Πανεπιστήμιο Ιωαννίνων
Σχολή Θετικών Επιστημών
Τμήμα Πληροφορικής

Μεταπτυχιακή Εργασία Ειδίκευσης

ΤΑΞΙΝΟΜΗΣΗ ΚΕΙΜΕΝΩΝ ΜΕ ΧΡΗΣΗ ΣΤΑΤΙΣΤΙΚΩΝ ΜΕΘΟΔΩΝ

ΣΤΥΛΙΑΝΟΣ ΜΑΣΤΡΟΓΙΑΝΝΑΚΗΣ

Ιωάννινα, Ιούνιος 2003

Πρόλογος

Σε αυτή την εργασία εξετάζονται στατιστικές μέθοδοι ταξινόμησης κειμένων. Οι μέθοδοι αυτές προσπαθούν να αναδείξουν την πιθανοτική φύση του προβλήματος ταξινόμησης κειμένων. Στόχος της εργασίας είναι να παρουσιάσει υπάρχουσες στατιστικές μεθόδους και να προτείνει νέες μεθόδους που προσπαθούν να αντιμετωπίσουν ικανοποιητικά τα προβλήματα των προηγουμένων μεθόδων.

Η παρούσα εργασία εκπονήθηκε στα πλαίσια του προγράμματος μεταπτυχιακών σπουδών του τμήματος Πληροφορικής του Πανεπιστημίου Ιωαννίνων. Πραγματοποιήθηκε υπό την επίβλεψη του Επ. Καθ. Α. Λύκα, η συμβολή του οποίου υπήρξε καθοριστική. Θα ήθελα να τον ευχαριστήσω για την ουσιώδη συνεργασία μας, το περιβάλλον εργασίας και την αμέριστη συμπαράστασή που μου πρόσφερε, τα οποία αποτέλεσαν αποφασιστικά στοιχεία για την ολοκλήρωση της εργασίας. Θα ήθελα επίσης να ευχαριστήσω τους καθηγητές του τμήματος Πληροφορικής κ.κ. Δ. Φωτιάδη και Κ. Μπλέκα για την βοήθεια και τις συμβουλές τους, τον συμφοιτητή μου Κ. Κωνσταντινόπουλο ο οποίος με βοήθησε στην πρώτη μου επαφή με το χώρο της στατιστικής αναγνώρισης προτύπων, καθώς και τον Κ. Βαλασούλη, του οποίου η βοήθεια σε θέματα τεχνικής φύσης ήταν πολύτιμη. Τέλος θα ήθελα να αφιερώσω αυτή την εργασία στην οικογένειά μου, και να τους ευχαριστήσω για την κατανόηση και συμπαράστασή τους.

Ιωάννινα, Ιούνιος 2003
Στυλιανός Μαστρογιαννάκης

Περιεχόμενα

Κεφάλαιο 1

Εισαγωγή

1.1 Το Πρόβλημα της Ταξινόμησης Κειμένου

Ο όρος ταξινόμηση κειμένου (Text Classification - TC) αναφέρεται στην εύρεση της κατηγορίας ενός κειμένου μέσα από ένα προκαθορισμένο σύνολο θεματικών κατηγοριών.

Ο κλάδος της ταξινόμησης κειμένου χρονολογείται από τις αρχές της δεκαετίας του 1960, ωστόσο μόλις την τελευταία δεκαετία επήλθε σημαντική εξέλιξη σε αυτόν. Σε αυτό συνέβαλλαν η ολοένα αυξανόμενη διαθεσιμότητα εγγράφων σε ψηφιακή μορφή καθώς και η ανάγκη που πρόεκυψε για την οργάνωσή τους. Ο κλάδος της ταξινόμησης κειμένου βρίσκει εφαρμογή σε πολλές περιοχές όπως, στην οργάνωση εγγράφων (document organization), στην ιεραρχική κατηγοριοποίηση ιστοσελίδων, στο φιλτράρισμα εγγράφων (text filtering), στην αποσαφήνιση της έννοιας μιας λέξης και γενικά σε εφαρμογές που απαιτούν την οργάνωση ενός πλήθους από έγγραφα ή την επιλεκτική και προσαρμοσμένη αποστολή εγγράφων.

Σε αυτό το εισαγωγικό κεφάλαιο περιγράφουμε έννοιες και θέματα που αφορούν τον κλάδο της ταξινόμησης κειμένου και οι οποίες θα χρησιμοποιηθούν στα επόμενα κεφάλαια.

1.2 Ορισμός του Προβλήματος της Ταξινόμησης Κειμένου

Σε αυτό το σημείο θα δώσουμε έναν ορισμό για το πρόβλημα μας: Ορίζουμε ως ταξινόμηση κειμένου (TC) την ανάθεση μιας λογικής (Boolean) τιμής σε κάθε ζεύγος $\langle d_j, c_i \rangle \in \mathcal{D} \times \mathcal{C}$, όπου $\mathcal{D}$ είναι ο χώρος των εγγράφων και $C = \{c_1, \ldots, c_{|C|}\}$ είναι ένα σύνολο από προκαθορισμένες κατηγορίες. Η

ανάθεση της τιμής $\mathcal{T}$ (αληθής) στο ζεύγος $\langle d_j, c_i \rangle$ δηλώνει ότι το έγγραφο d_j ανήκει στην κατηγορία c_i, ενώ η ανάθεση της τιμής $\mathcal{F}$ (ψευδής) δηλώνει ότι το έγγραφο d_j δεν ανήκει στην κατηγορία c_i.

Στόχος μας είναι να προσεγγίσουμε την άγνωστη συνάρτηση στόχο (target function) $\breve{\Phi} : \mathcal{D} \times \mathcal{C} \rightarrow \{\mathcal{T}, \mathcal{F}\}$, με μία συνάρτηση $\Phi : \mathcal{D} \times \mathcal{C} \rightarrow \{\mathcal{T}, \mathcal{F}\}$, η οποία ονομάζεται ταξινομητής (classifier). Η προσέγγιση πρέπει να γίνει κατά τέτοιο τρόπο ώστε ο ταξινομητής Φ και η άγνωστη συνάρτηση στόχος $\breve{\Phi}$ να συμπίπτουν όσο το δυνατόν περισσότερο.

1.3 Κατηγοριοποίηση των Μεθόδων Ταξινόμησης Κειμένου

1.3.1 Απλή και Πολλαπλή Ταξινόμηση Κειμένου

Ορίζουμε ως απλή (single-label) την περίπτωση της ταξινόμησης κειμένου κατά την οποία κάθε έγγραφο $d_j \in \mathcal{D}$ ανήκει σε μία μόνο κατηγορία $c_i, i = 1, \ldots, |\mathcal{C}|$. Οι μέθοδοι στις οποίες θα αναφερθούμε στα επόμενα κεφάλαια, υπάγονται σε αυτήν την περίπτωση.

Αντίθετα, στην πολλαπλή (multilabel) ταξινόμηση κειμένου μπορούμε να ταξινομήσουμε ένα έγγραφο d_j σε έναν αριθμό από κατηγορίες, από 0 έως $|\mathcal{C}|$.

Μία ειδική περίπτωση της απλής ταξινόμησης κειμένου είναι η δυαδική (binary), κατά την οποία κάθε έγγραφο $d_j \in \mathcal{D}$ ανήκει σε μία κατηγορία c_i ή στην συμπληρωματική της $\bar{c}_i$. Η ειδική αυτή περίπτωση βρίσκει εφαρμογή σε προβλήματα όπου μας ενδιαφέρει να μάθουμε αν ένα έγγραφο ανήκει σε μία κατηγορία ή όχι.

Το μεγαλύτερο τμήμα της υπάρχουσας βιβλιογραφίας στον τομέα της ταξινόμησης κειμένου διατυπώνεται με όρους της δυαδικής ταξινόμησης. Για το λόγο αυτό στις επόμενες ενότητες του εισαγωγικού κεφαλαίου θεωρούμε ότι το πρόβλημα της ταξινόμησης εγγράφων στο σύνολο των κατηγοριών $\mathcal{C} = \{c_1, \ldots, c_{|\mathcal{C}|}\}$ αποτελείται από $|\mathcal{C}|$ ανεξάρτητα προβλήματα ταξινόμησης στις κατηγορίες $c_i, i = 1, \ldots, |\mathcal{C}|$. Καθένας από τους $|\mathcal{C}|$ ταξινομητές αποφασίζει αν ένα έγγραφο d_j πρέπει να ταξινομηθεί στην αντίστοιχη κατηγορία c_i ή όχι, οπότε το ταξινομεί στην συμπληρωματική της $\bar{c}_i$. Αν ένα έγγραφο d_j ταξινομηθεί στην κατηγορία c_i τότε ονομάζεται θετικό παράδειγμα της κατηγορίας c_i, διαφορετικά αν ταξινομηθεί στην κατηγορία $\bar{c}_i$ ονομάζεται αρνητικό παράδειγμα της c_i.

1.3.2 Ταξινόμηση Κειμένου Οδηγούμενη από μία Κατηγορία και Ταξινόμηση Κειμένου Οδηγούμενη από ένα Έγγραφο

Γενικά υπάρχουν δύο τρόποι σύμφωνα με τους οποίους μπορούμε να χρησιμοποιήσουμε έναν ταξινομητή κειμένου (text classifier). Ανάλογα με τον τρόπο χρήσης οι τεχνικές διακρίνονται σε 2 κατηγορίες, τις οποίες και περιγράφουμε σε αυτήν την ενότητα.

Ας υποθέσουμε ότι δοθέντος ενός εγγράφου $d_j \in \mathcal{D}$, θέλουμε να βρούμε όλες τις κατηγορίες $c_i \in \mathcal{C}$ στις οποίες αυτό ανήκει. Η περίπτωση αυτή ταξινόμησης κειμένου ονομάζεται οδηγούμενη από έγγραφο (document-pivoted classification - DPC). Εναλλακτικά, ας θεωρήσουμε την περίπτωση κατά την οποία δοθείσης μίας κατηγορίας $c_i \in \mathcal{C}$ επιθυμούμε να βρούμε όλα τα έγγραφα $d_j \in \mathcal{D}$ που ανήκουν σε αυτήν, οπότε μιλάμε για ταξινόμηση κειμένου οδηγούμενη από κατηγορία (Class-pivoted classification - CPC). Ο διαχωρισμός αυτός είναι σημαντικός δεδομένου ότι τα σύνολα $\mathcal{C}$ και $\mathcal{D}$ μπορεί να μην είναι διαθέσιμα στο σύνολό τους από την αρχή.

Η οδηγούμενη από ένα έγγραφο ταξινόμηση κειμένου (DPC) είναι κατάλληλη όταν τα έγγραφα γίνονται διαθέσιμα σε διαφορετικές χρονικές στιγμές, εξαιτίας της φύσης της εφαρμογής που χρησιμοποιεί ταξινόμηση κειμένου. Παράδειγμα τέτοιας εφαρμογής αποτελεί το φιλτράρισμα ηλεκτρονικής αλληλογραφίας (e-mail filtering), κατά το οποίο ο χρήστης λαμβάνει ηλεκτρονική αλληλογραφία σε διάφορες χρονικές στιγμές και επιθυμεί να την διαχωρίσει σε θεματικές ενότητες ανάλογα με το περιεχόμενό της. Αντίθετα, η οδηγούμενη από μία κατηγορία ταξινόμηση κειμένου (CPC) χρησιμοποιείται κυρίως όταν επιθυμούμε να προσθέσουμε μία νέα κατηγορία $c_{|\mathcal{C}|+1}$ στο ήδη υπάρχον σύνολο από κατηγορίες $\mathcal{C} = \{c_1, \ldots, c_{|\mathcal{C}|}\}$, οπότε πρέπει να επαναληφθεί ταξινόμηση σε όλα τα έγγραφα λαμβάνοντας υπόψη και τη νέα κατηγορία $c_{|\mathcal{C}|+1}$.

1.3.3 Αυστηρή και Ιεραρχική Ταξινόμηση Κειμένου

Ο διαχωρισμός της ταξινόμησης σε αυστηρή (hard) και ιεραρχική (ranking) έχει να κάνει με την ελαστικότητα του συστήματος στην απόφαση για την ταξινόμηση ενός εγγράφου.

Σύμφωνα με τον ορισμό της ταξινόμησης κειμένου που δόθηκε στην ενότητα 1.2, στόχος αυτής είναι η ανάθεση μίας λογικής τιμής (αληθής ή ψευδής) σε κάθε ζεύγος $\langle d_j, c_i \rangle$, όπου $d_j \in \mathcal{D}$ είναι ένα έγγραφο και $c_i, i = 1, \ldots, C$ μία κατηγορία. Η ταξινόμηση αυτή ειναι γνωστή ως αυστηρή (hard), γιατί η απόφαση σχετικά με την κατηγορία που ανήκει το έγγραφο d_j, βασίζεται αυστηρά στις τιμές των αντίστοιχων ζευγών.

Εναλλακτικά για να αποφασίσουμε σε ποια κατηγορία ανήκει το έγγραφο

d_j μπορούμε να ιεραρχήσουμε τις κατηγορίες του συνόλου $C = \{c_1, \ldots, c_{|C|}\}$ με βάση την καταλληλότητά τους ως προς το έγγραφο d_j. Οπότε, η τελική απόφαση σχετικά με την κατηγορία του εγγράφου d_j λαμβάνεται εξετάζοντας τις πρώτες κατηγορίες με βάση την ιεράρχηση που πραγματοποιήθηκε. Συνήθως υπάρχει κάποιος ειδικός (human expert) που παίρνει την τελική απόφαση. Η παραπάνω διαδικασία ταξινόμησης κειμένου είναι γνωστή ως ιεραρχική (ranking text classification).

Η ιεραρχική ταξινόμηση βρίσκει εφαρμογή όταν η ποιότητα του συνόλου εκπαίδευσης δεν είναι καλή ή όταν τα έγγραφα που χρησιμοποιούνται για την εκπαίδευση του ταξινομητή δεν μπορούν να θεωρηθούν αντιπροσωπευτικό δείγμα των αγνώστων εγγράφων που θα εισέρθουν στον ταξινομητή. Επομένως, είναι επιθυμητή η παρέμβαση ενός ειδικού, ο οποίος θα πάρει την τελική απόφαση για την ταξινόμηση του εγγράφου.

1.4 Εφαρμογές Ταξινόμησης Κειμένου

1.4.1 Οργάνωση Εγγράφων

Τεχνικές ταξινόμησης κειμένου χρησιμοποιούνται σε εφαρμογές οργάνωσης και αρχειοθέτησης εγγράφων. Οι εφαρμογές αυτές στοχεύουν στο να οργανώσουν βάσεις δεδομένων οι οποίες αποτελούνται από έγγραφα κειμένου. Στη συνέχεια θα παρουσιαστούν παραδείγματα τέτοιων εφαρμογών.

Ας θεωρήσουμε ότι στα γραφεία μίας εφημερίδας εισέρχονται αγγελίες προς δημοσίευση. Οι αγγελίες αυτές πρέπει να ταξινομηθούν σε κατηγορίες όπως "Ενοικιάζεται", "Πωλείται", "Ζητείται εργασία" και άλλες. Επομένως, η χρησιμοποίηση ενός ταξινομητή κειμένου θα ωφελούσε σημαντικά την εφημερίδα που δέχεται τις αγγελίες ειδικά όταν το πλήθος αυτών είναι μεγάλο. Μία άλλη πιθανή εφαρμογή για μία εφημερίδα είναι η αρχειοθέτηση των άρθρων, που συντάσσουν οι δημοσιογράφοι της σε κατηγορίες όπως "Πολιτική", "Εσωτερικές ειδήσεις", "Διεθνείς ειδήσεις", "Αθλητισμός", "Χρηματιστήριο" και άλλες.

Μία άλλη σημαντική εφαρμογή που υπάγεται σε αυτήν την κατηγορία είναι η οργάνωση βάσεων δεδομένων από έγγραφα σε κατηγορίες, με στόχο την ευκολότερη και ταχύτερη αναζήτηση τους. Το όφελος που αποκομίζει κάποιος είναι μεγαλύτερο όσο αυξάνεται η πολυπλοκότητα της βάσης των εγγράφων.

1.4.2 Φιλτράρισμα Εγγράφων

Ας υποθέσουμε την ύπαρξη ενός παραγωγού πληροφοριών (producer) ο οποίος μεταδίδει έγγραφα σε έναν καταναλωτή πληροφοριών (consumer). Η ασύγχρονη διαδικασία με την οποία ταξινομούνται τα έγγραφα που αποστέλει ο παραγωγός στον καταναλωτή ονομάζεται φιλτράρισμα κειμένου (text filtering).

Μία τυπική περίπτωση φιλτραρίσματος εγγράφων αφορά την παροχή ειδήσεων (newsfeed), όπου παραγωγός είναι ένα πρακτορείο ειδήσεων και καταναλωτής μία εφημερίδα. Επιθυμούμε ένα σύστημα που να αποτρέπει την παράδοση στον καταναλωτή των εγγράφων – ειδήσεων που δεν τον ενδιαφέρουν. Για παράδειγμα, αν ο καταναλωτής είναι μία αθλητική εφημερίδα θα πρέπει να εμποδίζεται η παράδοση μη αθλητικών ειδήσεων. Επομένως, το φιλτράρισμα κειμένου μπορεί να θεωρηθεί ως μία ειδική περίπτωση απλής (single-label) ταξινόμησης κειμένου με δύο κατηγορίες: την κατηγορία των σχετικών εγγράφων και αυτή των μη σχετικών. Τα έγγραφα που ταξινομούνται στην κατηγορία των σχετικών παραδίδονται στον καταναλωτή, ενώ αυτά που ταξινομούνται στην κατηγορία των μη σχετικών αποδεσμεύονται από τον καταναλωτή. Στη συνέχεια, ο καταναλωτής μπορεί να ταξινομήσει σε θεματικές κατηγορίες τα έγγραφα που του παραδόθηκαν. Έτσι, για το παράδειγμα της αθλητικής εφημερίδας, όλα τα άρθρα που είναι σχετικά με αθλητισμό μπορούν να ταξινομηθούν ως προς το άθλημα στο οποίο αναφέρονται. Με τον τρόπο αυτό διευκολύνονται οι δημοσιογράφοι στο να έχουν πρόσβαση μόνο στα άρθρα που τους ενδιαφέρουν.

Μία άλλη περίπτωση φιλτραρίσματος εγγράφων αποτελεί το φιλτράρισμα ηλεκτρονικής αλληλογραφίας (e-mail filter) [16, 17], κατά την οποία αποδεσμεύεται η αλληλογραφία που δεν ενδιαφέρει το χρήστη. Το χρήσιμο τμήμα της αλληλογραφίας, δηλαδή εκείνο που δεν αποδεσμεύτηκε, μπορεί στη συνέχεια να διαχωριστεί σε κατηγορίες ανάλογα με τα ενδιαφέροντα του χρήστη που χρησιμοποιεί το φίλτρο.

1.4.3 Ιεραρχική Κατηγοριοποίηση Ιστοσελίδων

Η ιεραρχική κατηγοριοποίηση ιστοσελίδων είναι μία πολύ σημαντική εφαρμογή της ταξινόμησης κειμένου. Ουσιαστικά πρόκειται για εφαρμογή της ταξινόμησης κειμένου σε ιστοσελίδες, οι οποίες αποτελούν ειδική μορφή εγγράφων. Η πολυτιμότητά της οφείλεται σε μεγάλο βαθμό στην ευρεία διάδοση του διαδικτύου (internet).

Η ιεραρχική κατηγοριοποίηση ιστοσελίδων έχει δύο σημαντικές ιδιομορφίες σε σχέση με μία τυπική εφαρμογή ταξινόμησης κειμένου.

- Η πρώτη ιδιομορφία έχει να κάνει με την ύπαρξη συνδέσμων (links) σε

μία ιστοσελίδα [20]. Ένας σύνδεσμος (link) μεταξύ δύο ιστοσελίδων αποτελεί ένδειξη ότι τα περιεχόμενα των ιστοσελίδων αυτών συσχετίζονται. Το γεγονός αυτό αποτελεί πλούσια πηγή πληροφορίας για τη διαδικασία της ταξινόμησης μίας ιστοσελίδας που περιέχει συνδέσμους προς άλλες ιστοσελίδες. Υπάρχουν τεχνικές που εκμεταλεύονται την πιθανή συσχέτιση μεταξύ των ιστοσελίδων που συνδέονται και οι οποίες διευκολύνουν τη διαδικασία της ταξινόμησης μίας ιστοσελίδας.

- Η δεύτερη ιδιομορφία αφορά την ιεραρχική δομή του συνόλου των κατηγοριών [26]. Σε αντίθεση με ένα τυπικό πρόβλημα ταξινόμησης κειμένου όπου διαθέτουμε ένα σύνολο από κατηγορίες, εδώ κάθε κατηγορία μπορεί να διαχωρισθεί επιμέρους σε ένα σύνολο από υποκατηγορίες και ούτω καθεξής. Με τον τρόπο αυτό το αρχικό πρόβλημα της ταξινόμησης των ιστοσελίδων ανάγεται σε ένα πλήθος από μικρότερα αντίστοιχα προβλήματα ταξινόμησης.

Χρήση της παραπάνω εφαρμογής κάνουν οι μηχανές αναζήτησης του διαδικτύου (web search engines), όπως είναι το Yahoo! (http://www.yahoo.com). Το Yahoo! οργανώνει τις ιστοσελίδες σε θεματικές κατηγορίες όπως "Business and Economy", "Computers and Internet", "Entertainment", "Education" και άλλες. Οι κατηγορίες αυτές στη συνέχεια χωρίζονται σε υποκατηγορίες δημιουργώντας κατά αυτόν τον τρόπο μία ιεραρχία. Για παράδειγμα η κατηγορία "Computers and Internet" χωρίζεται σε υποκατηγορίες όπως "Internet", "WWW", "Software", "Games" και άλλες. Ο χρήστης του Yahoo! μπορεί να διασχίσει την ιεραρχία των κατηγοριών μέχρι να φτάσει στην κατηγορία που τον ενδιαφέρει και να υλοποιήσει την αναζήτησή του στις ιστοσελίδες που ανήκουν στη συγκεκριμένη κατηγορία. Ο χρήστης του Yahoo! κερδίζει σε ταχύτητα γιατί η αναζήτησή του εφαρμόζεται σε μικρότερο πλήθος ιστοσελίδων, σε ποιότητα αποτελεσμάτων γιατί ελαχιστοποιείται η εμφάνιση ιστοσελίδων όχι τόσο σχετικών με την αναζήτησή του και τέλος, σε ευκολία μέσω της διάσχισης της ιεραρχίας των κατηγοριών.

1.4.4 Εννοιολογική Αποσαφήνιση Λέξεων

Γενικά υπάρχουν λέξεις με περισσότερες από μία έννοιες. Οι λέξεις αυτές ονομάζονται πολύσημες (polysemous). Για να είναι σε θέση κάποιος να εντοπίσει την έννοια μίας πολύσημης λέξης κοιτάζει τα συμφραζόμενα και την ερμηνεύει δίνοντας της την κατάλληλη ερμηνεία με βάση τις λέξεις που την περιβάλλουν. Ορίζουμε ως εννοιολογική αποσαφήνιση μίας λέξης (word sense disambiguation) τη διαδικασία εύρεσης της έννοιας (από ένα πλήθος εννοιών

1.5 Αναπαράσταση Κειμένου

Ας θεωρήσουμε ένα κείμενο $d_j \in \mathcal{D}$, όπου $\mathcal{D}$ ένα σύνολο από κείμενα. Το κείμενο αυτό έχει μορφή που επιτρέπει στον άνθρωπο να το αναγνώσει και να επεξεργαστεί τα περιεχόμενά του. Όμως, η μορφή αυτή δεν είναι κατάλληλη για έναν ταξινομητή. Για τον λόγο αυτό είναι απαραίτητο να εφαρμοστεί μία διαδικασία αντιστοίχισης (indexing) του κειμένου d_j σε μία συμπαγή αναπαράσταση του περιεχομένου του, η οποία να είναι κατανοητή για τον ταξινομητή κειμένου που θα χρησιμοποιηθεί.

Υποθέτουμε ότι κάθε κείμενο αποτελείται από μονάδες οι οποίες ονομάζονται όροι (terms). Όροι μπορεί να είναι λέξεις ή φράσεις, ανάλογα με το τί είναι σημαντικό για τον ταξινομητή. Η διάσπαση του κειμένου σε όρους γίνεται για να μπορέσουμε να αναπαραστήσουμε το κείμενο ως διάνυσμα, όπου κάθε συνιστώσα του διανύσματος αντιστοιχεί σε έναν όρο.

Ένα κείμενο d_j αναπαρίσταται συνήθως ως διάνυσμα από βάρη όρων

$$d_j = \langle w_{1j}, \ldots, w_{|\mathcal{T}|j} \rangle, \tag{1.1}$$

όπου $\mathcal{T}$ είναι το σύνολο των όρων που εμφανίζονται τουλάχιστον μία φορά σε τουλάχιστον ένα έγγραφο του συνόλου εγγράφων $\mathcal{D}$, και w_{kj} είναι το βάρος (weight) του όρου t_k στο έγγραφο d_j. Η απόδοση βαρών γίνεται για να δωθεί έμφαση στο γεγονός ότι κάποιοι όροι είναι πιο σημαντικοί από άλλους όρους για ένα συγκεκριμένο έγγραφο.

