρακτηριστικών αναμεταξύ τους για τον σχηματισμό καινούριων γεγονότων.

8.2.Αναγνώριση Γεγονότων με την Μέθοδο Κατηγοριοποίησης Κειμένων με ΓΝΓ

Η μέθοδος κατηγοριοποίησης κειμένων με ΓΝΓ χρησιμοποιήθηκε για την αναγνώριση γεγονότων σε κοινωνικά δίκτυα. Τα γεγονότα ήταν καθορισμένα ακολουθώντας συγκεκριμένες περιγραφές και η μέθοδος υλοποιήθηκε ακολουθώντας την τεχνική επεξεργασίας κειμένου. Ένα σύνολο από προκαθορισμένα γεγονότα ορίστηκε στο σύστημα και καθώς νέα κείμενα εμφανιζόντουσαν ένας ταξινομητής τα συσχετίζει με τα αντίστοιχα γεγονότα. Στο τέλος παρήχθησαν τα γεγονότα που εμφανίστηκαν και την έκταση των κειμένων που αναφέρονται σε αυτά και συγκεκριμένα ποια κείμενα κατηγοριοποιούνται στα αντίστοιχα γεγονότα.

Η διαδικασία αναπαράστασης κειμένων σύμφωνα με το μοντέλο ΓΝΓ και για τα γεγονότα αλλά και για τα μεμονωμένα κείμενα έγινε όπως περιγράφεται στην ενότητα 4. Με μόνη προσθήκη την χρήση της μετρικής ομοιότητας Jaccard. Αντίστοιχα χρησιμοποιήθηκαν τεχνικές ομοιότητας γράφων και προεπεξεργασίας κειμένων. Για μια φορά ακόμη βλέπουμε πως η κατηγοριοποίηση κειμένων είναι μια τεχνική στην οποία μπορούν να αναχθούν πολλές εφαρμογές όπου πρέπει να εξαχθούν συμπεράσματα ή προβλέψεις.

Ο συντελεστής ομοιότητας Jaccard (JSC) [145] είναι μια μετρική που υπολογίζει την τομή ενός συνόλου χαρακτηριστικών προς την ένωση τους όπως φαίνεται στην εξίσωση 8.1

$$J(G_T, G_C) = \frac{G_T \cap G_C}{G_T \cup G_C} (8.1)$$

Το σύνολο χαρακτηριστικών που χρησιμοποιήσαμε είναι οι ακμές των γράφων που αναπαριστούν ένα γεγονός G_C και ένα κείμενο G_T.

8.3.Πειραματική Αξιολόγηση

MEdiaEval

Το μοντέλο ταξινόμησης ΓΝΓ εφαρμόστηκε στο σύνολο δεδομένων MEdiaEval [146] που αποτελείται από διευθύνσεις URL φωτογραφιών από το Instagram και το Flickr. Όλες οι φωτογραφίες συνοδεύονται από μεταδεδομένα και ταξινομούνται βάσει κοινωνικών εκδηλώσεων. Στόχος μας είναι να διερευνήσουμε εάν τα μεταδεδομένα κειμένου που περιγράφουν ένα αντικείμενο μπορούν να χρησιμοποιηθούν από το προτεινόμενο μοντέλο ταξινόμησης ΓΝΓ για την αναγνώριση γεγονότων. Το σύνολο δεδομένων περιέχει 201.602 φωτογραφίες με τα αντίστοιχα μεταδεδομένα τους και αντιστοιχούν σε 10.431 γεγονότα.

Εφαρμόσαμε την μέθοδο κατηγοριοποίησης κειμένων με ΓΝΓ όπως παρουσιάζεται στην ενότητα 4 με σταθμισμένους και μη σταθμισμένους γράφους και τις μετρικές περιεχομένου (CS) κατά αξία (VS) και κανονικοποιημένη αξία (NVS) για N-γράμματα από 2 έως 5 όπως φαίνονται στον πίνακα 8.1

Τα αποτελέσματα των πειραμάτων χωρίς καμία προεπεξεργασία αποκαλύπτουν ότι οι μη σταθμισμένοι γράφοι υπερβαίνουν σε ακρίβεια τους σταθμισμένους γράφους σε όλες τις μετρήσεις αξιολόγησης εκτός από την μετρική Macro Recall. Επίσης συμπεράνουμε ότι, μια υψηλότερη τάξη για το N δεν επιφέρει καλύτερα αποτελέσματα αξιολόγησης. Τα 3-γράμματα και τα 4-γράμματα παράγουν καλύτερα αποτελέσματα από τα 5-γράμματα για μη ζυγισμένους γράφους και μετρήσεις Micro, Macro F - Measure.

Επειδή το MEdiaEval είναι ένα σύνολο δεδομένων όπου υπάρχουν μεγάλες αποκλίσεις στο μέγεθος των κατηγοριών, μπορούμε να δούμε ότι η κανονικοποιημένη ομοιότητα τιμών στους σταθμισμένους γράφους παρουσιάζει καλά αποτελέσματα αξιολόγησης. Επίσης, παρατηρούνται πολύ καλά αποτελέσματα αξιολόγησης με τις Μάκρο μετρικές για την κανονικοποιημένη ομοιότητα τιμών και επιβεβαιώνει ότι ο κανονικοποιημένος παράγοντας ενισχύει την κατηγοριοποίηση όπου οι κατηγορίες παρουσιάζουν ανόμοιο μέγεθος.

	Macro Precision	Macro Recall	Macro F-Measure	Micro F-Measure
CS 2Grams	0.8222	0.8132	0.8027	0.7321
CS 3Grams	0.8495	0.8601	0.8387	0.7566
CS 4Grams	0.8442	0.8677	0.8387	0.7595
CS 5Grams	0.8385	0.8687	0.8351	0.7562
VS 2Grams	0.4834	0.6545	0.5135	0.4205
VS 3Grams	0.5215	0.6895	0.5504	0.4511
VS 4Grams	0.5579	0.7160	0.5828	0.4711
VS 5Grams	0.6005	0.7459	0.6211	0.5038
NVS 2Grams	0.7530	0.8544	0.7626	0.6791
NVS 3Grams	0.7692	0.8687	0.7788	0.6906
NVS 4Grams	0.7800	0.8754	0.7875	0.6953
NVS 5Grams	0.8032	0.8902	0.8090	0.7145

Πίνακας 8.1. χρήση του μοντέλου αναπαράστασης ΓΝΓ στο σύνολο δεδομένων MEdiaEval

Οι ακόλουθοι δύο πίνακες παρουσιάζουν τα αποτελέσματα χρήσης των προεπεξεργασμένων κειμένων με σταθμισμένους και μη σταθμισμένους γράφους. Το στάδιο προεπεξεργασίας είναι πιο σημαντικό σε αυτό το σύνολο δεδομένων, επειδή τα δεδομένα κειμένου που συνοδεύουν τις εικόνες είναι λίγα και η προεπεξεργασία μπορεί να φιλτράρει ένα μεγάλο ποσοστό των ΓΝΓ. Το μικρό μέγεθος των κειμένων δεδομένων, μας περιορίζει επίσης, στο να κάνουμε πειράματα με τον βαθμό Ν να υπερβαίνει το πέντε.

	Macro Precision	Macro Recall	Macro F-Measure	Micro F-Measure
NSW CS 2Grams	0.8199	0.8159	0.8022	0.7290
NSW CS 3Grams	0.8396	0.8586	0.8316	0.7515
NSW CS 4Grams	0.8317	0.8619	0.8280	0.7518
NSW CS 5Grams	0.8252	0.8597	0.8229	0.7471
NSW NVS 2Grams	0,7593	0,8605	0,7692	0,6813
NSW NVS 3Grams	0,7776	0,8757	0,7869	0,6955
NSW NVS 4Grams	0,7981	0,8896	0,8060	0,7112
NSW NVS 5Grams	0,8176	0,9004	0,8233	0,7253

Πίνακας 8.2 Αφαίρεση των κοινών λέξεων από το σύνολο δεδομένων MEdiaEval

Η λημματοποίηση των κειμένων αυξάνει την Μίκρο Φ-μετρική κατά 0,25% και την Μάκρο Φ-μετρική κατά 0,78%. Σε μεγαλύτερου μήκους κείμενα με 3-10 παραγράφους όπως είδαμε στην ενότητα 4, διαπιστώσαμε ότι η λημματοποίηση όταν εφαρμόστηκε μόνη της είχε αρνητική επίδραση σε αντίθεση με την τεχνική απομάκρυνσης των κοινών λέξεων που αύξησε την ακρίβεια της μεθόδου. Σε σύντομα κείμενα όπως αυτά του συνόλου δεδομένων MEdiaEval αυξήθηκε η ακρίβεια με χρήση της λημματοποίησης όπως βλέπουμε στον πίνακα 8.3.

	Macro Precision	Macro Recall	Macro F-Measure	Micro F-Measure
Stm CS 2Grams	0.8280	0.8189	0.8087	0.7355
Stm CS 3Grams	0.8530	0.8650	0.8430	0.7600
Stm CS 4Grams	0.8470	0.8725	0.8423	0.7621
Stm CS 5Grams	0.8408	0.8718	0.8378	0.7579
Stm NVS 2Grams	0,7524	0,8536	0,7621	0,6794
Stm NVS 3Grams	0,7694	0,8688	0,7791	0,6925
Stm NVS 4Grams	0,7911	0,8834	0,7987	0,7102
Stm NVS 5Grams	0,8107	0,8952	0,8164	0,7250

Πίνακας 8.3 εφαρμογή της λημματοποίησης στο σύνολο δεδομένων MEdiaEval

Η αφαίρεση των κοινών λέξεων ελαττώνει ελαφρώς τόσο το Μίκρο όσο και το Μάκρο Φ-Μέτρο. Τα μικρά κείμενα παράγουν μικρούς ΓΝΓ με αποτέλεσμα ακόμη και οι κοινές λέξεις να συμβάλουν στην αναπαράσταση και κατηγοριοποίηση.

Zubiaga

Η αναγνώριση γεγονότων με το μοντέλο κατηγοριοποίησης ΓΝΓ εφαρμόστηκε σε ένα σύνολο δεδομένων από tweets που αφορούν τους 26 αγώνες ποδοσφαίρου που διεξάχθηκαν μεταξύ 1ης και 24 Ιουλίου 2011 για το πρωτάθλημα Copa America 2011 στην Αργεντινή [147]. Τα tweets μαζεύτηκαν χρησιμοποιώντας τα hashtags #ca2011, #copaamerica και #copaamerica2011 και ως γεγονότα θεωρήθηκαν, όταν σημειώθηκε γκολ, κόκκινη κάρτα, ακύρωση γκολ, έναρξη αγώνα, τέλος αγώνα, διακοπή αγώνα και συνέχεια αγώνα. Ο σκοπός αυτής της έρευνας είναι από την ροή των ερχόμενων tweet να προβλέψουμε αν συνέβη κάποιο από τα αναφερθέντα γεγονότα.

Οι δημιουργία των γράφων η σύγκριση και η κατηγοριοποίηση πραγματοποιήθηκε όπως περιγράφεται στην ενότητα 4. Στον πίνακα 8.4 βλέπουμε τα αποτελέσματα χρησιμοποιώντας την μετρική περιεχομένου για Ν-γράμματα βαθμού από 2 έως 10. Παρατηρούμε ότι τα Ν-γράμματα βαθμού 3 ή 4 παρουσιάζουν τα καλύτερα συγκριτικά αποτελέσματα. Επιβεβαιώνεται ότι για μικρά κείμενα όπως tweets μια αναπαράσταση ΓΝΓ με μικρό βαθμό Ν παρουσιάζει καλύτερα αποτελέσματα. Ενώ για κείμενα που έχουν μεγαλύτερη έκταση όπως είδαμε στα πειράματα της ενότητας 4 μια αναπαράσταση με Ν-γράμματα βαθμού 6 είχε καλύτερα αποτελέσματα.

	Macro Precision	Macro Recall	Macro F-Measure	Micro F-Measure
CS 2Grams	0,8487	0,8068	0,8201	0,8552
CS 3Grams	0,8751	0,8442	0,8520	0,8725
CS 4Grams	0,8606	0,8448	0,8474	0,8668
CS 5Grams	0,8477	0,8335	0,8319	0,8589
CS 6Grams	0,8308	0,8228	0,8183	0,8491
CS 7Grams	0,8185	0,8078	0,7961	0,8359
CS 8Grams	0,8065	0,7935	0,7695	0,8228
CS 9Grams	0,7833	0,7851	0,7396	0,8062
CS 10Grams	0,7788	0,7743	0,7239	0,7878

Πίνακας 8.4 Αναγνώριση κοινοτήτων με Μετρική περιεχομένου

Τα επόμενα δύο σύνολα πειραμάτων πραγματοποιήθηκαν με την μετρική ομοιότητας που βασίζονται στον μέγιστο κοινό υπογράφο. Παρατηρήσαμε μια μείωση της ακρίβειας και στις δύο εκδοχές στον πίνακα 8.5 λαμβάνοντας υπόψη τους κόμβους που περιέχει ο MCS και στον πίνακα 8.6 τις κατευθυνόμενες ακμές.

	Micro F-Measure	Macro Precision	Macro Recall	Macro F-Measure
MCSN 2Grams	0,6190	0,6102	0,5917	0,5911
MCSN 3Grams	0,7117	0,7452	0,6536	0,6779
MCSN 4Grams	0,6997	0,7176	0,6966	0,7004
MCSN 5Grams	0,7324	0,7457	0,7306	0,7292
MCSN 6Grams	0,7636	0,7542	0,7657	0,7506

MCSN 7Grams	0,7838	0,7567	0,7760	0,7513
MCSN 8Grams	0,7888	0,7519	0,7820	0,7422
MCSN 9Grams	0,7803	0,7432	0,7737	0,7177
MCSN 10Grams	0,7693	0,7522	0,7641	0,7076

Πίνακας 8.5 Αναγνώριση κοινοτήτων με Μετρική ομοιότητας MCSN

Στα πειραματικά αποτελέσματα με τις μετρικές MCS, παρατηρήσαμε μια πτώση στις τιμές αξιολόγησης, που μας κάνουν να μην τις θεωρούμε την καλύτερη επιλογή σύγκρισης γράφων. Συνεχίσαμε με την μετρική απόστασης Jaccard όπου παρουσιάζονται τα πειραματικά αποτελέσματα στον πίνακα 8.7 όπου και πάλι η ακρίβεια που παρουσιάζουν είναι μικρότερη από ότι με την μετρική περιεχομένου που είδαμε στον πίνακα 8.4

	Micro F-Measure	Macro Precision	Macro Recall	Macro F-Measure
MCSDE 2Grams	0,8389	0,8151	0,7775	0,7862
MCSDE 3Grams	0,7530	0,7784	0,7064	0,7293
MCSDE 4Grams	0,7312	0,7541	0,7416	0,7396
MCSDE 5Grams	0,7599	0,7665	0,7545	0,7526
MCSDE 6Grams	0,7805	0,7720	0,7758	0,7652
MCSDE 7Grams	0,7895	0,7872	0,7731	0,7629
MCSDE 8Grams	0,7915	0,7785	0,7788	0,7508
MCSDE 9Grams	0,7817	0,7606	0,7703	0,7225
MCSDE 10Grams	0,7693	0,7630	0,7619	0,7105

Πίνακας 8.6 Αναγνώριση κοινοτήτων με Μετρική ομοιότητας MCSDE

Πίνακας 8.7 Αναγνώριση κοινοτήτων με Μετρική ομοιότητας Jaccard

First Story Detection

Το επόμενο σύνολο δεδομένων που εφαρμόσαμε την τεχνική ανάγνωσης γεγονότων με το μοντέλο ΓΝΓ είναι το First Story Detection (FSD) [148]. Το FSD είναι ένα σύνολο από tweets που μαζεύτηκαν από την 1η Απριλίου 2009 μέχρι την 14 Οκτωβρίου 2009 μέσω του Twitter streaming API και περιλαμβάνει γεγονότα, όπως τον θάνατο διασημοτήτων, φυσικές καταστροφές, αθλητικά γεγονότα, πολιτικά γεγονότα, ψυχαγωγίας, επαγγελματικά γεγονότα, πυροβολισμούς και αεροπορικά δυστυχήματα. Τα tweets κατηγοριοποιήθηκαν από δύο εξειδικευμένους εργαζόμενους στον χώρο των κοινωνικών δικτύων.

Πειράματα πραγματοποιήθηκαν όπως και στις προηγούμενες περιπτώσεις. Ο πίνακας 8.8 παρουσιάζει τα πειραματικά αποτελέσματα με την μετρική περιεχομένου όπου η ακρίβεια φτάνει το 98,14%.

	Macro Precision	Macro Recall	Macro F-Measure	Micro F-Measure
CS 2Grams	0,8967	0,8965	0,8929	0,9841
CS 3Grams	0,9112	0,9134	0,9099	0,9867
CS 4Grams	0,8989	0,9024	0,8973	0,9845
CS 5Grams	0,8848	0,8877	0,8817	0,9814
CS 6Grams	0,8767	0,8679	0,8670	0,9756
CS 7Grams	0,8790	0,8630	0,8652	0,9721
CS 8Grams	0,8731	0,8463	0,8524	0,9672
CS 9Grams	0,8758	0,8398	0,8491	0,9637
CS 10Grams	0,8651	0,8327	0,8400	0,9615

Πίνακας 8.8 FSD: Αναγνώριση κοινοτήτων με Μετρική περιεχομένου

Χρησιμοποιώντας τις μετρικές MCS βλέπουμε την ακρίβεια να έχει ελαττωθεί σε σχέση με την μετρική περιεχομένου. Φαινόμενο που το παρατηρήσαμε και στα προηγούμενα σύνολα δεδομένων.

	Macro Precision	Macro Recall	Macro F-Measure	Micro F-Measure
MCSN 2Grams	0,8075	0,7975	0,7936	0,9198
MCSN 3Grams	0,8508	0,8328	0,8328	0,9566
MCSN 4Grams	0,8419	0,8193	0,8222	0,9601
MCSN 5Grams	0,8305	0,8129	0,8140	0,9566
MCSN 6Grams	0,8495	0,8317	0,8330	0,9615
MCSN 7Grams	0,8449	0,8265	0,8277	0,9610
MCSN 8Grams	0,8450	0,8287	0,8289	0,9601
MCSN 9Grams	0,8528	0,8269	0,8322	0,9593
MCSN 10Grams	0,8507	0,8159	0,8241	0,9562

Πίνακας 8.9 FSD: Αναγνώριση κοινοτήτων με Μετρική ομοιότητας MCSN

	Macro Precision	Macro Recall	Macro F-Measure	Micro F-Measure
MCSDE 2Grams	0,8806	0,8727	0,8702	0,9721

MCSDE 3Grams	0,8560	0,8456	0,8417	0,9628
MCSDE 4Grams	0,8503	0,8224	0,8273	0,9597
MCSDE 5Grams	0,8472	0,8275	0,8297	0,9597
MCSDE 6Grams	0,8575	0,8330	0,8378	0,9606
MCSDE 7Grams	0,8574	0,8272	0,8347	0,9593
MCSDE 8Grams	0,8601	0,8236	0,8328	0,9579
MCSDE 9Grams	0,8480	0,8168	0,8237	0,9562
MCSDE 10Grams	0,8490	0,8148	0,8225	0,9553

Πίνακας 8.10 FSD: Αναγνώριση κοινοτήτων με Μετρική ομοιότητας MCSDE

Η ομοιότητα Jaccard χρησιμοποιήθηκε στο σύνολο δεδομένων FSD παρουσιάζοντας χειρότερα αποτελέσματα από ότι η μετρική περιεχομένου.

	Macro Precision	Macro Recall	Macro F-Measure	Micro F-Measure
JSC 2Grams	0,7410	0,6811	0,6916	0,8671
JSC 3Grams	0,8330	0,8262	0,8214	0,9376
JSC 4Grams	0,8393	0,8455	0,8341	0,9442
JSC 5Grams	0,8485	0,8572	0,8419	0,9446
JSC 6Grams	0,8427	0,8535	0,8350	0,9446
JSC 7Grams	0,8380	0,8549	0,8326	0,9500
JSC 8Grams	0,8266	0,8482	0,8237	0,9513
JSC 9Grams	0,8331	0,8481	0,8263	0,9539
JSC 10Grams	0,8337	0,8403	0,8218	0,9531

Πίνακας 8.11 FSD: Αναγνώριση κοινοτήτων με Μετρική ομοιότητας Jaccard

8.4.Συμπεράσματα

Τα κείμενα που αναρτώνται στα κοινωνικά δίκτυα είναι μια χρήσιμη πηγή πληροφορίας από όπου μπορούν να αναγνωριστούν γεγονότα σε πολύ μικρό χρονικό διάστημα από την στιγμή που θα συμβούν ή που θα ξεκινήσει η προετοιμασία για να συμβούν. Τεχνικές επεξεργασίας φυσικής γλώσσας προσφέρουν πολύτιμα εργαλεία για την αναγνώριση των γεγονότων.

Περισσότερο ενδιαφέρον παρουσιάζουν οι τεχνικές που όχι απλά αναγνωρίζουν ότι συμβαίνει ένα γεγονός αλλά καταλαβαίνουν και την φύση του γεγονότος αυτού. Αυτού του είδους οι τεχνικές ανάγονται σε μια μέθοδο επιβλεπόμενης μηχανικής μάθησης όπου οι κατηγορίες που αναμένουμε να αναγνωριστούν έχουν περιγραφεί από πριν.