Υπάρχουν διαφορετικές προσεγγίσεις στην αναπαράσταση ενός εγγράφου - κειμένου, οι οποίες οφείλονται

1. στους διαφορετικούς τρόπους ορισμού των όρων, δηλαδή των δομικών μονάδων ενός κειμένου,

2. στους διαφορετικούς τρόπους υπολογισμού των βαρών των όρων (term weights).

Όσον αφορά την πρώτη περίπτωση, ως όρος είθισται να θεωρείται μία λέξη. Ωστόσο, ένας όρος μπορεί να είναι μία φράση, δηλαδή ένα σύνολο από λέξεις το οποίο είτε εμφανίζεται αρκετά συχνά στο σύνολο των εγγράφων $\mathcal{D}$ (στατιστικός ορισμός της φράσης), είτε έχει επιλεχθεί με βάση τη γραμματική της γλώσσας (συντακτικός ορισμός της φράσης). Οι μέθοδοι ταξινόμησης κειμένου, οι οποίες θα παρουσιαστούν στα επόμενα κεφάλαια, χρησιμοποιούν την πρώτη προσέγγιση, θεωρούν δηλαδή ως όρους λέξεις.

Όσον αφορά τη δεύτερη περίπτωση, έστω w_{kj} το βάρος του όρου t_k στο έγγραφο d_j. Οι πιο διαδεδομένοι τρόποι υπολογισμού του βάρους w_{kj} είναι οι ακόλουθοι:

- *Υπολογισμός των βαρών με βάση την παρουσία ή την απουσία των όρων στο κείμενο:* Τα βάρη που προκύπτουν με χρήση αυτής της προσέγγισης είναι δυαδικά (binary).
  - $w_{kj} = 1$, αν ο όρος t_k υπάρχει στο έγγραφο d_j.
  - $w_{kj} = 0$, αν ο όρος t_k απουσιάζει από το έγγραφο d_j.
- *Υπολογισμός των βαρών με βάση της συχνότητας εμφάνισης των όρων στο κείμενο:* Το βάρος w_{kj} ισούται με το πλήθος των εμφανίσεων του όρου t_k στο έγγραφο d_j.

$$w_{kj} = \#(t_k, d_j) \tag{1.2}$$

- *Υπολογισμός των βαρών με βάση την συνάρτηση tfidf(term frequency inverse document frequency weighting):* Το βάρος w_{kj} υπολογίζεται με χρήση της συνάρτησης tfidf, η οποία ορίζεται ως εξής:

$$w_{kj} = tfidf(t_k, d_j) = \#(t_k, d_j) \cdot \log \frac{|\mathcal{D}|}{\#_{\mathcal{D}}(t_k)} \tag{1.3}$$

όπου $\#(t_k, d_j)$ είναι το πλήθος των εμφανίσεων του όρου t_k στο έγγραφο d_j, $|\mathcal{D}|$ είναι ο πληθικός αριθμός του συνόλου εγγράφων $\mathcal{D}$, και $\#_{\mathcal{D}}(t_k)$ είναι το πλήθος των εγγράφων του $\mathcal{D}$ στα οποία εμφανίζεται ο όρος t_k.

Η tfidf συνάρτηση μας εξασφαλίζει ότι όσο πιο συχνά εμφανίζεται ένας όρος σε ένα έγγραφο, τόσο πιο αντιπροσωπευτική είναι η τιμή που του προσδίδει και ότι σε όσο περισσότερα έγγραφα εμφανίζεται τόσο λιγότερο διαχωρίσιμος γίνεται. Η εξίσωση 1.3 είναι μία από τις πολλές παραλλαγές της συνάρτησης tfidf που υπάρχουν στη βιβλιογραφία.

Υπάρχουν τεχνικές ταξινόμησης κειμένου που απαιτούν τα βάρη w_{kj} να παίρνουν τιμές στο διάστημα [0 , 1]. Για το λόγο αυτό οι τιμές των βαρών που προκύπτουν από την εφαρμογή της συνάρτησης tfidf κανονικοποιούνται. Η κανονικοποίηση που χρησιμοποιείται συνήθως είναι η συνημιτοειδής (cosine), η οποία δίνεται από τη σχέση:

$$w_{kj} = \frac{tfidf(t_k, d_j)}{\sqrt{\sum_{s=1}^{|\mathcal{D}|}(tfidf(t_s, d_j))^2}} \tag{1.4}$$

Στο σημείο αυτό πρέπει να τονιστεί ότι πριν από τη διαδικασία αντιστοίχισης των εγγράφων d_j σε διανύσματα από βάρη όρων, πραγματοποιείται ένα

στάδιο προεπεξεργασίας των εγγράφων. Το στάδιο αυτό χωρίζεται σε φάσεις και έχει ως στόχο να μειώσει τον αριθμό των όρων που χρησιμοποιούνται για την αναπαράσταση ενός εγγράφου. Οι κυριότερες φάσεις του είναι οι ακόλουθες:

- Αφαίρεση συνηθισμένων λέξεων: πρόκειται για λέξεις οι οποίες δεν παρέχουν κάποια ιδιαίτερη πληροφορία για το περιέχόμενο των εγγράφων όπως άρθρα, προθέσεις, σύνδεσμοι κτλ. Η φάση αυτή λαμβάνει χώρα σχεδόν πάντα.

- Επιλογή της μορφολογικής ρίζας για κάθε λέξη (stemming): για κάθε λέξη του εγγράφου κρατάμε μόνο τη ρίζα με αποτέλεσμα τη μείωση του χώρου των όρων. Για παράδειγμα ας θεωρήσουμε τις λέξεις "test", "tests", "testing", "tested" οι οποίες έχουν ως ρίζα τη λέξη "test". Τότε, με εφαρμογή αυτής της φάσης διατηρούμε 1 όρο αντί για 4, μειώνοντας κατά πολύ τη διάσταση του χώρου των όρων. Ο αλγόριθμος για την επιλογή της ρίζας που χρησιμοποίηθηκε στην εργασία μας είναι αυτός του Porter [13].

- Επιλογή των όρων από όλο το κείμενο ή μέρος αυτού: ανάλογα με τη μορφή των εγγράφων κάποιος μπορεί να επιλέξει κάποιο τμήμα των εγγράφων από το οποίο θα προκύψουν οι όροι. Για παράδειγμα, στο σύνολο εγγράφων "20 Newsgroups", κάθε έγγραφο ξεκινάει με επικεφαλίδες (headers) και στη συνέχεια περιέχει το κείμενο που αποτελεί ουσιαστικά το περιεχόμενό του. Επομένως είναι επιθυμητό να επιλεχθούν οι όροι από το τμήμα του εγγράφου που προκύπτει με αφαίρεση των επικεφαλίδων που βρίσκονται στην αρχή των εγγράφων.

1.6 Μείωση της Διάστασης των Προτύπων Κειμένου

1.6.1 Το Πρόβλημα της Μεγάλης Διάστασης

Σε προβλήματα ταξινόμησης κειμένου ο χώρος των όρων έχει μεγάλη διάσταση. Το γεγονός αυτό μπορεί να προκαλέσει προβλήματα στη διαδικασία της ταξινόμησης, τα οποία οφείλονται στο φαινόμενο το οποίο είναι γνωστό ως "κατάρα της διάστασης" (curse of dimensionality). Σύμφωνα με το φαινόμενο αυτό, οι εφαρμογές όπου ο χώρος προτύπων έχει μεγάλη διάσταση, απαιτούν μεγάλο πλήθος από πρότυπα για την εκπαίδευσή τους. Επειδή στην πράξη η ποσότητα των δεδομένων που αποτελούν το σύνολο εκπαίδευσης είναι περιορισμένη, πρέπει να μειωθεί η διάσταση των προτύπων, στην περίπτωση της ταξινόμησης κειμένου ο χώρος των όρων.

Ορίζουμε ως μείωση της διάστασης (dimensionality reduction) των προτύπων κειμένου, τη διαδικασία με την οποία ο χώρος των προτύπων κειμένου (ο οποίος ισούται με τον χώρο των όρων) μειώνεται από $|\mathcal{T}|$ σε $|\mathcal{T}'|$, έτσι ώστε $|\mathcal{T}'| \ll |\mathcal{T}|$.

Η μείωση της διάστασης είναι επίσης ωφέλιμη γιατί περιορίζει το φαινόμενο της υπερεκπαίδευσης (overfitting), σύμφωνα με το οποίο ένας ταξινομητής παρουσιάζει μικρό σφάλμα στο σύνολο εκπαίδευσης και μεγάλο σφάλμα σε άγνωστα δεδομένα. Πειράματα έχουν δείξει ότι για την αποφυγή της υπερεκπαίδευσης απαιτείται ένας αριθμός από πρότυπα εκπαίδευσης ανάλογος με τον αριθμό των όρων που χρησιμοποιούνται. Επιπλέον, το υπολογιστικό κόστος της τεχνικής ταξινόμησης η οποία θα χρησιμοποιηθεί ελαττώνεται σημαντικά ως αποτέλεσμα της μείωσης.

Ωστόσο, η μείωση της διάστασης εγκυμονεί τον κίνδυνο να απωλεσθούν πολύτιμοι όροι για τη σημασία του κειμένου. Για το λόγο αυτό η διαδικασία της μείωσης της διάστασης πρέπει να εκτελείται με μεγάλη προσοχή.

Η μείωση της διάστασης διακρίνεται σε τοπική και καθολική, ανάλογα με το αν εφαρμόζεται τοπικά για κάθε κατηγορία ξεχωριστά ή καθολικά για όλες τις κατηγορίες αντίστοιχα:

- τοπική (local): για κάθε κατηγορία $c_i, i = 1, \ldots, |\mathcal{C}|$ επιλέγεται ένα σύνολο όρων $\mathcal{T}_i'$, τέτοιο ώστε $|\mathcal{T}_i'| \ll |\mathcal{T}|$, το οποίο χρησιμοποιείται για την ταξινόμηση των εγγράφων στην κατηγορία c_i. Επομένως, αν θεωρήσουμε ένα έγγραφο d_j τότε η αναπαράστασή του, μετά από εφαρμογή τοπικής μείωσης της διάστασής του, είναι διαφορετική για κάθε κατηγορία c_i, διότι για κάθε κατηγορία χρησιμοποιείται διαφορετικό υποσύνολο όρων $\mathcal{T}_i'$ ως προς το αρχικό σύνολο $\mathcal{T}$.

- καθολική (global): ένα σύνολο όρων $\mathcal{T}'$, τέτοιο ώστε $|\mathcal{T}'| \ll |\mathcal{T}|$, χρησιμοποιείται για την ταξινόμηση των εγγράφων σε όλες τις κατηγορίες $\mathcal{C} = \{c_1, \ldots, c_{|\mathcal{C}|}\}$.

Μία δεύτερη κατηγοριοποίηση έχει να κάνει με τη φύση των όρων που προκύπτουν μετά από την εφαρμογή μείωσης της διάστασης του χώρου των όρων $\mathcal{T}$ στο χώρο $\mathcal{T}'$. Η διαδικασία μείωσης της διάστασης διακρίνεται σε:

- μείωση της διάστασης με επιλογή όρων: ο χώρος των όρων $\mathcal{T}'$ είναι υποσύνολο του χώρου $\mathcal{T}$.

- μείωση της διάστασης με εξαγωγή όρων: οι όροι που συγκροτούν το σύνολο $\mathcal{T}'$ δεν έχουν τον ίδιο τύπο με τους όρους του συνόλου $\mathcal{T}$. Προκύπτουν με συνδυασμούς ή μετασχηματισμούς των αρχικών όρων.

Για παράδειγμα, οι όροι του συνόλου $\mathcal{T}$ είναι λέξεις ενώ, οι όροι του συνόλου $\mathcal{T}'$ μπορεί να είναι ομάδες από λέξεις.

Στις επόμενες ενότητες παρουσιάζουμε τεχνικές με τις οποίες μπορεί κάποιος να υλοποιήσει μείωση της διάστασης με επιλογή ή με εξαγωγή όρων.

1.6.2 Μείωση της Διάστασης με Επιλογή Όρων

Η τεχνική της επιλογής όρων (term selection) επιχειρεί να επιλέξει, από το αρχικό σύνολο όρων $\mathcal{T}$, το σύνολο των όρων $\mathcal{T}'$ για το οποίο ισχύει $|\mathcal{T}'| \ll |\mathcal{T}|$ και το οποίο αποφέρει την καλύτερη απόδοση ως προς την αναπαράσταση των εγγράφων. Υπάρχουν δύο προσεγγίσεις με τις οποίες μπορεί να γίνει η επιλογή των όρων και οι οποίες είναι γνωστές ως προσέγγιση του περιτυλίγματος (wrapper approach) και προσέγγιση του φιλτραρίσματος (filtering approach).

Στην πρώτη προσέγγιση, το σύνολο των όρων $\mathcal{T}'$ το οποίο προκύπτει μετά τη μείωση, χρησιμοποιείται κατά την κατασκευή του ταξινομητή (classifier). Ξεκινώντας από ένα αρχικό σύνολο όρων $\mathcal{T}$ δημιουργούμε ένα νέο σύνολο $\mathcal{T}'$ με προσθήκη ή αφαίρεση όρων. Στη συνέχεια, κατασκευάζουμε τον ταξινομητή βασιζόμενοι στο σύνολο $\mathcal{T}'$ και ελέγχουμε την απόδοσή του. Η διαδικασία αυτή επαναλαμβάνεται για ένα πλήθος από διαφορετικά σύνολα όρων, τα οποία προκύπτουν με κατάλληλη επιλογή. Τελικά, επιλέγουμε το σύνολο των όρων $\mathcal{T}'$ για το οποίο ο ταξινομητής επιτυγχάνει την καλύτερη απόδοση. Η προσέγγιση αυτή ονομάζεται προσέγγιση του περιτυλίγματος (wrapper approach) διότι, η επιλογή των όρων είναι άρρηκτα συνυφασμένη με την μεθοδο ταξινόμησης που χρησιμοποιούμε. Το πλεονέκτημα της έγκειται στο ότι είναι προσαρμοσμένη με τη μέθοδο ταξινόμησης κειμένου, ωστόσο παρουσιάζει ως μειονέκτημα το απαγορευτικό κόστος δεδομένου ότι ο χώρος των διαφορετικών συνόλων από όρους είναι πολύ μεγάλος.

Στη δεύτερη προσέγγιση, διατηρούμε το σύνολο των όρων $\mathcal{T}'$ για το οποίο ισχύει $|\mathcal{T}'| \ll |\mathcal{T}|$ και το οποίο προκύπτει με εφαρμογή μιας συνάρτησης για κάθε όρο του αρχικού συνόλου $\mathcal{T}$. Η συνάρτηση αυτή μετράει την "πολυτιμότητα" του όρου που δέχεται ως όρισμα για τη διαδικασία της ταξινόμησης. Ουσιαστικά λειτουργεί ως φίλτρο το οποίο επιτρέπει να περάσουν οι πολυτιμότεροι όροι. Για το λόγο αυτό και ονομάζεται προσέγγιση του φιλτραρίσματος (filtering approach). Σε αντίθεση με πριν, το υπολογιστικό κόστος είναι σαφώς μικρότερο. Στη συνέχεια παρουσιάζονται οι πιο διαδομένες συναρτήσεις που χρησιμοποιούνται σε αυτήν την προσέγγιση.

Συχνότητα Εμφάνισης του Όρου στο Έγγραφο:

Η συνάρτηση συχνότητας εμφάνισης (document frequency) ενός όρου t_k σε ένα σύνολο εγγράφων $\mathcal{D}$ συμβολίζεται με $\#_{\mathcal{D}}(t_k)$ και αντιστοιχεί στο πλήθος των εγγράφων του $\mathcal{D}$ στα οποία εμφανίζεται ο όρος t_k. Με τη χρήση της συνάρτησης $\#_{\mathcal{D}}(t_k)$ επιλέγονται οι όροι που εμφανίζονται στα περισσότερα έγγραφα του συνόλου $\mathcal{D}$. Οι Yang και Pedersen [40] έδειξαν με μια σειρά από πειράματα ότι με συνάρτηση φιλτραρίσματος τη $\#_{\mathcal{D}}(t_k)$, είναι δυνατή η μείωση της διάστασης του χώρου των όρων κατά ένα παράγοντα του 10 χωρίς απώλειες στην απόδοση, ενώ με μείωση κατά ένα παράγοντα του 100 επιφέρονται μόλις μικρές απώλειες. Το πλεονέκτημα της $\#_{\mathcal{D}}(t_k)$ είναι η απλοϊκή και αποδοτική της μορφή. Μία άλλη στρατηγική χρήσης της προκύπτει με αφαίρεση των όρων που εμφανίζονται το πολύ σε x έγγραφα, όπου x ένα κατώφλι (threshold) που ορίζουμε εμείς.

Άλλες Συναρτήσεις Επιλογής Όρων:

Εκτός από τη συνάρτηση συχνότητας εμφάνισης (document frequency) υπάρχουν κι άλλες συναρτήσεις με πιο πολύπλοκη μορφή. Οι πιο διαδεδομένες από αυτές είναι:

- Ο Παράγοντας Συσχέτισης DIA (DIA Association Factor)
- Η Συνάρτηση χ^2 (Chi-square)
- Ο Συντελεστής NGL (NGL Coefficient)
- Το Κέρδος Πληροφορίας (Information Gain)
- Η Αμοιβαία Πληροφορία (Mutual Information)
- Ο Λόγος Πιθανοτήτων (Odds Ratio)
- Ο Βαθμός Συσχέτισης (Relevancy Score)
- Ο Συντελεστής GSS (GSS Coefficient)

Οι ορισμοί των παραπάνω συναρτήσεων περιγράφονται στον πίνακα 1.1 και όπως μπορεί να παρατηρήσει κανείς, περιέχουν πιθανότητες στο χώρο των εγγράφων, οι οποίες υπολογίζονται μετρώντας τον αριθμό των εμφανίσεων των όρων στο σύνολο των εγγράφων. Για παράδειγμα, η $P(t_k, c_i)$ εκφράζει την πιθανότητα ο όρος t_k να εμφανίζεται σε ένα τυχαίο έγγραφο x της κατηγορίας c_i, ενώ η $P(\bar{t}_k, c_i)$ την πιθανότητα να μην εμφανίζεται, αντίστοιχα.

Όλες οι συναρτήσεις εφαρμόζονται σε έναν όρο t_k και "τοπικά" σε μία κατηγορία, ωστόσο υπάρχουν περιπτώσεις που επιθυμούμε να αναφερθούμε "καθολικά" στο σύνολο των κατηγοριών. Για το λόγο αυτό χρησιμοποιούμε είτε το άθροισμα $f_{sum}(t_k) = \sum_{i=1}^{|C|} f(t_k, c_i)$, είτε το άθροισμα με βάρη (weighted sum) $f_{wsum}(t_k) = \sum_{i=1}^{|C|} P(c_i) \cdot f(t_k, c_i)$, είτε το μέγιστο $f_{max}(t_k) = \max_{i=1}^{|C|} f(t_k, c_i)$ των τιμών $f(t_k, c_i)$, οι οποίες είναι τοπικά ορισμένες σε μία κατηγορία c_i.

Οι συναρτήσεις αυτές μετράνε την εξάρτηση μεταξύ ενός όρου t_k και μιας κατηγορίας c_i. Έστω f μία από αυτές τις συναρτήσεις. Όσο πιο μεγάλη είναι η τιμή $f(t_k, c_i)$, τόσο πιο εξαρτημένος είναι ο όρος t_k με την κατηγορία c_i. Επομένως, για την κατασκευή του νέου συνόλου όρων $\mathcal{T}'$ επιλέγονται οι όροι με τη μεγαλύτερη εξάρτηση, ή ισοδύναμα με τη μεγαλύτερη τιμή της συνάρτησης.

Έχουν πραγματοποιηθεί πειράματα με στόχο τη σύγκριση των διαφόρων συναρτήσεων μείωσης διάστασης [29, 40]. Τα πειράματα αυτά έδειξαν ότι οι συναρτήσεις που περιέχονται στον πίνακα 1.1 επιτυγχάνουν καλύτερα αποτελέσματα από τη συνάρτηση συχνότητας εμφάνισης ενός όρου στο έγγραφο (document frequency).

1.6.3 Μείωση της Διάστασης με Εξαγωγή Όρων

Κατά τη μείωση της διάστασης με εξαγωγή όρων προσπαθούμε ξεκινώντας από το αρχικό σύνολο όρων $\mathcal{T}$ να παράγουμε ένα σύνολο όρων $\mathcal{T}'$, το οποίο να αποτελείται από τεχνητούς όρους. Το σύνολο $\mathcal{T}'$ πρέπει να μεγιστοποιεί την αποδοτικότητα της αναπαράστασης και να ικανοποιεί τον περιορισμό $|\mathcal{T}'| \ll |\mathcal{T}|$. Η χρησιμοποίηση τεχνητών όρων έχει ως στόχο την επίλυση προβλημάτων πολυσημίας, ομωνυμίας και συνωνυμίας μεταξύ των όρων του αρχικού συνόλου $\mathcal{T}$. Κάθε μέθοδος εξαγωγής αποτελείται από το στάδιο κατά το οποίο εξάγονται οι νέοι όροι από τους όρους του αρχικού συνόλου και από το στάδιο κατά το οποίο τα έγγραφα του συνόλου των εγγράφων $\mathcal{D}$ μετατρέπονται στις νέες αναπαραστάσεις τους με βάση το νέο σύνολο όρων. Δύο μέθοδοι εξαγωγής όρων έχουν χρησιμοποιηθεί σε προβλήματα ταξινόμησης κειμένου, η ομαδοποίηση όρων και η ευρετηριοποίηση της κρυμμένης σημασιολογίας, οι οποίες παρουσιάζονται στη συνέχεια.

Ομαδοποίηση Όρων (Term Clustering):

Η μέθοδος της ομαδοποίησης όρων (term clustering) προσπαθεί να κατασκευάσει ομάδες (clusters) από λέξεις που "συσχετίζονται" σε μεγάλο βαθμό. Έπειτα από τη δημιουργία των ομάδων, παράγεται ένα νέο σύνολο το οποίο περιέχει έναν αντιπροσωπευτικό όρο για κάθε ομάδα λέξεων (για παράδειγμα

Συνάρτηση	Συμβολισμός	Μαθηματικός τύπος
DIA association factor	$z(t_k, c_i)$	$P(c_i\|t_k)$
Information gain	$IG(t_k, c_i)$	$\sum_{c\in\{c_i,\bar{c}_i\}} \sum_{t\in\{t_k,\bar{t}_k\}} P(t,c) \cdot \log \frac{P(t,c)}{P(t)\cdot P(c)}$
Mutual Information	$MI(t_k, c_i)$	$\log \frac{P(t_k,c_i)}{P(t_k)\cdot P(c_i)}$
Chi–square	$\chi^2(t_k, c_i)$	$\frac{\lvert\mathcal{D}\rvert\cdot[P(t_k,c_i)\cdot P(\bar{t}_k,\bar{c}_i)-P(t_k,\bar{c}_i)\cdot P(\bar{t}_k,c_i)]^2}{P(t_k)\cdot P(\bar{t}_k)\cdot P(c_i)\cdot P(\bar{c}_i)}$
NGL coefficient	$NGL(t_k, c_i)$	$\frac{\sqrt{\lvert\mathcal{D}\rvert}\cdot[P(t_k,c_i)\cdot P(\bar{t}_k,\bar{c}_i)-P(t_k,\bar{c}_i)\cdot P(\bar{t}_k,c_i)]}{\sqrt{P(t_k)\cdot P(\bar{t}_k)\cdot P(c_i)\cdot P(\bar{c}_i)}}$
Relevancy score	$RS(t_k, c_i)$	$\log \frac{P(t_k\|c_i)+d}{P(t_k\|\bar{c}_i)+d}$
Odds ratio	$OR(t_k, c_i)$	$\frac{P(t_k\|c_i)\cdot(1-P(t_k\|\bar{c}_i))}{(1-P(t_k\|c_i))\cdot P(t_k\|\bar{c}_i)}$
GSS coefficient	$GSS(t_k, c_i)$	$P(t_k, c_i) \cdot P(\bar{t}_k, \bar{c}_i) - P(t_k, \bar{c}_i) \cdot P(\bar{t}_k, c_i)$

Πίνακας 1.1: Οι βασικές συναρτήσεις που χρησιμοποιούνται για μείωση της διάστασης με επιλογή όρων.

το κέντρο της ομάδας ή μία από τις λέξεις της ομάδας), αντί όλων των λέξεων της κάθε ομάδας. Με τον τρόπο αυτό επιτυγχάνεται μείωση των όρων του αρχικού συνόλου $\mathcal{T}$.

Για τη δημιουργία των ομάδων (clusters) απαιτείται κάποιο μέτρο συσχέτισης μεταξύ των όρων. Συνήθως, το μέτρο αυτό είναι ο βαθμός με τον οποίο οι όροι εμφανίζονται στα ίδια έγγραφα. Η μέθοδος ομαδοποίησης (clustering) που χρησιμοποιεί αυτό το κριτήριο συσχέτισης υπάγεται στην κατηγορία των μεθόδων μάθησης χωρίς επίβλεψη (unsupervised learning), δεδομένου ότι δεν επηρεάζεται από την κατηγορία στην οποία ανήκει κάθε έγγραφο. Ωστόσο, υπάρχουν μέθοδοι ομαδοποίησης που ανήκουν στην κατηγορία των μεθόδων μάθησης με επίβλεψη (supervised learning). Οι μέθοδοι αυτές χρησιμοποιούν ως κριτήριο συσχέτισης το βαθμό με τον οποίο οι όροι εμφανίζονται σε έγγραφα της ίδιας κατηγορίας (ή συνόλου κατηγοριών).

Βασικά πλεονεκτήματα της μεθόδου ομαδοποίησης όρων είναι αφενός η μείωση του θορύβου που επιτυγχάνεται μέσω της αντικατάστασης περιττών όρων από την ομάδα στην οποία ανήκουν και αφετέρου η αύξηση της συχνότητας εμφάνισης των νέων όρων στα έγγραφα.

Ευρετηριοποίηση της Κρυμμένης Σημασιολογίας (Latent Semantic Indexing – LSI):

Η μέθοδος της ευρετηριοποίησης της κρυμμένης σημασιολογίας επιτυγχάνει μείωση της διάστασης των διανυσμάτων κειμένου εφαρμόζοντας ανάλυση ιδιοτιμών (singular value decomposition – SVD) στον πίνακα των εγγράφων, έστω M. Ο πίνακας M έχει διαστάσεις $|\mathcal{T}| \times |\mathcal{D}|$, όπου $\mathcal{T}$ το αρχικό σύνολο των όρων και $\mathcal{D}$ το σύνολο των εγγράφων προς αναπαράσταση. Κάθε στήλη του πίνακα M είναι ένα διάνυσμα κειμένου από το σύνολο εγγράφων $\mathcal{D}$. Η μέθοδος αυτή επιλύει τα προβλήματα που προκύπτουν από την ύπαρξη πολύσημων και συνώνυμων λέξεων. Βασική διαφορά με τις προηγούμενες μεθόδους μείωσης διάστασης είναι ότι παράγει όρους οι οποίοι διαισθητικά δεν έχουν νόημα.

Πλεονέκτημα της μεθόδου είναι ότι μπορεί να συνδυάσει όρους από το αρχικό σύνολο $\mathcal{T}$ οι οποίοι δε συσχετίζονται ιδιαίτερα με κάποια κατηγορία και να παράγει ένα νέο όρο αρκετά πολύτιμο για τη διαδικασία της ταξινόμησης. Αντίστοιχα, μειονέκτημα της μεθόδου είναι ότι όροι από το αρχικό σύνολο $\mathcal{T}$ με ιδιαίτερη συσχέτιση ως προς κάποια κατηγορία μπορεί να χαθούν με τη δημιουργία του νέου συνόλου όρων $\mathcal{T}'$.

if	*((wheat & farm)*	**or**	
	(wheat & commodity)	**or**	
	(bushels & export)	**or**	
	(wheat & tones)	**or**	
	(wheat & winter & ¬ soft))	**then**	WHEAT **else** ¬ WHEAT

Σχήμα 1.1: Κανόνας της μορφής 1.5. Με πλάγια γράμματα παριστάνονται οι λέξεις κλειδιά ενώ, με κεφαλαία οι κατηγορίες.

1.7 Προσεγγίσεις του Προβλήματος της Ταξινόμησης

Στη δεκαετία του 1980 οι τεχνικές που χρησιμοποιούνταν για προβλήματα ταξινόμησης κειμένου υπάγονταν στον κλάδο της μηχανικής γνώσης (knowledge engineering). Από τις αρχές της δεκαετίας του 1990 άρχισαν να γίνονται ιδιαίτερα δημοφιλής τεχνικές ενταγμένες στον κλάδο της μηχανικής μάθησης (machine learning), σε σημείο τέτοιο ώστε σήμερα να αποτελούν το κέντρο του ενδιαφέροντος στην ερευνητική κοινότητα η οποία ασχολείται με τον τομέα της ταξινόμησης κειμένου. Οι δύο αυτές προσεγγίσεις παρουσιάζονται στις ενότητες που ακολουθούν.

1.7.1 Μηχανική Γνώσης (Knowledge Engineering – KE)

Οι τεχνικές ταξινόμησης κειμένου, που ακολουθούν την προσέγγιση της μηχανικής γνώσης (knowledge engineering), κατασκευάζουν χειρωνακτικά ένα έμπειρο σύστημα ικανό να ταξινομεί κείμενα άγνωστης κατηγορίας. Το έμπειρο αυτό σύστημα αποτελείται από ένα σύνολο λογικών κανόνων, οι οποίοι κατασκευάζονται χειρωνακτικά, δηλαδή με τη βοήθεια ειδικών στο χώρο (human experts). Για κάθε κατηγορία καθορίζεται ένα σύνολο από λογικούς κανόνες, της μορφής

$$\textbf{if } <\textit{DNF formula}> \textbf{ then } <\textit{category}>, \tag{1.5}$$

όπου DNF (Disjunctive Normal Form) formula είναι ένας τύπος που περιέχει διαζεύξεις από συζευκτικές προτάσεις. Ένα έγγραφο $d_j \in \mathcal{D}$ ταξινομείται στην κατηγορία <*category*> αν και μόνο αν ικανοποιεί την <*DNF formula*> , δηλαδή αν και μόνο αν ικανοποιεί τουλάχιστον μία από τις συζευκτικές της προτάσεις, δεδομένου ότι διαθέτουμε διαζεύξεις μεταξύ των συζευκτικών προτάσεων.

Παράδειγμα της μηχανικής προσέγγισης αποτελεί το έμπειρο σύστημα CONSTRUE [30], το οποίο αναπτύχθηκε από την ερευνητική ομάδα "Carnegie Group" για λογαριασμό του πρακτορείου ειδήσεων Reuters. Στο σχήμα 1.1 παρουσιάζεται ένα δείγμα κανόνα το οποίο χρησιμοποιήθηκε στο σύστημα CONSTRUE.

Το μειονέκτημα της προσέγγισης είναι γνωστό από το πεδίο των έμπειρων συστημάτων ως συμφόρηση στην απόκτηση γνώσεων (knowledge acquisition bottleneck). Ένας μηχανικός γνώσης (knowledge engineer) σε συνεργασία με έναν ειδικό στο πεδίο ορισμού των εγγράφων προς ταξινόμηση (domain expert) καθορίζουν το σύνολο των κανόνων για κάθε κατηγορία. Σε περίπτωση που τροποποιηθεί κάποια από τις παραμέτρους-συνθήκες του συστήματος, θα πρέπει οι παραπάνω ειδικοί να παρέμβουν για να προσαρμόσουν το υπάρχον έμπειρο σύστημα στα νέα δεδομένα. Ενδεικτικά αναφέρουμε ότι αν ανανεωθεί το σύνολο των κατηγοριών (για παράδειγμα με προσθήκη νέας κατηγορίας), θα πρέπει οι δύο ειδικοί να επέμβουν, ενώ αν ο υπάρχον ταξινομητής εφαρμοσθεί σε διαφορετικό πεδίο (για παράδειγμα σε τελείως διαφορετικό σύνολο κατηγοριών) τότε ένας νέος ειδικός εξειδικευμένος στο διαφορετικό πεδίο (domain expert) πρέπει επίσης να παρέμβει, και σε συνεργασία με το μηχανικό γνώσης (knowledge engineer) να κατασκευάσουν το έμπειρο σύστημα από την αρχή. Όπως γίνεται αντιληπτό, πιθανές τροποποιήσεις επιφέρουν δαπανηρές επιπτώσεις, τόσο σε χρόνο όσο και σε κόστος.

1.7.2 Μηχανική Μάθηση (Machine Learning – ML)

Στην προσέγγιση της μηχανικής μάθησης [9], μία γενική επαγωγική διαδικασία κατασκευάζει αυτόματα έναν ταξινομητή για κάθε κατηγορία $c_i \in C$ "παρατηρώντας" τα χαρακτηριστικά ενός συνόλου εγγράφων, το οποίο έχει προηγουμένως ταξινομηθεί χειρωνακτικά στην κατηγορία c_i από έναν ειδικό στο πεδίο (domain expert). Από αυτά τα χαρακτηριστικά, η επαγωγική διαδικασία συγκεντρώνει εκείνα που πρέπει να διαθέτει ένα άγνωστο έγγραφο d_j για να ταξινομηθεί στην κατηγορία c_i. Με τον τρόπο αυτό, η κατασκευή ταξινομητών κειμένου για το σύνολο των κατηγοριών $C = \{c_1, \ldots, c_{|C|}\}$ μπορεί να μελετηθεί ως $|C|$ ανεξάρτητες εργασίες κατασκευής ενός ταξινομητή για κάθε κατηγορία $c_i \in C$. Καθένας από τους ταξινομητές αποτελεί ένα κανόνα που μας επιτρέπει να αποφασίσουμε αν το άγνωστο έγγραφο d_j πρέπει να ταξινομηθεί στην κατηγορία c_i που αντιπροσωπεύει ο αντίστοιχος ταξινομητής.

Τα πλεονεκτήματα αυτής της προσέγγισης σε σχέση με αυτήν της μηχανικής γνώσης είναι προφανή. Η προσέγγιση της μηχανικής μάθησης κατασκευάζει αυτόματα έναν ταξινομητή για κάθε κατηγορία c_i βασιζόμενη αποκλειστικά στο σύνολο των εγγράφων που ανήκουν στην κατηγορία c_i. Επο-

μένως, αν αναθεωθεί το αρχικό σύνολο των κατηγοριών ή αν επιθυμούμε το σύστημα μας να εφαρμοσθεί σε διαφορετικό πεδίο, το μόνο που χρειάζεται είναι η επαγωγική, αυτόματη κατασκευή ενός ταξινομητή για κάθε κατηγορία με βάση ένα διαφορετικό σύνολο από ταξινομημένα έγγραφα, χωρίς να απαιτείται η παρέμβαση είτε του μηχανικού γνώσης είτε του ειδικού στο πεδίο.

Όσον αφορά την αποδοτικότητα, οι τεχνικές ταξινόμησης που ακολουθούν την προσέγγιση της μηχανικής μάθησης επιτυγχάνουν εντυπωσιακά αποτελέσματα, καθιστώντας τη διαδικασία κατασκευής ταξινομητών κειμένου βιώσιμη τόσο σε χρόνο όσο και σε κόστος σε σχέση με τη χειρωνακτική ταξινόμηση, όπου χρειάζεται η παρέμβαση του ανθρώπινου παράγοντα.

Σύνολο Εκπαίδευσης, Σύνολο Ελέγχου και Σύνολο Επικύρωσης

Η προσέγγιση της μηχανικής μάθησης, όπως έχει τονισθεί, βασίζεται στην ύπαρξη ενός αρχικού συνόλου από έγγραφα $\mathcal{D} = \{\mathrm{d}_1, \ldots, \mathrm{d}_{|\mathcal{D}|}\}$, των οποίων η κατηγορία είναι γνωστή και ανήκει στο σύνολο $\mathcal{C} = \{\mathrm{c}_1, \ldots, \mathrm{c}_{|\mathcal{C}|}\}$. Για να μπορέσουμε να εκτιμήσουμε την αποδοτικότητα ενός ταξινομητή, χωρίζουμε το σύνολο των εγγράφων $\mathcal{D}$ σε δύο υποσύνολα, τα οποία δεν έχουν απαραίτητα το ίδιο μέγεθος:

- στο σύνολο εκπαίδευσης (training set) $TV = \{d_1, \ldots, d_{|TV|}\}$. Οι ταξινομητές κατασκευάζονται επαγωγικά για κάθε μία από τις κατηγορίες του συνόλου $\mathcal{C} = \{\mathrm{c}_1, \ldots, \mathrm{c}_{|\mathcal{C}|}\}$, παρατηρώντας τα χαρακτηριστικά των εγγράφων του συνόλου εκπαίδευσης.

- στο σύνολο ελέγχου (test set) $Te = \{d_{|TV|+1}, \ldots, d_{|\mathcal{D}|}\}$. Πρόκειται για το σύνολο των εγγράφων το οποίο χρησιμοποιείται για τον έλεγχο της αποδοτικότητας των ταξινομητών. Κάθε έγγραφο $d_j \in Te$ τροφοδοτείται σε καθέναν από τους $|\mathcal{C}|$ ταξινομητές ούτως ώστε να αποφασιστεί η κατηγορία του και στη συνέχεια η απόφαση αυτή των ταξινομητών συγκρίνεται με την πραγματική κατηγορία του εγγράφου $d_j, j \in \{|TV| + 1, \ldots, |\mathcal{D}|\}$, η οποία είναι γνωστή. Για τη μέτρηση της αποδοτικότητας του συστήματος βασιζόμαστε στο πόσο συχνά οι αποφάσεις του ταιριάζουν με τις πραγματικές κατηγορίες των εγγράφων του συνόλου ελέγχου Te.

Στο σημείο αυτό πρέπει να τονιστεί ότι τα έγγραφα του συνόλου ελέγχου Te δεν πρέπει με κανέναν τρόπο να συμμετέχουν κατά τη διάρκεια της επαγωγικής κατασκευής των ταξινομητών. Σε αντίθετη περίπτωση, τα πειραματικά αποτελέσματα που προκύπτουν είναι πλασματικά διότι το σύστημα θα έχει

εκπαιδευτεί χρησιμοποιώντας αυτά τα έγγραφα κι επομένως δεν του είναι άγνωστα.

Συχνά επιθυμούμε να βελτιστοποιήσουμε τις εσωτερικές παραμέτρους ενός ταξινομητή, δηλαδή να προσαρμόσουμε τις τιμές των παραμέτρων ώστε ο ταξινομητής να επιτυγχάνει τη μέγιστη απόδοση. Για το λόγο αυτό, το σύνολο $\{d_1, \ldots, d_{|TV|}\}$ μπορεί επιπλέον να χωριστεί στο σύνολο εκπαίδευσης (training set) $T_r = \{d_1, \ldots, d_{|T_r|}\}$, με βάση το οποίο θα κατασκευαστεί ο ταξινομητής και στο σύνολο επικύρωσης (validation set) $V_a = \{d_{|T_r|+1}, \ldots, d_{|TV|}\}$, πάνω στο οποίο θα επαναλαμβάνονται έλεγχοι που στοχεύουν στη βελτιστοποίηση των παραμέτρων του ταξινομητή.

Ενδεικτικά, αναφέρουμε ότι υπάρχουν κι άλλες στρατηγικές διαχωρισμού του αρχικού συνόλου των εγγράφων. Μία από αυτές είναι και η k-fold cross validation.

1.8 Μέθοδοι Ταξινόμησης Κειμένου

Η επαγωγική κατασκευή ταξινομητών κειμένου έχει αντιμετωπιστεί με πλήθος μεθόδων, εκ των οποίων οι σημαντικότερες πρόκειται να παρουσιαστούν σε αυτήν την ενότητα. Στόχος αυτών των μεθόδων είναι να μπορούν να αποφασίσουν την κατηγορία ενός άγνωστου εγγράφου d_j. Αυτό επιτυγχάνεται με τον ορισμό μιας συνάρτησης κατηγοριοποίησης CSV (Categorization Status Value). Ο ορισμός της συνάρτησης κατηγοριοποίησης διαφέρει ανάλογα με το αν επιθυμούμε να κατασκευάσουμε έναν αυστηρό (hard) ή έναν ιεραρχικό (ranking) ταξινομητή, όπως αυτοί ορίστηκαν στην ενότητα 1.3.3 .

Στην περίπτωση της ιεραρχικής ταξινόμησης κειμένου, η κατασκευή ενός ταξινομητή συνίσταται στον ορισμό της συνάρτησης κατηγοριοποίησης $CSV_i : \mathcal{D} \rightarrow [0, 1]$ για κάθε κατηγορία $c_i \in \mathcal{C}, \mathcal{C} = \{c_1, \ldots, c_{|\mathcal{C}|}\}$. Η συνάρτηση CSV για την κατηγορία c_i, CSV_i, παίρνει ως είσοδο ένα έγγραφο $d_j \in \mathcal{D}$ και επιστρέφει μία τιμή μεταξύ 0 και 1, η οποία εκφράζει κατά πόσο το έγγραφο d_j ανήκει στην κατηγορία c_i. Όσο μεγαλύτερη είναι η τιμή που επιστρέφει η συνάρηση κατηγοριοποίησης, τόσο αυξάνεται η πεποίθηση ότι το έγγραφο d_j ανήκει στην κατηγορία c_i. Τα έγγραφα ιεραρχούνται λαμβάνοντας υπόψη τις τιμές που επιστρέφει η CSV_i για κάθε κατηγορία c_i και στη συνέχεια κάποιος ειδικός (human expert) παίρνει την τελική απόφαση ταξινόμησης με βάση την ιεράρχηση που πραγματοποιήθηκε.

Στην περίπτωση της αυστηρής ταξινόμησης, η συνάρτηση κατηγοριοποίησης για την κατηγορία c_i, CSV_i, ορίζεται όπως στην ιεραρχική ταξινόμηση με τη διαφορά ότι ορίζεται επιπλέον ένα κατώφλι (threshold) τ_i πάνω από το οποίο θεωρούμε ότι ένα έγγραφο ανήκει στην κατηγορία c_i. Εναλλακτικά,

μπορεί να οριστεί ως $CSV_i : \mathcal{D} \rightarrow \{\mathrm{T}, \mathrm{F}\}$, δηλαδή να επιστρέφει την τιμή "αληθής" (True) όταν το έγγραφο ανήκει στην κατηγορία c_i ή την τιμή "ψευδής" (False) στην αντίθετη περίπτωση.

Ένα θέμα που εγείρεται είναι ο καθορισμός του κατωφλίου τ_i. Οι πολιτικές οι οποίες έχουν αναπτυχθεί διακρίνονται σε αναλυτικές και σε πειραματικές. Οι αναλυτικές εφαρμόζονται όταν υπάρχει κάποιο θεωρητικό αποτέλεσμα που να υποδεικνύει τον τρόπο υπολογισμού του κατωφλίου, ώστε να μεγιστοποιείται η απόδοση του συστήματος. Τυπικό παράδειγμα αποτελεί η πιθανοτική πολιτική καθορισμού κατωφλίου (probability thresholding). Οι πειραματικές πολιτικές δοκιμάζουν διάφορες τιμές για το κατώφλι τ_i και επιλέγουν τελικά την τιμή εκείνη που ικανοποιεί κάποιο κριτήριο. Οι πιο γνωστές πολιτικές είναι η CSV ή αλλιώς Scut, η αναλογική (proportional) ή αλλιώς Pcut και η καθορισμένη (fixed) ή αλλιώς Rcut.

Στις ενότητες που ακολουθούν παρουσιάζονται οι δημοφιλέστερες μέθοδοι ταξινόμησης κειμένου και ο τρόπος με τον οποίο ορίζεται η συνάρτηση κατηγοριοποίησης CSV_i σε κάθε μία από αυτές.

1.8.1 Στατιστικές Μέθοδοι

Οι στατιστικές μέθοδοι προσπαθούν να αναδείξουν την πιθανοτική φύση του προβλήματος της ταξινόμησης κειμένου. Ο αντίστοιχος τομέας της στατιστικής αναγνώρισης προτύπων είναι ο παλιότερος και καλύτερα θεμελιωμένος και βασίζεται σε έννοιες της θεωρίας στατιστικής και πιθανοτήτων.

Ας υποθέσουμε ότι έχουμε στη διάθεσή μας ένα σύνολο εγγράφων $\mathcal{D} = \{\mathrm{d}_1, \ldots, \mathrm{d}_{|\mathcal{D}|}\}$ το καθένα εκ των οποίων ανήκει σε μία από τις κατηγορίες του συνόλου $\mathcal{C} = \{\mathrm{c}_1, \ldots, \mathrm{c}_{|\mathcal{C}|}\}$. Κάθε έγγραφο είναι ένα διάνυσμα το οποίο έχει τη μορφή που δίνεται από τη σχέση 1.1. Επιδιώκουμε να κατασκευάσουμε ένα σύστημα ταξινόμησης το οποίο θα αποφασίζει σε ποια κατηγορία ανήκει ένα άγνωστο έγγραφο (αναγνώριση της κατηγορίας από την οποία προέρχεται). Υποθέτουμε ότι υπάρχει μία εκ των προτέρων (prior) πιθανότητα $P(c_i)$ σύμφωνα με την οποία ένα τυχαία επιλεγμένο έγγραφο του συνόλου $\mathcal{D}$ ανήκει στην κατηγορία c_i. Οι εκ των προτέρων (prior) πιθανότητες εκφράζουν την εκ των προτέρων γνώση ή πίστη μας για το ποια είναι η κατηγορία του εγγράφου, προτού αυτό εμφανιστεί στο σύστημα ταξινόμησης. Ένας απλός τρόπος για να υπολογίσουμε τις prior πιθανότητες $P(c_i), i = 1, \ldots, |\mathcal{C}|$ είναι να απαριθμήσουμε για το σύνολο $\mathcal{D}$ τα έγγραφα που ανήκουν σε κάθε κατηγορία και να εκφράσουμε έπειτα τις prior πιθανότητες με βάση το ακόλουθο κλάσμα:

$$P(c_i) = \frac{\text{αριθμός εγγράφων της κατηγορίας } c_i}{\text{αριθμός εγγράφων συνόλου } \mathcal{D}} \tag{1.6}$$

Προφανώς υπάρχουν κάποια τμήματα του χώρου των εγγράφων από τα οποία είναι πιθανότερο να προέρθουν έγγραφα της κατηγορίας c_1, επίσης κάποια αντίστοιχα τμήματα από τα οποία είναι πιθανότερο να προέρθουν έγγραφα της κατηγορίας c_2 και ούτω καθεξής. Αυτό σημαίνει ότι για ένα άγνωστο έγγραφο d_j εκφράζονται $|C|$ τιμές πιθανότητας έτσι ώστε η κάθε μία να προσδιορίζει το βαθμό "πίστης" μας το συγκεκριμένο έγγραφο d_j να ανήκει στην αντίστοιχη κατηγορία. Εισάγουμε την υπό συνθήκη συνάρτηση πυκνότητας πιθανότητας (probability density function - pdf) $p(d_j \mid c_i), i = 1, \ldots, |C|$ για κάθε κατηγορία έτσι ώστε να εκφράζει την κατανομή των εγγράφων που ανήκουν στην κατηγορία c_i.

Για ένα άγνωστο έγγραφο d_j που έχει καταφθάσει στο σύστημα, η πιθανότητα να ανήκει στην κατηγορία $c_i, i = 1, \ldots, |C|$ δίνεται από το **θεώρημα του Bayes** :

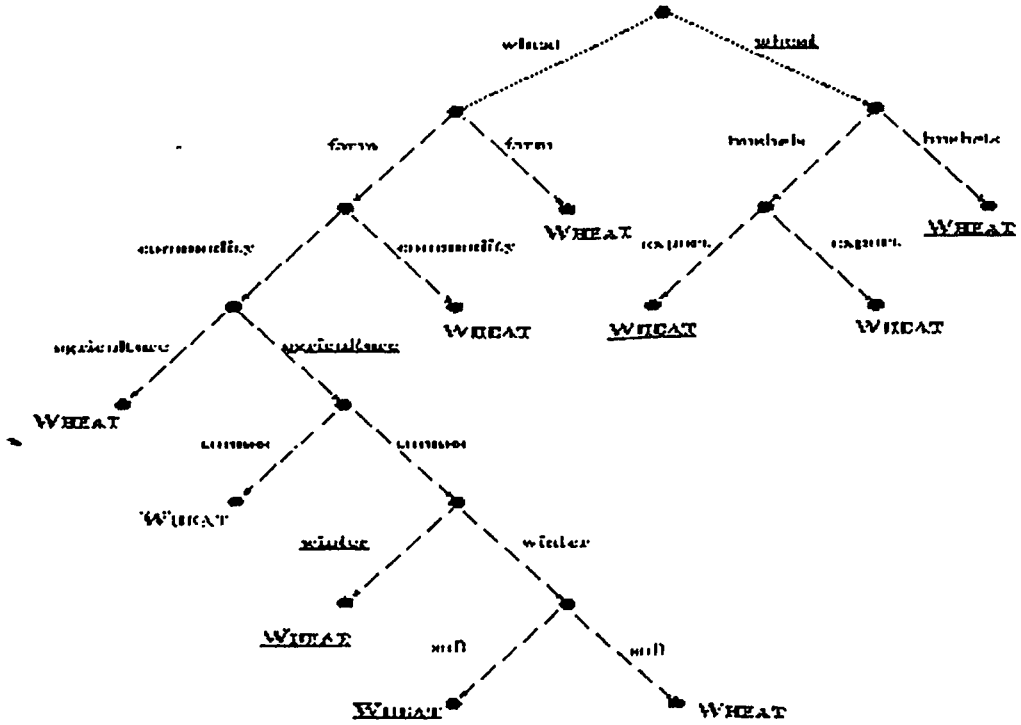
$$P(c_i \mid d_j) = \frac{P(c_i)p(d_j \mid c_i)}{p(d_j)} \tag{1.7}$$

όπου

$$p(d_j) = \sum_{i=1}^{|C|} P(c_i)p(d_j \mid c_i) \tag{1.8}$$

Η πιθανότητα $P(c_i \mid d_j)$ αντιπροσωπεύει την εκ των υστέρων (posterior) "πίστη" μας σχετικά με το ποια είναι η κατηγορία του εγγράφου d_j. Για να ταξινομήσουμε ένα άγνωστο έγγραφο d_j εφαρμόζουμε τον **κανόνα απόφασης του Bayes** , σύμφωνα με τον οποίο επιλέγουμε την κατηγορία με τη μεγαλύτερη εκ των υστέρων (posterior) πιθανότητα. Στο σημείο αυτό πρέπει να τονιστεί ότι όπως έχει αποδειχθεί, η διαδικασία απόφασης με βάση τον κανόνα του Bayes ελαχιστοποιεί την πιθανότητα λανθασμένης απόφασης.