Το μοντέλο κατηγοριοποίησης κειμένων με ΓΝΓ παρουσίασε πολύ καλά αποτελέσματα αναγνώρισης γεγονότων καθώς και κατηγοριοποίησης tweets στα γεγονότα που σχετίζονται. Η μετρική περιεχομένου σε όλα τα σύνολα δεδομένων είχε καλύτερα αποτελέσματα.

9. Εφαρμογή σε Συστήματα Συστάσεων σε Εμπορικές Πλατφόρμες

Στην παρούσα ενότητα κάνουμε πρώτα μια επισκόπηση στις μεθόδους κατασκευής συστημάτων συστάσεων που προσανατολίζονται στην σύσταση εμπορικών προϊόντων. Θα παρουσιάσουμε συνοπτικά τις τεχνικές που χρησιμοποιούν φίλτρα που βασίζονται στο περιεχόμενο, συνεργατικά φίλτρα, εναλλακτικές, καθώς και υβριδικές τεχνικές. Έπειτα θα αναφερθούμε στον τρόπο που τα συστήματα συστάσεων χρησιμοποιούν τεχνικές ανάκτησης πληροφορίας. Τέλος υλοποιούμε ένα καινοτόμο σύστημα συστάσεων που χρησιμοποιεί το μοντέλο ΓΝΓ όπως το παρουσιάσαμε στην ενότητα 4 και έρχεται να λύσει το cold start problem και το profiling, καθώς βασίζεται μόνο στο κείμενο που συνοδεύει και παρουσιάζει κάθε προϊόν στον τελικό καταναλωτή.

9.1.Εισαγωγή στα Συστήματα Συστάσεων

Τα Συστήματα Συστάσεων είναι μοντέλα και αλγόριθμοι που μπορούν να κάνουν συστάσεις σε χρήστες. Δηλαδή μέθοδοι που μελετούν την συμπεριφορά ενός χρήστη μέσα σ' ένα πληροφοριακό σύστημα και μπορούν να του προτείνουν κάποιο προϊόν, υπηρεσία, δημοσιογραφικό άρθρο, έναν άλλον χρήστη και γενικώς οτιδήποτε μπορεί να τον ενδιαφέρει. Στην παρούσα εργασία μας ενδιαφέρει να προσφέρουμε συστάσεις για ένα εμπορικό προϊόν ή υπηρεσία σε έναν υποψήφιο πελάτη με σκοπό να ταιριάζει καλύτερα στις ανάγκες του. Το αντικείμενο που πρόκειται να συσταθεί από εδώ και πέρα θα το αποκαλούμε προϊόν.

Η χρήση συνεργατικών φίλτρων είναι μια από τις δύο πιο διαδεδομένες τεχνικές κατασκευής συστημάτων συστάσεων. Στην μέθοδο αυτή μαζεύονται πληροφορίες για το τι προτιμήσεις έχει ένα μεγάλο πλήθος χρηστών. Για κάθε χρήστη βρίσκουμε τους χρήστες που έχουν όμοια προφίλ με αυτόν και με βάση τις προτιμήσεις τους προτείνονται προϊόντα στον αρχικό χρήστη.

Τα συστήματα συστάσεων που χρησιμοποιούν φίλτρα που βασίζονται στο περιεχόμενο χρησιμοποιούν την περιγραφή ενός αντικειμένου αλλά και το προφίλ ενδιαφερόντων που έχει κάθε χρήστης. Η ιδέα βασίζεται στο ότι ένας χρήστης έχει ένα πλήθος ενδιαφερόντων και κάθε αντικείμενο ικανοποιεί ένα πλήθος ενδιαφερόντων. Όσο πιο πολύ ταυτίζονται αυτά τα δύο, τόσο πιο μεγάλο θα είναι το ενδιαφέρον του χρήστη προς το αντικείμενο αυτό.

Υπάρχουν υβριδικές μέθοδοι που συνδυάζουν τα συνεργατικά φίλτρα με τα φίλτρα που βασίζονται στο περιεχόμενο, καθώς και άλλες που χρησιμοποιούν εναλλακτικά κριτήρια όπως τον χρόνο παρατήρησης κάθε προϊόντος. Εμείς θα ασχοληθούμε αναλυτικά με το να εφαρμόσουμε τεχνικές ανάκτησης πληροφορίας που επεξεργάζονται τα κείμενα περιγραφής κάθε προϊόντος για να παράγουν συστάσεις.

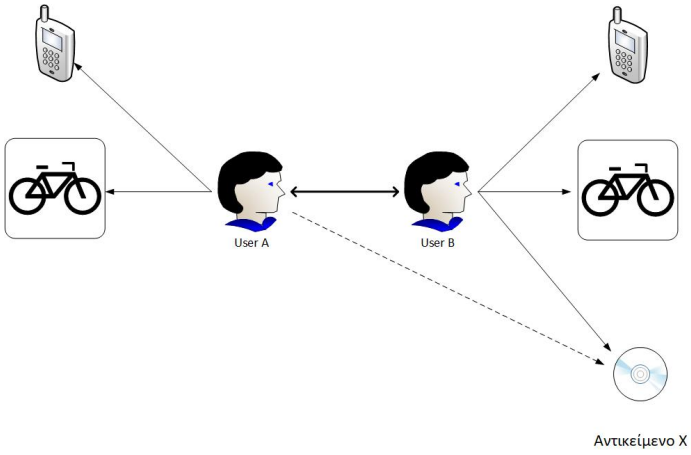
Κάθε προϊόν σε ένα εμπορικό πληροφοριακό σύστημα συνοδεύεται από ένα κείμενο που το περιγράφει. Σκοπός μας είναι με βάση αυτό το κείμενο να προταθούν στον χρήστη όλα τα άλλα προϊόντα που μοιάζουν πιο πολύ με αυτό που επέλεξε. Μοντέλα όπως τα Standard και Extended Boolean, Vector Space, Latent Semantic Analysis, Probabilistic Semantic Analysis, Binary Independence τα οποία θα περιγραφούν, μπορούν να χρησιμοποιηθούν.

Μια καινοτόμος μέθοδος του πεδίου της ανάκτησης πληροφορίας είναι αυτή που χρησιμοποιεί τα N-γράμματα γράφων για να ανακτήσει και να κατηγοριοποιήσει κείμενα. Με βάση αυτή την μέθοδο κατασκευάσαμε ένα σύστημα συστάσεων στα πρότυπα των φίλτρων που βασίζονται στο περιεχόμενο.

Το σύστημα συστάσεων που κάνει χρήση γράφων Ν-γραμμάτων μοντελοποιεί κάθε κείμενο από τα προϊόντα που επέλεξε ένας χρήστης με την μορφή ενός γράφου και κάθε θεματική ενότητα προϊόντων επίσης ως έναν γράφο. Στη συνέχεια με μια διαδοχική σειρά συγκρίσεων γράφων είναι σε θέση να προτείνει τα βέλτιστα προϊόντα για κάθε χρήστη που εισέρχεται σ' ένα εμπορικό πληροφοριακό σύστημα.

9.2.Συνεργατικά Φίλτρα

Η χρήση συνεργατικών φίλτρων (collaborative filtering) [149] σκοπό έχει να δημιουργήσει προφίλ χρηστών για τις προτιμήσεις τους καθώς και σχέσεις μεταξύ τους. Η ιδέα βασίζεται στην αρχή ότι "το προϊόν που ενδιέφερε τους φίλους μου αναμένεται να ενδιαφέρει και εμένα". Οπότε έχουμε ένα σύνολο χρηστών και ξέρουμε ότι ο κάθε ένας ενδιαφέρεται ή δεν ενδιαφέρεται για ένα σύνολο συγκεκριμένων αντικειμένων. Θέλοντας να μάθουμε αν ένας χρήστης Α ενδιαφέρεται για το αντικείμενο Χ κοιτάζουμε το σύνολο χρηστών που έχουν πιο κοινά ενδιαφέροντα με τον χρήστη Α και έχουν εκφράσει άποψη για το αντικείμενο Χ. Με βάση την άποψη αυτών των χρηστών αποφασίζουμε για το αν θα συστήσουμε το αντικείμενο Χ στον χρήστη Α ή όχι.



Σχήμα 9.1 Σύσταση προϊόντος μέσω συνεργατικών φίλτρων

Για να γίνει πιο κατανοητή η μέθοδος θα περιγράψουμε το ακόλουθο παράδειγμα. Σε μια πλατφόρμα που εμπορεύεται δίσκους μουσικής είναι εγγεγραμμένο ένα πλήθος χρηστών. Για κάθε χρήστη μπορεί να έχει φτιαχτεί ένα προφίλ με βάση τους δίσκους που έχει αγοράσει. Οι χρήστες που έχουν αγοράσει τους περισσότερους κοινούς δίσκους αναμεταξύ τους θα έχουν και παρόμοια προφίλ. Οπότε το προϊόν που θα προτείνουμε στον χρήστη Χ θα είναι ένας δίσκος που δεν τον έχει αγοράσει αλλά τον έχουν αγοράσει οι άλλοι χρήστες που έχουν παρόμοιο προφίλ με αυτόν. Για την κατασκευή των προφίλ μπορούμε να περιλάβουμε και άλλες πληροφορίες όπως το σε ποιους δίσκους έχει γίνει review, έχουν προστεθεί στο wantlist ακόμη και πληροφορίες όπως ο τόπος διαμονής του χρήστη.

Η λειτουργία των συνεργατικών φίλτρων δεν γίνεται αντιληπτή από τον χρήστη, αφού συνήθως δεν του ζητούνται ρητά πληροφορίες να εισαγάγει στο σύστημα, απλώς παρακολουθείται η κίνηση του και οι επιλογές του. Αυτό από την άλλη μπορεί να έχει το δυσάρεστο αποτέλεσμα οι χρήστες να είναι δυσαρεστημένοι λόγω του ότι δεν τους αρέσει να παρακολουθούνται και να καταγράφονται οι κινήσεις τους. Το ζήτημα των προσωπικών δεδομένων είναι πάντα ένα ευαίσθητο θέμα που οποιοδήποτε εμπορικό πληροφοριακό σύστημα θα πρέπει να λάβει υπόψη του. Άλλες μέθοδοι όπως η χρήση των φίλτρων που βασίζονται στο περιεχόμενο που θα διαπραγματευθούμε στην επόμενη ενότητα δεν χρειάζονται την κατασκευή προφίλ χρηστών, οπότε ξεπερνιέται αυτός ο περιορισμός.

Η μέθοδος των συνεργατικών φίλτρων βασίζεται στα ακόλουθα τρία βήματα.

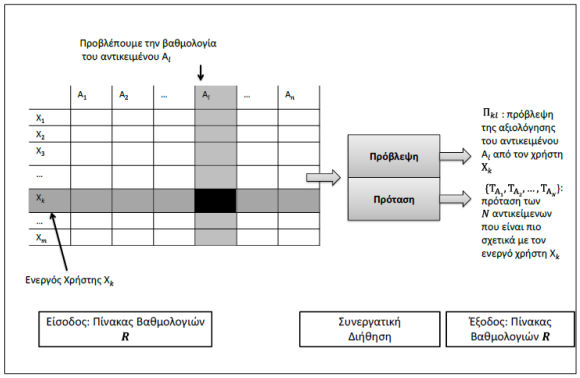
1. Κατασκευή ενός rating για κάθε χρήστη που προσδιορίζει ποια προϊόντα του αρέσουν, ποια όχι και για ποια δεν έχει εκφράσει ακόμη άποψη.

2. Εύρεση των χρηστών που έχουν παρόμοια μοτίβα rating με τον χρήστη Χ που εξετάζουμε.

3. Χρήση των ratings των χρηστών που βρήκαμε στο δεύτερο βήμα για να προταθούν προϊόντα για τα οποία ο χρήστης X που μας ενδιαφέρει δεν έχει εκφράσει άποψη.

Το βήμα 2 μπορεί να υλοποιηθεί με την χρήση του αλγορίθμου Nearest Neighbor Search (k-NN) [150]. Ο k-NN είναι ο πιο διαδεδομένος αλγόριθμος που χρησιμοποιείται για την επίλυση του προβλήματος βελτιστοποίησης δοθέντος ενός συνόλου S διανυσμάτων και ενός διανύσματος q. Ο k-nn μπορεί να βρει με καλή προσέγγιση τα k πιο κοντινά διανύσματα στο q. Στην περίπτωση μας κάθε προφίλ χρήστη θα παρουσιάζεται ως ένα διάνυσμα όπου η κάθε διάσταση θα αντιστοιχεί σ' ένα προϊόν. Αυτό μπορεί να το πετύχει είτε με κατακερματισμό του χώρου και αποκλεισμό των υποχώρων που δεν οδηγούν σε βέλτιστα αποτελέσματα (Branch and Bound), είτε με το να ομαδοποιεί τα πιο κοινά διανύσματα μαζί, είτε ακόμη και με απλή σειριακή αναζήτηση.

Μια άλλη μορφή συνεργατικών φίλτρων που χρησιμοποιείται από την Amazon [151] είναι η κατασκευή ενός δισδιάστατου πίνακα που στις γραμμές και τις στήλες του έχει εξίσου όλα τα προϊόντα. Αυτός ο πίνακας αντικειμένου προς αντικείμενο προσδιορίζει ότι συνήθως ένας χρήστης που επιλέγει ένα προϊόν το επιλέγει μαζί με κάποια άλλα που προσδιορίζονται από την αντίστοιχη γραμμή του πίνακα. Ο πίνακας αυτός συνοψίζει τις επιλογές που έκαναν οι χρήστες στο παρελθόν και με την επιλογή ενός ή περισσότερων προϊόντων από έναν χρήστη είναι σε θέση να χρησιμοποιηθεί για να γίνουν οι κατάλληλες συστάσεις που αντιστοιχούν στις γραμμές των προϊόντων που επέλεξε.



Σχήμα 9.2 Πίνακας προτιμήσεων

Όπως φαίνεται στο Σχήμα 9.2 Ο χρήστης δεν έχει εκφράσει άποψη για το αντικείμενο A_I αλλά μπορεί να θεωρηθεί ότι η προτίμηση του είναι ίση με τον μέσο όρο της βαθμολογίας που έχουν οι χρήστες για αυτό το αντικείμενο που έχουν πιο όμοια προφίλ μαζί του.

Τα συστήματα συστάσεων που κάνουν χρήση συνεργατικών φίλτρων μπορούν περαιτέρω να χωριστούν σε δύο κατηγορίες στα Memory-based (Βασισμένα στην μνήμη) Modeled-based (Βασισμένα στο μοντέλο). Στις επόμενες δύο υποενότητες τα παρουσιάζουμε αναλυτικά.

9.2.1.Βασισμένα στην μνήμη

Στα συνεργατικά φίλτρα που βασίζονται στην μνήμη, αρχικά επιλέγουμε ένα σύνολο U από χρήστες που είναι πιο όμοιοι με τον χρήστη που μας ενδιαφέρει και έπειτα με βάση τις προτιμήσεις αυτών των χρηστών βαθμολογούμε κατάλληλα τα προϊόντα τα οποία δεν έχει ακόμη προσπελάσει ο αρχικός μας χρήστης.

Η εύρεση του συνόλου U των πιο όμοιων χρηστών μπορεί να γίνει με το να συγκρίνουμε τον αρχικό μας X χρήστη με τους υπόλοιπους του συστήματος και να κρατήσουμε αυτούς που έχουν την υψηλότερη ομοιότητα. Η ομοιότητα μεταξύ δύο χρηστών μπορεί να ποσοτικοποιηθεί με το να πάρουμε τα βάρη για όλα τα προϊόντα που έχουν προσπελάσει οι δύο χρήστες x, y. Αυτό μπορούμε να το πετύχουμε με την εξίσωση Pearson Correlation 9.1

$$simil(x,y) = \frac{\sum_{i \in I_{xy}} (r_{x,i} - \bar{r}_x) \cdot (r_{y,i} - \bar{r}_y)}{\sqrt{\sum_{i \in I_{xy}} (r_{x,i} - \bar{r}_x)^2 \cdot (r_{y,i} - \bar{r}_y)^2}} \quad (9.1)$$

Όπου I_{xy} είναι το σύνολο αντικειμένων που έχουν προσπελαστεί και βαθμολογηθεί από αμφοτέρους τους χρήστες και $\bar{r}_x$ ο μέσος όρος βαρύτητας που δίνει ο χρήστης X στα προϊόντα που προσπέλασε.

Αν θεωρήσουμε το προφίλ κάθε χρήστη ως ένα διάνυσμα αντί να χρησιμοποιήσουμε την εξίσωση 9.1 μπορούμε να χρησιμοποιήσουμε την 9.2 που εκφράζει την ομοιότητα μεταξύ δύο χρηστών με βάση τον συντελεστή συνημίτονου. Οι όροι ορίζονται όπως και πιο πάνω.

$$\cos\left(\vec{x}, \vec{y}\right) = \frac{\vec{x} \cdot \vec{y}}{\left\|\vec{x}\right\| \cdot \left\|\vec{y}\right\|} = \frac{\sum_{i \in I_{xy}} r_{x,i} \cdot r_{y,i}}{\sqrt{\sum_{i \in I_x} {r_{x,i}}^2} \cdot \sqrt{\sum_{i \in y} {r_{y,i}}^2}} \quad (9.2)$$

Αφού βρούμε το σύνολο U των πιο όμοιων χρηστών προς τον χρήστη που μας ενδιαφέρει είμαστε σε θέση να εκτιμήσουμε το rating των προϊόντων που ακόμη δεν έχει προσπελάσει με βάση έναν από τους τύπους 9.3 έως 9.5

$$r_{u,i} = \frac{1}{N} \sum_{u' \in U} r_{u',i} \quad (9.3)$$

$$r_{u,i} = k \sum_{u' \in U} simil(u,u') r_{u',i} \quad (9.4)$$

$$r_{u,i} = \bar{r}_u + k \sum_{u' \in U} simil(u,u')(r_{u',i} - \bar{r}_{u'}) \quad (9.5)$$

Όπου $\bar{r}_u$ είναι ο μέσο όρος των βαθμών που δίνει ο χρήστης u για τα προϊόντα που έχει προσπελάσει και το k δίνεται από τον τύπο 9.6

$$\text{k} = \frac{1}{\sum_{u' \in U} |simil(u,u')|} \quad (9.6)$$

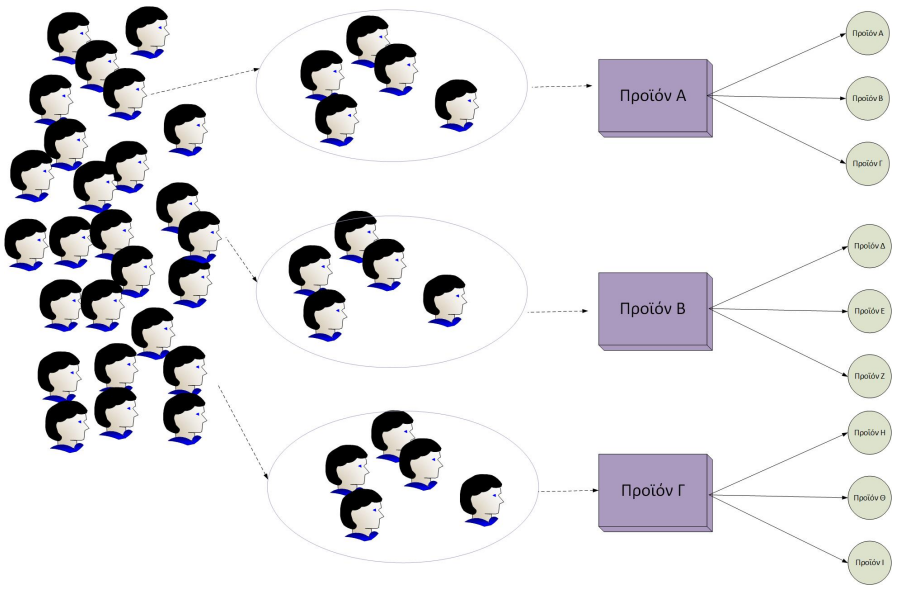
Ένα μεγάλο πλεονέκτημα των συνεργατικών φίλτρων που είναι βασισμένα στην μνήμη είναι ότι τα αποτελέσματα των συστάσεων μπορούν να ερευνηθούν και να εξηγηθούν άμεσα εφόσον έχουμε κατανοήσει την μέθοδο.

Νέα δεδομένα μπορούν να προστεθούν και να γίνουν συστάσεις για αυτά, εφόσον βέβαια μια αρχική κρίσιμη μάζα χρηστών τα προσπελάσει και εκφράσει άποψη για αυτά. Σε νέους χρήστες δεν θα μπορεί να προσφέρει συστάσεις γιατί προϋποθέτει ο χρήστης να έχει ήδη κάποια δραστηριότητα στο σύστημα ούτως ώστε να παρέχει συστάσεις.