Ωστόσο, η εκτίμηση της συνάρτησης πυκνότητας πιθανότητας $p(d_i \mid c_j)$ παρουσιάζει προβλήματα, δεδομένου ότι το πλήθος των πιθανών διανυσμάτων d_j είναι εξαιρετικά μεγάλο. Για το λόγο αυτό έχει επικρατήσει στην πλειοψηφία των στατιστικών ταξινομητών κειμένου να γίνεται η υπόθεση ότι οι συνιστώσες του διανύσματος που αναπαριστά ένα έγγραφο είναι στατιστικά ανεξάρτητες μεταξύ τους κι επομένως η συνάρτηση πυκνότητας πιθανότητας $p(d_i \mid c_j)$ μπορεί να υπολογιστεί ως:



Σχήμα 1.2: Δέντρο απόφασης ισοδύναμο με τον DNF κανόνα του σχήματος 1.1 . Οι ακμές του δέντρου έχουν ετικέτες όρους και τα φύλλα κατηγορίες. Η υπογράμμιση μίας κατηγορίας δηλώνει άρνηση (¬).

$$p(d_j \mid c_i) = \prod_{k=1}^{|T|} P(w_{kj} \mid c_i) \tag{1.9}$$

Οι κατανομές που ορίζονται με βάση την εξίσωση 1.9 είναι γνωστές ως πολυωνυμικές (multinomial) [15]. Οι στατιστικοί ταξινομητές που κάνουν χρήση της παραπάνω υπόθεσης ονομάζονται "αφελείς" (naive), διότι η τελευταία συνήθως δεν ανταποκρίνεται στην πραγματικότητα.

Στην περίπτωση των στατιστικών ταξινομητών κειμένου, η συνάρτηση κατηγοριοποίησης $CSV_i(d_j)$ για το έγγραφο d_j ορίζεται ως η εκ των υστέρων πιθανότητα το έγγραφο d_j να ανήκει στην κατηγορία c_i, $P(c_i \mid d_j)$.

1.8.2 Δέντρα Απόφασης

Ένα δέντρο απόφασης (decision tree) [9] είναι ένα δέντρο του οποίου οι εσωτερικοί κόμβοι έχουν ετικέτα κάποιον όρο, οι κλάδοι που ξεκινούν από έναν

εσωτερικό κόμβο αντιστοιχούν σε ελέγχους στην τιμή-βάρος που έχει ο όρος κατά την αναπαράσταση ενός εγγράφου του συνόλου ελέγχου και τα φύλλα είναι κατηγορίες. Παράδειγμα δέντρου απόφασης παρουσιάζεται στο σχήμα 1.2 .

Η κατασκευή του ταξινομητή κειμένου συνίσταται στη δημιουργία του δέντρου απόφασης. Για να καθορίσουμε την κατηγορία ενός αγνώστου εγγράφου d_j, δηλαδή ενός εγγράφου του συνόλου ελέγχου (test set) Te, αρκεί να διασχίσουμε το δέντρο απόφασης. Ξεκινώντας από τη ρίζα, ελέγχουμε αναδρομικά τις τιμές–βάρη που έχουν οι όροι, οι οποίοι αντιπροσωπεύουν τους εσωτερικούς κόμβους, στην αναπαράσταση του εγγράφου d_j και ανάλογα μεταβαίνουμε στο κατάλληλο παιδί. Η διαδικασία αυτή επαναλαμβάνεται μέχρι να φτάσουμε σε φύλλο, οπότε ταξινομούμε το έγγραφο στην κατηγορία που αντιπροσωπεύει το φύλλο αυτό. Οι περισσότεροι ταξινομητές αυτής της μορφής, χρησιμοποιούν δυαδική αναπαράσταση για τα έγγραφα (βλ. ενότητα 1.5) κι επομένως τα δέντρα απόφασης που προκύπτουν είναι δυαδικά δεδομένου ότι κάθε όρος (εσωτερικός κόμβος) μπορεί να πάρει δύο τιμές (0 ή 1).

Ας υποθέσουμε το δέντρο απόφασης του σχήματος 1.2, το οποίο χρησιμοποιεί δυαδική αναπαράσταση εγγράφων. Έστω ένα έγγραφο το οποίο περιέχει τους όρους *wheat* και *farm* . Προκειμένου να το ταξινομήσουμε ξεκινάμε από τη ρίζα του δέντρου κι ελέγχουμε την παρουσία (τιμή 1) ή απουσία (τιμή 0) του όρου *wheat* στο έγγραφο. Δεδομένου ότι ο όρος υπάρχει στο έγγραφο, μεταβαίνουμε στο αριστερό παιδί και κατόπιν ελέγχουμε με τον ίδιο τρόπο την παρουσία του όρου *farm*. Ομοίως μεταβαίνουμε στο δεξί παιδί, το οποίο είναι φύλλο. Επομένως ταξινομούμε το έγγραφο στην κατηγορία *WHEAT*, η οποία είναι η κατηγορία που αντιστοιχεί στο φύλλο αυτό. Όπως μπορεί κανείς να παρατηρήσει κάθε μονοπάτι (path) του δέντρου αντιστοιχεί σε μία σύζευξη όρων κι επομένως όλο το δέντρο αποτελεί μία διάζευξη από συζεύξεις.

Υπάρχουν διάφοροι αλγόριθμοι για την εκπαίδευση ενός δέντρου εκπαίδευσης, εκ των οποίων οι πιο γνωστοί είναι ο ID3, ο C4.5 και ο C5. Οι περισσότεροι αλγόριθμοι που έχουν αναπτυχθεί για την εκπαίδευση ενός δέντρου απόφασης είναι παραλλαγές ενός κεντρικού (core) αλγορίθμου, ο οποίος υλοποιεί από-πάνω-προς-τα-κάτω λάιμαργη αναζήτηση (top–down greedy search) στο χώρο των πιθανών δέντρων απόφασης. Τα βήματα του κεντρικού αλγορίθμου για την εκπαίδευση του δέντρου απόφασης για την κατηγορία c_i είναι τα ακόλουθα:

1. Ανέθεσε όλα τα έγγραφα του συνόλου εκπαίδευσης (training set - T_r) στη ρίζα του δέντρου.

2. Αν όλα τα έγγραφα στον τρέχων κόμβο έχουν την ίδια ετικέτα (δηλαδή ανήκουν στην ίδια κατηγορία, c_i ή $\bar{c}_i$), τερμάτισε και χαρακτήρισε τον

κόμβο με αυτήν την ετικέτα–κατηγορία. Διαφορετικά επέλεξε έναν όρο t_k με βάση κάποιο κριτήριο. Τα κριτήρια που χρησιμοποιούνται σύνήθως είναι το κέρδος (gain) και η εντροπία (entropy).

3. Χώρισε το σύνολο εκπαίδευσης σε υποσύνολα από έγγραφα σύμφωνα με την τιμή του όρου t_k, έτσι ώστε κάθε υποσύνολο να περιέχει έγγραφα που έχουν την ίδια τιμή για τον όρο t_k.

4. Δημιούργησε έναν κόμβο παιδί για κάθε πιθανή τιμή του όρου t_k και επανέλαβε αναδρομικά τη διαδικασία που ξεκινάει από το βήμα 2 για κάθε κόμβο παιδί, μέχρις ότου κάθε φύλλο του δέντρου να περιέχει έγγραφα της ίδιας κατηγορίας.

Ένα δέντρο απόφασης που παράγεται με βάση τον παραπάνω αλγόριθμο συχνά παρουσιάζει προβλήματα υπερεκπαίδευσης (overfitting), τα οποία οφείλονται στο γεγονός ότι αν επιτρέψουμε την πλήρη ανάπτυξη του δέντρου πολλές φορές προκύπτουν κλάδοι που αντιστοιχούν σε συγκεκριμένα έγγραφα του συνόλου εκπαίδευσης. Για το λόγο αυτό, μετά την εκπαίδευσή του δέντρου απόφασης εφαρμόζεται μία διαδικασία κλαδέματος (pruning) των κλάδων που δημιουργούν προβλήματα υπερεκπαίδευσης, η οποία αποσκοπεί στην αποφυγή του φαινομένου της υπερεκπαίδευσης.

1.8.3 Κανόνες Απόφασης

Η μέθοδος των κανόνων απόφασης (decision rules) [18] κατασκευάζει έναν κανόνα σε DNF μορφή για κάθε κατηγορία $c_i, i = 1, \ldots, |\mathcal{C}|$. Παράδειγμα κανόνα απόφασης παρουσιάζεται στο σχήμα 1.1 . Ένας DNF κανόνας αποτελείται από διαζεύξεις μεταξύ συζευκτικών προτάσεων. Κάθε συζευκτική πρόταση περιέχει τις λέξεις οι οποίες πρέπει να υπάρχουν ή να απουσιάζουν από ένα έγγραφο d_j, ώστε να ταξινομηθεί στην κατηγορία c_i. Ένα άγνωστο έγγραφο d_j ταξινομείται στην κατηγορία c_i αν και μόνο αν ικανοποιεί τουλάχιστον μία από τις συζευκτικές προτάσεις του DNF κανόνα απόφασης που αντιστοιχεί στην κατηγορία c_i.

Ας θεωρήσουμε τον κανόνα απόφασης για την κατηγορία "WHEAT" του σχήματος 1.1 . Έστω ένα έγγραφο το οποίο περιέχει τις λέξεις *"wheat"* και *"winter"* αλλά δεν περιέχει τη λέξη *"soft"* . Για να αποφασίσουμε την κατηγορία του, εξετάζουμε αν ικανοποιείται τουλάχιστον μία από τις συζευκτικές προτάσεις του DNF κανόνα. Όπως εύκολα μπορεί να παρατηρήσει κανείς ικανοποιείται η τελευταία κι επομένως το έγγραφο ταξινομείται στην κατηγορία WHEAT.

Η στρατηγική με την οποία εκπαιδεύουμε ένα κανόνα μάθησης για μία κατηγορία c_i είναι η εξής: θεωρούμε ότι κάθε έγγραφο d_j του συνόλου εκπαίδευσης (training set – Te) αποτελεί μία συζευκτική πρόταση, η οποία είναι της μορφής

$$term_1 \& term_2 \& \ldots \& term_n \rightarrow category \tag{1.10}$$

όπου $term_1, \ldots, term_n$ είναι οι όροι που περιέχει το έγγραφο d_j και category είναι η κατηγορία του, δηλαδή c_i αν πρόκειται για θετικό παράδειγμα ή $\bar{c}_i$ αν πρόκειται για αρνητικό παράδειγμα της c_i. Το σύνολο των συζευκτικών προτάσεων το οποίο προκύπτει με εφαρμογή της παραπάνω αναπαράστασης για όλα τα έγγραφα του συνόλου εκπαίδευσης αποτελεί έναν κανόνα απόφασης. Ο κανόνας αυτός ωστόσο παρουσιάζει προβλήματα υπερεκπαίδευσης γιατί είναι εξειδικευμένος στα έγγραφα του συνόλου εκπαίδευσης. Για το λόγο αυτό προσπαθούμε να τον γενικεύσουμε μέσα από μία σειρά από τροποποιήσεις, όπως συγχώνευση μεταξύ δύο συζευκτικών προτάσεων, αφαίρεση όρων από μία συζευκτική πρόταση κ.α. Οι τροποποιήσεις αυτές γίνονται με κάποιο κριτήριο ελαχιστοποίησης, το οποίο εξαρτάται από τον εκάστοτε αλγόριθμο εκπαίδευσης, και κατά τέτοιο τρόπο ώστε ο τελικός ταξινομητής για την κατηγορία c_i να είναι σε θέση να ταξινομεί σωστά όλα τα έγγραφα του συνόλου εκπαίδευσης, όπως και ο αρχικός κανόνας. Η διαδικασία κατασκευής του κανόνα απόφασης ολοκληρώνεται με την εφαρμογή "κλαδέματος" (pruning) στις συζευκτικές προτάσεις του, όπως ακριβώς στα δέντρα απόφασης (decision trees).

1.8.4 Προσεγγιστικές Μέθοδοι

Οι προσεγγιστικές μέθοδοι (regression methods) προσπαθούν να προσεγγίσουν μία πραγματική συνάρτηση στόχο $\breve{\Phi} : \mathcal{D} \times \mathcal{C} \rightarrow [0, 1]$ για το πρόβλημα της ταξινόμησης κειμένου μέσω μίας συνάρτησης ταξινομητή $\Phi : \mathcal{D} \times \mathcal{C} \rightarrow [0, 1]$. Ο ταξινομητής που κατασκευάζουν οι μέθοδοι αυτές διαφέρει από τον ορισμό της ενότητας 1.2 στο γεγονός ότι ο πρώτος επιστρέφει μία τιμή στο διάστημα $[0, 1]$.

Τυπικό παράδειγμα των προσεγγιστικών μεθόδων αποτελεί η μέθοδος LLSF (Linear Least-Squares Fit) [41]. Ας θεωρήσουμε ένα έγγραφο d_j κι ας συμβολίσουμε με $I(d_j)$ το διάνυσμα με το οποίο αναπαρίσταται, δηλαδή $I(d_j) = \langle w_{1j}, \ldots, w_{|\mathcal{T}|j} \rangle$. Στόχος της μεθόδου είναι να παράγει ένα διάνυσμα εξόδου για ένα άγνωστο έγγραφο d_j με τόσες συνιστώσες όσες και οι κατηγορίες, $O(d_j) = \langle c_{1j}, \ldots, c_{|\mathcal{C}|j} \rangle$ δοθέντος του διανύσματος $I(d_j)$. Οι τιμές των συνιστωσών του διανύσματος εξόδου καθορίζουν την κατηγορία του εγ-

γράφου. Επομένως η κατασκευή του ταξινομητή κειμένου ισοδυναμεί με τον υπολογισμό ενός $|\mathcal{C}| \times |\mathcal{T}|$ πίνακα $\hat{M}$ τέτοιου ώστε $\hat{M}I(d_j) = O(d_j)$.

Η μέθοδος LLSF υπολογίζει τον $\hat{M}$ ως τον πίνακα εκείνο που ελαχιστοποιεί το σφάλμα ταξινόμησης στο σύνολο εκπαίδευσης, δηλαδή

$$\hat{M} = argmin_M \|MI - O\|_F \tag{1.11}$$

όπου

- $\|V\|_F = \sqrt{\sum_{i=1}^{|\mathcal{C}|} \sum_{j=1}^{|\mathcal{T}|} v_{ij}^2}$ είναι η Frobenius νόρμα ενός $|\mathcal{C}| \times |\mathcal{T}|$ πίνακα V.
- I είναι ένας $|\mathcal{T}| \times |\mathrm{T_r}|$ πίνακας, ο οποίος έχει στήλες τα διανύσματα $I(d_j), d_j \in T_r$ αναπαράστασης των εγγράφων του συνόλου εκπαίδευσης T_r.
- O είναι ένας $|\mathcal{C}| \times |\mathrm{T_r}|$ πίνακας, ο οποίος έχει στήλες τα διανύσματα εξόδου $O(d_j), d_j \in T_r$ που αντιστοιχούν στα έγγραφα του συνόλου εκπαίδευσης T_r.

Ο πίνακας $\hat{M}$ υπολογίζεται συνήθως με ανάλυση ιδιοτιμών (singular value decomposition) στο σύνολο εκπαίδευσης. Το στοιχείο m_{ik} του πίνακα $\hat{M}$ αναπαριστά το βαθμό συσχέτισης μεταξύ της κατηγορίας c_i και του όρου t_k.

Η μέθοδος LLSF κατασκευάζει αρκετά αποδοτικούς ταξινομητές, ωστόσο έχει υψηλό κόστος, το οποίο οφείλεται στον υπολογισμό του πίνακα $\hat{M}$.

1.8.5 Γραμμικές Μέθοδοι

Ας θεωρήσουμε ένα έγγραφο $d_j = \langle w_{1j}, \ldots, w_{|\mathcal{T}|j} \rangle$. Οι γραμμικές (linear) μέθοδοι ταξινόμησης κατασκευάζουν για κάθε κατηγορία $c_i, i = 1, \ldots, |\mathcal{C}|$ ένα διάνυσμα $y_i = \langle y_{1i}, \ldots, y_{|\mathcal{T}|i} \rangle$, το οποίο έχει την ίδια διάσταση με το διάνυσμα αναπαράστασης του εγγράφου d_j. Τα διάνυσμα y_i ονομάζεται προφίλ (profile) της κατηγορίας c_i, ενώ ο επαγόμενος ταξινομητής γραμμικός (linear) ή βασισμένος σε προφίλ (profile–based).

Στην περίπτωση των γραμμικών μεθόδων, η συνάρτηση κατηγοριοποίησης (CSV) της κατηγορίας c_i για το έγγραφο d_j, $CSV_i(d_j)$, ορίζεται ως το εσωτερικό γινόμενο των διανυσμάτων d_j και c_i, δηλαδή

$$CSV_i(d_j) = \sum_{k=1}^{|\mathcal{T}|} y_{ki} \cdot w_{kj}, \tag{1.12}$$

Για να εξετάσουμε αν ένα έγγραφο d_j ταξινομείται στην κατηγορία c_i αρκεί να υπολογίσουμε την τιμή κατηγοριοποίησης $CSV_i(d_j)$ και στη συνέχεια να εφαρμόσουμε τη μεθοδολογία που περιγράφεται στην αρχή της ενότητας 1.8 .

Οι διάφορες γραμμικές μέθοδοι διαφέρουν στον τρόπο υπολογισμού των διανυσμάτων – προφίλ y_i των κατηγοριών. Μπορούν να διαχωριστούν σε δύο μεγάλες κλάσεις, στις ομαδικές και στις σειριακές μεθόδους:

- Οι ομαδικές (batch) μέθοδοι κατασκευάζουν ένα προφίλ χρησιμοποιώντας όλα τα έγγραφα του συνόλου εκπαίδευσης σε ένα βήμα. Τυπικό παράδειγμα μεθόδου αυτής της κλάσης αποτελεί η μέθοδος Rocchio, η οποία θα αναλυθεί σε επόμενη ενότητα.

- Οι σειριακές (on–line) ή αυξητικές (incremental) μέθοδοι κατασκευάζουν ένα προφίλ από το πρώτο κι όλας έγγραφο του συνόλου εκπαίδευσης που επεξεργάζονται και το βελτιώνουν σταδιακά καθώς εξετάζουν νέα έγγραφα από το σύνολο εκπαίδευσης.

  Παράδειγμα αυξητικής μεθόδου είναι το perceptron [24], το οποίο αποτελεί το απλούστερο νευρωνικό δίκτυο. Τα νευρωνικά δίκτυα ως μέθοδοι κατασκευής ταξινομητών κειμένου πρόκειται να μελετηθούν σε επόμενη ενότητα. Για κάθε κατηγορία c_i υλοποιούμε ένα perceptron, το οποίο θα αντιστοιχεί στον ταξινομητή κειμένου της c_i. Το perceptron έχει $|\mathcal{T}|$ εισόδους, όσες και η διάσταση ενός εγγράφου, και μία έξοδο, της οποίας η τιμή καθορίζει την κατηγορία του εγγράφου. Αν η έξοδος λαμβάνει τιμή 1, τότε το έγγραφο το οποίο εισήχθη στο perceptron ανήκει στην κατηγορία c_i, διαφορετικά αν λαμβάνει τιμή 0 τότε ανήκει στην $\bar{c}_i$. Επομένως, για να εξετάσουμε αν ένα άγνωστο έγγραφο ανήκει στην κατηγορία c_i, αρκεί να το τροφοδοτήσουμε ως είσοδο στο perceptron της c_i και να υπολογίσουμε την έξοδό του. Έχει ως παραμέτρους τα βάρη $w_i, i = 1, \ldots, |\mathcal{T}|$, τα οποία χαρακτηρίζουν την συνεισφορά της αντίστοιχης εισόδου (όρου) στην έξοδο του νευρώνα. Το διάνυσμα των βαρών του perceptron χρησιμοποιείται ως προφίλ της κατηγορίας c_i.

  Η εκπαίδευση του perceptron, δηλαδή ο καθορισμός των βαρών του w_i, μπορεί να γίνει με αυξητικό τρόπο ως εξής: κάθε φορά που εξετάζεται ένα νέο έγγραφο d_j του συνόλου εκπαίδευσης T_r , ταξινομείται με βάση το τρέχον προφίλ (διάνυσμα των βαρών w_i). Αν το αποτέλεσμα της ταξινόμησης είναι σωστό, δεν υπεισέρχεται καμία αλλαγή στις τιμές των βαρών, σε αντίθετη περίπτωση τα βάρη τροποποιούνται. Αν το έγγραφο d_j είναι θετικό παράδειγμα της κατηγορίας c_i τότε τα βάρη w_i, τα οποία αντιστοιχούν σε μη μηδενική είσοδο δηλαδή σε όρο που υπάρχει στο d_j, αυξάνονται κατά μία σταθερή ποσότητα $\alpha > 0$, η οποία ονομάζεται

ρυθμός μάθησης (learning rate). Αν το έγγραφο d_j είναι αρνητικό παράδειγμα της κατηγορίας c_i τότε τα ίδια βάρη μειώνονται κατά α. Επειδή η τροποποίηση των τιμών των βαρών γίνεται μόνο στην περίπτωση όπου έχουμε λάθος στην απόφαση ταξινόμησης, η μέθοδος εκπαίδευσης του perceptron είναι γνωστή ως οδηγούμενη από σφάλμα (mistake-driven).

Άλλες παραλλαγές του perceptron που έχουν μελετηθεί στα πλαίσια του προβλήματος της ταξινόμησης κειμένου είναι οι μέθοδοι Positive Winnow Balanced Winnow.

1.8.6 Η Μέθοδος Rocchio

Η μέθοδος Rocchio [31] κατασκευάζει γραμμικούς ταξινομητές κειμένου. Οι συνιστώσες του προφίλ της κατηγορίας c_i, δηλαδή του διανύσματος y_i, υπολογίζονται με βάση τη σχέση:

$$y_{ki} = \beta \cdot \sum_{\{d_j \in POS_i\}} \frac{w_{kj}}{|POS_i|} - \gamma \cdot \sum_{\{d_j \in NEG_i\}} \frac{w_{kj}}{|NEG_i|} \qquad (1.13)$$

όπου w_{kj} είναι το βάρος του όρου t_k στο έγγραφο d_j, POS_i είναι το σύνολο των εγγράφων του συνόλου εκπαίδευσης T_r τα οποία ανήκουν στην κατηγορία c_i (θετικά - positive παραδείγματα της κατηγορίας c_i) δηλαδή $POS_i = \{d_j \in T_r \mid d_j \in c_i\}$ και NEG_i είναι το σύνολο των εγγράφων του συνόλου εκπαίδευσης T_r τα οποία δεν ανήκουν στην κατηγορία c_i ή ισοδύναμα ανήκουν στην κατηγορία $\bar{c}_i$ (αρνητικά - negative παραδείγματα της κατηγορίας c_i) δηλαδή $NEG_i = \{d_j \in Tr \mid d_j \in \bar{c}_i\}$.

Τα β και γ είναι παράμετροι της μεθόδου και καθορίζουν πόσο σημαντικά είναι τα θετικά ή τα αρνητικά παραδείγματα, αντίστοιχα, για τη μέθοδο. Τυπικές επιλογές για τις παραμέτρους β και γ είναι οι εξής:

- $\beta = 16$ και $\gamma = 4$: δίνεται έμφαση στο ρόλο που διαδραματίζουν για τη μέθοδο τα θετικά παραδείγματα της κατηγορίας c_i και ταυτόχρονα ατονίζει ο αντίστοιχος ρόλος των αρνητικών παραδειγμάτων.

- $\beta = 1$ και $\gamma = 0$: το προφίλ της κατηγορίας c_i υπολογίζεται λαμβάνοντας υπόψη μόνο τα θετικά παραδείγματα της ($\gamma = 0$) και αντιστοιχεί στο κέντρο των θετικών αυτών παραδειγμάτων ($\beta = 1$).

Γενικά, οι ταξινομητές που προκύπτουν με τη μέθοδο Rocchio ανταμοίβουν την εγγύτητα ενός αγνώστου εγγράφου στο κέντρο των θετικών παραδειγμάτων και την απόσταση από το κέντρο των αρνητικών. Τα πλεονεκτήματα της μεθόδου είναι η εύκολη υλοποίηση, η ταχύτητα εκπαίδευσης των ταξινομητών, λόγω της απλής μορφής που έχει ο τύπος υπολογισμού των

συνιστωσών του προφίλ μιας κατηγορίας, και το γεγονός ότι είναι εύκολα κατανοητή από τον άνθρωπο. Αντίστοιχα, το βασικό μειονέκτημα της μεθόδου εστιάζεται στην απόδοση του παραγόμενου ταξινομητή κειμένου. Ο τελευταίος λόγω της γραμμικής του φύσης χωρίζει το χώρο των εγγράφων γραμμικά σε δύο υποχώρους, στον υπόχωρο των θετικών και στον υπόχωρο των αρνητικών παραδειγμάτων – εγγράφων μίας κατηγορίας. Επομένως σε περιπτώσεις όπου τα έγγραφα μίας κατηγορίας είναι διασκορπισμένα σε δύο ή περισσότερες ομάδες (clusters), ο ταξινομητής ταξινομεί λανθασμένα ένα πλήθος από αυτά λόγω του ότι αδυνατεί να διαμερίσει το χώρο των εγγράφων σωστά και αποδοτικά. Προκειμένου να βελτιωθεί η αποδοτικότητα των ταξινομητών που παράγονται με χρήση αυτής της μεθόδου έχουν προταθεί διάφορες τροποποιήσεις.

1.8.7 Νευρωνικά Δίκτυα

Η μέθοδος των νευρωνικών δικτύων (neural networks) [33, 34] κατασκευάζει ένα νευρωνικό δίκτυο, το οποίο είναι προσαρμοσμένο στο πρόβλημα της ταξινόμησης κειμένου. Το νευρωνικό δίκτυο αποτελείται από το επίπεδο εισόδου (input layer), ένα ή περισσότερα "κρυμμένα" επίπεδα (hidden layers) και από το επίπεδο εξόδου (output layer). Το επίπεδο εισόδου περιέχει $|\mathcal{T}|$ νευρώνες (neurons), όπου $\mathcal{T}$ είναι το σύνολο των όρων που χρησιμοποιείται για την αναπαράσταση των εγγράφων. Στο επίπεδο εξόδου υπάρχουν $|\mathcal{C}|$ νευρώνες, όσοι και οι κατηγορίες του προβλήματος. Οι νευρώνες των κρυμμένων επιπέδων συνήθως υλοποιούν σιγμοειδής συναρτήσεις και χαρακτηριζονται από μία πόλωση. Το πλήθος των κρυμμένων επιπέδων καθώς επίσης και το πλήθος των κρυμμένων νευρώνων σε καθένα από τα κρυμμένα επίπεδα επηρεάζει τη γενικευτική ικανότητα του συστήματος ταξινόμησης και συνήθως προκύπτει εμπειρικά έπειτα από εφαρμογή πειραμάτων. Η διασύνδεση μεταξύ των νευρώνων δύο διαδοχικών επιπέδων συνήθως είναι πλήρης, δηλαδή κάθε νευρώνας ενός επιπέδου συνδέεται με όλους τους νευρώνες που υπάρχουν στα γειτονικά του επίπεδα. Κάθε σύνδεση μεταξύ δύο νευρώνων χαρακτηρίζεται από ένα βάρος. Το σύνολο των βαρών των συνδέσεων μεταξύ των νευρώνων καθώς και των πολώσεων των νευρώνων αποτελούν τις παραμέτρους του δικτύου.

Για να ταξινομήσουμε ένα άγνωστο έγγραφο d_j, αρκεί να το τροφοδοτήσουμε στο επίπεδο εισόδου και να υπολογίσουμε τις εξόδους των νευρώνων κάθε επιπέδου μέχρι το επίπεδο εξόδου. Η διαδικασία αυτή ονομάζεται ευθύς υπολογισμός (forward computation) ή αλλιώς πέρασμα προς τα εμπρός (forward pass). Οι έξοδοι των νευρώνων του επιπέδου εξόδου καθορίζουν την κατηγορία στην οποία ταξινομείται το άγνωστο έγγραφο. Ανάλογα με το είδος της ταξινόμησης κειμένου, οι έξοδοι των νευρώνων του επιπέδου εξόδου

παίρνουν δυαδικές (0 ή 1) ή πραγματικές τιμές μέσα στο διάστημα [0, 1]. Στην πρώτη περίπτωση (δυαδική ταξινόμηση), αν η i-έξοδος του δικτύου για το έγγραφο d_j παίρνει τιμή 1 τότε ταξινομείται στην κατηγορία c_i, διαφορετικά στην $\bar{c}_i$. Στη δεύτερη περίπτωση, το έγγραφο ταξινομείται στην κατηγορία που αντιστοιχεί στη μεγαλύτερη έξοδο του δικτύου.

Η εκπαίδευση του νευρωνικού δικτύου, δηλαδή ο καθορισμός των παραμέτρων του, γίνεται με χρήση της μεθόδου οπισθοδρομικής διάδοσης ή αλλιως διάδοσης προς τα πίσω (backpropagation). Σύμφωνα με τη μέθοδο αυτή, για κάθε έγγραφο του συνόλου εκπαίδευσης εφαρμόζεται ευθύς υπολογισμός προκειμένου να υπολογιστούν οι έξοδοι του δικτύου και στη συνέχεια το παραγόμενο σφάλμα διαδίδεται προς τα πίσω με σκοπό την τροποποίηση των βαρών του δικτύου και την ελαχιστοποίηση του σφάλματος.

1.8.8 Μέθοδοι Βασισμένες σε Παραδείγματα

Οι μέθοδοι οι οποίες βασίζονται σε παραδείγματα (example-based ή instance-based ή memory-based methods) δεν κατασκευάζουν κάποιο καθολικό (global) μοντέλο ταξινομητή κειμένου, το οποίο θα είναι σε θέση να ταξινομεί οποιοδήποτε άγνωστο έγγραφο. Αντίθετα, κάθε φορά που εξετάζεται ένα άγνωστο έγγραφο, χρησιμοποιούν όλα τα παραδείγματα-έγγραφα του συνόλου εκπαίδευσης και αποφασίζουν την κατηγορία του. Ανήκουν σε μία ευρύτερη κατηγορία μεθόδων οι οποίες είναι γνωστές ως "τεμπέλικες" (lazy), διότι δεν "κουράζονται" να κατασκευάσουν κάποιον ταξινομητή εκ των προτέρων αλλά περιμένουν την είσοδο ενός αγνώστου εγγράφου (ερώτησης) για να επεξεργαστούν τα έγγραφα του συνόλου εκπαίδευσης και να το ταξινομήσουν. Η φάση της εκπαίδευσης αυτών των μεθόδων συνίσταται μόνο στην αποθήκευση των εγγράφων του συνόλου εκπαίδευσης. Το υπολογιστικό κόστος επιβαρύνει σχεδόν αποκλειστικά τη φάση της ταξινόμησης, όπου χρησιμοποιούνται όλα τα έγγραφα του συνόλου εκπαίδευσης προκειμένου να αποφασιστεί η κατηγορία του αγνώστου εγγράφου.

Η πιο αντιπροσωπευτική από τις μεθόδους που βασίζονται σε παραδείγματα είναι η μέθοδος των Κ κοντινότερων γειτόνων (K-NN, K-Nearest Neighbor) [41]. Η μέθοδος αυτή υποθέτει ότι όλα τα έγγραφα αντιστοιχούν σε σημεία στο χώρο διάστασης $|\mathcal{T}|$, όπου είναι το σύνολο των όρων βάση του οποίου γίνεται η αναπαράσταση των εγγράφων. Οι κοντινότεροι γείτονες ενός εγγράφου ορίζονται με κάποιο μέτρο κοντινότητας-ομοιότητας (για παράδειγμα την ευκλείδια απόσταση).

Ας υποθέσουμε ένα άγνωστο έγγραφο d_j, για το οποίο επιθυμούμε να εξετάσουμε αν ανήκει στην κατηγορία c_i. Τα βήματα που ακολουθεί η μέθοδος των Κ κοντινότερων γειτόνων είναι τα εξής:

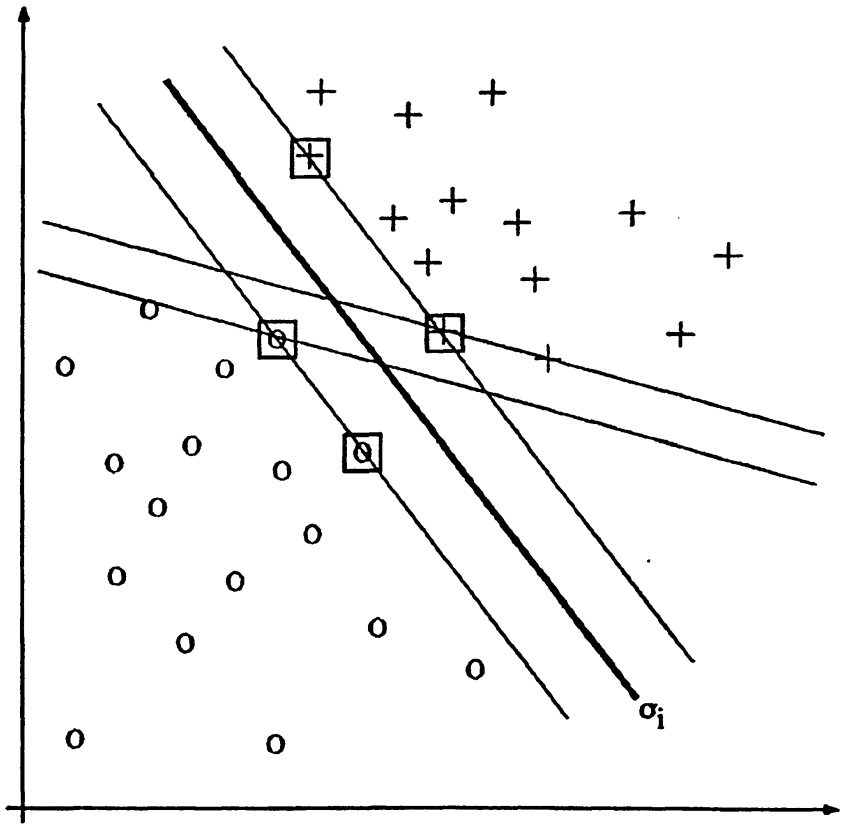
- Εύρεση των Κ κοντινότερων γειτόνων του εγγράφου d_j από τα έγγραφα του συνόλου εκπαίδευσης, με βάση κάποιο προκαθορισμένο μέτρο ομοιότητας. Οι Κ κοντινότεροι γείτονες συγκροτούν τη γειτονιά (neighborhood) του d_j.
- Ψηφοφορία (voting): το έγγραφο d_j ταξινομείται στην πλειοψηφούσα κατηγορία (c_i ή $\bar{c}_i$) ανάμεσα στα Κ κοντινότερα έγγραφα.

Τα πλεονεκτήματα της μεθόδου είναι ότι δεν χρειάζεται εκπαίδευση και επιπλέον είναι απλή στην υλοποίηση (απαιτείται μόνο ο υπολογισμός του μέτρου κοντινότητας μεταξύ δύο εγγράφων). Βασικά μειονεκτήματα της αποτελούν, η εξάρτηση από το Κ (το Κ καθορίζεται πειραματικά), το υψηλό κόστος ταξινόμησης των αγνώστων εγγράφων (δεδομένου ότι σχεδόν όλοι οι υπολογισμοί λαμβάνουν χώρα κατά τη φάση της ταξινόμησης), η χρήση επιπρόσθετης μνήμης για την αποθήκευση και επεξεργασία των εγγράφων του συνόλου εκπάιδευσης (για το λόγο αυτό στην πράξη είθισθαι να εφαρμόζεται κάποια τεχνική ευρετηριοποίησης, όπως το kd-δέντρο του οποίου τα φύλλα αποθηκεύουν κοντινά έγγραφα) και η ευαισθησία στην "κατάρα της μεγάλης διάστασης" (curse of dimensionality) (όταν η διάσταση είναι μεγάλη, υπάρχει μεγάλο πλήθος όρων οι οποίοι δεν έχουν ουσιαστική συνεισφορά στο πρόβλημα της ταξινόμησης, ωστόσο επηρεάζουν την κοντινότητα μεταξύ δύο εγγράφων).

Μία βελτίωση του K–NN είναι η μέθοδος των Κ κοντινότερων γειτόνων με βάρη εξαρτώμενα από την απόσταση (distance–weighted K–NN), κατά την οποία αποδίδεται ένα βάρος σε καθέναν από τους Κ κοντινότερους γείτονες ενός αγνώστου εγγράφου d_j σύμφωνα με την απόστασή τους από το τελευταίο. Η μέθοδος αυτή δίνει μεγαλύτερο βάρος στους γείτονες του d_j που βρίσκονται καντύτερα ως προς αυτό. Για να ταξινομήσουμε το έγγραφο d_j στην κατηγορία c_i ή στην $\bar{c}_i$, αρκεί να αθροίσουμε τα βάρη για κάθε κατηγορία και να βρούμε την κατηγορία με το μεγαλύτερο άθροισμα βαρών.

1.8.9 Η Μέθοδος SVM

Η μέθοδος SVM (Support Vector Machine) [32] κατασκευάζει έναν ταξινομητή για την κατηγορία c_i μέσω της εύρεσης ενός υπερεπιπέδου, το οποίο διαχωρίζει κατά βέλτιστο τρόπο τα θετικά από τα αρνητικά παραδείγματα της c_i. Κάθε πιθανό υπερεπίπεδο για το διαχωρισμό των θετικών από τα αρνητικά παραδείγματα ονομάζεται υπερεπίπεδο απόφασης (decision hyperplane) διότι, βάση αυτού μπορούμε να αποφανθούμε για την κατηγορία ενός άγνωστου εγγράφου d_j.



Σχήμα 1.3: Κατασκευή ταξινομητών κειμένου με χρήση της μεθόδου SVM (Support Vector Machines) σε 2 διαστάσεις. Οι σταυροί αντιστοιχούν στα θετικά παραδείγματα, ενώ οι κύκλοι στα αρνητικά. Οι γραμμές αντιστοιχούν σε υπερεπίπεδα απόφασης. Το υπερεπίπεδο απόφασης σ_i είναι το καλύτερο δυνατό, διότι αποτελεί το μεσαίο στοιχείο του ευρύτερου συνόλου από παράλληλα υπερεπίπεδα. Με τετράγωνα περικλείονται τα παραδείγματα (support vectors) που χρησιμοποιούνται στον καθορισμό των υπερεπιπέδων απόφασης.

Ας θεωρήσουμε όλα τα υπερεπίπεδα απόφασης $\sigma_1, \sigma_2, \ldots$ στον χώρο διάστασης $|\mathcal{T}|$, δηλαδή στον χώρο αναπαράστασης των εγγράφων, τα οποία διαχωρίζουν τα θετικά από τα αρνητικά παραδείγματα του συνόλου εκπαίδευσης. Η μέθοδος SVM προσπαθεί να εντοπίσει το υπερεπίπεδο απόφασης σ_i που επιτρέπει το μεγαλύτερο-ευρύτερο δυνατό περιθώριο μεταξύ των θετικών και αρνητικών παραδειγμάτων. Για να γίνει ευκολότερα αντιληπτή η λογική της μεθόδου, θα μελετηθεί η διδιάστατη περίπτωση του σχήματος 1.3 . Στις δύο διαστάσεις τα υπερεπίπεδα απόφασης είναι ευθείες. Στο παράδειγμα αυτό, το σύνολο των εγγράφων του συνόλου εκπαίδευσης είναι γραμμικά διαχώρισιμο. Η μέθοδος SVM αναζητά το σύνολο των παράλληλων ευθειών με το μεγαλύτερο εύρος (δηλαδή με τη μεγαλύτερη απόσταση μεταξύ δύο ευθειών του συνόλου) κι από αυτό το σύνολο επιλέγει την ευθεία που βρίσκεται στο μέσον. Η εύρεση όλων των πιθανών υπερεπιπέδων απόφασης γίνεται με χρήση ενός μέρους από τα έγγραφα του συνόλου εκπαίδευσης, τα οποία ονομάζονται διανύσματα υποστήριξης (support vectors).

Η μέθοδος SVM αποτελεί μία από τις σημαντικότερες μεθόδους κατασκευής ταξινομητών κειμένου, λόγω των πολύ καλών ιδιοτήτων που διαθέτει. Εφαρμόζεται και στην περίπτωση που το σύνολο των εγγράφων του συνόλου εκπαίδευσης δεν είναι γραμμικά διαχωρίσιμο, σε αντίθεση με τις γραμμικές μεθόδους. Δεν προϋποθέτει τη μείωση της διάστασης των εγγράφων, δεδομένου ότι ανταποκρίνεται εξίσου ικανοποιητικά και σε προβλήματα μεγάλης διάστασης. Επιπλέον, είναι από τις αποδοτικότερες μεθόδους που έχουν μελετηθεί.

1.8.10 Επιτροπές Ταξινομητών Κειμένου

Ας υποθέσουμε ότι θέλουμε να εξετάσουμε αν ένα άγνωστο έγγραφο d_j ταξινομείται στην κατηγορία c_i. Η μέθοδος των επιτροπών από ταξινομητές κειμένου (classifier committees) [36] εφαρμόζει k διαφορετικούς ταξινομητές $\Phi_1, \ldots, \Phi_k$, καθένας από τους οποίους αποφασίζει αν το έγγραφο d_j ανήκει στην κατηγορία c_i, και στη συνέχεια συνδυάζει κατάλληλα τις αποφάσεις τους. Μία επιτροπή ταξινομητών (classifier committee) χαρακτηρίζεται από την επιλογή των k ταξινομητών $\Phi_1, \ldots, \Phi_k$ και από τη στρατηγική με την οποία συνδυάζονται οι αποφάσεις τους προκειμένου να ληφθεί η τελική απόφαση ταξινόμησης.

Οσον αφορά την επιλογή των k ταξινομητών, αυτοί πρέπει να είναι όσο το δυνατόν πιο ανεξάρτητοι προκειμένου η μέθοδος να επιτυγχάνει ικανοποιητικά αποτελέσματα. Οι ταξινομητές μπορεί να διαφέρουν μεταξύ τους είτε στην αναπαράσραση των εγγράφων, είτε στη μέθοδο ταξινόμησης βάση της οποίας κατασκευάζονται είτε και στα δύο.

Οι πιο δημιφιλείς στρατηγικές για το συνδυασμό των k ταξινομητών κειμένου είναι οι ακόλουθες:

- Πλειοψηφική ψηφοφορία (Majority Voting – MV): ένα έγγραφο d_j ταξινομείται στην κατηγορία c_i, όταν η πλειοψηφία των k ταξινομητών $\Phi_1, \ldots, \Phi_k$ το ταξινομεί στη c_i.

- Γραμμικός συνδυασμός με βάρη (Weighted Linear Combination – WLC): η τελική συνάρτηση κατηγοριοποίησης της κατηγορίας c_i για το έγγραφο d_j, $CSV_i(d_j)$, προκύπτει από το ζυγισμένο άθροισμα (weighted sum) των συναρτήσεων κατηγοριοποίησης των k ταξινομητών την κατηγορίας c_i για το d_j, δηλαδή

$$CSV_i(d_j) = \sum_{m=1}^{k} w_m \cdot CSV_{mi}(d_j) \tag{1.14}$$

όπου $CSV_{mi}(d_j)$ είναι η συνάρτηση κατηγοριοποίησης του m–ταξινομητή κειμένου Φ_m της επιτροπής, της κατηγορίας c_i για το έγγραφο d_j και w_m το βάρος που αντιστοιχεί στον m–ταξινομητή Φ_m, $m = 1, \ldots, k$. Το βάρος w_m εκφράζει την αναμενόμενη αποδοτικότητα του ταξινομητή Φ_m σε σχέση με τους υπόλοιπους ταξινομητές. Αυτό σημαίνει ότι ταξινομητές που επιφέρουν καλύτερα αποτελέσματα διαδραματίζουν σημαντικότερο ρόλο στην απόφαση που θα λάβει η επιτροπή. Τα βάρη ικανοποιούν τη σχέση

$$\sum_{m=1}^{k} w_m = 1 \tag{1.15}$$

και υπολογίζονται συνήθως με χρήση των εγγράφων του συνόλου επικύρωσης (validation set).

- Δυναμική επιλογή ταξινομητή (Dynamic Classifier Selection – DCS): για να ταξινομηθεί ένα άγνωστο έγγραφο d_j επιλέγεται ο ταξινομητής Φ_t με τη μεγαλύτερη απόδοση στα l πιο κοντινά έγγραφα του d_j (γειτονιά του d_j). Τα l πιο κοντινά έγγραφα ως προς το d_j εντοπίζονται συνήθως με τον αλγόριθμο K-NN (K–Κοντινότεροι Γείτονες) και προέρχονται από το σύνολο επικύρωσης (validation set). Η επιτροπή των ταξινομητών κειμένου υιοθετεί την απόφαση ταξινόμησης του ταξινομητή Φ_t για το έγγραφο d_j.

- Προσαρμοσμένος συνδυασμός ταξινομητών (Adaptive Classifier Combination – ACC): πρόκειται για τροποποίηση της στρατηγικής του γραμμικού συνδυασμού με βάρη (WLC) με χρήση στοιχείων από τη δυναμική επιλογή ταξινομητή (DCS). Αυτό που διαφοροποιείται σε αυτή την

στρατηγική είναι ο τρόπος υπολογισμού των βαρών w_m. Το βάρος $w_m, m = 1, \ldots, k$ υπολογίζεται με βάση την απόδοση του m–ταξινομητή της επιτροπής στα l πιο κοντινά έγγραφα του συνόλου επικύρωσης ως προς το άγνωστο έγγραφο d_j. Τα l πιο κοντινά έγγραφα εντοπίζονται όπως ακριβώς και στη δυναμική επιλογή ταξινομητή (DCS).

Η μέθοδος των επιτροπών από ταξινομητές κειμένου αποτελεί προφανώς μία αρκετά ακριβή μέθοδο ταξινόμησης, διότι απαιτεί την υλοποίηση k διαφορετικών μεθόδων ταξινόμησης. Για το λόγο αυτό και χρησιμοποιείται μόνο σε περιπτώσεις όπου επιδιώκεται όσο το δυνατόν μεγαλύτερη απόδοση.

1.9 Ανασκόπηση της Εργασίας

Στα επόμενα κεφάλαια της εργασίας μελετάμε στατιστικές μεθόδους για το πρόβλημα της ταξινόμησης κειμένου. Στο δέυτερο κεφάλαιο, περιγράφεται η μέθοδος Naive Bayes, η οποία βασίζεται σε ένα συγκεκριμένο μοντέλο παραγωγής κειμένων, από το οποίο προκύπτουν και οι παράμετροι τις οποίες υπολογίζει. Στο τρίτο κεφάλαιο, περιγράφουμε μία επέκταση του μοντέλου παραγωγής των εγγράφων ώστε να εκμεταλλεύεται την πιθανή ανάλυση μίας κατηγορίας σε θεματικές υποκατηγορίες. Η μέθοδος που παρουσιάζουμε για τον υπολογισμό των παραμέτρων του βελτιωμένου στατιστικού μοντέλου παραγωγής των εγγράφων ονομάζεται Subtopic μέθοδος ταξινόμησης. Χρησιμοποιεί τον αλγόριθμο EM για την εκτίμηση των παραμέτρων του ταξινομητή.

Στα επόμενα δύο κεφάλαια εισάγουμε τις μεθόδους που προτείνουμε ώστε να βελτιώσουμε την απόδοση του Subtopic ταξινομητή. Στο τέταρτο κεφάλαιο προτείνουμε τον Kd–Subtopic ταξινομητή, ο οποίος χρησιμοποιεί μια ειδική δομή δεδομένων γνωστή ως kd–δέντρο, προκειμένου να επιλύσει το πρόβλημα της αρχικοποίησης του αλγορίθμου EM, ο οποίος χρησιμοποιείται από την Subtopic μέθοδο ταξινόμησης. Στο πέμπτο κεφάλαιο εισάγουμε έναν αυξητικό (greedy) αλγόριθμο εκπαίδευσης του Subtopic ταξινομητή προκειμένου να βελτιώσουμε την απόδοσή του.

Ανακεφαλαιώνοντας τις προηγούμενες μεθόδους ταξινόμησης κειμένων, το έκτο κεφάλαιο περιέχει τα πειράματα τα οποία πραγματοποιήθηκαν για τη μελέτη της απόδοσης των επαγόμενων ταξινομητών κειμένου.

Κεφάλαιο 2

Η Μέθοδος Naive Bayes

2.1 Γενικά

Η μέθοδος ταξινόμησης κειμένου Naive Bayes ανήκει στην κατηγορία των στατιστικών μεθόδων. Υποθέτει ένα συγκεκριμένο στατιστικό μοντέλο για την παραγωγή των εγγράφων, το οποίο ενσωματώνει 3 υποθέσεις:

1. τα έγγραφα παράγονται με χρήση μικτού μοντέλου,
2. υπάρχει ένα–προς–ένα αντιστοιχία μεταξύ των μικτών συνιστωσών (πυρήνων) του μικτού μοντέλου και των κατηγοριών του προβλήματος ταξινόμησης,
3. οι μικτές συνιστώσες είναι πολυωνυμικές (multinomial) κατανομές ορισμένες πάνω σε ένα λεξικό που αποτελείται από τις λέξεις των εγγράφων.

Η μέθοδος Naive Bayes κατασκευάζει έναν ταξινομητή κειμένου υπολογίζοντας τις παραμέτρους του στατιστικού μοντέλου παραγωγής των εγγράφων. Στη συνέχεια του κεφαλαίου παρουσιάζεται η διαδικασία παραγωγής των εγγράφων και κατόπιν η μέθοδος ταξινόμησης Naive Bayes, η οποία βασίζεται στο μοντέλο παραγωγής.