9.2.2.Βασισμένα στο μοντέλο

Στην Model-Based αρχιτεκτονική αναπτύσσουμε μοντέλα που χρησιμοποιούν τεχνικές εξόρυξης δεδομένων, αλγόριθμους μηχανικής μάθησης και αναγνωρίσεις κοινών μοτίβων (patterns) με βάση δεδομένα στα οποία τα εκπαιδεύσαμε. Αυτά τα μοντέλα ενσωματώνουν τεχνικές από Bayesian Networks, clustering models και latent semantic models[152]. Πιο συγκεκριμένα μέθοδοι όπως οι singular value decomposition[20], probabilistic latent semantic analysis[24], Multiple Multiplicative Factor[153], Latent Dirichlet allocation[28] καθώς και μοντέλα markov decision process based προσαρμόζονται για τις ανάγκες κατασκευής ενός συστήματος συστάσεων.

Αυτές οι μέθοδοι έχουν την ικανότητα να βρίσκουν κείμενα, αντικείμενα, προϊόντα που έχουν κοινά χαρακτηριστικά μεταξύ τους. Είναι φανερό πως προσφέρουν βασικά εργαλεία για την παραγωγή συστάσεων. Οι τεχνικές κατηγοριοποίησης (classification) και συσταδοποίησης (clustering) ενδείκνυται γιατί είναι σε θέση να διαχειρίζονται μεγάλο πλήθος δεδομένων με βάση τα χαρακτηριστικά που έχει κάθε αντικείμενο ή χρήστης. Αυτές οι μέθοδοι είναι πιο ολιστικές και καταφέρνουν να αποκαλύψουν λανθάνοντα (latent) και κρυμμένα στοιχεία.



Σχήμα 9.3 Συστάσεις με βάση την συστάδα που ανήκουν οι χρήστες

Στο Σχήμα 9.3 φαίνεται πως μετά την συσταδοποίηση των χρηστών, χρήστες από διαφορετικές συστάδες ακόμη και αν επιλέξουν το ίδιο προϊόν μπορεί να λάβουν διαφορετικές συστάσεις. Η συστάδα που βρίσκεται ένας χρήστης καθορίζει το ποια προϊόντα θα του συσταθούν.

Τα περισσότερα από τα μοντέλα βασίζονται στο να δημιουργήσουν μια ταξινόμηση και ομαδοποίηση χαρακτηριστικών για να ανακαλύψουμε την συμπεριφορά του χρήστη με βάση τα δεδομένα που έχουμε για αυτόν. Κοινές συμπεριφορές μεταξύ χρηστών μας επιτρέπουν να κάνουμε προβλέψεις για περαιτέρω συστάσεις σε προϊόντα ή υπηρεσίες που μπορεί να μην έχουν ακόμη επιλέξει οι ίδιοι αλλά έχουν επιλέξει άτομα τα οποία βρίσκονται μαζί στην ίδια ομαδοποίηση και κατάταξη.

Στα θετικά των συνεργατικών φίλτρων που είναι βασισμένα στο μοντέλο είναι ότι μπορεί να διαχειριστεί αποδοτικά το γεγονός ότι οι πληροφορίες για τα δεδομένα που έχουμε από τους χρήστες προς τα προϊόντα είναι αραιές. Δηλαδή το ότι οι χρήστες έχουν εκφράσει άποψη για λίγα προϊόντα σε σχέση με το πλήθος των προϊόντων που είναι διαθέσιμα στο πληροφοριακό σύστημα.

Από την άλλη πλευρά η υψηλή ακρίβεια της μεθόδου μειώνεται όσο το πλήθος τον προϊόντων αυξάνεται λόγω του ότι η χρήση των άδηλων μεταβλητών θα γίνεται μεγαλύτερη.

9.2.3.Υβριδικά συνεργατικά φίλτρα

Έχουν υλοποιηθεί συνεργατικά φίλτρα που συνδυάζουν τις τεχνικές που είναι βασισμένες στην μνήμη με τις τεχνικές που είναι βασισμένες στο μοντέλο. Αυτά τα φίλτρα έχουν δημιουργηθεί με σκοπό να ξεπεράσουν τους περιορισμούς που προκύπτουν από την κάθε μία τεχνική ξεχωριστά.

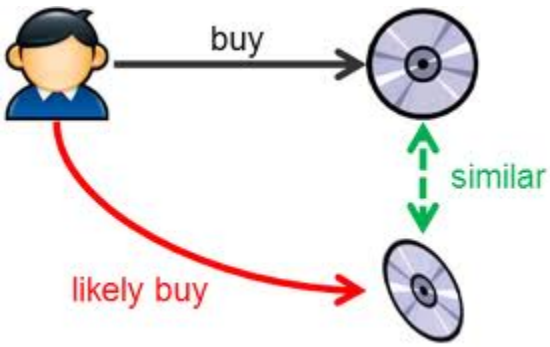
Για τις ανάγκες του Google news προτάθηκε ένα συνεργατικό φίλτρο [154] που συνδυάζει σ' ένα γραμμικό μοντέλο την συσταδοποίηση MinHash [155], το Probabilistic Latent Sematic Indexing (PLSI) με το ποιες ειδήσεις και πόσες φορές της έχει επισκεφτεί ένας χρήστης.

Τέτοιες υβριδικές μέθοδοι έχουν πολύ καλή ακρίβεια στις συστάσεις τους αλλά είναι πιο δύσκολο να υλοποιηθούν και η πολυπλοκότητα των υπολογισμών που έχουν να κάνουν είναι αισθητά μεγαλύτερη σε σχέση με μια απλή μέθοδο.

9.3.Φίλτρα που Βασίζονται στο Περιεχόμενο

Τα φίλτρα που βασίζονται στο περιεχόμενο (Content Based filtering)[156] είναι συστήματα που βασίζονται στην περιγραφή ενός αντικειμένου και στο προφίλ ενδιαφερόντων που έχει κάθε χρήστης. Το προφίλ του χρήστη περιγράφει τον τύπο των αντικειμένων που ενδιαφέρουν τον χρήστη. Έπειτα υπάρχει μια μέθοδος όπου συγκρίνει όλα τα αντικείμενα με το προφίλ του κάθε χρήστη και αποφασίζει ποια τον ενδιαφέρουν για να του συσταθούν.

Αυτό που κάνουν τα Content Based Recommendation Systems είναι να αναλύουν τις περιγραφές των αντικειμένων που έχει δει ο χρήστης καθώς και αυτά που δεν έχει δει, με σκοπό να αναγνωρίζει από αυτά που δεν έχει δει ακόμη αυτά που τον ενδιαφέρουν.

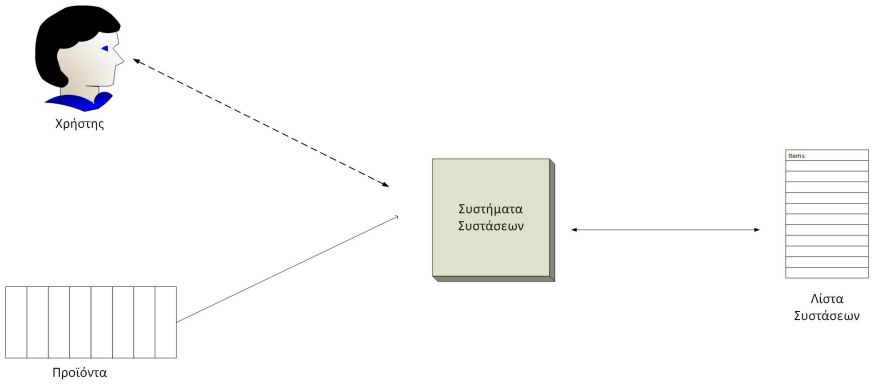


Σχήμα 9.4 Συστάσεις που βασίζονται στην ομοιότητα των προϊόντων

Το πρώτο θέμα που θα μας απασχολήσει είναι ο τρόπος που αναπαριστάται η περιγραφή καθενός αντικειμένου εσωτερικά στο σύστημα μας. Αυτό που θέλουμε είναι να έχουμε τα δεδομένα μας σε μια

δομημένη μορφή. Ο καλύτερος τρόπος είναι να τα αποθηκεύσουμε σε tables μιας βάσης δεδομένων. Οι στήλες θα περιγράφουν τις ιδιότητες του αντικειμένου. Κάθε αντικείμενο θα περιγράφεται από μια εγγραφή και θα ξεχωρίζονται αναμεταξύ τους από ένα μοναδικό ID.

Στο Σχήμα 9.5 φαίνεται πως κάθε χρήστης έχει ένα προφίλ που βασίζεται στα χαρακτηριστικά των αντικειμένων που έχει βαθμολογήσει. Υπάρχει μια βάση δεδομένων όπου για κάθε προϊόν αναφέρεται ποια χαρακτηριστικά έχει. Με βάση το προφίλ του χρήστη ανακτούμε από την βάση και κατατάσσουμε τα προϊόντα που ταιριάζουν καλύτερα σε αυτά που έχει προσπελάσει ούτως ώστε να του συσταθούν.



Σχήμα 9.5 Περιγραφή χρήστη και προϊόντος για την παραγωγή λίστας συστάσεων

Πολλές φορές δεν μας δίνονται οι περιγραφές των δεδομένων σε δομημένη μορφή. Μπορεί να είναι απλώς ένα κείμενο που περιγράφει ένα αντικείμενο. Τότε καλούμαστε εμείς να μετατρέψουμε το απλό κείμενο σε μια δομημένη αναπαράσταση. Πολλές λέξεις κυρίως αυτές που μας ενδιαφέρουν μπορούμε να τις αντιμετωπίζουμε ως ιδιότητες. Η ιδιότητα θα είναι boolean και θα υποδηλώνει αν υπάρχει μέσα στα κείμενο ή θα είναι κάποιος ακέραιος και θα υποδηλώνει το πόσες φορές συναντήσαμε την λέξη αυτή στο κείμενο.

Ένα επόμενο θέμα που προκύπτει στην διαδικασία αυτή είναι ότι λέξεις όπως "επεξεργάζομαι", "επεξεργαστής", "επεξεργασία" θα ξεχωρίσουν ως διαφορετικές έννοιες. Θα ήταν προτιμότερο να της ενοποιήσουμε σε έναν όρο. Αυτό επιτυγχάνεται με το να εργαστούμε και να θεωρήσουμε ως όρους που μας ενδιαφέρουν τις ρίζες των λέξεων [157]. Έπειτα μπορούμε να κάνουμε χρήση μιας μεθόδου που μετρά την συχνότητα που συναντήθηκε ένας όρος αντιστρόφως ανάλογα με τον αριθμό των κειμένων [158]. Όσο πιο μεγάλος είναι ο αριθμός w(t,d) σημαίνει πως τόσο πιο συχνά συναντήθηκε ο αντίστοιχος όρος και τόσο πιο πολύ χαρακτηρίζει το αντικείμενο. Ο τύπος δίνεται από την σχέση 9.7

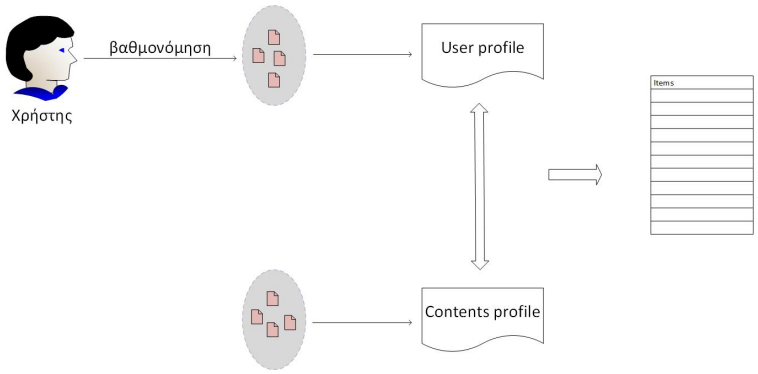
$$w(t,d) = \frac{tf_{t,d}\log\left(\frac{N}{df_t}\right)}{\sqrt{\sum_i \left(tf_{t,d}\right)^2 log\left(\frac{N}{tf_{t_i}}\right)^2}} \quad (9.7)$$

Όπου d είναι ένα κείμενο, t είναι ένας όρος μέσα σε ένα κείμενο, N είναι το πλήθος των κειμένων, $tf_{t,d}$ είναι η συχνότητα που υπάρχει ο όρος t στο συγκεκριμένο κείμενο d και df_t είναι ο αριθμός των κειμένων στην συλλογή των N κειμένων που περιέχουν τον t όρο.

Αφού αναπαραστήσουμε την περιγραφή καθενός αντικειμένου, μας ενδιαφέρει να δημιουργήσουμε προφίλ που θα αντιπροσωπεύουν τα ενδιαφέροντα που μπορεί να έχει κάθε χρήστης. Τα προφίλ μπορούν να κατασκευάζονται κυρίως από δύο τύπους πληροφοριών. Τις προτιμήσεις των χρηστών και το ιστορικό των χρηστών.

Τα προφίλ που βασίζονται στις προτιμήσεις των χρηστών περιγράφουν τους τύπους των αντικειμένων που ενδιαφέρουν τον χρήστη. Με αυτό τον τρόπο θα έχουν τα n αντικείμενα που ενδιαφέρουν περισσότερο κάθε χρήστη και θα είναι σε θέση να προβλέψουν κατά πόσο ο χρήστης ενδιαφέρεται για κάποιο άλλο αντικείμενο που υπάρχει στο σύστημα.

Ο καλύτερος τρόπος για την κατασκευή του προφίλ του χρήστη είναι να καταγράφουμε όλες τις αλληλεπιδράσεις του χρήστη με το σύστημα όπως όλα τα αντικείμενα που επισκέφτηκε, αγόρασε και αναζήτησε. Υπάρχουν πολλές χρήσεις του ιστορικού των αλληλεπιδράσεων του χρήστη. Το σύστημα μπορεί να έχει καταγεγραμμένα τα αντικείμενα που έχει δει και ενδιαφέρουν αυτόν τον καιρό τον χρήστη. Όσα αντικείμενα έχει ήδη επισκεφτεί ή έχει αγοράσει ο χρήστης θα φιλτραριστούν και δεν θα προταθούν για να συσταθούν αλλά μπορούν να είναι τα δεδομένα που θα εκπαιδεύσουν έναν αλγόριθμο μηχανικής μάθησης.

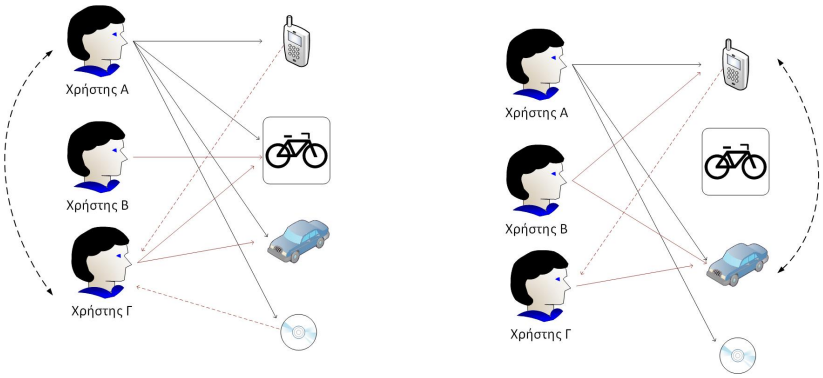


Σχήμα 9.6 Βαθμονόμηση ομοιότητας του προφίλ προϊόντος με το προφίλ χρήστη

Το ιστορικό ενός χρήστη μπορεί να χρησιμοποιηθεί από ένα recommendation σύστημα βασισμένο σε συγκεκριμένους κανόνες (ruled-based) για να παράγει συστάσεις. Αυτό σημαίνει πως οι συστάσεις που θα γίνουν βασίζονται σε συγκεκριμένους κανόνες που εφαρμόζονται πάνω στα ιστορικά δεδομένα του χρήστη. Παράδειγμα κανόνα για να διευκρινίσουμε τι εννοούμε είναι για κάθε βιβλίο ή ταινία που έχει δει ο χρήστης να συστήνουμε την συνέχεια της. Αν έχει δει την ταινία ο Άρχοντας των δακτυλιδιών το πρώτο μέρος είναι πολύ πιθανόν έπειτα να ενδιαφερθεί και για τον Άρχοντα των δακτυλιδιών το δεύτερο μέρος.

Αφού μαζέψουμε τα δεδομένα που κρίνουμε για κάθε χρήστη περνάμε στην επόμενη φάση όπου θα πρέπει να τα μοντελοποιήσουμε και να τα επεξεργαστούμε. Για κάθε χρήστη θα πρέπει να παράγουμε ένα μοντέλο των προτιμήσεων του, έπειτα να χρησιμοποιήσουμε αλγόριθμους κατηγοριοποίησης με σκοπό κάθε φορά που δίνεται ένα καινούριο αντικείμενο το σύστημα να είναι σε θέση να προβλέψει τον βαθμό που τον ενδιαφέρει. Υπάρχουν πολλοί αλγόριθμοι που μπορούν να ολοκληρώσουν το content based recommendation σύστημα μας. Ο κάθε ένας από τους αλγόριθμους αυτούς χρειάζεται μια εκτενή παρουσίαση για να κατανοηθεί και να υλοποιηθεί. Παρακάτω θα περιγράψουμε συνοπτικά τους βασικότερους.

Στο σχήμα 9.7 φαίνεται η διαφορά μεταξύ των φίλτρων που βασίζονται στο περιεχόμενο με τα συνεργατικά φίλτρα. Στα συνεργατικά φίλτρα βρίσκουμε ποιοι χρήστες μοιάζουν περισσότερο και αναλόγως με το τι έχει επιλέξει ο καθένας γίνονται συστάσεις στους υπόλοιπους. Στα φίλτρα που βασίζονται στο περιεχόμενο βρίσκουμε ποια προϊόντα είναι όμοια. Αν κάποιος χρήστης επιλέξει ένα προϊόν θα του συσταθούν τα υπόλοιπα προϊόντα που είναι περισσότερο όμοια με αυτό που επέλεξε.



Σχήμα 9.7 Φίλτρα που βασίζονται στο περιεχόμενο και συνεργατικά φίλτρα

9.3.1.Δέντρα απόφασης

Τα δεδομένα που έχουμε μαζέψει μπορούμε με μια αναδρομική μέθοδο να τα διαμερίζουμε σε υπό - ομάδες δεδομένων έως ότου να φτάσουμε στο σημείο κάθε υπό - ομάδα να αντιστοιχεί σε μια κλάση. Με αυτόν τον τρόπο θα φτιαχτεί ένα δέντρο αποφάσεων (decision tree) [159]. Η διαμέριση θα πραγματοποιηθεί με κριτήριο την ύπαρξη ενός συγκεκριμένου όρου. Αφού φτιαχτεί το δέντρο απόφασης από τα δεδομένα του προφίλ του χρήστη, μπορούμε για κάθε αντικείμενο να αποφανθούμε κατά πόσο αξίζει ή όχι να συσταθεί.

9.3.2.Μέθοδοι κοντινότερου γείτονα

Με την χρήση του αλγόριθμου κοντινότερου γείτονα (nearest neighbor) [160] συγκρίνουμε κάθε αντικείμενο που έχουμε στο σύστημα μας με τα αντικείμενα που υπάρχουν στο προφίλ του χρήστη. Τα αντικείμενα που είναι στο προφίλ του χρήστη ξέρουμε σε τι βαθμό τον ενδιαφέρουν. Από αυτή την σύγκριση βρίσκουμε τα k πιο όμοια αντικείμενα του συστήματος σε σχέση με αυτά που έχει επιλέξει ο χρήστης. Τέλος αφού γνωρίζουμε το πόσο ενδιαφέρεται ο χρήστης για αυτά τα k αντικείμενα θα έχουμε επίσης μια ένδειξη για το πόσο ενδιαφέρεται για κάθε άλλο αντικείμενο.

9.3.3.Ανατροφοδότηση συνάφειας

Στην μέθοδο της ανατροφοδότησης συνάφειας (relevance feedback) [161] κάθε φορά προσφέρεται στο σύστημα μια λίστα με τα χαρακτηριστικά που ενδιαφέρουν τον χρήστη και δέχεται ως απάντηση τα αντικείμενα που ταιριάζουν πιο καλά σε αυτά τα χαρακτηριστικά. Αυτό που γίνεται στην συνέχεια είναι να μπορεί ο χρήστης να ανατροφοδοτεί το σύστημα με το να δηλώνει κατά πόσο ήταν επιτυχημένη και τον ενδιέφερε κάθε σύσταση που πήρε. Αν τον ενδιέφεραν τα κριτήρια που έγινε η επιλογή ενισχύονται αν όχι τα κριτήρια πρέπει να αλλάξουν.

9.3.4.Άλλες μέθοδοι

Υπάρχουν και άλλες μέθοδοι που ενσωματώνουν αλγόριθμους γραμμικής ταξινόμησης (Linear Classifiers) [162], πιθανολογικές μέθοδοι (Probablistic methods)[24], υβριδικές που συνδυάζονται με άλλες μεθόδους και πολλές άλλες. Το κοινό όλων αυτών είναι ότι προσπαθούν να βρουν από τα χαρακτηριστικά που έχουν τα αντικείμενα, πια είναι παρόμοια. Έπειτα γνωρίζοντας από το προφίλ του χρήστη ποια αντικείμενα του άρεσαν να του συσταθούν παρόμοια αντικείμενα.

9.4.Τεχνικές Ανάκτησης Πληροφορίας σε Συστήματα Συστάσεων

Τεχνικές εξόρυξης πληροφορίας μπορούν να χρησιμοποιηθούν για την δημιουργία συστημάτων συστάσεων όπου, είτε προσπαθούμε να συσταδοποιήσουμε σύνολα αντικειμένων, είτε να βρούμε συχνά επαναλαμβανόμενα μοτίβα συμπεριφοράς που θα μας βοηθήσουν να προβλέψουμε ποιο θα είναι το επόμενο αντικείμενο που θα ενδιαφέρει έναν χρήστη.

Για παράδειγμα, αν ένα μεγάλο σύνολο χρηστών βάζει στο ηλεκτρονικό καλάθι του τα προϊόντα Α, Β, Γ, τότε το πρόγραμμα αν δει ότι ο χρήστης έχει βάλει στο καλάθι του τα Α και Β, αυτομάτως θα του προτείνει και το Γ.

Οι τεχνικές ανάκτησης πληροφορίας μπορούν να χρησιμοποιηθούν και στα φίλτρα που βασίζονται στο περιεχόμενο και στα συνεργατικά φίλτρα. Είναι εργαλεία που μπορούν να εφαρμοστούν σε εντελώς διαφορετικά συστήματα.

Στα συνεργατικά φίλτρα μια τεχνική ανάκτησης πληροφορίας μπορεί να συσταδοποιήσει και να βρει παρόμοιους χρήστες με βάση το σημασιολογικό νόημα των κειμένων που σχετίζονται μαζί του.

Στα φίλτρα που βασίζονται στο περιεχόμενο, αυτό που συσταδοποιείται είναι τα ίδια τα προϊόντα και όχι οι χρήστες. Με δεδομένο ότι έχουμε ένα προϊόν, είμαστε σε θέση να βρούμε ποια είναι τα προϊόντα που έχουν πιο πολλά κοινά χαρακτηριστικά.

Οι περισσότερες τεχνικές είναι σε θέση να επεξεργάζονται απλό κείμενο που μπορεί να σχετίζεται με ένα προϊόν ή χρήστη. Το κείμενο αυτό μετατρέπεται αυτοματοποιημένα σ' ένα πλήθος μεταβλητών και έπειτα εφαρμόζοντας ένα μαθηματικό μοντέλο όπως της συνολοθεωρίας, της γραμμικής άλγεβρας και των πιθανοτήτων, είναι σε θέση να βρει τα πιο όμοια προϊόντα και χρήστες.

9.5.Σύστημα Συστάσεων που κάνει χρήση των Γράφων Ν-γραμμάτων

Μια καινοτόμος μέθοδος ανάκτησης πληροφορίας και κατηγοριοποίησης κειμένων προϊόντων είναι η κατηγοριοποίηση που κάνει χρήση του γράφου Ν-γραμμάτων (N-Grams Graph Classification) [49]. Η κατηγοριοποίηση μέσω του γράφου Ν-γραμμάτων θα αποτελέσει την βάση για να κατασκευάσουμε ένα πρωτότυπο σύστημα συστάσεων.

Στις παρακάτω παραγράφους περιγράφουμε το θεωρητικό μέρος της μεθόδου, τις διάφορες παραλλαγές που μπορούν να γίνουν, αναλύουμε πλήρως την αρχιτεκτονική του συστήματος συστάσεων που υλοποιήσαμε, περιγράφουμε μελλοντικές έρευνες που μπορούν να γίνουν για την βελτίωση της ακρίβειας και της πολυπλοκότητας. Τέλος δοκιμάζουμε πραγματικά δεδομένα εμπορικών προϊόντων και υπολογίζουμε την ακρίβεια των αποτελεσμάτων της.

9.5.1.Γράφος Ν-γραμμάτων

Όπως έχουμε αναφέρει και στην ενότητα 4 κάθε κείμενο μπορεί να αναπαρασταθεί ως ένας γράφος Ν-γραμμάτων. Τα Ν-γράμματα είναι όλοι οι πιθανοί συνδυασμοί Ν συνεχόμενων χαρακτήρων που μπορούν να προκύψουν από ένα κείμενο. Αυτά τα Ν-γράμματα αναπαριστώνται ως κόμβοι στον γράφο. Αν δύο Ν-γράμματα βρίσκονται κοντά στο αρχικό κείμενο τότε θα πρέπει να προστεθεί μια ακμή που να τα ενώνει.

Στην υλοποίηση για το σύστημα συστάσεων επιλέξαμε το Ν ίσο με το τρία. Οπότε έχουμε τριγράμματα και για κάθε τρίγραμμα σχηματίζουμε τρεις ακμές με τα τρία επόμενα τριγράμματα, εκτός βέβαια αν

βρισκόμαστε στο τρίτο από το τέλος, δεύτερο από το τέλος και τελευταίο τρίγραμμα όπου εκεί θα σχηματίζονται δύο, μία και καμία ακμή αντίστοιχα.

Ο σχηματισμός των κόμβων και των ακμών του γράφου καθώς και οι μετρικές ομοιότητας γράφων είναι όπως περιγράφονται στην ενότητα 4.

9.5.2.Κατασκευή του γράφου τριγραμμάτων που αντιστοιχεί σ' ένα θέμα

Κάθε κείμενο μπορεί να αναπαριστάται μέσω του μοντέλου αναπαράστασης του γράφου τριγραμμάτων, ως ένας γράφος. Σε μια επέκταση της διαδικασίας αυτής μπορούμε να αναπαριστούμε ως γράφο τριγραμμάτων όχι μόνο ένα κείμενο αλλά και ένα θέμα από τα θέματα που υπάρχουν μια συλλογή κειμένων.

Μια συλλογή κειμένων περιλαμβάνει κείμενα που το κάθε ένα ιδανικά ανήκει σε μία κατηγορία που αντιστοιχεί σ' ένα θέμα. Αν το κάθε κείμενο ανήκει σε μία κατηγορία τότε έχουμε την περίπτωση του hard classification. Αν κάθε κείμενο ανήκει σε περισσότερες από μία κατηγορίες τότε έχουμε την περίπτωση του soft classification.

Θα ξεκινήσουμε να περιγράψουμε την μέθοδο κατασκευής του γράφου τριγραμμάτων στην περίπτωση που κάθε κείμενο ανήκει σε μια κατηγορία και έπειτα θα την επεκτείνουμε και στην περίπτωση που κάθε κείμενο ανήκει σε πολλές κατηγορίες.

Έχοντας μια συλλογή κειμένων που περιλαμβάνει ένα πλήθος κατηγοριών που στην κάθε κατηγορία περιέχονται ένα πλήθος κειμένων μπορούμε να κατασκευάσουμε έναν γράφο τριγραμμάτων που να αντιπροσωπεύει την κάθε κατηγορία. Αυτό μπορεί να γίνει μέσω των δύο ακόλουθων μεθόδων:

- Ενώνουμε όλα τα κείμενα που ανήκουν στην ίδια κατηγορία σαν να είναι ένα κείμενο και εφαρμόζουμε την μέθοδο υλοποίησης γράφων Ν-γραμμάτων.
- Έχουμε σχηματίσει τους γράφους τριγραμμάτων για κάθε κείμενο και τους ενώνουμε σ' έναν μεγαλύτερο. Αν έχουμε επιλέξει μια υλοποίηση με βάρη στις ακμές, τότε κάθε ακμή του καινούριου γράφου θα έχει βάρος όσο είναι το άθροισμα όλων των βαρών των ακμών.

Στην περίπτωση που έχουμε soft classification εφαρμόζουμε μια από τις προηγούμενες δύο μεθόδους με την διαφορά ότι κάθε κείμενο θα ενσωματωθεί σε όλους τους γράφους θεμάτων που ανήκει.

Η κατασκευή ενός γράφου που αντιστοιχεί σ' ένα θέμα προϋποθέτει ότι έχουμε ένα σύνολο κειμένων εκπαίδευσης. Από την στιγμή που έχουν φτιαχτεί οι γράφοι είμαστε σε θέση να τους εφαρμόσουμε σε τεχνικές κατηγοριοποίησης και ανάκτησης πληροφορίας σύμφωνα με τα πρότυπα που αναφέραμε στην ενότητα 9.4.

9.5.3.Αρχιτεκτονική του συστήματος συστάσεων που χρησιμοποιεί τον γράφο τριγραμμάτων.

Έχουμε επεξηγήσει τις έννοιες του συστήματος συστάσεων, γράφου τριγραμμάτων που αντιπροσωπεύει ένα κείμενο, γράφου τριγραμμάτων που αντιπροσωπεύει ένα θέμα, σύγκρισης γράφων, στις προηγούμενες ενότητες. Τώρα είμαστε σε θέση να παρουσιάσουμε ένα σύστημα συστάσεων που χρησιμοποιεί όλες αυτές τις τεχνικές για να προτείνει εμπορικά προϊόντα σε χρήστες που χρησιμοποιούν ένα εμπορικό πληροφοριακό σύστημα.

Το σύστημα συστάσεων που προτείνουμε ανήκει στην κατηγορία των συστημάτων συστάσεων που χρησιμοποιούν φίλτρα που βασίζονται στο περιεχόμενο και χρησιμοποιεί την καινοτόμο μέθοδο ανάκτησης πληροφορίας που αναπαριστά τα κείμενα ως ΓΝΓ.

Ο χρήστης με το που μπαίνει σε ένα πληροφοριακό σύστημα εμπορικών προϊόντων και επιλέγει να δει ή βάζει στο καλάθι του ένα ή περισσότερα προϊόντα του συστήνονται τα R πιο όμοια προϊόντα σε σχέση με αυτά που επέλεξε σύμφωνα με τον αλγόριθμο συστήματος συστάσεων 9.1

Η μέθοδος αρχικά έχει φτιάξει έναν γράφο τριγραμμάτων για κάθε κατηγορία προϊόντων με βάση τα κείμενα που περιγράφουν τα προϊόντα της κατηγορίας. Έπειτα, φτιάχνει ένα γράφο τριγραμμάτων για το προϊόν που επιλέχθηκε από τον χρήστη. Στην συνέχεια, συγκρίνει κάθε γράφο τριγραμμάτων των κατηγοριών με τον γράφο τριγραμμάτων του προϊόντος. Η σύγκριση που έχει τον μεγαλύτερο συντελεστή ομοιότητας υποδεικνύει σε ποια κατηγορία ανήκει το επιλεγμένο προϊόν. Τέλος, γίνεται μια σύγκριση του γράφου τριγραμμάτων του προϊόντος με όλους τους γράφους τριγραμμάτων προϊόντων που ανήκουν στην κατηγορία που κατηγοριοποιήθηκε το επιλεγμένο προϊόν. Τα R πιο όμοια προϊόντα είναι αυτά που θα συσταθούν στον χρήστη.

Αλγόριθμος 9.1 Σύστασης Προϊόντων	
	Είσοδος: Κείμενο περιγραφής του κάθε προϊόντος, Κ κατηγορίες με ένα σύνολο κειμένων προϊόντων που ανήκει στην κάθε μία. Το τρέχον προϊόν-προϊόντα που επέλεξε ο χρήστης. Έξοδος: Τα R καλύτερα προϊόντα που θα συσταθούν.
1:	Για κάθε κατηγορία φτιάχνουμε τον γράφο 3-γραμμάτων που την αναπαριστά από τα κείμενα προϊόντων που αποτελείται.
2:	Για το τρέχον προϊόν-προϊόντα που επέλεξε ο χρήστης φτιάχνουμε τον γράφο τριγραμμάτων του από το κείμενο που το συνοδεύει.
3:	Συγκρίνουμε τον γράφο που παράχθηκε στο βήμα 2 με όλους του γράφους που παράχθηκαν στο βήμα1.
4	Το προϊόν ανήκει στην κατηγορία με την οποία είχαμε το μεγαλύτερο συντελεστή σύγκρισης γράφων κατηγορίας – προϊόντος.
5	Συγκρίνουμε τον γράφο του προϊόντος που επέλεξε ο χρήστης με όλους τους γράφους των προϊόντων στην κατηγορία που βρίσκεται.
6	Τα R προϊόντα που έχουν τον καλύτερο συντελεστή σύγκρισης του τρέχοντος προϊόντος με τα προϊόντα κατηγορίας, θα είναι αυτά που θα συσταθούν.

Στην ειδική περίπτωση που ο αριθμός R των προϊόντων που καλούμαστε να ανακτήσουμε είναι μεγαλύτερος από το πλήθος των προϊόντων που υπάρχουν στην κατηγορία που ανήκει το προϊόν που επέλεξε ο χρήστης, μπορούμε να ανατρέξουμε στις επόμενες κατηγορίες που έχουν τον μεγαλύτερο συντελεστή ομοιότητας. Το ότι η ομοιότητα ενός προϊόντος με μια κατηγορία είναι ένας βαθμονομημένος αριθμός, μας δίνει την δυνατότητα να ξέρουμε και σε ποιες άλλες κατηγορίες ανήκει το προϊόν σε μια φθίνουσα σειρά.

9.5.4.Επιλογή γράφου τριγραμμάτων και μεθόδου συγκρίσεων γράφων

Στην παρούσα εφαρμογή των ΓΝΓ επιλέξαμε να έχουμε έναν κατευθυνόμενο γράφο για να προσδιορίζει ποιο τρίγραμμα προηγείται και ποιο έπεται. Δεν χρησιμοποιήσαμε βάρη στης ακμές, οπότε αν δύο τριγράμματα συνορεύουν συχνά ή μόνο μια φορά, είναι μια πληροφορία που δεν αποδίδεται από τον γράφο.

Για την σύγκριση της ομοιότητας των γράφων χρησιμοποιήσαμε την μέθοδο ομοιότητας περιεχομένου. Σε κάθε ένα από τα δύο στάδια σύγκρισης γράφων μπορούν να χρησιμοποιηθούν διαφορετικής μορφής γράφοι και μέθοδοι σύγκρισης. Θα μπορούσαμε όταν προσπαθούμε να κατατάξουμε ένα προϊόν σε ποια κατηγορία ανήκει, να χρησιμοποιήσουμε γράφους με κανονικοποιημένα βάρη και κανονικοποιημένη ομοιότητα κατά αξία, ενώ όταν ψάχνουμε το πιο όμοιο προϊόν μέσα στην κατηγορία να χρησιμοποιήσουμε γράφους χωρίς βάρη και ομοιότητα περιεχομένου.

9.5.5.Πλεονεκτήματα του συστήματος συστάσεων που κάνει χρήση του γράφου τριγραμμάτων

Παρακάτω παραθέτουμε τα πλεονεκτήματα του συστήματος συστάσεων που κάνει χρήση του γράφου τριγραμμάτων, έναντι των περισσότερων διαδεδομένων συστημάτων συστάσεων

- Δεν χρειάζεται η αποθήκευση και χρήση προφίλ χρηστών
- Μπορεί να συστήσει καινούρια προϊόντα που μπαίνουν στο σύστημα (cold start problem)
- Η μέθοδος δεν αποδίδει μόνο αν ένα προϊόν προτείνεται για να συσταθεί ή όχι αλλά και σε τι ποσοστό αξίζει να συσταθεί
- Γρηγόρη εύρεση των πιο όμοιων προϊόντων με αυτό που επέλεξε ο χρήστης. Δεν χρειάζεται να διατρέξουμε και να συγκρίνουμε όλα τα προϊόντα, αλλά μόνο το προϊόν με τον γράφο αναπαράστασης κάθε κατηγορίας συν το πόσα προϊόντα έχει η κάθε κατηγορία.
- Ένα προϊόν μπορεί να ανήκει σε πολλές κατηγορίες σε διαφορετικά ποσοστά στην κάθε μια.
- Καταγράφεται εν μέρει η πληροφορία της διάταξης των λέξεων σε ένα κείμενο.

9.5.6.Σύστημα συστάσεων συνεργατικών φίλτρων που κάνει χρήση γράφων τριγραμμάτων

Μια εναλλακτική υλοποίηση του συστήματος συστάσεων που κάνει χρήση του γράφου τριγραμμάτων θα ήταν η κατασκευή του σύμφωνα με τα πρότυπα των συνεργατικών φίλτρων.

Σε αυτή την περίπτωση θα μπορούσαμε να κρατήσουμε ένα προφίλ για κάθε χρήστη που θα αποτελείται από τα προϊόντα που επέλεξε ή αγόρασε συνοδευόμενο με τα αντίστοιχα κείμενα των προϊόντων. Από αυτά τα κείμενα θα φτιάξουμε τον γράφο τριγραμμάτων που αντιστοιχούν σε κάθε χρήστη. Συγκρίνοντας τους γράφους τριγραμμάτων μεταξύ χρηστών, ξέρουμε ποιοι χρήστες έχουν παρόμοιες προτιμήσεις.

Κάθε φορά που ένας χρήστης θα συνδέεται στο πληροφοριακό σύστημα εμπορικών προϊόντων θα αναγνωρίζεται το προφίλ, ποιοι άλλοι χρήστες έχουν παρόμοιο προφίλ και θα του συστήνονται τα προϊόντα που επέλεξαν οι χρήστες που έχουν παρόμοιο προφίλ μαζί του, αλλά δεν επέλεξε αυτός. Η υλοποίηση αυτού του συστήματος συστάσεων διαφοροποιείται σε αρκετά σημεία από το σύστημα συστάσεων που φτιάχτηκε σύμφωνα με τα πρότυπα των φίλτρων που βασίζονται στο περιεχόμενο.

9.5.7.Πειραματικά αποτελέσματα

Το σύστημα συστάσεων που βασίζεται στην κατηγοριοποίηση με γράφους τριγραμμάτων υλοποιήθηκε προγραμματιστικά στην γλώσσα προγραμματισμού java και δοκιμάστηκε σε πραγματικά κείμενα από κριτικές προϊόντων της Amazon. Το σύνολο δεδομένων της Amazon που μας δόθηκε, παρουσιάζεται στην μελέτη που πραγματοποίησε ο dr J. McAuley και είχε τίτλο Hidden Factors and Hidden Topics:Understanding Rating Dimensions with Review Text [163].

Το σύνολο δεδομένων περιείχε ένα σύνολο από κατηγορίες προϊόντων και για κάθε προϊόν χρησιμοποιήσαμε ως περιγραφή του τα reviews που έκαναν οι χρήστες για αυτό. Υπήρχαν προϊόντα που είχαν πολλά reviews και προϊόντα που είχαν λίγα. Το συνολικό πλήθος των reviews ήταν 34.686.770 τα οποία γράφτηκαν από 6.643.669 χρήστες για 2.441.053 προϊόντα.

Χρήστες αρχικά επέλεξαν να δούνε ένα προϊόν από τα αναφερόμενα στο σύνολο δεδομένων της Amazon και έπειτα τους προβλήθηκε μια σύσταση να δουν ένα επόμενο προϊόν. Η αξιολόγηση των συστάσεων πραγματοποιήθηκε μέσω ερωτήσεων των χρηστών αν θεωρούν ενδιαφέρον το προϊόν που τους συστήνεται και το βρίσκουν σχετικό με τις προτιμήσεις τους. Τα πειραματικά αποτελέσματα έδειξαν ένα ποσοστό επιτυχημένων συστάσεων της τάξης του 79% το οποίο είναι ένα ποσοστό που φανερώνει την καταλληλότητα του μοντέλου κατηγοριοποίησης κειμένων που κάνει χρήση των ΓΝΓ για την παραγωγή συστάσεων σε εμπορικές εφαρμογές.

9.6.Συμπεράσματα

Στην ενότητα αυτή είδαμε τις βασικές τεχνικές με τις οποίες κατασκευάζονται τα συστήματα συστάσεων. Αναλύσαμε τα συνεργατικά φίλτρα και τα φίλτρα που βασίζονται στο περιεχόμενο, παρουσιάσαμε πώς οι τεχνικές ανάκτησης πληροφορίας μπορούν να είναι ο πυρήνας ή τα επιμέρους εργαλεία που θα δημιουργήσουν ένα σύστημα συστάσεων.

Δεν υπάρχει σωστή και λάθος μέθοδος παραγωγής συστάσεων. Το ποια μέθοδο συστάσεων θα χρησιμοποιήσουμε, εξαρτάται από τις ανάγκες κάθε φορά που υπάρχουν στο πληροφοριακό σύστημα εμπορικών προϊόντων που θέλουμε να το ενσωματώσουμε. Κριτήρια μας είναι η μορφή των δεδομένων που περιγράφουν τα προϊόντα, καθώς και τους χρήστες, η αποδοτικότητα του συστήματος, ακόμη και νομικά ζητήματα, όπως το κατά πόσο μπορούμε να παρακολουθούμε και να επεξεργαζόμαστε τα προσωπικά δεδομένα των χρηστών που παράγονται όταν χρησιμοποιούν το σύστημα μας.

Τέλος, κατασκευάσαμε την τεχνική σύστασης προϊόντων που κατηγοριοποιεί τα προϊόντα, χρησιμοποιώντας γράφους τριγραμμάτων. Είδαμε ότι αυτή η μέθοδος μπορεί να ικανοποιήσει τις απαιτήσεις ενός συστήματος συστάσεων εμπορικών προϊόντων. Καθώς και το ότι ένα σύνολο κειμένων που αποτελούν κριτική σε εμπορικά προϊόντα μπορεί να είναι μια πηγή πληροφορίας με την οποία εκπαιδεύεται ένα σύστημα συστάσεων.