2.2 Περιγραφή του Μοντέλου Παραγωγής των Εγγράφων

Θεωρούμε ότι έχουμε στη διάθεση μας ένα σύνολο εγγράφων $\mathcal{D} = \{d_1, \ldots, d_{|\mathcal{D}|}\}$. Υποθέτουμε ότι κάθε έγγραφο $d \in \mathcal{D}$ παράγεται σύμφωνα με μία μικτή

κατανομή με παραμέτρους θ *[υπόθεση 1]* . Μία μικτή κατανομή ορίζεται ως μία ειδική περίπτωση κυρτού γραμμικού συνδυασμού ενός πεπερασμένου αριθμού συναρτήσεων πυκνότητας πιθανότητας που ονομάζονται συνιστώσες ή πυρήνες. Δηλαδή η πιθανότητα μίας τυχαίας μεταβλητής d που ακολουθεί μικτή κατανομή γράφεται ως άθροισμα συναρτήσεων πυκνότητας με βάρη. Η μικτή κατανομή που αποτελείται από $|C|$ τέτοιες συναρτήσεις, όπου $C = \{c_1, \ldots, c_{|C|}\}$, δίνεται από την ακόλουθη σχέση:

$$p(d) = \sum_{j=1}^{|C|} P(c_j)p(d \mid c_j) \tag{2.1}$$

όπου με c_j παριστάνουμε τον πυρήνα j ή αλλιώς τη συστατική συνιστώσα j, ενώ την αντίστοιχη κατανομή $p(d \mid c_j)$ του μικτού μοντέλου την ονομάζουμε συστατική συνάρτηση πυκνότητας πιθανότητας της ολικής κατανομής. Το βάρος $P(c_j)$ αποτελεί την εκ των προτέρων (prior) πιθανότητα σύμφωνα με την οποία η παραγωγή ενός εγγράφου οφείλεται στον συστατικό πυρήνα c_j. Οι prior πιθανότητες υπόκεινται στους εξής περιορισμούς:

$$\sum_{j=1}^{|C|} P(c_j) = 1, P(c_j) \geq 0 \tag{2.2}$$

επίσης οι συναρτήσεις $p(d \mid c_j)$ προφανώς ικανοποιούν τη σχέση:

$$\int p(x \mid c_j)dx = 1 \tag{2.3}$$

Η συστατική κατανομή $p(d \mid c_j)$ εκφράζει την υπό συνθήκη πιθανότητα βάση της οποίας ο πυρήνας c_j παράγει το έγγραφο d. Προκειμένου να παράγουμε ένα έγγραφο που ακολουθεί μικτή κατανομή της μορφής 2.1 επιλέγουμε, καταρχήν, έναν πυρήνα c_j από το σύνολο C των $|C|$ πυρήνων με πιθανότητα $P(c_j)$ και στη συνέχεια παράγουμε το έγγραφο με βάση τη συστατική κατανομή $p(d \mid c_j)$.

Κάθε έγγραφο $d \in \mathcal{D}$ ανήκει σε μία κατηγορία. Υποθέτουμε ότι υπάρχει ένα–προς–ένα αντιστοιχία μεταξύ των μικτών συνιστωσών (πυρήνων) και των κατηγοριών *[υπόθεση 2]* . Επομένως, χρησιμοποιούμε το συμβολισμό c_j για να αναφερόμαστε στον j πυρήνα καθώς επίσης και στην j κατηγορία. Στην περίπτωση αυτή το πρόβλημα αναδιατυπώνεται ως πρόβλημα ταξινόμησης κειμένου σε $|C|$ κατηγορίες, όσες και το πλήθος των πυρήνων που χρησιμοποιείται από τη μικτή κατανομή για την παραγωγή των εγγράφων του συνόλου $\mathcal{D}$. Στη συνέχεια, θα καταλήξουμε με τη βοήθεια επιπλέον υποθέσεων σε μία

ικανοποιητική εκτίμηση των συστατικών κατανομών της μικτής κατανομής για το μοντέλο παραγωγής των εγγράφων.

Ορίζουμε ως λεξικό (vocabulary) $V = \langle w_1, w_2, \ldots, w_{|V|} \rangle$ το σύνολο των λέξεων που χρησιμοποιούνται για την αναπαράσταση των εγγράφων του συνόλου $\mathcal{D}$. Θεωρούμε κάθε έγγραφο d ως μία διατεταγμένη λίστα από λέξεις $\langle w_{d_1}, w_{d_2}, \ldots, w_{d_{|d|}} \rangle$. Συμβολίζουμε με w_{d_k} τη λέξη w_t στη θέση k του εγγράφου d, όπου w_t είναι μία λέξη του λεξικού V. Οπότε η συστατική κατανομή του πυρήνα j γράφεται ως:

$$p(d \mid c_j) = p(\langle w_{d_1}, \ldots, w_{d_{|d|}} \rangle \mid c_j) \tag{2.4}$$

Κατά τη διαδικασία παραγωγής ενός εγγράφου d από έναν πυρήνα c_j, θεωρούμε ότι επιλέγεται πρώτα το μήκος $|d|$ (αριθμός λέξεων) που θα έχει το έγγραφο d. Στη συνέχεια ο πυρήνας c_j παράγει $|d|$ λέξεις (όσες και το μήκος του εγγράφου που επιλέχθηκε) με βάση τη συστατική του κατανομή $p(d \mid c_j)$. Υποθέτουμε ότι κάθε λέξη παράγεται ανεξάρτητα από το μήκος του εγγράφου. Επειδή το έγγραφο d θεωρείται ως διατεταγμένη λίστα από λέξεις, η παραγωγή μίας νέας λέξης εξαρτάται από όλες τις λέξεις που έχουν δημιουργηθεί προηγουμένως. Επομένως, ισχύει

$$p(\langle w_{d_1}, \ldots, w_{d_{|d|}} \rangle \mid c_j) = P(|d|) \prod_{k=1}^{|d|} P(w_{d_k} \mid c_j; w_{d_q}, q < k) \tag{2.5}$$

Στο σημείο αυτό κάνουμε την αφελή (naive) υπόθεση ότι οι λέξεις του εγγράφου d παράγονται ανεξάρτητα από τις υπόλοιπες λέξεις που περιέχονται στο ίδιο έγγραφο. Επιπλέον υποθέτουμε ότι η παραγωγή μίας λέξης δεν εξαρτάται από τη θέση της στο έγγραφο. Μπορούμε να εκφράσουμε τις παραπάνω υποθέσεις με την ακόλουθη σχέση:

$$P(w_{d_k} \mid c_j; w_{d_q}, q < k) = P(w_{d_k} \mid c_j) \tag{2.6}$$

Συνδυάζοντας τις σχέσεις 2.4, 2.5 και 2.6 καταλήγουμε στην εξής μορφή για τη συνάρτηση πυκνότητας πιθανότητας του πυρήνα c_j:

$$p(d \mid c_j) = P(|d|) \prod_{k=1}^{|d|} P(w_{d_k} \mid c_j) \tag{2.7}$$

Η κατανομή της μορφής 2.6 ονομάζεται πολυωνυμική (multinomial) και έχει παραμέτρους τις πιθανότητες των λέξεων $P(w_t \mid c_j), t = \{1, \ldots, |V|\}$. Οι παράμετροι της πολυωνυμικής κατανομής του πυρήνα c_j συμβολίζονται με $\theta_{w_t|c_j}$, δηλαδή $\theta_{w_t|c_j} \equiv P(w_t \mid c_j), t = \{1, \ldots, |V|\}$, και ικανοποιούν τη σχέση:

$$\sum_t P(w_t \mid c_j) = 1 \tag{2.8}$$

Πρέπει να σημειωθεί ότι το μήκος του εγγράφου δεν αποτελεί παράμετρο της πολυωνυμικής κατανομής δεδομένου ότι θεωρείται ομοιόμορφα κατανεμημένο.

Το μικτό μοντέλο, το οποίο περιγράφεται από την 2.1, έχει ως παραμέτρους τα βάρη (εκ των προτέρων πιθανότητες) $P(c_j)$, τα οποία θα συμβολίζουμε με θ_{c_j}, καθώς επίσης και τις παραμέτρους των συστατικών κατανομών $\theta_{w_t|c_j}, c_j \in C$. Επομένως, το σύνολο των παραμέτρων του μικτού μοντέλου είναι

$$\theta = \{\theta_{w_t|c_j} : w_t \in V, c_j \in C; \theta_{c_j} : c_j \in C\} \tag{2.9}$$

2.3 Εκπαίδευση του Ταξινομητή Naive Bayes

Θεωρούμε ότι έχουμε στη διάθεση μας ένα σύνολο εκπαίδευσης $\mathcal{D} = \{\mathrm{d}_1, \ldots, \mathrm{d}_{|\mathrm{D}|}\}$, το οποίο αποτελείται από έγγραφα για τα οποία υποθέτουμε το μοντέλο παραγωγής της ενότητας 2.1 . Η εκπαίδευση του ταξινομητή Naive Bayes συνίσταται στον υπολογισμό των παραμέτρων θ του μοντέλου παραγωγής.

Η μέθοδος Naive Bayes εφαρμόζει μέγιστη εκ των υστέρων (MAP – maximum a posterior) εκτίμηση στο σύνολο των παραμέτρων θ. Προσπαθεί να εντοπίσει την τιμή του θ για την οποία μεγιστοποιείται η εκ των υστέρων πιθανότητα $P(\theta \mid D)$. Για να διευκολυνθεί ο υπολογισμός του ζητούμενου θ, αντικαθιστούμε την εκ των υστέρων πιθανότητα $P(\theta \mid D)$ με χρήση του θεωρήματος του Bayes και στη συνέχεια αγνοούμε την ποσότητα $P(D)$, διότι είναι ανεξάρτητη του θ. Δηλαδή:

$$\begin{aligned} \theta_{MAP} &\equiv \arg\max_{\theta} P(\theta|D) \\ &= \arg\max_{\theta} \frac{P(D \mid \theta)P(\theta)}{P(D)} \\ &= \arg\max_{\theta} P(D \mid \theta)P(\theta) \end{aligned} \tag{2.10}$$

Στην εξίσωση 2.10, ο όρος $P(D \mid \theta)$ αποτελεί την πιθανοφάνεια του συνόλου εκπαίδευσης D και υπολογίζεται ως

$$P(D \mid \theta) = \prod_{i=1}^{|D|} P(d_i \mid \theta) \tag{2.11}$$

όπου τα $P(d_i \mid \theta)$ υπολογίζονται από τη σχέση 2.1. Επιπλέον, ο όρος $P(\theta)$ αναπαρίσταται από μία Dirichlet κατανομή:

$$P(\theta) \propto \prod_{cj \in C} ((\theta_{c_j})^{\alpha-1} \prod_{w_t \in V} (\theta_{w_t|c_j})^{\alpha-1}) \tag{2.12}$$

όπου α είναι μία θετική σταθερά (για τον υπολογισμό των παραμέτρων του μοντέλου έχουμε θέσει $\alpha = 2$). Στη συνέχεια μεγιστοποιούμε την εκ των υστέρων πιθανότητα $P(\theta \mid D)$ (όπως αυτή εκφράζεται μετά την αντικατάσταση των $P(D \mid \theta)$ και $P(\theta)$ από τις σχέσεις 2.11 και 2.12 αντίστοιχα) λύνοντας το σύστημα των μερικών παραγώγων του $\log P(\theta \mid D)$. Δεδομένου ότι πρόκειται για σύστημα με περιορισμούς (βλ. 2.8) χρησιμοποιούμε συντελεστές Lagrange για να βρούμε τη λύση του. Η παραπάνω διαδικασία οδηγεί στις εκτιμήσεις $\hat{\theta}$ των παραμέτρων θ, των οποίων οι αναλυτικές εκφράσεις παρουσιάζονται ακολούθως.

Η εκτίμηση για τις παραμέτρους $\hat{\theta}_{w_t|c_j}, t = 1, \ldots, |V|$ της πολυωνυμικής κατανομής $p(d \mid c_j)$ του πυρήνα c_j περιγράφεται από τη σχέση:

$$\hat{\theta}_{w_t|c_j} \equiv P(w_t \mid c_j) = \frac{1 + \sum_{i=1}^{|D|} N(w_t, d_i) P(c_j \mid d_i)}{|V| + \sum_{s=1}^{|V|} \sum_{i=1}^{|D|} N(w_s, d_i) P(c_j \mid d_i)} \tag{2.13}$$

όπου με $N(w_t, d_i)$ παριστάνουμε τον αριθμό των εμφανίσεων της λέξης w_t στο έγγραφο d_i. Η πιθανότητα ένα έγγραφο d_i να ανήκει στην κατηγορία c_j, $P(c_j \mid d_i)$ λαμβάνει την τιμή 0 ή 1, ανάλογα με την κατηγορία που αντιστοιχεί στο έγγραφο d_i και η οποία είναι γνωστή για τα έγγραφα του συνόλου εκπαίδευσης. Επομένως, οι παράμετροι $\hat{\theta}_{w_t|c_j}$ υπολογίζονται ως το πηλίκο του αριθμού των εμφανίσεων της λέξης w_t στα έγγραφα της κατηγορίας c_j προς τον αριθμό των εμφανίσεων όλων των λέξεων του λεξικού V στα έγγραφα της ίδιας κατηγορίας. Προκειμένου οι εκτιμήσεις $\hat{\theta}_{w_t|c_j}$ να μην λαμβάνουν μηδενικές τιμές για λέξεις w_t που δεν εμφανίζονται συχνά στα έγγραφα του συνόλου εκπαίδευσης, προσαυξάνουμε τον αριθμό των εμφανίσεων της λέξης w_t στα έγγραφα της κατηγορίας c_j με έναν "ψευδομετρητή" (στην περίπτωση μας ο ψευδομετρητής ισούται με 1). Η διαδικασία αυτή είναι γνωστή ως Laplace smoothing. Όπως προκύπτει από την εξίσωση 2.13, για τον υπολογισμό των παραμέτρων της πολυωνυμικής κατανομής του πυρήνα c_j χρησιμοποιούνται μόνο τα έγγραφα της κατηγορίας c_j (ένα–προς–ένα αντιστοιχία κατηγοριών και πυρήνων).

Η εκτίμηση για τις εκ των προτέρων (prior) πιθανότητες $P(c_j), c_j \in C$ της μικτής πολυωνυμικής κατανομής περιγράφεται από τη σχέση:

$$\hat{\theta}_{c_j} \equiv P(c_j) = \frac{1 + \sum_{i=1}^{|D|} P(c_j \mid d_i)}{|C| + |D|} \tag{2.14}$$

όπου $|C|$ είναι το πλήθος των κατηγοριών του προβλήματος ταξινόμησης και $|D|$ το πλήθος των εγγράφων του συνόλου εκπαίδευσης. Όπως και προηγουμένως έχει εφαρμοσθεί Laplace smoothing.

Η μέθοδος Naive Bayes κατασκευάζει έναν ταξινομητή κειμένου υπολογίζοντας τις παραμέτρους των πυρήνων $\hat{\theta}_{w_t|c_j}$, με $t = 1, \ldots, |V|$ και $j = 1, \ldots, |C|$ (από την εξίσωση 2.13) και τα βάρη της μικτής πολυωνυμικής κατανομής $\hat{\theta}_{c_j}, j = 1, \ldots, |C|$ (από την εξίσωση 2.14. Το πλήθος των παραμέτρων του ταξινομητή είναι $|C| + |C||V|$.

2.4 Ταξινόμηση Αγνώστου Εγγράφου

Έστω ότι έχουμε εκπαιδεύσει έναν ταξινομητή κειμένου για ένα πρόβλημα ταξινόμησης με $|C|$ κατηγορίες με τη μέθοδο Naive Bayes. Υποθέτουμε ότι καταφθάνει στον ταξινομητή μας ένα άγνωστο έγγραφο d, για το οποίο επιθυμούμε να αποφασίσουμε την κατηγορία του.

Η πιθανότητα το έγγραφο d να ανήκει στην κατηγορία $c_j, j = 1, \ldots, |C|$ (ή ισοδύναμα να έχει παραχθεί από τον πυρήνα c_j) δίνεται από το θεώρημα του Bayes:

$$P(c_j \mid d) = \frac{P(c_j)p(d \mid c_j)}{p(d)} \tag{2.15}$$

ή ισοδύναμα με χρήση των εξισώσεων 2.1 και 2.7

$$P(c_j \mid d) = \frac{P(c_j)\prod_{k=1}^{|d|} P(w_{d_k} \mid c_j)}{\sum_{r=1}^{|C|} P(c_r)\prod_{k=1}^{|d|} P(w_{d_k} \mid c_r)} \tag{2.16}$$

Η πιθανότητα $P(c_j \mid d)$ αντιπροσωπεύει την εκ των υστέρων (posterior) "πεποίθηση" μας σχετικά με το ποια είναι η κατηγορία του εγγράφου d. Για να ταξινομήσουμε το άγνωστο έγγραφο d εφαρμόζουμε τον κανόνα απόφασης του Bayes, σύμφωνα με τον οποίο επιλέγουμε την κατηγορία με τη μεγαλύτερη εκ των υστέρων πιθανότητα. Επομένως η κατηγορία c στην οποία ταξινομούμε το έγγραφο d είναι εκείνη για την οποία ισχύει

$$c = \arg\max_{c_j} P(c_j \mid d) \tag{2.17}$$

Στο σημείο αυτό πρέπει να επισημανθεί το γεγονός ότι οι υποθέσεις του μοντέλου παραγωγής (ύπαρξη μικτής κατανομής που παράγει τα έγγραφα, ένα-προς-ένα αντιστοιχία πυρήνων και κατηγοριών, ανεξαρτησία παραγωγής μεταξύ των λέξεων) παραβιάζονται στα προβλήματα ταξινόμησης εγγράφων τα οποία συναντούμε στην πράξη. Ωστόσο, η μέθοδος Naive Bayes είναι εξαιρετικά απλή και επιτυγχάνει ικανοποιητικά αποτελέσματα. Στη συνέχεια της εργασίας επεκτείνουμε το μοντέλο παραγωγής και μελετάμε μεθόδους που αυξάνουν την απόδοση του Naive Bayes ταξινομητή κειμένου.

Κεφάλαιο 3

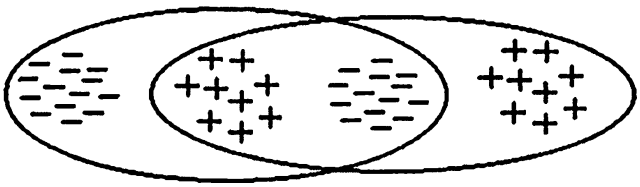
Η Μέθοδος Subtopic

3.1 Γενικά

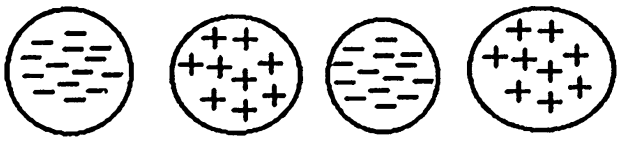
Η μέθοδος Naive Bayes για το πρόβλημα της ταξινόμησης κειμένου βασίζεται στο μοντέλο παραγωγής το οποίο παρουσιάστηκε στην ενότητα 2.2 . Όταν οι υποθέσεις του μοντέλου αυτού παραβιάζονται σε μεγάλο βαθμό, τότε ο παραγόμενος ταξινομητής κειμένου δεν αποδίδει ικανοποιητικά. Για το λόγο αυτό, έχουν γίνει προσπάθειες για να βελτιωθεί το συγκεκριμένο στατιστικό μοντέλο παραγωγής κειμένων. Στο κεφάλαιο αυτό, θα μελετήσουμε τη μέθοδο ταξινόμησης subtopic, η οποία εκμεταλλεύεται την πιθανή ανάλυση μίας κατηγορίας σε θεματικές υποκατηγορίες. Η μέθοδος ταξινόμησης subtopic δεν υιοθετεί την υπόθεση ότι υπάρχει ένα–προς–ένα αντιστοιχία μεταξύ των μικτών συνιστωσών (πυρήνων) του μικτού μοντέλου και των κατηγοριών του προβλήματος ταξινόμησης. Αντίθετα, την αντικαθιστά με την υπόθεση ότι υπάρχει πολλά–προς–ένα αντιστοιχία μεταξύ μικτών συνιστωσών και κατηγοριών. Στη συνέχεια, θα μελετήσουμε πως τροποποιείται το μοντέλο παραγωγής καθώς και πως γίνεται η εκπαίδευση της μεθόδου subtopic, η οποία βασίζεται στο νέο μοντέλο παραγωγής.

3.2 Βελτίωση του Μοντέλου Παραγωγής

Η δεύτερη υπόθεση του μοντέλου παραγωγής εγγράφων, το οποίο παρουσιάστηκε στην ενότητα 2.2, δηλώνει ότι υπάρχει ένα–προς–ένα αντιστοιχία μεταξύ των μικτών συνιστωσών (πυρήνων) του μικτού μοντέλου (το οποίο παράγει τα έγγραφα) και των κατηγοριών του προβλήματος ταξινόμησης. Όταν αυτή η υπόθεση παραβιάζεται σε μεγάλο βαθμό, τότε οι ταξινομητές κειμένου που βασίζονται σε αυτό το μοντέλο παραγωγής δεν αποδίδουν ικανοποιητικά.



Σχήμα 3.1: Παράδειγμα προτύπων το οποίο παραβιάζει την υπόθεση ότι υπάρχει ένα–προς-ένα αντιστοιχία μεταξύ μικτών κατανομών του μικτού μοντέλου και κατηγοριών του προβλήματος ταξινόμησης. Τα πρότυπα ανήκουν σε δυο κατηγορίες (την + και την -), ενώ υποθέτουμε ότι οι μικτές συνιστώσες είναι κανονικές (Gaussian).



Σχήμα 3.2: Το ίδιο παράδειγμα προτύπων με το σχήμα 3.1 . Υποθέτουμε δύο μικτές κανονικές συνιστώσες ανά κατηγορία. Στην περίπτωση αυτή, το μοντέλο παραγωγής των προτύπων είναι αρκετά πιο αντιπροσωπευτικό από το αντίστοιχο του σχήματος 3.1 .

Ας θεωρήσουμε το παράδειγμα του σχήματος 3.1, το οποίο παριστάνει ένα σύνολο προτύπων τα οποία ανήκουν σε δύο κατηγορίες (την + και την -). Το μοντέλο παραγωγής αυτών των προτύπων είναι το ίδιο με αυτό της ενότητας 2.2, με τη διαφορά ότι οι μικτές συνιστώσες είναι κανονικές (Gaussian). Όπως μπορεί εύκολα να παρατηρήσει κανείς, καμία από τις δύο κατηγορίες δεν μοντελλοποιείται καλά με τη χρήση ενός μόνο πυρήνα. Αντίθετα, αν το μοντέλο παραγωγής τροποποιηθεί έτσι ώστε τα πρότυπα κάθε κατηγορίας να παράγονται από μικτή κατανομή αποτελούμενη από δύο πυρήνες ανά κατηγορία του προβλήματος ταξινόμησης, τότε το νέο μοντέλο είναι πολύ πιο αντιπροσωπευτικό, όπως φαίνεται στο σχήμα 3.2 .

Το νέο μοντέλο παραγωγής εγγράφων ενσωματώνει τις εξής υποθέσεις, εκ των οποίων μόνο η δεύτερη διαφοροποιείται από το προηγούμενο μοντέλο της ενότητας 2.2:

1. τα έγγραφα παράγονται με χρήση μικτού μοντέλου,

2. υπάρχει πολλά–προς–ένα αντιστοιχία μεταξύ των μικτών συνιστωσών (πυρήνων) του μικτού μοντέλου και των κατηγοριών του προβλήματος ταξινόμησης. Αυτό σημαίνει ότι για κάθε κατηγορία υπάρχει ένα μικτό μοντέλο το οποίο παράγει τα έγγραφα αυτής της κατηγορίας,

3. οι μικτές συνιστώσες του μικτού μοντέλου μίας κατηγορίας είναι πολυωνυμικές (multinomial) κατανομές ορισμένες στις λέξεις των εγγράφων.

Στη συνέχεια, παρουσιάζουμε το νέο μοντέλο με έμφαση στις νέες παραμέτρους του, οι οποίες πρέπει να υπολογιστούν προκειμένου να κατασκευαστεί ο ταξινομητής κειμένου.