10. Τεχνικές Βελτιστοποίησης Λειτουργιών Υπολογιστικού Νέφους που Διαχειρίζονται την Υπηρεσία Κατηγοριοποίησης Κειμένων με Γράφους N-γραμμάτων

Στην ενότητα 4 είδαμε τον σχεδιασμό και την υλοποίηση ενός κατανεμημένου μοντέλου κατηγοριοποίησης ροών από κείμενα που βασίζεται στην αναπαράσταση ΓΝΓ. Οι ροές κειμένων μπορούν να έχουν μια κυμαινόμενη συχνότητα ενώ θα πρέπει η απόκριση του συστήματος να είναι σε πραγματικό χρόνο και να επηρεάζεται από το φόρτο εργασίας. Για να το πετύχουμε αυτό χρησιμοποιήσαμε το προγραμματιστικό μοντέλο BEAM και τις υποδομές υπολογιστικού νέφους της Google.

Στην παρούσα ενότητα θα παρουσιάσουμε το μοντέλο πρόβλεψης πόρων εφαρμογής και κινητικότητας χρηστών (Predictor of Application Resources and User Mobility PARUM) που έχει ως σκοπό να συμβάλει στην διαχείριση υπολογιστικών πόρων με βέλτιστο τρόπο. Συγκεκριμένα μια υπηρεσία όπως η κατηγοριοποίηση κειμένων ή οι ειδικές εφαρμογές της: αναγνώριση κοινοτήτων, συναισθηματική ανάλυση, αναγνώριση γεγονότων, σύσταση προϊόντων μπορεί φιλοξενηθεί και να εξυπηρετηθεί από δημόσιες ή ιδιωτικές υποδομές υπολογιστικού νέφους. Καθώς η ροή κειμένων θα αυξάνεται ή θα ελαττώνεται θα πρέπει αντίστοιχα να δεσμεύονται ή να αποδεσμεύονται υπολογιστικοί πόροι ικανοί να επεξεργαστούν και να παράγουν συμπεράσματα από τα κείμενα.

Το PARUM που παρουσιάζουμε αποτελείται από τρία βασικά εργαλεία που πραγματοποιούν τις ακόλουθες τρεις λειτουργίες. Προβλέπει το πλήθος των χρηστών που θα χρησιμοποιήσουν μια εφαρμογή ανά χρονική περίοδο και τοποθεσία. Αναγνωρίζει τις απαιτήσεις σε υπολογιστικούς πόρους που θα έχει κάθε εφαρμογή και αξιολογεί ένα πλάνο ανάθεσης μιας εφαρμογής σε εικονικές μηχανές.

10.1.Βέλτιστη Διαχείριση Υποδομών Υπολογιστικού Νέφους για την Εξυπηρέτηση Εφαρμογών Κατηγοριοποίησης Κειμένων

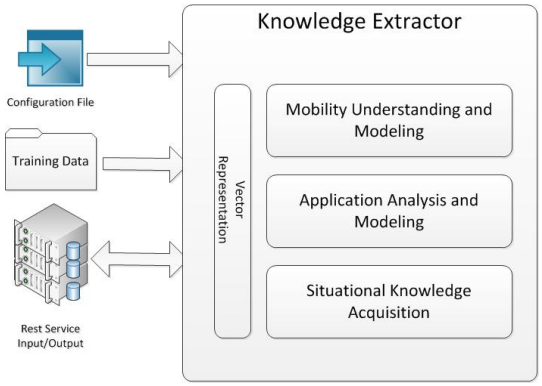
Οι μεγάλες απαιτήσεις στην επεξεργασία δεδομένων έχουν ως αποτέλεσμα τις αντίστοιχες απαιτήσεις σε υπολογιστικούς πόρους οι οποίοι συνήθως προσφέρονται από υποδομές υπολογιστικού νέφους. Οι πόροι συνεπάγονται ένα αντίστοιχο οικονομικό κόστος και απαιτείται ένας σχολαστικός σχεδιασμός για την αποτελεσματική χρήση τους. Έτσι, η επιλογή ενός σχεδίου ανάπτυξης μιας εφαρμογής θα πρέπει να γίνεται με εξισορρόπηση μεταξύ των πόρων που διατίθενται και την διασφάλιση της ποιότητας υπηρεσίας που απολαμβάνει ο χρήστης.

Οι υπηρεσίες πρέπει να εξυπηρετούνται όσο το δυνατόν πιο κοντά στον τόπο όπου προσπελαύνονται και χρησιμοποιούνται. Ταυτόχρονα, η κυμαινόμενη ποσότητα χρηστών που χρησιμοποιούν μια εφαρμογή απαιτεί μια δυναμική δέσμευση πόρων που λαμβάνει υπόψη την ανάλυση επιδόσεων και επιτρέπει μια αποδοτική χρήση των πόρων. Η έγκαιρη πρόβλεψη του όγκου των δεδομένων και της λειτουργίας των εφαρμογών ενισχύουν την ικανότητα μιας εφαρμογής να αναπτυχθεί και να εκτελεσθεί με αποτελεσματικότερο τρόπο. Για να προσφέρουμε αυτές τις εκτιμήσεις σχεδιάσαμε μια ενοποιημένη αναπαράσταση δεδομένων, δύο μεταμοντέλα πρόβλεψης και έναν αξιολογητή του σχεδίου ανάπτυξης.

Η ενοποιημένη αντιπροσώπευση περιλαμβάνει τη ενοποίηση δεδομένων, την τεχνική επιλογής χαρακτηριστικών και μια τεχνική κανονικοποίησης για ετερογενείς τύπους δεδομένων. Προτείνεται ένα μετα-μοντέλο πρόβλεψης κινητικότητας που περιέχει ένα σύνολο μεθόδων και επιλέγεται δυναμικά το πιο ακριβές. Με παρόμοιο τρόπο προτείνεται ένα μετα-μοντέλο πρόγνωσης για την εκτίμηση των απαιτήσεων πόρων της εφαρμογής που πρέπει να αναπτυχθεί. Στο τέλος, μια μέθοδος αξιολογεί τα υποψήφια σχέδια ανάπτυξης με βάση την εξαγόμενη γνώση. Το μοντέλο που προτείνουμε είναι πολύπλευρο για να καλύψει ένα ευρύ φάσμα περιπτώσεων χρήσης.

10.2.Μοντέλο Πρόβλεψης Πόρων Εφαρμογής και Κινητικότητας Χρηστών

Η αρχιτεκτονική του μοντέλου συνιστώσας PARUM παρουσιάζεται στο Σχήμα 10.1 και περιλαμβάνει τρία μοντέλα απόκτησης γνώσης και από ένα βοηθητικό υποσύστημα προεπεξεργασίας δεδομένων. Τα τρία μοντέλα απόκτησης γνώσης είναι η μοντελοποίηση κινητικότητας χρηστών (Mobility Understanding and Modeling), η μοντελοποίηση χρήσης εφαρμογών (Application Analysis and Modeling) και η αξιολόγηση πλάνων ανάθεσης μιας εφαρμογής σε εικονικές μηχανές με βάση την γνώση που έχει αποκτηθεί από προηγούμενες εφαρμογές (Situational Knowledge Acquisition). Κάθε ένα από τα μετά-μοντέλα παραγωγής προβλέψεων βασίζεται σε τεχνικές μηχανικής μάθησης που περιγράφονται στις επόμενες παραγράφους. Η προεπεξεργασία των δεδομένων πραγματοποιείται στο υποσύστημα Vector Representation και χρησιμοποιεί τις τεχνικές ενοποίησης δεδομένων και επιλογής χαρακτηριστικών.



Σχήμα 10.1 μοντέλο πρόβλεψης πόρων εφαρμογής και κινητικότητας χρηστών

Στο σχήμα 10.2 απεικονίζεται το πρόβλημα που επιλύει το PARUM. Αρχικά προσφέρονται πληροφορίες που αφορούν μια εφαρμογή όσων αφορά τους χρήστες που την προσπελαύνουν, την τοποθεσία και το πλήθος, χαρακτηριστικά που αφορούν την φύση της εφαρμογής, χαρακτηριστικά που αφορούν τις υπολογιστικές υποδομές που πρόκειται να την εξυπηρετήσουν, καθώς και διάφορα περιβαλλοντικά χαρακτηριστικά που ενδέχεται να την επηρεάσουν. Ο Vector Representation συλλέγει αυτές τις περιγραφές και τις διοχετεύει με έναν ενοποιημένο και προκαθορισμένο τρόπο στα επόμενα μετα-μοντέλα που τις συνδυάζουν με μια βάση γνώσης και παράγουν προβλέψεις όσον αφορά τους υπολογιστικούς πόρους που θα χρειαστούν και την τοποθεσία - αριθμό χρηστών που χρησιμοποιούν την εφαρμογή.

Αρχείο διαμόρφωσης

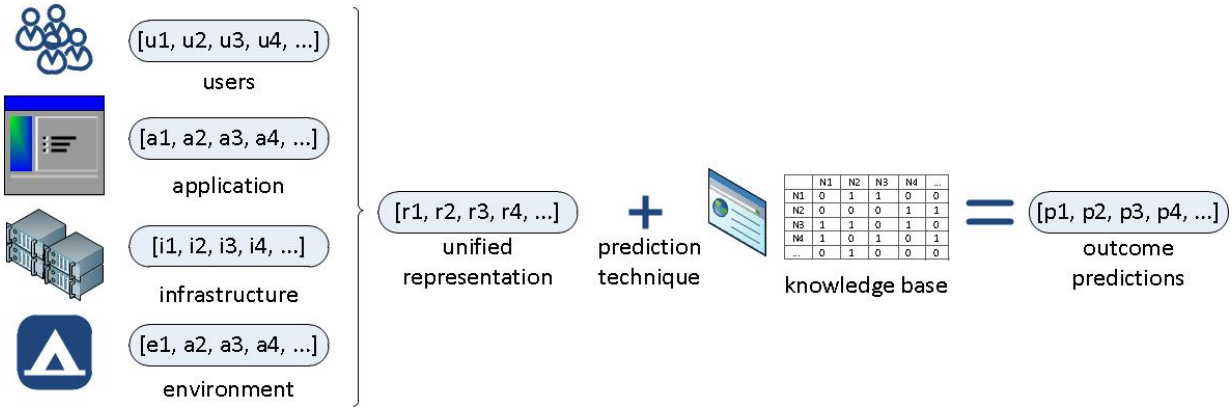
Οι επιμέρους λειτουργίες του PARUM συντονίζονται σύμφωνα με το αρχείο διαμόρφωσης (Configuration file) που καθορίζει ποιες τεχνικές πρόβλεψης και μορφές δεδομένων θα χρησιμοποιηθούν.

Δεδομένα Εκπαίδευσης

Ακολουθώντας το μοντέλο εποπτευόμενης μηχανικής μάθησης οι τρεις βασικές λειτουργίες του PARUM χρησιμοποιούν δεδομένα εκπαίδευσης (Training Data) για να δημιουργήσουν τις εσωτερικές αναπαραστάσεις γνώσης.

Είσοδος / Έξοδος

Η είσοδος και η έξοδος του PARUM πραγματοποιείται μέσω ενός RESTful API και JSON αρχείων.



Σχήμα 10.2 Ροή πληροφορίας και εξαγωγή προβλέψεων εφαρμογών που εξυπηρετούνται από υπηρεσίες νέφους

10.3.Μοντελοποίηση Συμπεριφοράς Κινητικότητας Χρηστών

Η κατανόηση και η μοντελοποίηση της κινητικότητας περιλαμβάνει την ανάλυση και πρόβλεψη της συμπεριφοράς των χρηστών των οποίων τα δεδομένα επεξεργάζονται. Στην δική μας περίπτωση τα κείμενα που γράφουν οι χρήστες μιας εφαρμογής πρέπει να κατηγοριοποιούνται σε θεματικές ενότητες. Είναι σημαντικό οι εφαρμογές να εξυπηρετούνται όσο το δυνατόν πιο κοντά στους χρήστες που τις προσπελαύνουν καθώς επίσης και να έχουμε μια πρόβλεψη του πλήθος των χρηστών για κάθε χρονική περίοδο.

Οι ενέργειες για την μοντελοποίηση της κινητικότητας σε μια νέα συγκεκριμένη εφαρμογή είναι οι εξής δύο: πρώτον, θα πρέπει να διαβιβαστεί στο PARUM ένα συμβατό σύνολο δεδομένων βασισμένο σε προηγούμενες παρατηρήσεις, προκειμένου να δημιουργηθεί η βάση γνώσεων σύμφωνα με το μοντέλο της εποπτευόμενης μηχανικής μάθησης. Στη συνέχεια, στο αρχείο διαμόρφωσης (configuration file) θα πρέπει να προσδιοριστεί η τεχνική πρόβλεψης που θα χρησιμοποιηθεί και ο τύπος των δεδομένων εισόδου.

Οι προβλέψεις του PARUM μπορούν να πραγματοποιηθούν με μεθόδους ταξινόμησης, ομαδοποίησης ή παλινδρόμησης. Η απόφαση σχετικά με την προσέγγιση πρόβλεψης που χρησιμοποιείται εξαρτάται από τον τύπο και την ποσότητα των παρεχόμενων δεδομένων και τις παραμέτρους που πρέπει να προβλεφθούν και ποικίλλουν σε κάθε περίπτωση χρήσης. Η ακόλουθη προτεινόμενη μέθοδος μπορεί επίσης να λάβει υπόψη της την χρονική εξέλιξη των παρατηρήσεων και των προβλεπόμενων παραμέτρων.

10.3.1.Επεξεργασία διανυσμάτων

Η ταξινόμηση, η ομαδοποίηση και η παλινδρόμηση των διανυσμάτων έχουν διερευνηθεί εκτενώς και εφαρμόζονται σε πολλούς τομείς. Χρησιμοποιούμε δύο από τις βασικές μεθόδους πρόβλεψης διανυσμάτων ως βασικά μοντέλα του PARUM: το μοντέλο Μηχανές Διανυσμάτων Υποστήριξης (SVM) και την Ταξινόμηση Bayes. Το μοντέλο SVM [9] αντιπροσωπεύει τις παρατηρήσεις ως σημεία σε έναν χώρο N-διαστάσεων. Στη συνέχεια, υπολογίζεται ένα υπερ-επίπεδο για να διαιρέσει τις παρατηρήσεις που ανήκουν σε διαφορετικές κατηγορίες. Το χάσμα μεταξύ παρατηρήσεων που ανήκουν σε διαφορετικές κατηγορίες πρέπει να είναι όσο το δυνατόν ευρύτερο. Ο ταξινομητής Gaussian Bayes [138] χρησιμοποιεί ένα μοντέλο με πιθανότητες υπό συνθήκη, στο οποίο οι τιμές των διανυσμάτων είναι οι ανεξάρτητες μεταβλητές και η κατηγορία, είναι η εξαρτημένη μεταβλητή y.

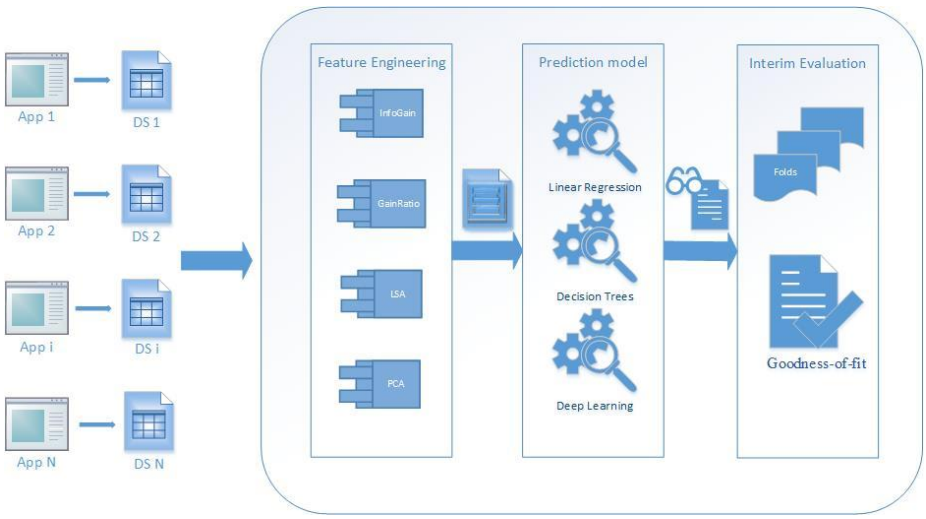
10.3.2.Διαμερισμός γράφων

Για να προσδιοριστούν οι κατηγορίες χρηστών ή η σχέση μεταξύ χρηστών και εφαρμογών, μπορεί να εφαρμοστεί μια μέθοδος διαμερισμού γράφων. Συνήθως τα προβλήματα διαμερισμού γράφων είναι NP-δύσκολα. Έχουν προταθεί ευριστικοί και προσεγγιστικοί αλγόριθμοι που παράγουν επαρκή αποτελέσματα. Ο αλγόριθμος Kernighan Lin [56] είναι ένας αλγόριθμος διαμερισμού γράφων ο οποίος εφαρμόζεται αποτελεσματικά σε έναν πυκνό γράφο με λιγότερους από 10000 κόμβους. Χρησιμοποιεί μια τεχνική που ανταλλάσσει κόμβους μεταξύ των διαμερισμάτων χρησιμοποιώντας μια μετρική μεταξύ εσωτερικού και εξωτερικού κόστους ακμών. Οι Girvan και Newman [102] πρότειναν τον αλγόριθμο κατανομής K-Means. Ο K-Means παράγει καλά αποτελέσματα με τον περιορισμό ότι οι παρατηρήσεις θα πρέπει να είναι γραμμικές. Σε αυτή τη μέθοδο, κάθε παρατήρηση αντιστοιχεί στη συστάδα με την πλησιέστερη μέση τιμή.

10.4.Μοντελοποίηση Χρήσης Εφαρμογών

Μια υπηρεσία όπως αυτή της κατηγοριοποίησης κειμένων με ΓΝΓ που εξυπηρετείται από υποδομές υπολογιστικού νέφους έχει κυμαινόμενες ανάγκες σε υπολογιστικούς πόρους οι οποίες έχουν ως αποτέλεσμα να χρειάζονται προβλέψεις του φόρτου εργασίας για μια έγκαιρη δυναμική δέσμευση και αποδέσμευση εικονικών μηχανών. Οι προβλέψεις σε συνδυασμό με την ποιότητα υπηρεσίας θα πρέπει να εξετάζονται για την αποφυγή των σημείων συμφόρησης και την ομαλή λειτουργία της εφαρμογής. Οι παράμετροι που αναλύονται και προβλέπονται από το μετά-μοντέλο Application Analysis and Modeling είναι η επεξεργαστική ισχύ, η μνήμη, το εύρος ζώνης που χρησιμοποιείται, το μέγεθος των αρχείων που ανταλλάσσεται, ο χρόνος απόκρισης και το χρονικό διάστημα μεταξύ δύο αιτήσεων.

Η πρόβλεψη απαιτήσεων πόρων αποτελεί ένα μετα-μοντέλο που περιλαμβάνει ένα σύνολο τεχνικών παλινδρόμησης για την εκτίμηση των αναγκών πόρων μιας εφαρμογής. Πριν από την έναρξη λειτουργίας του PARUM, θα πρέπει να δημιουργηθεί ένα σύνολο δεδομένων εκπαίδευσης που περιλαμβάνουν τις παρατηρήσεις της χρήσης της CPU, της κατανομής μνήμης, της απόδοσης και του χρόνου απόκρισης σε συνδυασμό με διαφορετικές κλίμακες φόρτου εργασίας και παραμέτρους περιβάλλοντος, όπως η χρονική περίοδος ή η περιοχή στην οποία εκτελείται η εφαρμογή.



Σχήμα 10.3 Μετα-μοντέλο πρόβλεψης υπολογιστικών πόρων

Η εφαρμογή που πρόκειται να αναπτυχθεί συνοδεύεται από ένα σύνολο δεδομένων που παρέχεται ως είσοδος στο PARUM. Το ενιαίο επίπεδο αναπαράστασης φέρνει τα δεδομένα σε μια προκαθορισμένη μορφή. Στη συνέχεια, το σύνολο δεδομένων χωρίζεται και επεξεργάζεται χρησιμοποιώντας την μέθοδο bootstrap [164] με παρατηρήσεις που χωρίζονται σε 65% για εκπαίδευση και 35% για αξιολόγηση.

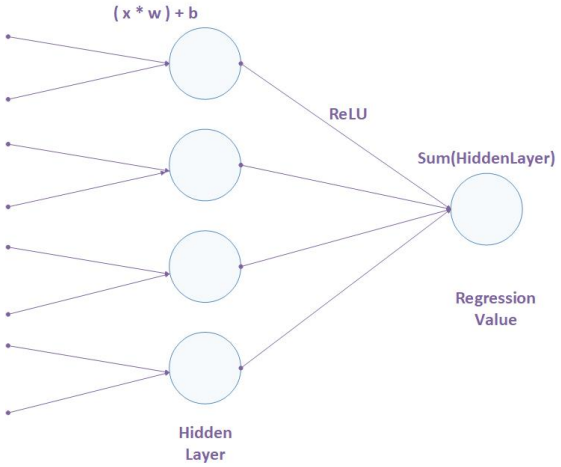
Ανάλογα με τον τύπο δεδομένων που περιλαμβάνονται στις ενοποιημένες δομές, τα αντίστοιχα πρότυπα πρόγνωσης ενεργοποιούνται και εκπαιδεύονται. Στη συνέχεια, το τμήμα δεδομένων αξιολόγησης παρέχεται στα εκπαιδευμένα προγνωστικά μοντέλα, προκειμένου να αξιολογηθούν και να επιλεγεί αυτό που παρουσιάζει την καλύτερη ακρίβεια. Χρησιμοποιούνται τέσσερις ξεχωριστές τεχνικές παλινδρόμησης για την πρόβλεψη των απαιτήσεων πόρων.

10.4.1.Πολλαπλή γραμμική παλινδρόμηση

Η πολλαπλή γραμμική παλινδρόμηση [165] χρησιμοποιείται για να εκφράσει τη σχέση μεταξύ μίας συνεχούς εξαρτώμενης μεταβλητής και ενός συνόλου ανεξάρτητων μεταβλητών. Οι ανεξάρτητες μεταβλητές μπορούν να είναι συνεχείς ή κατηγορηματικές. Υπολογίζεται η γραμμική σχέση μεταξύ της εξαρτημένης μεταβλητής και των ανεξάρτητων μεταβλητών. Η πολλαπλή γραμμική παλινδρόμηση εφαρμόστηκε με το κριτήριο Akaike για επιλογή μοντέλου.

10.4.2. Βαθιά εκμάθηση

Η βαθιά εκμάθηση (Deep Learning DL) χρησιμοποιείται συνήθως για σκοπούς ταξινόμησης και ομαδοποίησης, αλλά έχει επίσης εφαρμοστεί για προβλέψεις σε συνεχείς τιμές σύμφωνα με την προσέγγιση της παλινδρόμησης [166]. Μια πρόβλεψη παλινδρόμησης με DL μπορεί να εκτελεστεί χρησιμοποιώντας ένα πολλαπλών επιπέδων νευρωνικό δίκτυο και μια λειτουργία ενεργοποίησης γραμμικής μονάδας (ReLU) για κάθε κρυφό κόμβο. Οι τιμές ενεργοποίησης αθροίζονται και πολλαπλασιάζονται με τις τιμές βάρους για τους κόμβους των κρυφών επιπέδων. Στο επίπεδο εξόδου υπάρχει ένα απλό άθροισμα που συγκεντρώνει τις τιμές από το τελευταίο στρώμα που έχει αποκαλυφθεί όπως απεικονίζεται στο σχήμα 10.4.



Σχήμα 10.4 Παλινδρόμηση με Βαθιά εκμάθηση

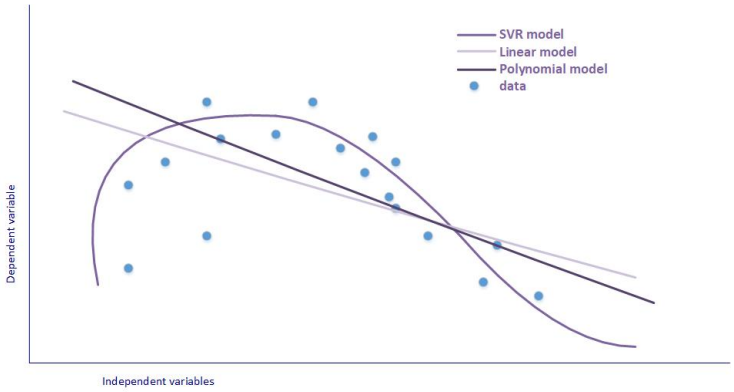
10.4.3.Υποστήριξη παλινδρόμησης διανύσματος

Η παλινδρόμηση διανυσμάτων υποστήριξης (SVR) [167] είναι μια μη παραμετρική τεχνική που ακολουθεί την ίδια βασική ιδέα του αλγόριθμου ταξινόμησης Μηχανές Διανυσμάτων Υποστήριξης προσαρμοσμένου να προβλέψει συνεχείς τιμές χρησιμοποιώντας την αρχή του μέγιστου περιθωρίου, το οποίο διαχειρίζεται η

SVR ως ένα κυρτό πρόβλημα βελτιστοποίησης. Η SVR μπορεί να υποστηρίξει γραμμικούς, Γκαουσιανούς και πολυωνυμικούς πυρήνες.

10.4.4.Bayesian regression

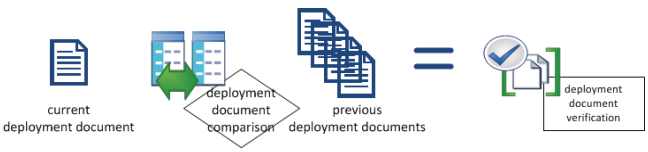
Η Bayesian Regression είναι μια πιθανοτική προσέγγιση για γραμμικές και πολυπαραγοντικές τεχνικές γραμμικής παλινδρόμησης που χρησιμοποιούν τα Bayesian συμπεράσματα. Η μέθοδος αυτή επιτρέπει την τοποθέτηση προκαταρκτικών τιμών στους συντελεστές και στον θόρυβο σε περίπτωση αραιού συνόλου δεδομένων.



Σχήμα 10.5 Τεχνικές παλινδρόμησης

10.5.Αξιολόγηση Σχέδιων Ανάθεσης

Η αξιολόγηση πλάνων ανάθεσης μιας εφαρμογής με βάση την γνώση που έχουμε αποκτήσει από προηγούμενες εφαρμογές της (Knowledge Acquisition) είναι η τρίτη κύρια λειτουργία του PARUM. Η αξιολόγηση του πλάνου ανάθεσης πραγματοποιείται συγκρίνοντας τις καταγεγραμμένες παρατηρήσεις εφαρμογών, το φόρτο εργασίας και τις συσκευές ακροδεκτών με τα προηγούμενα πλάνα ανάθεσης εφαρμογών. Μια υλοποίηση του k-nn αλγορίθμου υπολογίζει τα πιο όμοια πλάνα ανάθεσης. Ο βαθμός που οι προηγούμενες k-πιο όμοιες αναθέσεις παραβίασαν την συμφωνία σε επίπεδο υπηρεσιών (SLA) αποτελεί πρόβλεψη για την καταλληλότητα της τρέχουσας επιλογής.



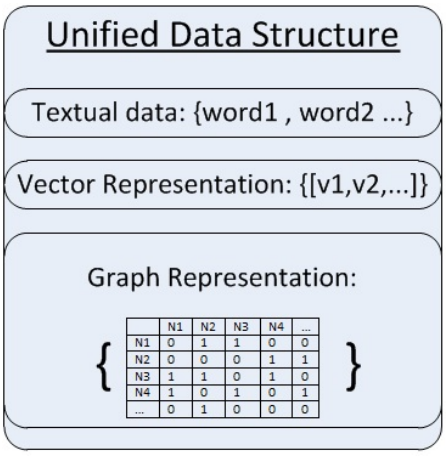
Σχήμα 10.6 Αξιολόγηση σχέδιων ανάθεσης μέσω αποκτηθείσας γνώσης

Τα σχέδια ανάπτυξης ακολουθούν μια διανυσματική αναπαράσταση με συνιστώσες χαρακτηριστικά όπως οι ανάγκες των εφαρμογών σε υπολογιστικούς πόρους: CPU, μνήμη, διακίνηση δεδομένων, απόκριση χρόνου και η προηγούμενη παραβίαση των SLA.

Χρησιμοποιώντας μια μέτρηση όπως η απόσταση Euclidean, Manhatan ή Minkowski, μετράμε τα πιο κοντινά σχέδια ανάπτυξης. Η αξιολόγηση ενός υποψήφιου σχεδίου ανάπτυξης είναι η αρμονική μέση τιμή των παραβιάσεων SLA.

10.6.Ενοποιημένη Αναπαράσταση

Η ενοποιημένη αναπαράσταση εισάγει ένα ενδιάμεσο στρώμα μεταξύ των πηγών εισερχομένων δεδομένων και των των μετα-μοντέλων πρόβλεψης PARUM. Ο σκοπός αυτού του στρώματος είναι να ενοποιήσει και να μετασχηματίσει τα δεδομένα εισόδου σε μορφή συμβατή και αναγνώσιμη από τους αλγόριθμους πρόβλεψης.

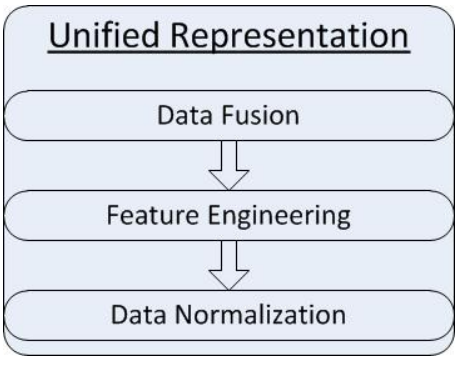


	N1	N2	N3	N4	...
N1	0	1	1	0	0
N2	0	0	0	1	1
N3	1	1	0	1	0
N4	1	0	1	0	1
...	0	1	0	0	0

Σχήμα 10.7 Ενοποιημένη αναπαράσταση δεδομένων

Οι ετερογενείς πηγές δεδομένων τροφοδοτούν συνεχώς το PARUM με παρατηρήσεις. Αυτές μπορεί να περιλαμβάνουν ανθρώπινες τροχιές, μετα-δεδομένα υπηρεσιών, δεδομένα σχετικά με τα συμφραζόμενα των χρηστών, χρήση πόρων ανά εφαρμογή, κλπ. Ακόμη και αν τα δεδομένα προέρχονται από διαφορετικές πηγές, ενδέχεται να αφορούν τις ίδιες οντότητες και μπορούν να αναλυθούν με ενοποιημένο τρόπο για να παρέχουν εκτιμήσεις.

Μια ενοποιημένη δομή αναπαράστασης έχει σχεδιαστεί για τη διατήρηση διανυσμάτων, γράφων και κειμένων μετά-δεδομένων που αφορούν μια εφαρμογή, έναν χρήστη εφαρμογής και την περιοχή ενδιαφέροντος. Η ενοποιημένη δομή αναπαράστασης απεικονίζεται στο σχήμα 10.7.



Σχήμα 10.8 Στάδια ενοποίησης δεδομένων

Από τεχνική άποψη, κλήσεις Rest μπορούν να παρέχουν αυτά τα δεδομένα στην ενιαία αναπαράσταση του PARUM συνοδευόμενες από ορισμένα προκαθορισμένα πεδία που καθορίζουν τις οντότητες με τις οποίες σχετίζονται.

Ακολουθείται το μοντέλο σύντηξης δεδομένων βασισμένο στη γνώση της συμπεριφοράς [168], στο οποίο η διαδικασία ενοποιημένης αντιπροσώπευσης δεδομένων βελτιώνει τα παρεχόμενα δεδομένα χρησιμοποιώντας τεχνικές μηχανικής χαρακτηριστικών [169], τεχνικές μείωσης διαστάσεων, τεχνικές ομαλοποίησης. Η επεξεργασία ροής δεδομένων απεικονίζεται στο Σχήμα 10.8.

10.6.1.Προεπεξεργασία δεδομένων

Η προεπεξεργασία δεδομένων πραγματοποιείται για να ταιριάζει τα δεδομένα στο σχήμα ενοποιημένης δομής αναπαράστασης και περιλαμβάνει τρία επακόλουθα βήματα. Την σύντηξη δεδομένων (Data Fusion) που συγκεντρώνει τα δεδομένα από τις πηγές δεδομένων, τη μηχανική των χαρακτηριστικών (Feature Engineering) που παράγουν τα πιο αντιπροσωπευτικά χαρακτηριστικά και την κανονικοποίηση των τιμών. Στόχος του σταδίου προεπεξεργασίας των δεδομένων είναι να συνδυάσουν τις σχετικές πληροφορίες από διάφορες πηγές σε μια ενιαία δομή που παρέχει μια περιγραφή σύμφωνη με προκαθορισμένα πρότυπα σε αντίθεση με τις μεμονωμένες πηγές δεδομένων.

Οι διαφορετικές πηγές εισαγωγής δεδομένων εικονίζονται στις ενοποιημένες δομές αναπαράστασης. Τα αρχεία βάσης γνώσεων χρησιμοποιούνται από τις τεχνικές πρόβλεψης σε συνδυασμό με την αναπαράσταση των στιγμιότυπων εισόδου έτσι ώστε να παράγουν τις ζητούμενες προβλέψεις. Δεν υπάρχει ανάγκη αποθήκευσης και ανάκτησης των δεδομένων εισόδου σε οποιοδήποτε είδος βάσης δεδομένων. Το υποσύνολο ενοποιημένης αναπαράστασης είναι υπεύθυνο για την αντιμετώπιση των διαφορετικών δεδομένων εισόδου και η πρόσθετη προσπάθεια μιας βάσης δεδομένων πρέπει να αποφεύγεται. Επιπλέον, τα αρχεία βάσης γνώσεων αποθηκεύονται σε αρχεία όπως JSON και arff. Οι τεχνικές πρόβλεψης δεν χρειάζονται συγκεκριμένα τμήματα των αποθηκευμένων δεδομένων, αλλά χρειάζονται ολόκληρη τη βάση γνώσεων για τη διεξαγωγή των διαδικασιών τους.

10.6.2.Σύντηξη δεδομένων

Η σύντηξη δεδομένων είναι ένα πολύ ισχυρό εργαλείο και βοηθάει τις μεθόδους πρόβλεψης. Για τις ανάγκες επεξεργασίας του PARUM τα δεδομένα αποτελούνται από δεδομένα προφίλ χρηστών και δεδομένα χρήσης εφαρμογών. Αυτά τα δεδομένα ενσωματώνονται στην Ενοποιημένη αναπαράσταση δεδομένων, χρησιμοποιώντας το μοντέλο σύντηξης δεδομένων βασισμένο στη γνώση της συμπεριφοράς.

Η σύντηξη δεδομένων από πολλούς αισθητήρες παίρνει υπόψη της τη συχνότητα και το μέγεθος του φορτίου δεδομένων από τις πηγές δεδομένων. Το μοντέλο που βασίζεται στη γνώση της συμπεριφοράς [168] αποτελείται από μια σειρά σταδίων. Το πρώτο στάδιο ανακτά τα δεδομένα από όλους τους αισθητήρες δεδομένων. Στο επόμενο στάδιο εξάγεται ένα διάνυσμα χαρακτηριστικών από τα ανακτημένα δεδομένα. Το τρίτο στάδιο συσχετίζει μια δομή δεδομένων με τις προκαθορισμένες ανάγκες. Στο τελευταίο στάδιο, εφαρμόζεται ένα σύνολο κανόνων σύμφωνα με τον φορμαλισμό της εκπροσώπησης. Αυτή η μέθοδος χρησιμοποιείται σε περίπτωση συχνής ενημέρωσης δεδομένων.

10.6.3.Μηχανική χαρακτηριστικών

Η εγγενής δομή των διαθέσιμων δεδομένων και οι ανάγκες των τελικών χρηστών μπορεί να συνιστούν πρόβλημα εποπτευόμενης μάθησης που περιλαμβάνει χαρακτηριστικά που δεν έχουν θετική συμβολή στην ακρίβεια μιας μεθόδου πρόβλεψης. Η ιδέα, ότι τα περισσότερα δεδομένα οδηγούν σε καλύτερα αποτελέσματα δεν εφαρμόζονται σε όλες τις περιπτώσεις. Για αυτό στις μεθόδους πρόβλεψης τα μεγάλα ποσά δεδομένων μπορεί να οδηγούν σε χαμηλή ακρίβεια αν δεν έχουν πρώτα προεπεξεργαστεί κατάλληλα.

Δύο διαφορετικά σύνολα τεχνικών χρησιμοποιούνται για να μετριάσουν το ζήτημα της επιλογής και βελτίωσης χαρακτηριστικών των συνόλων δεδομένων: η εξαγωγή χαρακτηριστικών και η επιλογή

χαρακτηριστικών. Και οι δύο μειώνουν την αναπαράσταση δεδομένων χρησιμοποιώντας λιγότερα χαρακτηριστικά. Χαρακτηριστικά, μεταβλητές, όροι, διαστάσεις είναι εναλλάξιμες έννοιες για τις ανάγκες του PARUM. Οι μέθοδοι εξαγωγής χαρακτηριστικών αναπαριστούν χαρακτηριστικά σε ένα νέο χώρο διαστάσεων που δημιουργεί συγχώνευση ή μετασχηματισμό των διαστάσεων του προβλήματος. Αντίθετα, οι μέθοδοι επιλογής χαρακτηριστικών δεν μετασχηματίζουν τις διαστάσεις. Επιλέγουν τις διαστάσεις που περιέχουν περισσότερης σημασίας πληροφορία βάσει μιας συγκεκριμένης αντικειμενικής λειτουργίας.

10.6.4.Εξαγωγή χαρακτηριστικών

Οι μέθοδοι εξαγωγής χαρακτηριστικών εισάγουν έναν νέο λιγότερων διαστάσεων χώρο χαρακτηριστικών που συνδυάζει τις αρχικές δυνατότητες δεδομένων. Τα νέα χαρακτηριστικά που σχηματίζονται πρέπει να ικανοποιούν τις ακόλουθες τρεις ιδιότητες. Θα πρέπει να προσφέρουν χρήσιμη πληροφορία χωρίς πλεονασμό, να συμβάλουν στις μεθόδους μηχανικής εκμάθησης και σε ορισμένες περιπτώσεις να παρέχουν καλύτερη παρουσίαση του προβλήματος. Δύο τεχνικές εξαγωγής χαρακτηριστικών που ικανοποιούν τα προαναφερθέντα κριτήρια και εφαρμόζονται στην ενοποιημένη αντιπροσώπευση είναι οι ακόλουθες.

Η Ανάλυση Βασικών Στοιχείων (Principal Component Analysis PCA) [170] είναι μια στατιστική μέθοδος που απεικονίζει τις ορθογώνιες διαστάσεις σε ένα νέο σύνολο συνιστωσών διαστάσεων. Το νέο σύνολο διαστάσεων καλείται κύρια συστατικά (principal component), είναι μικρότερο και διατηρεί την αρχική πληροφορία. Η βασική ιδέα είναι να μετατρέψουμε τα χαρακτηριστικά που σχετίζονται σε ένα νέο σύνολο χαρακτηριστικών που είναι γραμμικά μη συσχετισμένα. Η PCA είναι μια επαναληπτική διαδικασία. Το πρώτο κύριο στοιχείο θα πρέπει να έχει τη μεγαλύτερη διακύμανση και τα ακόλουθα στοιχεία θα πρέπει να έχουν μια φθίνουσα διακύμανση υπό τον περιορισμό να είναι ορθογώνια με τα προηγούμενα στοιχεία.

Η Ανάλυση Ανεξάρτητων Συστατικών (Independent Component Analysis ICA) [171] είναι μια μέθοδος εξαγωγής στατιστικών και υπολογιστικών χαρακτηριστικών που ανιχνεύει τα λανθάνοντα χαρακτηριστικά στα σύνολα τυχαίων παρατηρήσεων. Η ICA βασίζεται σε ένα γενετικό μοντέλο για πολυπαραγοντικά δεδομένα. Οι περιπτώσεις μπορεί να είναι γραμμικοί ή μη γραμμικοί συνδυασμοί των άγνωστων λανθάνοντων μεταβλητών, ενώ ο τρόπος με τον οποίο αναμειγνύονται είναι άγνωστος. Οι λανθάνουσες μεταβλητές που θα υπολογιστούν από την ICA είναι μη-γκαουσιανές, γραμμικές και αμοιβαία ανεξάρτητες και ονομάζονται ανεξάρτητα συστατικά των παρατηρούμενων δεδομένων. Η διαδικασία εξαγωγής χαρακτηριστικών δεν είναι αναστρέψιμη, επειδή κάποια στοιχεία χάνονται στη διαδικασία μετασχηματισμού.

10.6.5.Επιλογή χαρακτηριστικών

Η επιλογή χαρακτηριστικών είναι η διαδικασία επιλογής ενός υποσυνόλου χαρακτηριστικών από ένα σύνολο υποψήφιων χαρακτηριστικών βάσει μιας στατιστικής βαθμολογίας, όπως η μεταβολή της μεταβλητής συσχέτισης. Τα χαρακτηριστικά που θα επιλεγούν είναι τα πιο σημαντικά και αντιπροσωπευτικά των διαθέσιμων λειτουργιών. Αυτές οι μέθοδοι απαιτούν καλή κατανόηση του προβλήματος πρόβλεψης.

Οι τρεις βασικές προσεγγίσεις επιλογής χαρακτηριστικών είναι η μέθοδος περιτυλίγματος (wrapper method), η μέθοδος φίλτρου (filter method) και η ενσωματωμένη μέθοδος (embedded method). Η μέθοδος περιτυλίγματος και οι ενσωματωμένες τεχνικές επιλογής χαρακτηριστικών δεν μπορούν να εφαρμοστούν για τις ανάγκες του PARUM λόγω των απαιτήσεων υπολογισμού και της μη ικανότητας τους να διακρίνουν τη φάση επιλογής χαρακτηριστικών με το στάδιο πρόβλεψης. Από την άλλη πλευρά, οι μέθοδοι φίλτρων παράγουν καλά αποτελέσματα με χαμηλές απαιτήσεις υπολογισμού. Γνωστές τεχνικές φίλτρων είναι το κέρδος πληροφοριών (Information Gain), Chi-square, Αμοιβαία πληροφορία (Mutual Information), βαθμολογία Fisher και το κριτήριο χαμηλής διακύμανσης.