Υποθέτουμε ότι έχουμε στη διάθεση μας το σύνολο εκπαίδευσης $D = \{(d_i, t_{d_i}), i = 1, \ldots, |D|\}$, όπου d_i είναι το i–έγγραφο του συνόλου και $t_{d_i} \in T = \{t_1, \ldots, t_{|T|}\}$ είναι η κατηγορία του εγγράφου d_i. Λόγω του ότι ένα έγγραφο ανήκει σε μία κατηγορία, το σύνολο εκπαίδευσης $\mathcal{D}$ διαχωρίζεται σε $|T|$ ανεξάρτητα υποσύνολα (όσα και οι κατηγορίες του προβλήματος ταξινόμησης), έτσι ώστε το κάθε υποσύνολο D_{t_i} να περιέχει τα έγγραφα τα οποία ανήκουν στην κατηγορία t_i. Προφανώς ισχύει $|D_{t_1}| + \ldots + |D_{t_{|T|}}| = |D|$. Σε κάθε κατηγορία $t_i \in T$ αντιστοιχεί μία εκ των προτέρων (prior) πιθανότητα $P(t_i)$ και μία υπό συνθήκη συνάρτηση πυκνότητας πιθανότητας $p(d \mid t_i)$, ο καθορισμός των οποίων αποτελεί το ζητούμενο της διαδικασίας εκπαίδευσης του ταξινομητή κειμένου. Η συνάρτηση $p(d \mid t_i)$ επιθυμούμε να προσεγγίζει την κατανομή των εγγράφων της κατηγορίας t_i, πράγμα που σημαίνει ότι για τον προσδιορισμό της δεν είναι απαραίτητο να χρησιμοποιήσουμε ολόκληρο

το σύνολο εκπαίδευσης $\mathcal{D}$, αρκεί να χρησιμοποιήσουμε το σύνολο D_{t_i} των εγγράφων που ανήκουν στην κατηγορία t_i.

Για την εκτίμηση των κατανομών $p(d \mid t_i), i = 1, \ldots, |T|$ θα χρησιμοποιήσουμε $|T|$ μικτές πολυωνυμικές κατανομές, έτσι ώστε η καθεμία να έχει το δικό της σύνολο πυρήνων $C_{t_i} = \{c_1, \ldots, c_{|C_{t_i}|}\}$. Δηλαδή οι συναρτήσεις $p(d \mid t_i)$ ορίζονται από τις σχέσεις:

$$p(d \mid t_i) = \sum_{j=1}^{|C_{t_i}|} P(c_j) p(d \mid c_j), \text{ για } i = 1, \ldots, |T| \tag{3.1}$$

όπου με $|C_{t_i}|$ συμβολίζουμε τον αριθμό των πυρήνων της αντίστοιχης κατανομής. Το βάρος $P(c_j)$ αποτελεί την εκ των προτέρων (prior) πιθανότητα σύμφωνα με την οποία η παραγωγή ενός εγγράφου που ανήκει στην κατηγορία t_i οφείλεται στον πυρήνα $c_j \in C_{t_i}$, όπου C_{t_i} είναι το σύνολο των πυρήνων της κατηγορίας t_i. Οι prior πιθανότητες υπόκεινται στους εξής περιορισμούς:

$$\sum_{j=1}^{|C_{t_i}|} P(c_j) = 1, P(c_j) \geq 0, \text{ για } i = 1, \ldots, |T| \tag{3.2}$$

Οι συναρτήσεις $p(d \mid c_j)$ είναι πολυωνυμικές (multinomial) κι επομένως περιγράφονται από τη σχέση 2.7. Έχουν ως παραμέτρους τις πιθανότητες των λέξεων $P(w_t \mid c_j), t = 1, \ldots, |V|$ οι οποίες συμβολίζονται (όπως και στο κεφάλαιο 2) με $\theta_{w_t|c_j}$, ενώ τα αντίστοιχα βάρη τους συμβολίζονται με θ_{c_j}.

Προκειμένου να παράγουμε ένα έγγραφο d, επιλέγουμε καταρχήν μία κατηγορία t_i από το σύνολο $\mathcal{T}$ των κατηγοριών με πιθανότητα $P(t_i)$ και στη συνέχεια παράγουμε το έγγραφο με βάση τη μικτή κατανομή $p(d \mid t_i)$. Δηλαδή επιλέγουμε έναν πυρήνα c_j από το σύνολο των πυρήνων C_{t_i} της κατηγορίας t_i με πιθανότητα $P(c_j)$ και κατόπιν παράγουμε το έγγραφο με βάση τη συστατική πολυωνυμική κατανομή $p(d \mid c_j)$.

Συμβολίζουμε το σύνολο των παραμέτρων της μικτής πολυωνυμικής κατανομής $p(d \mid t_i)$ (η οποία χρησιμοποιείται για την παραγωγή των εγγράφων της κατηγορίας t_i) με θ_{t_i}, το οποίο ορίζεται ως

$$\theta_{t_i} = \{\theta_{w_t|c_j} : w_t \in V, c_j \in C_{t_i}, \theta_{c_j} : c_j \in C_{t_i}\} \tag{3.3}$$

ενώ το σύνολο των παραμέτρων θ του μοντέλου παραγωγής προκύπτει από τη συγκέντρωση των παραμέτρων θ_{t_i} για κάθε μικτή πολυωνυμική κατανομή $p(d \mid t_i)$ καθώς επίσης και των prior πιθανοτήτων $P(t_i)$. Δηλαδή,

$$\theta = \{P(t_1), \theta_{t_1}, \ldots, P(t_{|T|}), \theta_{t_{|T|}}\} \tag{3.4}$$

3.3 Εκπαίδευση του Subtopic Ταξινομητή

Θεωρούμε ότι έχουμε στη διάθεσή μας το σύνολο εκπαίδευσης $D = \{(d_i, t_{d_i}), i = 1, \ldots, |D|\}$, όπου d_i είναι το i–έγγραφο του συνόλου και $t_{d_i} \in T = \{t_1, \ldots, t_{|T|}\}$ είναι η κατηγορία του εγγράφου d_i. Υποθέτουμε ότι τα έγγραφα του συνόλου $\mathcal{D}$ δημιουργήθηκαν με βάση το στατιστικό μοντέλο της ενότητας 3.2 . Η εκπαίδευση του subtopic ταξινομητή συνίσταται στον υπολογισμό των παραμέτρων θ του μοντέλου παραγωγής, όπως αυτές περιγράφονται από τη σχέση 3.4. Ισοδύναμα, αρκεί να υπολογίσουμε για κάθε κατηγορία t_i την prior πιθανότητα $P(t_i)$ και το σύνολο των παραμέτρων θ_{t_i} της υπό συνθήκη κατανομής $p(d \mid t_i)$.

Δεδομένου ότι για κάθε έγγραφο $d \in D_{t_i}$ της κατηγορίας t_i δεν γνωρίζουμε τον πυρήνα έστω $c_j \in C_{t_i}$ από τον οποίο προέρχεται, το σύνολο εκπαίδευσης D_{t_i} για την κατηγορία t_i είναι ελλιπές και επομένως χρησιμοποιούμε τον αλγόριθμο EM (Expectation–Maximization) για να υπολογίσουμε τις παραμέτρους θ_{t_i} της υπό συνθήκη κατανομής $p(d \mid t_i)$. Στη συνέχεια, θα δώσουμε τη γενική περιγραφή του αλγορίθμου EM και κατόπιν θα μελετήσουμε πως εφαρμόζεται στην περίπτωση των μικτών πολυωνυμικών μοντέλων για τον υπολογισμό των παραμέτρων τους.

3.3.1 Ο Αλγόριθμος EM

Ο αλγόριθμος EM (Expectation–Maximization) ορίζεται ως μία γενική διαδικασία εύρεσης εκτιμητών μεγίστης πιθανοφάνειας (likelihood) σε προβλήματα όπου κάποιες μεταβλητές δεν έχουν παρατηρηθεί (unobserved or hidden variables). Θα δώσουμε, καταρχήν, ένα γενικό ορισμό του αλγορίθμου και εν συνεχεία θα δούμε τη μορφή που παίρνει στο πρόβλημα εκτίμησης πυκνότητας πιθανότητας υποθέτοντας ένα μικτό πολυωνυμικό μοντέλο.

Η λειτουργία του EM βασίζεται στη σχέση μεταξύ δύο συνόλων. Το πρώτο σύνολο το ονομάζουμε ελλιπές σύνολο (incomplete set) και το δεύτερο πλήρες σύνολο (complete set). Ελλιπή σύνολα προτύπων είναι συνήθως δείγματα δεδομένων που παίρνουμε από πειράματα ή στατιστικές μετρήσεις, για αυτό το λόγο και τέτοιου είδους σύνολα αποτελούν πραγματικά δεδομένα. Αντιθέτως σύνολα προτύπων που έχουν παρατηρηθεί πλήρως είναι συνήθως υποθετικά σύνολα και εκφράζουν την μορφή που θα θέλαμε να έχουν τα δεδομένα μας σε ένα πείραμα, αλλά που στην πράξη μία τέτοια μορφή δεν είναι διαθέσιμη, δηλαδή τα σύνολα αυτά είναι μη παρατηρήσιμα.

Υποθέτουμε ότι έχουμε ένα ελλιπές σύνολο προτύπων $X = \{x_1, x_2, \ldots\}$ για το οποίο ορίζεται η από κοινού κατανομή $g(\theta; X)$, η οποία εξαρτάται από το άγνωστο διάνυσμα παραμέτρων θ. Υποθέτουμε επίσης ένα πλήρες σύνολο

$Y = \{y_1, y_2, \ldots\}$ του οποίου η κατανομή $g_c(\theta; Y)$ εξαρτάται από το ίδιο διάνυσμα παραμέτρων θ. Οι λογαριθμικές πιθανοφάνειες των δύο συνόλων θα είναι αντίστοιχα:

$$L(\theta) = \log g(\theta; X) \tag{3.5}$$

$$L_c(\theta) = \log g_c(\theta; Y) \tag{3.6}$$

Το πλήρες σύνολο αφού είναι υποθετικό ορίζεται πάντα έτσι ώστε να υπάρχει μια πολλά–προς–ένα απεικόνιση μεταξύ των συνόλων Y και X. Η απεικόνιση υποδηλώνει το γεγονός ότι ένα ελλιπές σύνολο σχετίζεται με πολλά πλήρη σύνολα, με την έννοια ότι τα σύνολα αυτά είναι υποψήφια να εκφράσουν τη μορφή που θα θέλαμε να είχε το ελλιπές σύνολο.

Το πρόβλημά μας είναι να βρούμε εκείνο το διάνυσμα παραμέτρων για το οποίο μεγιστοποιείται η λογαριθμική πιθανοφάνεια του ελλιπούς συνόλου. Ο αλγόριθμος EM προσπαθεί να μεγιστοποιήσει την ποσότητα αυτή ($L(\theta)$) αναδεικνύοντας τη σχέση μεταξύ των δύο συνόλων. Συγκεκριμένα ο EM προσεγγίζει το πρόβλημα μεγιστοποίησης έμμεσα εφαρμόζοντας μία επαναληπτική διαδικασία για τη λογαριθμική πιθανοφάνεια $L_c(\theta)$ του πλήρους συνόλου. Επειδή όμως το σύνολο Y είναι μη παρατηρήσιμο και επομένως η λογαριθμική πιθανοφάνεια $L_c(\theta)$ είναι ακαθόριστη, ο EM την λαμβάνει ως τυχαία μεταβλητή και υπολογίζει την αναμενόμενη τιμή της. Σε κάθε επανάληψη το διανυσμα θ έχει μία καθορισμένη τιμή και η αναμενόμενη τιμή της $L_c(\theta)$ υπολογίζεται με βάση το σύνολο X καθώς και την τρέχουσα τιμή του θ. Εάν βρισκόμαστε στην t+1 επανάληψη του αλγορίθμου και το τρέχον διάνυσμα είναι το $\theta^{(t)}$, η προηγούμενη ποσότητα ορίζεται ως εξής:

$$Q(\theta; \theta^{(t)}) = E\{L_c(\theta) \mid X, \theta^{(t)}\} \tag{3.7}$$

Η παραπάνω σχέση ορίζει την $Q(\theta; \theta^{(t)})$ ως την υπό συνθήκη αναμενόμενη τιμή της $L_c(\theta)$ δεδομένου του X και του $\theta^{(t)}$. Κάθε επανάληψη του αλγορίθμου EM αποτελείται από δύο βήματα: το E–βήμα (Expectation–step) στο οποίο καθορίζεται η $Q(\theta; \theta^{(t)})$, και το M–βήμα (Maximization–step) στο οποίο μεγιστοποιείται η ποσότητα αυτή ως προς το διάνυσμα παραμέτρων. Πιο συγκεκριμένα τα βήματα στην t+1 επανάληψη ορίζονται ως εξής:

- E–βήμα: Υπολογισμός της ποσότητας $Q(\theta; \theta^{(t)})$
- M–βήμα: Καθορισμός του $\theta^{(t+1)}$ από τη σχέση $\theta^{(t+1)} = argmax_\theta Q(\theta; \theta^{(t)})$

Ο αλγόριθμος ξεκινά από μιά αρχική τιμή των παραμέτρων $\theta^{(0)}$ και συνεχίζει έως ότου η διαφορά

$$L(\theta^{(t+1)}) - L(\theta^{(t)}) \tag{3.8}$$

γίνει μικρότερη από μία καθορισμένη τιμή, πράγμα που σημαίνει ότι έχει επιτευχθεί η σύγκλιση. Σύμφωνα με τις ιδιότητες του αλγορίθμου, η λογαριθμική πιθανοφάνεια του ελλιπούς συνόλου δεν μειώνεται μετά από μία επανάληψη του αλγορίθμου, δηλαδή ισχύει:

$$L(\theta^{(t+1)}) \geq L(\theta^{(t)}) \tag{3.9}$$

Ο τρόπος που ορίζεται ο αλγόριθμος είναι τόσο γενικός που δεν είναι ξεκάθαρο για το αν πρόκειται για αλγόριθμο. Για κάποιον που βλέπει τον αλγόριθμο πρώτη φορά αναμφίβολα του δημιουργούνται ερωτήματα σχετικά με το πως ορίζεται το πλήρες σύνολο ή ποια είναι ακριβώς η απεικόνιση μεταξύ των δύο συνόλων. Επίσης δεν είναι ξεκάθαρος ο τρόπος με τον οποίο η μεγιστοποίηση της ποσότητας $Q(\theta; \theta^{(t)})$ σε κάθε επανάληψη έχει ως αποτέλεσμα την αύξηση της $L(\theta)$. Για αυτά τα ζητήματα ο αναγνώστης μπορεί να ανατρέξει στο άρθρο του Dempster [44] ή σε ένα πρόσφατο βιβλίο των McLachlan και Krishnan [47], το οποίο είναι αφιερωμένο στον αλγόριθμο EM. Στην επόμενη ενότητα θα εφαρμόσουμε τον αλγόριθμο στο πρόβλημα εύρεσης του εκτιμητή μεγίστης πιθανοφάνειας στην περίπτωση της μικτής πολυωνυμικής κατανομής.

3.3.2 Ο Αλγόριθμος EM για Μικτές Πολυωνυμικές Κατανομές

Στην ενότητα αυτή θα μελετήσουμε τον τρόπο υπολογισμού των παραμέτρων θ_{t_a} της μικτής πολυωνυμικής κατανομής $p(d \mid t_a)$, η οποία είναι υπεύθυνη για την παραγωγή των εγγράφων της κατηγορίας $t_a \in T$. Η εκτίμηση των παραμέτρων γίνεται με βάση το σύνολο εκπαίδευσης D_{t_a} της κατηγορίας t_a.

Κάθε πρότυπο στο πρόβλημα ταξινόμησης αποτελεί ένα ζεύγος της μορφής (d, t_a), όπου το t_a υποδεικνύει την κατηγορία του εγγράφου d, δηλαδή το σύνολο εκπαίδευσης αποτελείται, όπως έχουμε δει, από ανεξάρτητα σύνολα προτύπων ένα για κάθε κατηγορία. Αντιθέτως, αν και το μικτό μοντέλο θεωρεί το σύνολο προτύπων ως ένα μίγμα ανεξάρτητων υποσυνόλων, ένα για κάθε συστατικό πυρήνα, τα υποσύνολα αυτά στην πράξη είναι ακαθόριστα δεδομένου ότι τα έγγραφα δεν συνοδεύονται από κάποια πληροφορία υπόδειξης σχετικά με τον πυρήνα που ανήκουν. Υπό αυτήν την έννοια το σύνολο εκπαίδευσης D_{t_a} θεωρείται πως είναι ελλιπές. Προφανώς, θα θέλαμε τα πρότυπα εκπαίδευσης να ορίζονται όπως στο πρόβλημα ταξινόμησης, δηλαδή να είναι της μορφής (d, z) όπου z είναι ένας ακέραιος, $z = 1, \ldots, |C_{t_a}|$, που υποδεικνύει τον πυρήνα από τον οποίο έχει προέλθει το έγγραφο d. Εάν τα πρότυπά μας είχαν αυτή τη μορφή η εύρεση των παραμέτρων της μικτής κατανομής θα ήταν εύκολη. Για παράδειγμα, στην περίπτωση των πολυωνυμικών πυρήνων θα υπολογίζαμε τις παραμέτρους της κάθε πολυωνυμικής κατανομής χρησι-

μοποιώντας το αντίστοιχο σύνολο προτύπων, που όπως δείξαμε γίνεται με αναλυτικό τρόπο.

Με βάση τα προηγούμενα αν το σύνολο εκπαίδευσης είναι το $D_{t_a} = \{d_1, \ldots, d_{|D_{t_a}|}\}$ τότε, για κάθε έγγραφο d_n εισάγουμε τη μεταβλητή z_n ως ένα ακέραιο στο $\{1, |C_{t_a}|\}$ που υποδεικνύει τον πυρήνα που δημιούργησε το d_n. Ορίζουμε το πλήρες σύνολο ως εξής:

$$Y = \{(d_1, z_1), \ldots, (d_{|D_{t_a}|}, z_{|D_{t_a}|})\} \tag{3.10}$$

Η λογαριθμική πιθανοφάνεια του παραπάνω συνόλου θα είναι η ακόλουθη:

$$L_c(\theta_{t_a}) = \sum_{n=1}^{|D_{t_a}|} \log(P(c_{z_n}) p(d_n \mid c_{z_n})) \tag{3.11}$$

Η παραπάνω σχέση προκύπτει ως εξής: το έγγραφο d_n γνωρίζουμε ότι ανήκει στον πυρήνα z_n, πράγμα που σημαίνει ότι παράγεται με βάση την πιθανότητα $P(c_{z_n})p(d_n \mid c_{z_n})$ (κι όχι με την $p(d \mid t_a)$). Επομένως η από κοινού συνάρτηση κατανομής όλων των εγγράφων είναι $g_c(\theta_{t_a}; Y) = \prod_{n=1}^{|D_{t_a}|} P(c_{z_n}) p(d_n \mid c_{z_n})$, από την οποία συνεπάγεται η σχέση 3.11.

Στην πραγματικότητα οι μεταβλητές z_n είναι άγνωστες, κάτι που σημαίνει ότι το πλήρες σύνολο είναι ακαθόριστο (όπως προαναφέραμε κάθε τέτοιο σύνολο είναι υποθετικό). Υπάρχουν πολλές επιλογές για την μορφή του πλήρους συνόλου (αυτό είναι ένα παράδειγμα της πολλά προς ένα απεικόνισης μεταξύ των Y και D_{t_a} στον ορισμό του EM) που προκύπτουν αν σκεφτούμε ότι για κάθε έγγραφο d_n η μεταβλητή z_n μπορεί να πάρει τιμές από 1 έως $|C_{t_a}|$. Συγκεκριμένα υπάρχουν $|C_{t_a}|^{|D_{t_a}|}$ διαφορετικές επιλογές του πλήρους συνόλου καθώς και τόσες εκδοχές της αντίστοιχης λογαριθμικής πιθανοφάνειας $L_c(\theta_{t_a})$. Προκειμένου να υπολογίσουμε την αναμενόμενη τιμή όλων αυτών των εναλλακτικών εκδοχών, δηλαδή την $Q(\theta_{t_a}; \theta_{t_a}^{(t)})$, αρκεί να βρούμε την πιθανότητα με την οποία συμβαίνει κάθε μία. Η πιθανότητα με βάση την οποία το έγγραφο d_n ανήκει στατιστικά στον πυρήνα c_{z_n} δίνεται από το θεώρημα του Bayes:

$$P(c_{z_n} \mid d_n) = \frac{P(c_{z_n}) p(d_n \mid c_{z_n})}{\sum_{j=1}^{|C_{t_a}|} P(c_j) p(d_n \mid c_j)} \tag{3.12}$$

Αντικαθιστώντας από τον ορισμό της πολυωνυμικής κατανομής (2.7) καταλήγουμε στην ποσότητα που υπολογίζεται στο E-βήμα του αλγορίθμου EM:

$$P(c_j \mid d) = \frac{P(c_j) \prod_{k=1}^{|d|} P(w_{d_k} \mid c_j)}{\sum_{r=1}^{|C_{t_a}|} P(c_r) \prod_{k=1}^{|d|} P(w_{d_k} \mid c_r)} \tag{3.13}$$

όπου $j = 1, \ldots, |C_{t_a}|$ και d ένα έγγραφο της κατηγορίας t_a.

Σύμφωνα με τα παραπάνω, r_i από κοινού πιθανότητα με την οποία συμβαίνει μια συγκεκριμένη επιλογή των μεταβλητών z_n θα ισούται με $\prod_{n=1}^{|D_{t_a}|} P(c_{z_n} \mid d_n)$. Οπότε r_i ποσότητα $Q(\theta_{t_a}; \theta_{t_a}^{(t)})$ της t+1 επανάληψης του αλγορίθμου θα δίνεται από τη σχέση:

$$Q(\theta_{t_a}; \theta_{t_a}^{(t)}) = \sum_{z_1=1}^{|C_{t_a}|} \cdots \sum_{z_n=1}^{|C_{t_a}|} \{L_c(\theta_{t_a}) \prod_{n=1}^{|D_{t_a}|} P^{(t)}(c_{z_n} \mid d_n)\} \tag{3.14}$$

όπου χρησιμοποιείται ο δείκτης "t" σημαίνει ότι η αντίστοιχη ποσότητα έχει καθοριστεί με βάση την τρέχουσα τιμή των παραμέτρων $\theta_{t_a}^{(t)}$. Από την παραπάνω σχέση με λίγες πράξεις και χρησιμοποιώντας τη σχέση 3.11 προκύπτει:

$$Q(\theta_{t_a}; \theta_{t_a}^{(t)}) = \sum_{j=1}^{|C_{t_a}|} \sum_{n=1}^{|D_{t_a}|} P^{(t)}(c_j \mid d_n)\{\log P(c_j) + \log p(d_n \mid c_j)\} \tag{3.15}$$

Στο M–βήμα υπολογίζεται το σύνολο των παραμέτρων $\theta_{t_a}^{(t+1)}$ (όπως αυτές ορίζονται στη σχέση 3.3), το οποίο μεγιστοποιεί την ποσότητα 3.15. Οι νέες τιμές των παραμέτρων θα δίνονται από τις σχέσεις:

$$\theta_{c_j} \equiv P(c_j \mid t_a) = \frac{1 + \sum_{i=1}^{|D_{t_a}|} P(c_j \mid d_i)}{|C_{t_a}| + |D_{t_a}|} \tag{3.16}$$

$$\theta_{w_t|c_j} \equiv P(w_t \mid c_j) = \frac{1 + \sum_{i=1}^{|D_{t_a}|} N(w_t, d_i) P(c_j \mid d_i)}{|V| + \sum_{s=1}^{|V|} \sum_{i=1}^{|D_{t_a}|} N(w_s, d_i) P(c_j \mid d_i)} \tag{3.17}$$

Ο αλγόριθμος EM είναι ιδιαίτερα δημοφιλής διότι είναι απλός στην υλοποίηση και επίσης διότι μας εγγυάται την μονότονη αύξηση της πιθανοφάνειας του συνόλου εκπαίδευσης. Ωστόσο, βασικά μειονεκτήματα του είναι τα ακόλουθα:

1. ο πραγματικός αριθμός των πυρήνων είναι συνήθως άγνωστος.
2. δεν υπάρχει γενικά αποδεκτή μέθοδος για την αρχικοποίηση των παραμέτρων του μοντέλου. Διαφορετικές αρχικοποιήσεις οδηγούν σε διαφορετικές εκτιμήσεις των παραμέτρων.
3. ο αλγόριθμος μπορεί να παγιδευτεί σε κάποιο από τα πολλά τοπικά μέγιστα της συνάρτησης της πιθανοφάνειας.

3.3.3 Περιγραφή της Μεθόδου Subtopic

Η μέθοδος subtopic υπολογίζει τις παραμέτρους θ, όπως εκείνες ορίζονται στη σχέση 3.4, εφαρμόζοντας κατάλληλα τον αλγόριθμο EM για μικτές πολυωνυμικές κατανομές, ο οποίος παρουσιάστηκε στην ενότητα 3.3.2 . Στη συνέχεια παρουσιάζουμε τα βήματα της μεθόδου.