10.6.6.Κανονικοποίηση δεδομένων

Οι πηγές δεδομένων μπορούν να παρέχουν τιμές σε διαφορετική κλίμακα από αυτές της εσωτερικής αναπαράστασης γνώσης των μοντέλων. Μια διαδικασία κανονικοποίησης γεφυρώνει αυτές τις διαφορές των

χαρακτηριστικών ώστε να βρίσκονται οι παρατηρήσεις μεταξύ μιας καθορισμένης ελάχιστης και μέγιστης τιμής.

10.7.Εφαρμογή του PARUM

Η υλοποίηση του PARUM πραγματοποιήθηκε με χρήση της γλώσσας προγραμματισμού Java και ακολουθεί τις προδιαγραφές του Open Cloud Computing Interface (OCCI) [172]. Το PARUM χρησιμοποιεί μια ποικιλία τεχνικών πρόβλεψης, μερικές από αυτές υλοποιούνται με βάση τις διαθέσιμες βιβλιοθήκες Weka [173] και deeplearning4java [174]. Ενώ κάποιες άλλες μέθοδοι όπως αυτές που κάνουν χρήση αλυσίδων Markov υλοποιήθηκαν από την αρχή.

10.8.Συμπεράσματα

Η πρόβλεψη των απαιτήσεων σε υπολογιστικούς πόρους μια εφαρμογής που εξυπηρετείται από υποδομές υπολογιστικού νέφους συμβάλει στην βέλτιστη διαχείριση των εικονικών μηχανών όπου εξυπηρετούν την κατηγοριοποίηση ροών κειμένων με ΓΝΓ. Το PARUM παρέχει την εκπαίδευση διάφορων μοντέλων πρόβλεψης και την δυναμική επιλογή με τον πιο ακριβή τρόπο βάσει μιας μέτρησης αξιολόγησης.

Το PARUM είναι ένα μοντέλο που υποστηρίζει τις υποδομές υπολογιστικού νέφους για την εξυπηρέτηση μεγάλων δεδομένων κλιμακούμενης συχνότητας, χρησιμοποιώντας τεχνικές που παρέχουν προβλέψεις για την κινητικότητα των χρηστών σε μια περιοχή ενδιαφέροντος και τις ανάγκες πόρων των εφαρμογών.

Προτείνεται ένα ενοποιημένο επίπεδο αναπαράστασης για τη λήψη ετερογενών τύπων δεδομένων, την επιλογή χαρακτηριστικών και την ομαλοποίηση τους. Παρουσιάσαμε δύο μετα-μοντέλα πρόγνωσης που περιλαμβάνουν ένα σύνολο τεχνικών παλινδρόμησης. Το πρώτο την κατανομή των χρηστών γύρω από σημεία ενδιαφέροντος. Το δεύτερο αφορά την πρόβλεψη των υπολογιστικών πόρων που χρειάζεται μια εφαρμογή για να εξασφάλιση την ποιότητα υπηρεσίας της.

Οι περισσότερες λειτουργίες του PARUM [175] έχουν δοκιμαστεί και αξιολογηθεί κατά τη διάρκεια του ερευνητικού έργου Basmati [176] [177] και μας ενθαρρύνουν να το θεωρήσουμε ως ένα σημαντικό εργαλείο για τις υποδομές του cloud.

Συντομογραφίες

API	Application Programming Interface	Διεπαφή Προγραμματισμού Εφαρμογών
BoW	Bag of Words	Σάκος Λέξεων
CS	Containment Similarity	Ομοιότητα Περιεχομένου
DL	Deep Learning	Βαθιά Εκμάθηση
DV	Discrimination Value	Βαθμός Διακριτικότητας
FSD	First Story Detection	Πρώτη ανίχνευση Ιστορίας
GVSM	Generalized Vector Space Model	Γενικευμένο Μοντέλο Διανυσματικού Χώρου
ICA	Independent Component Analysis	Ανάλυση Ανεξάρτητων Συστατικών
IDF	Inverse Document Frequency	Αντίστροφη Συχνότητα Εγγράφων
LDA	Latent Dirichlet Allocation	Λανθάνουσα Κατανομή του Dirichlet
LR	Logistic Regression	Γραμμική Παλινδρόμηση
LSA	Latent Semantic Analysis	Λανθάνουσα Σημασιολογική Ανάλυση
MCS	Maximum Common Subgraph	Μέγιστος Κοινός Υπογράφος
NLP	Natural Language Processing	Φυσική Επεξεργασία Γλώσσας
NVS	Normalized Value Similarity	Κανονικοποιημένης Ομοιότητας Αξίας
OCCI	Open Cloud Computing Interface	Ανοιχτή Διεπαφή Υπολογιστικού Νέφους
PARUM	Predictor of Application Resources and User Mobility	Πρόβλεψη Πόρων Εφαρμογής & Κινητικότητας Χρηστών
PCA	Principal Component Analysis	Ανάλυση Βασικών Στοιχείων
PLSA	Probabilistic latent sematic analysis	Πιθανολογική λανθάνουσα σημασιολογική ανάλυση
SA	Sentiment Analysis	Συναισθηματική ανάλυση
SLA	Service Level Agreement	Συμφωνία σε Επίπεδο Υπηρεσιών
SN	Social Network	Κοινωνικό Δίκτυο
SVD	Singular Value Decomposition	Ανάλυση Πίνακα σε Ιδιάζουσες Τιμές
SVM	Support Vector Machines	Μηχανές Διανυσμάτων Υποστήριξης
SWR	Stop Words Removal	Αφαίρεση Κοινών Λέξεων
TCD	Topic communities Detection	Ανίχνευση Θεματικών Κοινοτήτων
TF	Term Frequency	Συχνότητα Όρων
VS	Value Similarity	Μέτρο Ομοιότητας Αξίας
VSM	Vector Space Model	Μοντέλο Διανυσματικού Χώρου
TEM	Tempered Expectation Maximization	Μετριασμένης Μεγιστοποίησης Αναμενόμενης Τιμής
ΓΝΓ	N-Gram Graphs	Γράφοι Ν-Γραμμάτων

Βιβλιογραφικές Αναφορές

[1] Gao, Y., Zhang, H., Zhao, X., Yan, S., 2017. Event Classification in Microblogs via Social Tracking. ACM Trans Intell Syst Technol 8, 35:1–35:14. https://doi.org/10.1145/2967502

[2] Sara Rosenthal, Noura Farra, Preslav Nakov, 2017. Sentiment Analysis in Twitter. E 11th Int. Workshop Semantic Eval. SemEval-2017.

[3] Violos, John, Konstantinos Tserpes, Iraklis Varlamis, and Theodora Varvarigou. "Text Classification Using the N-Gram Graph Representation Model over High Frequency Data Streams." Frontiers in Applied Mathematics and Statistics 4 (2018). https://doi.org/10.3389/fams.2018.00041.

[4] Nguyen, H.-L., Woon, Y.-K., Ng, W.-K., 2015. A survey on data stream clustering and classification. Knowl. Inf. Syst. 45, 535–569. https://doi.org/10.1007/s10115-014-0808-1

[5] Xu, S., Wang, J., 2016. A fast incremental extreme learning machine algorithm for data streams classification. Expert Syst. Appl. 65, 332–344. https://doi.org/10.1016/j.eswa.2016.08.052

[6] Akidau, T., Bradshaw, R., Chambers, C., Chernyak, S., Fernández-Moctezuma, R.J., Lax, R., McVeety, S., Mills, D., Perry, F., Schmidt, E., Whittle, S., 2015. The Dataflow Model: A Practical Approach to Balancing Correctness, Latency, and Cost in Massive-scale, Unbounded, Out-of-order Data Processing. Proc VLDB Endow 8, 1792–1803. https://doi.org/10.14778/2824032.2824076

[7] Chen, Z., Ma, N., Liu, B., 2018. Lifelong Learning for Sentiment Classification. ArXiv180102808 Cs.

[8] Jiang, M., Liang, Y., Feng, X., Fan, X., Pei, Z., Xue, Y., Guan, R., 2018. Text classification based on deep belief network and softmax regression. Neural Comput. Appl. 29, 61–70. https://doi.org/10.1007/s00521-016-2401-x

[9] Xing, Y., Shen, F., Luo, C., Zhao, J., 2015. L3-SVM: a lifelong learning method for SVM, in: 2015 International Joint Conference on Neural Networks (IJCNN). Presented at the 2015 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. https://doi.org/10.1109/IJCNN.2015.7280379

[10] Wallach, H.M., 2006. Topic Modeling: Beyond Bag-of-words, in: Proceedings of the 23rd International Conference on Machine Learning, ICML '06. ACM, New York, NY, USA, pp. 977–984. https://doi.org/10.1145/1143844.1143967

[11] Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., Demirbas, M., 2010. Short Text Classification in Twitter to Improve Information Filtering, in: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '10. ACM, New York, NY, USA, pp. 841–842. https://doi.org/10.1145/1835449.1835643

[12] Krishnan, S.P.T., Gonzalez, J.L.U., 2015. Google Cloud Pub/Sub, in: Building Your Next Big Thing with Google Cloud Platform. Apress, Berkeley, CA, pp. 277–292. https://doi.org/10.1007/978-1-4842-1004-8_12

[13] G. Salton, A. Wong, and C. S. Yang, "A Vector Space Model for Automatic Indexing," Commun ACM, vol. 18, no. 11, pp. 613–620, Nov. 1975.

[14] E. Amigó, J. Gonzalo, J. Artiles, and F. Verdejo, "A comparison of extrinsic clustering evaluation metrics based on formal constraints," Inf. Retr., vol. 12, no. 4, pp. 461–486, Aug. 2009.

[15] M. Popovič and P. Willett, "The effectiveness of stemming for natural-language access to Slovene textual data," J. Am. Soc. Inf. Sci., vol. 43, no. 5, pp. 384–390, Jun. 1992.

[16] Y. Zhang, R. Jin, and Z.-H. Zhou, “Understanding bag-of-words model: a statistical framework,” Int. J. Mach. Learn. Cybern., vol. 1, no. 1–4, pp. 43–52, Dec. 2010

[17] S. K. M. Wong, W. Ziarko, and P. C. N. Wong, “Generalized Vector Spaces Model in Information Retrieval,” in Proceedings of the 8th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, NY, USA, 1985, pp. 18–25.

[18] T. K. Landauer, P. W. Foltz, and D. Laham, “An introduction to latent semantic analysis,” Discourse Process., vol. 25, no. 2–3, pp. 259–284, 1998.

[19] G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais, “The Vocabulary Problem in Human-system Communication,” Commun ACM, vol. 30, no. 11, pp. 964–971, Nov. 1987.

[20] P. G. H. Golub and D. C. Reinsch, “Singular value decomposition and least squares solutions,” Numer. Math., vol. 14, no. 5, pp. 403–420, Apr. 1970.

[21] Kanungo, T. et al. 2002. An efficient k-means clustering algorithm: analysis and implementation. IEEE Transactions on Pattern Analysis and Machine Intelligence. 24, 7 (Jul. 2002), 881–892.

[22] G. H. Golub, F. T. Luk, and M. L. Overton, “A Block Lanczos Method for Computing the Singular Values and Corresponding Singular Vectors of a Matrix,” ACM Trans Math Softw, vol. 7, no. 2, pp. 149–169, Jun. 1981.

[23] Clauset, A., Shalizi, C., Newman, M. “Power-Law Distributions in Empirical Data,” SIAM Rev., vol. 51, no. 4, pp. 661–703, Nov. 2009.

[24] T. Hofmann, “Probabilistic Latent Semantic Indexing,” in Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, NY, USA, 1999, pp. 50–57.

[25] T. K. Moon, “The expectation-maximization algorithm,” IEEE Signal Process. Mag., vol. 13, no. 6, pp. 47–60, Nov. 1996.

[26] G. J. McLachlan, K. E. Basford, and G. J. McLachlan, Mixture Models. New York, N.Y: CRC Press, 1987.

[27] K. Rose, E. Gurewwitz, and G. Fox, “A Deterministic Annealing Approach to Clustering,” Pattern Recogn Lett, vol. 11, no. 9, pp. 589–594, Sep. 1990.

[28] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet Allocation,” J Mach Learn Res, vol. 3, pp. 993–1022, Mar. 2003.

[29] M. Kuczma, An Introduction to the Theory of Functional Equations and Inequalities: Cauchy’s Equation and Jensen’s Inequality. Springer, 2008.

[30] J. M. Joyce, “Kullback-Leibler Divergence,” in International Encyclopedia of Statistical Science, M. Lovric, Ed. Springer Berlin Heidelberg, 2011, pp. 720–722.

[31] Gomes, Heitor Murilo, Barddal, J.P., Enembreck, F., Bifet, A., 2017. A Survey on Ensemble Learning for Data Stream Classification. ACM Comput Surv 50, 23:1–23:36. https://doi.org/10.1145/3054925

[32] Bertini Junior, J.R., Nicoletti, M. do C., 2018. An iterative boosting-based ensemble for streaming data classification. Inf. Fusion 45, 66–78. https://doi.org/10.1016/j.inffus.2018.01.003

[33] Wang, L., Shen, H., Tian, H., 2017. Weighted Ensemble Classification of Multi-label Data Streams, in: Advances in Knowledge Discovery and Data Mining, Lecture Notes in Computer Science. Presented at the

Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, Cham, pp. 551–562. https://doi.org/10.1007/978-3-319-57529-2_43

[34] Al-Khateeb, T., Masud, M.M., Al-Naami, K.M., Seker, S.E., Mustafa, A.M., Khan, L., Trabelsi, Z., Aggarwal, C., Han, J., 2016. Recurring and Novel Class Detection Using Class-Based Ensemble for Evolving Data Stream. IEEE Trans. Knowl. Data Eng. 28, 2752–2764. https://doi.org/10.1109/TKDE.2015.2507123

[35] Bifet, A., Zhang, Jiajin, Fan, W., He, C., Zhang, Jianfeng, Qian, J., Holmes, G., Pfahringer, B., 2017. Extremely Fast Decision Tree Mining for Evolving Data Streams, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '17. ACM, New York, NY, USA, pp. 1733–1742. https://doi.org/10.1145/3097983.3098139

[36] Gomes, Heitor M., Bifet, A., Read, J., Barddal, J.P., Enembreck, F., Pfharinger, B., Holmes, G., Abdessalem, T., 2017. Adaptive random forests for evolving data stream classification. Mach. Learn. 106, 1469–1495. https://doi.org/10.1007/s10994-017-5642-8

[37] Data Science Platform [WWW Document], 2018. . RapidMiner. URL https://rapidminer.com/ (accessed 3.21.18).

[38] Machine Learning for Data Streams [WWW Document], 2018. . MIT Press. URL https://mitpress.mit.edu/books/machine-learning-data-streams (accessed 3.21.18).

[39] Zhang, X., Zhao, J., LeCun, Y., 2015. Character-level Convolutional Networks for Text Classification. ArXiv150901626 Cs.

[40] Voulodimos, A., Doulamis, N., Doulamis, A., Protopapadakis, E., 2018. Deep Learning for Computer Vision: A Brief Review [WWW Document]. Comput. Intell. Neurosci. https://doi.org/10.1155/2018/7068349

[41] Glorot, X., Bordes, A., Bengio, Y., 2011. Domain Adaptation for Large-scale Sentiment Classification: A Deep Learning Approach, in: Proceedings of the 28th International Conference on International Conference on Machine Learning, ICML'11. Omnipress, USA, pp. 513–520.

[42] Hassan, A., Mahmood, A., 2017. Efficient Deep Learning Model for Text Classification Based on Recurrent and Convolutional Layers, in: 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA). Presented at the 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 1108–1113. https://doi.org/10.1109/ICMLA.2017.00009

[43] Sima, A.-C., Stockinger, K., Affolter, K., Braschler, M., Monte, P., Kaiser, L., 2018. A hybrid approach for alarm verification using stream processing, machine learning and text analytics. Presented at the International Conference on Extending Database Technology (EDBT), March 26-29, 2018, ACM. https://doi.org/10.21256/zhaw-3487

[44] Read, J., Perez-Cruz, F., Bifet, A., 2015. Deep Learning in Partially-labeled Data Streams, in: Proceedings of the 30th Annual ACM Symposium on Applied Computing, SAC '15. ACM, New York, NY, USA, pp. 954–959. https://doi.org/10.1145/2695664.2695871

[45] Belcastro, L., Marozzo, F., Talia, D., Trunfio, P., 2017. Big Data Analysis on Clouds, in: Handbook of Big Data Technologies. Springer, Cham, pp. 101–142. https://doi.org/10.1007/978-3-319-49340-4_4

[46] Dias de Assunção, M., da Silva Veith, A., Buyya, R., 2018. Distributed data stream processing and edge computing: A survey on resource elasticity and future directions. J. Netw. Comput. Appl. 103, 1–17. https://doi.org/10.1016/j.jnca.2017.12.001

[47] Google Cloud Big Data and Machine Learning Blog, 2016. Comparing Cloud Dataflow autoscaling to Spark and Hadoop [WWW Document]. Google Cloud Platf. URL https://cloud.google.com/blog/big-data/2016/03/comparing-cloud-dataflow-autoscaling-to-spark-and-hadoop (accessed 3.8.18).

[48] Tyler Akidau, Frances Perry, 2016. Dataflow/Beam & Spark: A Programming Model Comparison [WWW Document]. Google Cloud Platf. URL https://cloud.google.com/dataflow/blog/dataflow-beam-and-spark-comparison (accessed 3.8.18).

[49] Giannakopoulos, G., Karkaletsis, V., Vouros, G., Stamatopoulos, P., 2008. Summarization System Evaluation Revisited: N-gram Graphs. ACM Trans Speech Lang Process 5, 5:1–5:39. https://doi.org/10.1145/1410358.1410359

[50] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," Proc. Natl. Acad. Sci., vol. 95, no. 25, pp. 14863–14868, Dec. 1998.

[51] P. Haldar, I. D. Pavord, D. E. Shaw, M. A. Berry, M. Thomas, C. E. Brightling, A. J. Wardlaw, and R. H. Green, "Cluster Analysis and Clinical Asthma Phenotypes," Am. J. Respir. Crit. Care Med., vol. 178, no. 3, pp. 218–224, Aug. 2008.

[52] M. J. Berry and G. Linoff, Data Mining Techniques: For Marketing, Sales, and Customer Support. New York, NY, USA: John Wiley & Sons, Inc., 1997.

[53] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," Phys. Rev. E, vol. 69, no. 2, p. 026113, Feb. 2004.

[54] J. Shi and J. Malik, "Normalized cuts and image segmentation," IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, no. 8, pp. 888–905, Aug. 2000.

[55] A. Shepitsen, J. Gemmell, B. Mobasher, and R. Burke, "Personalized Recommendation in Social Tagging Systems Using Hierarchical Clustering," in Proceedings of the 2008 ACM Conference on Recommender Systems, New York, NY, USA, 2008, pp. 259–266.

[56] B. W. Kernighan and S. Lin, "An Efficient Heuristic Procedure for Partitioning Graphs," Bell Syst. Tech. J., vol. 49, no. 2, pp. 291–307, Feb. 1970.

[57] R. Kuhn and D. P. Culhane, "Applying Cluster Analysis to Test a Typology of Homelessness by Pattern of Shelter Utilization: Results from the Analysis of Administrative Data," Am. J. Community Psychol., vol. 26, no. 2, pp. 207–232, Apr. 1998.

[58] V. Estivill-Castro, "Why So Many Clustering Algorithms: A Position Paper," SIGKDD Explor Newsl, vol. 4, no. 1, pp. 65–75, Jun. 2002.

[59] B. E. Dom, "An Information-theoretic External Cluster-validity Measure," in Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence, San Francisco, CA, USA, 2002, pp. 137–145.

[60] J. B. Hirschberg and A. Rosenberg, "V-Measure: A conditional entropy-based external cluster evaluation," 2007.

[61] M. Meilă, "Comparing Clusterings: An Axiomatic View," in Proceedings of the 22Nd International Conference on Machine Learning, New York, NY, USA, 2005, pp. 577–584.

[62] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "On Clustering Validation Techniques," J. Intell. Inf. Syst., vol. 17, no. 2–3, pp. 107–145, Dec. 2001.

[63] W. M. Rand, "Objective Criteria for the Evaluation of Clustering Methods," J. Am. Stat. Assoc., vol. 66, no. 336, pp. 846–850, 1971.

[64] P. Jaccard, "The Distribution of the Flora in the Alpine Zone.1," New Phytol., vol. 11, no. 2, pp. 37–50, Feb. 1912.

[65] E. B. Fowlkes and C. L. Mallows, “A Method for Comparing Two Hierarchical Clusterings,” J. Am. Stat. Assoc., vol. 78, no. 383, pp. 553–569, Sep. 1983.

[66] M. S. George Karypis and Vipin Kumar, “A Comparison of Document Clustering Techniques,” KDD Workshop Text Min. 2000.

[67] G. Forman, “An Extensive Empirical Study of Feature Selection Metrics for Text Classification,” J Mach Learn Res, vol. 3, pp. 1289–1305, Mar. 2003.

[68] P. Pantel and D. Lin, “Efficiently Clustering Documents with Committees,” in PRICAI 2002: Trends in Artificial Intelligence, M. Ishizuka and A. Sattar, Eds. Springer Berlin Heidelberg, 2002, pp. 424–433.

[69] A. Bagga and B. Baldwin, “Entity-based Cross-document Coreferencing Using the Vector Space Model,” in Proceedings of the 17th International Conference on Computational Linguistics - Volume 1, Stroudsburg, PA, USA, 1998, pp. 79–85.

[70] C. J. V. RIJSBERGEN, “FOUNDATION OF EVALUATION,” J. Doc., vol. 30, no. 4, pp. 365–373, Dec. 1974.

[71] S. V. Stehman, “Selecting and interpreting measures of thematic classification accuracy,” Remote Sens. Environ., vol. 62, no. 1, pp. 77–89, Oct. 1997.

[72] D. L. Davies and D. W. Bouldin, “A Cluster Separation Measure,” IEEE Trans. Pattern Anal. Mach. Intell., vol. PAMI-1, no. 2, pp. 224–227, Apr. 1979.

[73] J. C. Dunn, “A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters,” J. Cybern., vol. 3, no. 3, pp. 32–57, 1973.

[74] P. J. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,” J. Comput. Appl. Math., vol. 20, pp. 53–65, Nov. 1987.

[75] Forman, G., Scholz, M., 2010. Apples-to-apples in Cross-validation Studies: Pitfalls in Classifier Performance Measurement. SIGKDD Explor Newsl 12, 49–57. doi:10.1145/1882471.1882479

[76] D’Addabbo, A., Maglietta, R., 2015. Parallel selective sampling method for imbalanced and large data classification. Pattern Recognit. Lett. 62, 61–67. https://doi.org/10.1016/j.patrec.2015.05.008

[77] Psomakelis, E., Tserpes, K., Anagnostopoulos, D., Varvarigou, T., 2015. Comparing methods for Twitter Sentiment Analysis. ArXiv150502973 Cs.

[78] Raymond, J.W., Gardiner, E.J., Willett, P., 2002. RASCAL: Calculation of graph similarity using maximum common edge subgraphs. Comput. J. 45, 2002.

[79] Nikolić, M., 2012. Measuring similarity of graph nodes by neighbor matching. Intell. Data Anal. 16, 865–878. https://doi.org/10.3233/IDA-2012-00556

[80] Fotis Aisopos, Dimitrios Tzannetos, John Violos, Theodora Varvarigou, 2016. Using N-Gram Graphs for Sentiment Analysis: An Extended Study on Twitter - IEEE Conference Publication [WWW Document]. URL http://ieeexplore.ieee.org/abstract/document/7474354/ (accessed 3.8.18).

[81] Jain, D.S. and S., 2015. Evaluation of Stemming and Stop Word Techniques on Text Classification Problem [WWW Document]. Int. J. Sci. Res. Comput. Sci. Eng. URL http://www.isroset.org/journal/IJSRCSE/ (accessed 5.1.18).

[82] Van Asch, V., 2013. Macro- and micro-averaged evaluation measures.

[83] Joachims, T., 1998. Text categorization with Support Vector Machines: Learning with many relevant features, in: Nédellec, C., Rouveirol, C. (Eds.), Machine Learning: ECML-98, Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 137–142.

[84] Porter, M. f., 2006. An algorithm for suffix stripping. Program 40, 211–218. https://doi.org/10.1108/00330330610681286

[85] Li, B., Vogel, C., 2010. Improving Multiclass Text Classification with Error-Correcting Output Coding and Sub-class Partitions, in: Farzindar, A., Kešelj, V. (Eds.), Advances in Artificial Intelligence, Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 4–15.

[86] Lan, M., Tan, C.-L., Low, H.-B., 2006. Proposing a New Term Weighting Scheme for Text Categorization, in: Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1, AAAI'06. AAAI Press, Boston, Massachusetts, pp. 763–768.

[87] Larochelle, H., Bengio, Y., 2008. Classification Using Discriminative Restricted Boltzmann Machines, in: Proceedings of the 25th International Conference on Machine Learning, ICML '08. ACM, New York, NY, USA, pp. 536–543. https://doi.org/10.1145/1390156.1390224

[88] Jason DM Rennie, 2003. On The Value of Leave-One-Out Cross-Validation Bounds.

[89] Dhillon, I. et al. 2005. A Fast Kernel-based Multilevel Algorithm for Graph Clustering. Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining (New York, NY, USA, 2005), 629–634.

[90] Violos, John, Konstantinos Tserpes, Athanasios Papaoikonomou, Magdalini Kardara, and Theodora Varvarigou. "Clustering Documents Using the 3-Gram Graph Representation Model." In Proceedings of the 18th Panhellenic Conference on Informatics, 29:1–29:5. PCI '14. New York, NY, USA: ACM, 2014. https://doi.org/10.1145/2645791.2645812.

[91] Jianbo Shi and Jitendra Malik, Member, IEEE; Normalized Cuts and Image Segmentation; IEEE Transactions on pattern analysis and machine intelligence, vol. 22, no. 8, august 2000

[92] Stanley Wasserman and Katherine Faust, Social Network Analysis: Methods and Applications, Cambridge University

[93] Tero Aittokallio and Benno Schwikowski; Graph-based methods for analyzing networks in cell biology; BRIEFINGS IN BIOINFORMATICS. VOL 7. NO 3. 243^255

[94] S. N. Dorogovtsev and J. F. F. Mendes, Evolution of Networks: From Biological Nets to the Internet and WWW ~Oxford UniversityPress, Oxford, 2003

[95] Ronald L. Breiger, Scott A. Boorman and Phipps Arabie; An algorithm for blocking relational data, with applications to social network analysis and comparison with multidimensional scaling; technical report no. 244

[96] Path Similarity Skeleton Graph Matching; Xiang Bai and Longin Jan Latecki, Senior Member, IEEE

[97] J. M. Anthonisse, Technical Report BN 9/71, Stichting Mathematicsh Centrum, Amsterdam ~1971! ~unpublished!.

[98] L. C. Freeman, Sociometry 40, 35 ~1977

[99] Andrew Y. Wu, Michael Garland, Jiawei Han Mining Scale-free Networks using Geodesic Clustering; Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining

[100] J. Scott, Social Network Analysis: A Handbook, Sage, London, 2nd edition, 2000

[101] Mark Newman, Fast algorithm for detecting community structure in networks, Phys. Rev., E, 2004

[102] M. Girvan and M. Newman, Community structure in social and biological networks. In Proceedings of National Academic Science, USA 99, 7821{7826, 2002.

[103] Mark Newman, Detecting community structure in networks, Eur. Phys. 38, 321-330, 2004.

[104] Aaron Clauset, et al., Finding community structure in very large networks, Phys. Rev. E 70, 066111, 2004

[105] Norman L. Johnson, Adrienne W. Kemp , Samuel Kotz; Univariate Discrete Distributions; Wiley Series in Probability and Statistics

[106] Norman L. Johnson, Samuel Kotz, N. Balakrishnan; Discrete Multivariate Distributions; Wiley Series in Probability and Statistics

[107] D. Zhou, E. Manavoglu, J. Li, C.L. Giles, and H. Zha. Probabilistic models for discovering E-communities. In WWW ’06: Proceedings of the 15th international conference onWorldWideWeb, page 182. ACM, 2006

[108] S. Asur, S. Parthasarathy, and D. Ucar. An event-based framework for characterizing the evolutionary behavior of interaction graphs. In Proc. of the 13th ACM SIGKDD Conference, 2007.

[109] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. On the bursty evolution of blogspace. In Proc. of the 12th WWW Conference, 2003.

[110] R. Kumar, J. Novak, and A. Tomkins. Structure and evolution of online social networks. In Proc. of the 12th ACM SIGKDD Conference, 2006.

[111] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In Proc. of the 11th ACM SIGKDD Conference, 2005

[112] Y. Lin, H. Sundaram, Y. Chi, J. Tatemura, and B. L.Tseng. Blog community discovery and evolution based on mutual awareness expansion. In Proc. of the Int. Conf. on Web Intelligence, 2007.

[113] G. Palla, A.-L. Barabasi, and T. Vicsek. Quantifying social group evolution. Nature, 446, 2007.

[114] M. Spiliopoulou, I. Ntoutsi, Y. Theodoridis, and R. Schult. Monic: modeling and monitoring cluster transitions. In Proc. of the 12th ACM SIGKDD Conference, 2006

[115] M. Toyoda and M. Kitsuregawa. Extracting evolution of web communities from a series of web archives. In HYPERTEXT ’03: Proc. of the 14th ACM conference on hypertext and hypermedia, 2003.

[116] Y.R. Lin, Y. Chi, S. Zhu, H. Sundaram, and B.L. Tseng. Facetnet: a framework for analyzing communities and their evolutions in dynamic networks. In WWW ’08: Proceeding of the 17th international conference on World Wide Web, pages 685–694, New York, NY, USA, 2008. ACM

[117] D. Chakrabarti, R. Kumar, and A. Tomkins.Evolutionary clustering. In Proc. of the 12th ACM SIGKDD Conference, 2006

[118] K. Yu, S. Yu, and V. Tresp. Soft clustering on graphs. In NIPS, 2005.

[119] Notes on Kullback-Leibler Divergence and Likelihood Theory Jonathon Shlens Systems Neurobiology Laboratory, Salk Insitute for Biological Studies, La Jolla, CA 92037

[120] Wei Chen, Chi Wang, Yajun Wang ;Scalable Influence Maximization for Prevalent Viral Marketing in Large-Scale Social Networks; Data Mining (ICDM), 2010 IEEE 10th International Conference on 13-17 Dec. 2010

[121] William Thomas Tutte; Graph Theory, Cambridge University Press

[122] Dijkstra, E. W. (1959). "A note on two problems in connexion with graphs". Numerische Mathematik 1: 269–271

[123] “Benevento, la denuncia:,” YouReporter. [Online]. Available: http://www.youreporter.it/video_Benevento_la_denuncia_l_alveo_del_fiume_mai_pulito. [Accessed: 03-Apr-2016].

[124] “Storms cause floods in southern Italy (2 - English - ANSA.it.” [Online]. Available: http://www.ansa.it/english/news/vatican/2015/10/16/storms-cause-floods-in-southern-italy-2_eb4519f6-5bc6-4f94-b397-b56e93fe7f01.html. [Accessed: 03-Apr-2016].

[125] Aisopos, F. et al. 2011. Sentiment analysis of social media content using N-Gram graphs. Proceedings of the 3rd ACM SIGMM international workshop on Social media (2011), 9–14.

[126] Huang, S. et al. 2013. Sentiment and Topic Analysis on Social Media: A Multi-task Multi-label Classification Approach. Proceedings of the 5th Annual ACM Web Science Conference (New York, NY, USA, 2013), 172–181

[127] Liu, S.M. and Chen, J.-H. 2015. A Multi-label Classification Based Approach for Sentiment Classification. Expert Syst. Appl. 42, 3 (Feb. 2015), 1083–1093.

[128] Wilson, T. et al. 2005. Recognizing Contextual Polarity in Phrase-level Sentiment Analysis. Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (Stroudsburg, PA, USA, 2005), 347–354.

[129] Taboada, M. et al. 2011. Lexicon-Based Methods for Sentiment Analysis. Computational Linguistics. 37, 2 (Apr. 2011), 267–307.

[130] Khan, A.Z. et al. 2015. Combining lexicon-based and learning-based methods for Twitter sentiment analysis. International Journal of Electronics, Communication and Soft Computing Science & Engineering (IJECSCSE). (2015), 89.

[131] Baccianella, S. et al. 2010. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. (May 2010).

[132] Pak, A. and Paroubek, P. 2010. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta (2010).

[133] IEEE Xplore Abstract - An ontology based sentiment analysis for mobile products using tweets: http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=6921974&url=http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Fabs_all.jsp%3Farnumber%3D6921974. Accessed: 2016-04-13.

[134] Kontopoulos, E. et al. 2013. Ontology-based sentiment analysis of twitter posts. Expert Systems with Applications. 40, 10 (Aug. 2013), 4065–4074.

[135] Sam, K. M and Chatwin, C. R. Ontology-Based Sentiment Analysis Model of Customer Reviews for Electronic Products. International Journal of e-Education, e-Business, e-Management and e-Learning.

[136] Violos, John, Konstantinos Tserpes, Evangelos Psomakelis, Konstantinos Psychas, and Theodora Varvarigou. “Sentiment Analysis Using Word-Graphs.” In Proceedings of the 6th International Conference on Web Intelligence, Mining and Semantics, 22:1–22:9. WIMS ’16. New York, NY, USA: ACM, 2016. https://doi.org/10.1145/2912845.2912863.

[137] Gunn, S.R. 1998. Support Vector Machines for Classification and Regression.

[138] John, G. and Langley, P. 1995. Estimating Continuous Distributions in Bayesian Classifiers. In Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence (1995), 338–345.

[139] Chagheri, S. et al. 2012. Feature vector construction combining structure and content for document classification. 2012 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT) (Mar. 2012), 946–950.

[140] Xu, Y. et al. 2007. A study on mutual information-based feature selection for text categorization. Journal of Computational Information Systems. 3, 3 (2007), 1007–1012.

[141] Narr, S. et al. 2012. Language-independent twitter sentiment analysis. Knowledge Discovery and Machine Learning (KDML), LWA. (2012), 12–14.

[142] Maglogiannis, I.G. 2007. Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in EHealth, HCI, Information Retrieval and Pervasive Technologies. IOS Press.

[143] Atefeh, Farzindar, and Wael Khreich. "A Survey of Techniques for Event Detection in Twitter." Computational Intelligence 31, no. 1 (February 1, 2015): 132–64. https://doi.org/10.1111/coin.12017.

[144] Brants, Thorsten, and Francine Chen. "A System for New Event Detection." In SIGIR Forum (ACM Special Interest Group on Information Retrieval), 330–37, 2003. https://doi.org/10.1145/860435.860495.

[145] Niwattanakul, S., Singthongchai, J., Naenudorn, E., Wanapu, S., 2013. Using of Jaccard Coefficient for Keywords Similarity. Lect. Notes Eng. Comput. Sci. 2202, 380–384.

[146] Georgios Petkos, Symeon Papadopoulos, Vasileios Mezaris, Raphael Troncy, Philipp Cimiano, Timo Reuter, Yiannis Kompatsiaris, 2014. Social Event Detection at MediaEval: a three-year retrospect of tasks and results. Proc. 2014 Workshop Soc. Events Web Multimed. Conjuction ICMR.

[147] Zubiaga, Arkaitz, Damiano Spina, Enrique Amigó, and Julio Gonzalo. "Towards Real-Time Summarization of Scheduled Events from Twitter Streams." ArXiv:1204.3731 [Cs], April 17, 2012. http://arxiv.org/abs/1204.3731.

[148] Petrović, Saša, Miles Osborne, and Victor Lavrenko. "Streaming First Story Detection with Application to Twitter." In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 181–189. HLT '10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010. http://dl.acm.org/citation.cfm?id=1857999.1858020.

[149] J. L. Herlocker, J. A. Konstan, and J. Riedl, "Explaining Collaborative Filtering Recommendations," in Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work, New York, NY, USA, 2000, pp. 241–250.

[150] P. N. Yianilos, "Data Structures and Algorithms for Nearest Neighbor Search in General Metric Spaces," in Proceedings of the Fourth Annual ACM-SIAM Symposium on Discrete Algorithms, Philadelphia, PA, USA, 1993, pp. 311–321.

[151] G. Linden, B. Smith, and J. York, "Amazon.com recommendations: item-to-item collaborative filtering," IEEE Internet Comput., vol. 7, no. 1, pp. 76–80, Jan. 2003.

[152] T. Hofmann, "Latent Semantic Models for Collaborative Filtering," ACM Trans Inf Syst, vol. 22, no. 1, pp. 89–115, Jan. 2004.

[153] B. Marlin and R. S. Zemel, "The Multiple Multiplicative Factor Model for Collaborative Filtering," in Proceedings of the Twenty-first International Conference on Machine Learning, New York, NY, USA, 2004, p. 73–.

[154] A. S. Das, M. Datar, A. Garg, and S. Rajaram, “Google News Personalization: Scalable Online Collaborative Filtering,” in Proceedings of the 16th International Conference on World Wide Web, New York, NY, USA, 2007, pp. 271–280.

[155] A. Z. Broder, “On the resemblance and containment of documents,” in Compression and Complexity of Sequences 1997. Proceedings, 1997, pp. 21–29.

[156] M. J. Pazzani and D. Billsus, “Content-Based Recommendation Systems,” in The Adaptive Web, P. Brusilovsky, A. Kobsa, and W. Nejdl, Eds. Springer Berlin Heidelberg, 2007, pp. 325–341.

[157] J. Xu and W. B. Croft, “Corpus-based Stemming Using Cooccurrence of Word Variants,” ACM Trans Inf Syst, vol. 16, no. 1, pp. 61–81, Jan. 1998.

[158] H. C. Wu, R. W. P. Luk, K. F. Wong, and K. L. Kwok, “Interpreting TF-IDF Term Weights As Making Relevance Decisions,” ACM Trans Inf Syst, vol. 26, no. 3, pp. 13:1–13:37, Jun. 2008.

[159] S. R. Safavian and D. Landgrebe, “A survey of decision tree classifier methodology,” IEEE Trans. Syst. Man Cybern., vol. 21, no. 3, pp. 660–674, May 1991.

[160] L. Peterson, “K-nearest neighbor,” Scholarpedia, vol. 4, no. 2, p. 1883, 2009.

[161] J. Rocchio, “Relevance feedback in information retrieval,” in The SMART Retrieval System: Experiments in Automatic Document Processing, G. Salton, Ed. Prentice-Hall, Englewood Cliffs NJ, 1971, pp. 313–323.

[162] T. Zhang and V. S. Iyengar, “Recommender Systems Using Linear Classifiers,” J Mach Learn Res, vol. 2, pp. 313–334, Mar. 2002.

[163] J. McAuley and J. Leskovec. Hidden factors and hidden topics: understanding rating dimensions with review text. RecSys, 2013.

[164] Braga-Neto, U.M., Dougherty, E.R., 2004. Is cross-validation valid for small-sample microarray classification? Bioinformatics 20, 374–380. https://doi.org/10.1093/bioinformatics/btg419

[165] Eberly, L.E., 2007. Multiple Linear Regression, in: Topics in Biostatistics, Methods in Molecular BiologyTM. Humana Press, pp. 165–187. https://doi.org/10.1007/978-1-59745-530-5_9

[166] Enzo Busseti, 2012. Deep Learning for Time Series Modeling | Deep Learning | Artificial Neural Network [WWW Document]. Scribd. URL https://www.scribd.com/document/293198006/Deep-Learning-for-Time-Series-Modeling (accessed 12.8.17).

[167] Chang, C.-C., Lin, C.-J., 2011. LIBSVM: A Library for Support Vector Machines. ACM Trans Intell Syst Technol 2, 27:1–27:27. https://doi.org/10.1145/1961189.1961199

[168] Hall, D.L., McMullen, S.A.H., 2004. Mathematical Techniques in Multisensor Data Fusion. Artech House. Reid Turner, C., Fuggetta, A., Lavazza, L., Wolf, A.L., 1999. A conceptual basis for feature engineering. J. Syst. Softw. 49, 3–15. https://doi.org/10.1016/S0164-1212(99)00062-X

[169] Reid Turner, C., Fuggetta, A., Lavazza, L., Wolf, A.L., 1999. A conceptual basis for feature engineering. J. Syst. Softw. 49, 3–15. https://doi.org/10.1016/S0164-1212(99)00062-X

[170] Jolliffe, Ian. “Principal Component Analysis.” In International Encyclopedia of Statistical Science, 1094–96. Springer, Berlin, Heidelberg, 2011. https://doi.org/10.1007/978-3-642-04898-2_455.

[171] Hyvärinen, Aapo. Survey on Independent Component Analysis, 1999.

[172] “Open Cloud Computing Interface – Open Standard | Open Community.” Accessed August 12, 2018. http://occi-wg.org/.

[173] Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. “The WEKA Data Mining Software: An Update.” SIGKDD Explor. Newsl. 11, no. 1 (November 2009): 10–18. https://doi.org/10.1145/1656274.1656278.

[174] “Deeplearning4j.” Accessed August 12, 2018. https://deeplearning4j.org/.

[175] Violos, John, Vinicius Monteiro de Lira, Patrizio Dazzi, Jörn Altmann, Baseem Al-Athwari, Antonia Schwichtenberg, Young-Woo Jung, Theodora Varvarigou, and Konstantinos Tserpes. “User Behavior and Application Modeling in Decentralized Edge Cloud Infrastructures.” In Economics of Grids, Clouds, Systems, and Services, 193–203. Lecture Notes in Computer Science. Springer, Cham, 2017. https://doi.org/10.1007/978-3-319-68066-8_15.

[176] Santoso, G. Z., Y. W. Jung, S. W. Seok, E. Carlini, P. Dazzi, J. Altmann, J. Violos, and J. Marshall. “Dynamic Resource Selection in Cloud Service Broker.” In 2017 International Conference on High Performance Computing Simulation (HPCS), 233–35, 2017. https://doi.org/10.1109/HPCS.2017.43.

[177] Altmann, Jörn, Baseem Al-Athwari, Emanuele Carlini, Massimo Coppola, Patrizio Dazzi, Ana Juan Ferrer, Netsanet Haile, et al. “BASMATI: An Architecture for Managing Cloud and Edge Resources for Mobile Users.” In Economics of Grids, Clouds, Systems, and Services, 56–66. Lecture Notes in Computer Science. Springer, Cham, 2017. https://doi.org/10.1007/978-3-319-68066-8_5.