Μέθοδος Subtopic

- *Είσοδος:* το σύνολο εκπαίδευσης $\mathcal{D} = \{(\mathrm{d_i}, \mathrm{t_{d_i}}), \mathrm{i} = 1, \ldots, |\mathrm{D}|\}$, το οποίο χωρίζεται στα ανεξάρτητα υποσύνολα $D_{t_a}, a = 1, \ldots, |T|$, όπου T το σύνολο των κατηγοριών του προβλήματος ταξινόμησης.
- Καθορισμός του πλήθους $|C_{t_a}|$ των πυρήνων ανά κατηγορία.
- Για κάθε κατηγορία $t_a \in T$ κάνε τα εξής:
  - Αρχικοποίησε τυχαία τις prior πιθανότητες $P(c_j \mid d)$ των πυρήνων c_j της κατηγορίας t_a για κάθε έγγραφο d του συνόλου εκπαίδευσης D_{t_a} της κατηγορίας t_a. Η αρχικοποίηση πρέπει να τηρεί τον περιορισμό $\sum_{j=1}^{|C_{t_a}|} P(c_j \mid d) = 1$.
  - Αρχικοποίησε τις παραμέτρους θ_{t_a} της μικτής πολυωνυμικής κατανομής $p(d \mid t_a)$ με βάση τις σχέσεις 3.16 και 3.17.
  - Υπολόγισε την prior πιθανότητα $P(t_a)$ ένα τυχαίο έγγραφο να ανήκει στην κατηγορία t_a, από τη σχέση:

$$P(t_a) = \frac{1 + |D_{t_a}|}{|T| + |D|} \tag{3.18}$$

  - Εφαρμογή EM μέχρις ότου η μεταβολή στην λογαριθμική πιθανοφάνεια $L(\theta_{t_a}^{(t+1)}) - L(\theta_{t_a}^{(t)})$ να γίνει μικρότερη από μία καθορισμένη τιμή:
    - **Ε–Βήμα :** Υπολόγισε τις posterior $P(c_j \mid d)$ με βάση τη σχέση 3.13.
    - **Μ–Βήμα :** Υπολόγισε τις παραμέτρους θ_{t_a} της μικτής πολυωνυμικής κατανομής $p(d \mid t_a)$ από τις σχέσεις 3.16 και 3.17.
- *Έξοδος:* Ο subtopic ταξινομητής ο οποίος ορίζεται από το σύνολο των παραμέτρων θ (σχέση 3.4).

Η λογαριθμική πιθανοφάνεια της μικτής πολυωνυμικής κατανομής $p(d \mid t_a)$ περιγράφεται από τη σχέση:

$$\tilde{L}(\theta_{t_a}) = \sum_{d \in D_{t_a}} \log \sum_{j=1}^{|C_{t_a}|} P(c_j) \prod_{k=1}^{|d|} P(w_{d_k} \mid c_j) \tag{3.19}$$

3.4 Ταξινόμηση Αγνώστου Εγγράφου

Έστω ότι έχουμε εκπαιδεύσει έναν ταξινομητή κειμένου για ένα πρόβλημα ταξινόμησης με $|\mathcal{T}|$ κατηγορίες με τη μέθοδο subtopic. Υποθέτουμε ότι καταφθάνει στον ταξινομητή μας ένα άγνωστο έγγραφο d, για το οποίο επιθυμούμε να αποφασίσουμε την κατηγορία του.

Η πιθανότητα το έγγραφο d να ανήκει στην κατηγορία $t_a, a = 1, \ldots, |T|$ δίνεται από το θεώρημα του Bayes:

$$P(t_a \mid d) = \frac{P(t_a) p(d \mid t_a)}{p(d)} \tag{3.20}$$

ή ισοδύναμα με χρήση των εξισώσεων 3.1 και 2.7:

$$P(t_a \mid d) = \frac{P(t_a) \sum_{j=1}^{|C_{t_a}|} P(c_j) \prod_{k=1}^{|d|} P(w_{d_k} \mid c_j)}{\sum_{b=1}^{|T|} P(t_b) \sum_{r=1}^{|C_{t_b}|} P(c_r) \prod_{k=1}^{|d|} P(w_{d_k} \mid c_r)}, \tag{3.21}$$

όπου η prior πιθανότητα $P(t_a)$ δίνεται από την εξίσωση 3.18. Η πιθανότητα $P(t_a \mid d)$ αντιπροσωπεύει την εκ των υστέρων (posterior) "πεποίθησή" μας σχετικά με το ποια είναι η κατηγορία του εγγράφου d. Για να ταξινομήσουμε το άγνωστο έγγραφο d εφαρμόζουμε τον κανόνα του Bayes, σύμφωνα με τον οποίο επιλέγουμε την κατηγορία με τη μεγαλύτερη εκ των υστέρων πιθανότητα. Επομένως η κατηγορία t στην οποία ταξινομούμε το έγγραφο d είναι εκείνη για την οποία ισχύει:

$$t = \arg\max_{t_k} P(t_k \mid d) \tag{3.22}$$

Κεφάλαιο 4

Η Μέθοδος Kd–Subtopic

4.1 Γενικά

Η μέθοδος ταξινόμησης subtopic, η οποία μελετήθηκε στο κεφάλαιο 3, εφαρμόζει τον αλγόριθμο EM για μικτές πολυωνυμικές κατανομές προκειμένου να υπολογίσει τις παραμέτρους θ του μοντέλου παραγωγής εγγράφων, όπως αυτό παρουσιάστηκε στην ενότητα 3.2 . Ο αλγόριθμος EM είναι ιδιαίτερα δημοφιλής γιατί είναι απλός στην υλοποίηση και εγγυάται την μονότονη αύξηση της πιθανοφάνειας του συνόλου εκπαίδευσης. Ωστόσο, βασικό του μειονέκτημα αποτελεί η εξάρτησή του από την αρχικοποίηση των παραμέτρων θ του μοντέλου. Διαφορετικές αρχικοποιήσεις οδηγούν σε διαφορετικές εκτιμήσεις των παραμέτρων και επομένως σε διαφορετικούς ταξινομητές κειμένου. Για το λόγο αυτό προτείνουμε τη μέθοδο kd–subtopic για την αντιμετώπιση του προβλήματος της αρχικοποίησης του EM με χρήση μίας δομής δεδομένων, γνωστής ως kd–δέντρο (kd–tree). Στη συνέχεια, θα μελετήσουμε τα kd–δέντρα καθώς επίσης τους τρόπους με τους οποίους τα εκμεταλλευόμαστε για την επίλυση του προβλήματος της αρχικοποίησης του EM.

4.2 Η Δομή Δεδομένων Kd–δέντρο (Kd–tree)

4.2.1 Περιγραφή ενός Kd–δέντρου

Η δομή δεδομένων kd–δέντρο (kd–tree) αποτελεί μία ειδική κατηγορία δυαδικού δέντρου. Χρησιμοποιείται σε προβλήματα ομαδοποίησης (clustering) προτύπων. Κάθε κόμβος του kd–δέντρου περιέχει μία ομάδα από πρότυπα. Στη ρίζα του kd–δέντρου περιέχεται ολόκληρο το σύνολο των προτύπων, τα οποία επιθυμούμε να διαχωρίσουμε σε ομάδες (clusters) από πρότυπα, ξένες

μεταξύ τους. Κάθε εσωτερικός κόμβος διασπάται σε δύο παιδιά, τα οποία μοιράζονται τα πρότυπα του πατέρα, συγκροτώντας με αυτόν τον τρόπο δύο νέες ομάδες προτύπων (clusters). Τα φύλλα του δέντρου συνιστούν τους κόμβους που δεν έχουν διασπαστεί. Στην περίπτωση του προβλήματος της ταξινόμησης κειμένου τα πρότυπα που περιέχουν οι κόμβοι του kd–δέντρου είναι έγγραφα του συνόλου εκπαίδευσης.

Ας θεωρήσουμε ένα σύνολο από έγγραφα $\mathcal{D} = \{\mathrm{d}_1, \ldots, \mathrm{d}_{|\mathrm{D}|}\}$. Κάθε έγγραφο $d_i \in \mathcal{D}$ αναπαρίσταται ως ένα διάνυσμα στήλη της μορφής

$$d_i = (d_{i1}, \ldots, d_{i|V|})^T \tag{4.1}$$

όπου $\mathcal{V}$ είναι το λεξικό (vocabulary), δηλαδή το σύνολο των λέξεων το οποίο χρησιμοποιείται για την αναπαράσταση των εγγράφων. Η συνιστώσα d_{ij} εγγράφου d_i αντιστοιχεί στον αριθμό των εμφανίσεων της λέξης w_j στο d_i. Ωστόσο η τελική μορφή των διανυσμάτων d_i προκύπτει ύστερα από κανονικοποίησή τους με βάση τη σχέση:

$$d_{ij} = \frac{d_{ij}}{\sum_{k=1}^{|V|} d_{ik}} \tag{4.2}$$

Η παραπάνω σχέση μας εξασφαλίζει ότι το επαγόμενο διάνυσμα για την αναπαράσταση του εγγράφου d_i είναι διάνυσμα πιθανοτήτων, ιδιότητα την οποία θα χρησιμοποιήσουμε κατά την αρχικοποίηση του EM όπως θα δούμε αργότερα.

Κάθε κόμβος (cluster) C του Kd–δέντρου αποτελείται από τα εξής πεδία:

- K_C : πλήθος των εγγράφων του κόμβου (cluster) C
- M_C : πίνακας ο οποίος αποθηκεύει στις στήλες του τα έγγραφα που περιέχει ο κόμβος C. Η διάστασή του είναι $|V| \times K_C$ και μαθηματικά περιγράφεται ως:
$$M_C = \{d_1, \ldots, d_{K_C}\} \tag{4.3}$$
- μ_C : διάνυσμα το οποίο αναπαριστά το κέντρο (centroid) του κόμβου (cluster) C. Αντιστοιχεί στη μέση τιμή (mean) των εγγράφων που περιέχονται στο C και ορίζεται ως:
$$\mu_C = M_C \cdot e / K_C \tag{4.4}$$
όπου με $e = (1, \ldots, 1)^T$ θεωρούμε από εδώ και στο εξής ένα διάνυσμα κατάλληλης διάστασης ώστε να ορίζονται οι αντίστοιχες πράξεις μεταξύ πινάκων και διανυσμάτων.

- u_C : διάνυσμα το οποίο αντιστοιχεί στην κατεύθυνση διάσπασης (splitting direction). Ως άξονας διάσπασης θεωρείται ο ορθογώνιος άξονας στον οποίο σημειώνεται η μέγιστη διακύμανση (variance), δηλαδή η μέγιστη διασπορά των εγγράφων από τη μέση τιμή. Η διακύμανση για τον ορθογώνιο άξονα j ορίζεται ως:

$$var_j = \frac{1}{|D|}\sum_{i=1}^{|D|}(d_{ij} - \mu_j)^2 \tag{4.5}$$

Οπότε αν ο ορθογώνιος άξονας στην k–διάσταση επιτυγχάνει τη μέγιστη διακύμανση για τα έγγραφα του συνόλου $\mathcal{D}$, δηλαδή ισχύει

$$k = argmax_j var_j \tag{4.6}$$

τότε το διάνυσμα u_c θα είναι το $(0, \ldots, 1, \ldots, 0)$, με το 1 να εντοπίζεται στην k–θέση.

- $scatter_C$: τιμή η οποία χρησιμοποιείται για την επιλογή του κόμβου ο οποίος πρόκειται να διασπαστεί. Ορίζεται ως το άθροισμα των τετραγώνων των αποστάσεων των εγγράφων του κόμβου (cluster) C από το κέντρο του C (μεταβλητότητα). Η τιμή *scatter* για τον κόμβο C περιγράφεται από τη σχέση:

$$scatter_C = \sum_{i=1}^{K_C} \| d_i - \mu_C \|_2^2 \tag{4.7}$$

όπου

$$\| x \|_2 = \sqrt{\sum_i x_i^2} \tag{4.8}$$

η Ευκλείδια νόρμα του διανύσματος x.

- L : δείκτης στο αριστερό παιδί του κόμβου C.
- R : δείκτης στο δεξί παιδί του κόμβου C.

Στην επόμενη ενότητα παρουσιάζουμε τον αλγόριθμο για την κατασκευή ενός kd–δέντρου.

4.2.2 Αλγόριθμος Κατασκευής ενός Kd–δέντρου

Ο αλγόριθμος, ο οποίος χρησιμοποιήθηκε για την κατασκευή των kd–δέντρων, δέχεται ως είσοδο ένα σύνολο $\mathcal{D}$ από έγγραφα και ένα επιθυμητό πλήθος από

ομάδες εγγράφων (clusters) c_{max}, στις οποίες καταλήγει. Ο αλγόριθμος ξεκινάει με τη δημιουργία της ρίζας (root) του kd–δέντρου, η οποία περιέχει όλα τα έγγραφα του συνόλου $\mathcal{D}$. Στη συνέχεια διασπάει κάθε φορά ένα κατάλληλα επιλεγμένο φύλλο σε δύο παιδιά και επαναλαμβάνει αυτήν τη διαδικασία μέχρις ότου το πλήθος των φύλλων να γίνει ίσο με c_{max}, οπότε και τερματίζει.

Ο αλγόριθμος για την κατασκευή ενός kd–δέντρου εμπεριέχει δύο βασικά στάδια:

1. Επιλογή του κόμβου (φύλλου), έστω C, προς διάσπαση:

   Επιλέγουμε το φύλλο C του υπάρχοντος kd–δέντρου που διαθέτει τη μεγαλύτερη τιμή scatter, όπως αυτή έχει οριστεί από τη σχέση 4.7. Η τιμή scatter ενός κόμβου εκφράζει την απόσταση που έχουν τα έγγραφα του από το κέντρο του, ως ένα μέτρο για τη συνοχή (cohesiveness) του κόμβου.

2. Διάσπαση του κόμβου C σε δύο παιδιά:

   Προκειμένου να διασπάσουμε τον κόμβο C, τον οποίο επιλέξαμε στο προηγούμενο στάδιο, αρκεί να δημιουργήσουμε τα δύο παιδιά του L και R και να μοιράσουμε σε αυτά τα έγγραφα που υπάρχουν στον κόμβο C. Ορίζουμε τη γραμμική συνάρτηση διαχωρισμού

   $$g_C(d) = u_C^T(d - \mu_C) \tag{4.9}$$

   βάση της οποίας γίνεται ο διαχωρισμός του συνόλου των εγγράφων του κόμβου σε δύο υποσύνολα, ένα για κάθε παιδί του C. Αν $g_C(d) \leq 0$, τότε το έγγραφο d τοποθετείται στο αριστερό παιδί L του C, διαφορετικά τοποθετείται στο δεξί παιδί R.

Στη συνέχεια παρουσιάζουμε τα βήματα του αλγορίθμου κατασκευής ενός kd–δέντρου:

- Είσοδος: ένα σύνολο εγγράφων $\mathcal{D}$ (στα οποία έχει εφαρμοστεί κανονικοποίηση με χρήση της σχέσης 4.2) και ένας επιθυμητός αριθμός από φύλλα–ομάδες (clusters) c_{max} για το kd–δέντρο που θέλουμε να κατασκευάσουμε.
- Αρχικοποίησε το kd–δέντρο δημιουργώντας τη ρίζα του. Υπολόγισε τα πεδία της ρίζας, όπως αυτά περιγράφονται στην ενότητα 4.2.1.
- Για $c = 2, 3, \ldots, c_{max}$ κάνε τα εξής:
  - Επέλεξε το φύλλο C με τη μεγαλύτερη scatter τιμή.

- Δημιούργησε τα παιδιά L και R του κόμβου C.
- Για κάθε έγγραφο $d \in C$ του κόμβου προς διάσπαση C, αν $g_C(d) \leq 0$ τότε τοποθέτησε το d στο αριστερό παιδί L του C, διαφορετικά τοποθέτησέ το στο δεξί παιδί R.
- Υπολόγισε τα υπόλοιπα πεδία, όπως αυτά περιγράφονται στην ενότητα 4.2.1, για τους κόμβους L και R.

• Έξοδος: ένα kd–δέντρο το οποίο διαμερίζει το σύνολο εγγράφων $\mathcal{D}$ σε c_{max} ομάδες, όσες και τα φύλλα του.

4.3 Εκπαίδευση του Kd–Subtopic Ταξινομητή

Η μέθοδος kd–subtopic υπολογίζει τις παραμέτρους θ, όπως εκείνες ορίζονται στη σχέση 3.4, με τρόπο ίδιο με αυτόν της μεθόδου subtopic. Η διαφορά των δύο μεθόδων έγκειται στον τρόπο με τον οποίο αρχικοποιούν τις παραμέτρους θ. Η μέθοδος kd–subtopic υλοποιεί την αρχικοποίηση εκμεταλλευόμενη τη δομή δεδομένων kd–δέντρο, την οποία περιγράψαμε στις προηγούμενες ενότητες.

Θεωρούμε ότι έχουμε στη διάθεσή μας το σύνολο εκπαίδευσης $D = \{(d_i, t_{d_i}), i = 1, \ldots, |D|\}$, όπου d_i είναι το i–έγγραφο του συνόλου και $t_{d_i} \in T = \{t_1, \ldots, t_{|T|}\}$ είναι η κατηγορία του εγγράφου d_i. Προκειμένου να υπολογίσουμε τις παραμέτρους θ του ταξινομητή κειμένου, αρκεί να υπολογίσουμε για κάθε κατηγορία $t_a \in T$ την prior πιθανότητα $P(t_a)$ και το σύνολο των παραμέτρων θ_{t_a} της υπό συνθήκη κατανομής $p(d \mid t_a)$. Το σύνολο εκπαίδευσης για την κατηγορία t_a είναι το D_{t_a}. Για κάθε κατηγορία t_a κατασκευάζουμε ένα kd–δέντρο, σύμφωνα με τον αλγόριθμο της ενότητας 4.2.2, χρησιμοποιώντας το σύνολο εκπαίδευσης D_{t_a}. Το kd–δέντρο για την κατηγορία t_a ομαδοποιεί τα έγγραφα του συνόλου D_{t_a} σε $|C_{t_a}|$ ομάδες (φύλλα), όσες και το πλήθος των πυρήνων της μικτής πολυωνυμικής κατανομής $p(d \mid t_a)$, δηλαδή $c_{max} = |C_{t_a}|$. Με τον τρόπο αυτό αντιστοιχούμε κάθε πυρήνα c_j της κατανομής $p(d \mid t_a)$ με ένα φύλλο ϕ_j του kd–δέντρου. Επομένως, η αρχικοποίηση των παραμέτρων των πυρήνων της κατανομής $p(d \mid t_a)$ μπορεί να γίνει με κατάλληλη χρησιμοποίηση των αντίστοιχων φύλλων του kd–δέντρου της κατηγορίας t_a.

Στη συνέχεια προτείνουμε δύο τρόπους για την αρχικοποίηση των παραμέτρων του kd–subtopic ταξινομητή κειμένου.

4.3.1 Αρχικοποίηση των Παραμέτρων θ_{t_a}

Υποθέτουμε ότι έχουμε κατασκευάσει το kd–δέντρο της κατηγορίας t_a, με τον τρόπο που μελετήσαμε προηγουμένως. Στόχος μας είναι η αρχικοποίηση των

παραμέτρων θ_{t_a} της μικτής πολυωνυμικής κατανομής $p(d \mid t_a)$. Ισοδύναμα, επιθυμούμε να αρχικοποιήσουμε τις παραμέτρους $\theta_{w_t|c_j}$ των πυρήνων $c_j \in C_{t_a}$ καθώς και τα αντίστοιχα βάρη τους θ_{c_j}.

Θέτουμε τις παραμέτρους $\theta_{w_t|c_j}$ του πυρήνα c_j ίσες με τις αντίστοιχες συνιστώσες του κέντρου μ_{ϕ_j} του φύλλου ϕ_j.

$$\theta_{w_t|c_j} = \mu_{\phi_j t}, \text{ για } t = 1, \ldots, |V| \text{ και } j = 1, \ldots, |C_{t_a}| \tag{4.10}$$

όπου μ_{ϕ_j} είναι το κέντρο του φύλλου ϕ_j, το οποίο αντιστοιχεί στον πυρήνα c_j της μικτής κατανομής $p(d \mid t_a)$.

Επιπλέον, υπολογίζουμε το βάρος θ_{c_j} του πυρήνα c_j με βάση το σύνολο των εγγράφων που περιέχονται στο φύλλο ϕ_j, το οποίο αντιστοιχεί στον c_j.

$$\theta_{c_j} = \frac{1 + K_{\phi_j}}{|C_{t_a}| + |D_{t_a}|} \tag{4.11}$$

οπου K_{ϕ_j} είναι το πλήθος των εγγράφων που περιέχονται στο φύλλο ϕ_j.

Αναλυτικά, τα βήματα του αλγορίθμου kd–subtopic είναι τα ακόλουθα:

Μέθοδος Kd–Subtopic Version 1

- *Είσοδος:* το σύνολο εκπαίδευσης $\mathcal{D} = \{(\mathrm{d_i}, \mathrm{t_{d_i}}), \mathrm{i} = 1, \ldots, |\mathrm{D}|\}$, το οποίο χωρίζεται στα ανεξάρτητα υποσύνολα $D_{t_a}, a = 1, \ldots, |T|$, όπου T το σύνολο των κατηγοριών του προβλήματος ταξινόμησης.
- Καθορισμός του πλήθους $|C_{t_a}|$ των πυρήνων ανά κατηγορία.
- Για κάθε κατηγορία $t_a \in T$ κάνε τα εξής:
  - **Κατασκεύασε το kd–δέντρο για την κατηγορία t_a με $|C_{t_a}|$ φύλλα.**
  - **Αρχικοποίησε τις παραμέτρους θ_{t_a} της μικτής πολυωνυμικής κατανομής $p(d \mid t_a)$ με βάση τις σχέσεις 4.10 και 4.11.**
  - Υπολόγισε την prior πιθανότητα $P(t_a)$ ένα τυχαίο έγγραφο να ανήκει στην κατηγορία t_a, από τη σχέση 3.18.
  - Εφαρμογή EM μέχρις ότου η μεταβολή στην λογαριθμική πιθανοφάνεια $L(\theta_{t_a}^{(t+1)}) - L(\theta_{t_a}^{(t)})$ να γίνει μικρότερη από μία καθορισμένη τιμή:
    * **E–Βήμα :** Υπολόγισε τις posterior $P(c_j \mid d)$ με βάση τη σχέση 3.13.

* **M–Βήμα :** Υπολόγισε τις παραμέτρους θ_{t_a} της μικτής πολυωνυμικής κατανομής $p(d \mid t_a)$ από τις σχέσεις 3.16 και 3.17.

- *Έξοδος:* Ο kd–subtopic ταξινομητής ο οποίος ορίζεται από το σύνολο των παραμέτρων θ (σχέση 3.4).

4.3.2 Αρχικοποίηση των Prior Πιθανοτήτων $P(c_j \mid d)$

Ένας εναλλακτικός τρόπος αρχικοποίησης των παραμέτρων θ_{t_a} της μικτής πολυωνυμικής κατανομής $p(d \mid t_a)$ ή ισοδύναμα, των παραμέτρων $\theta_{w_t|c_j}$ των πυρήνων $c_j \in C_{t_a}$ καθώς και τα αντίστοιχων βαρών τους θ_{c_j} μπορεί να γίνει μέσω της αρχικοποίησης των prior πιθανοτήτων $P(c_j \mid d)$ για όλα τα έγγραφα του συνόλου εκπαίδευσης $d \in D_{t_a}$.

Εκμεταλλευόμαστε την αντιστοιχία μεταξύ των πυρήνων c_j και των φύλλων του kd–δέντρου ϕ_j. Ισχυριζόμαστε ότι τα έγγραφα που περιέχονται στο φύλλο ϕ_j έχουν μεγαλύτερη πιθανότητα να προέρχονται από τον πυρήνα c_j (ο οποίος αντιστοιχεί στο φύλλο ϕ_j) από ότι στους υπόλοιπους πυρήνες $c_r, r \neq j$. Επομένως για κάθε έγγραφο $d \in D_{t_a}$ βρίσκουμε το φύλλο ϕ_j στο οποίο περιέχεται κι επομένως τον πυρήνα c_j και κάνουμε την εξής ανάθεση:

$$P(c_j \mid d) = \begin{cases} B, & \text{αν d περιέχεται στο φύλλο } \phi_j \\ \frac{1-B}{|C_{t_a}|-1}, & \text{αν d δεν περιέχεται στο φύλλο } \phi_j \end{cases} \quad (4.12)$$

όπου B ένας αριθμός στο [0.5 , 1], ο οποίος εκφράζει πόσο ισχυρή είναι η πεποίθηση μας ότι το έγγραφο d το οποίο περιέχεται στο φύλλο ϕ_j προέρχεται από τον πυρήνα c_j.

Αναλυτικά, τα βήματα της εναλλακτικής αυτής μεθόδου είναι τα ακόλουθα:

Μέθοδος Kd–Subtopic Version 2

- *Είσοδος:* το σύνολο εκπαίδευσης $\mathcal{D} = \{(\mathrm{d_i}, \mathrm{t_{d_i}}), \mathrm{i} = 1, \ldots, |\mathrm{D}|\}$, το οποίο χωρίζεται στα ανεξάρτητα υποσύνολα $D_{t_a}, a = 1, \ldots, |T|$, όπου T το σύνολο των κατηγορών του προβλήματος ταξινόμησης.
- Καθορισμός του πλήθους $|C_{t_a}|$ των πυρήνων ανά κατηγορία.
- Για κάθε κατηγορία $t_a \in T$ κάνε τα εξής:
  - **Κατασκεύασε το kd–δέντρο για την κατηγορία t_a με $|C_{t_a}|$ φύλλα.**
  - **Αρχικοποίησε τις prior πιθανότητες $P(c_j \mid d)$ των πυρήνων c_j της κατηγορίας t_a για κάθε έγγραφο d του συνόλου εκπαίδευσης D_{t_a} της κατηγορίας t_a με βάση τη σχέση 4.12.**

- **Υπολόγισε τις παραμέτρους θ_{t_a} της μικτής πολυωνυμικής κατανομής $p(d \mid t_a)$ με βάση τις σχέσεις 3.16 και 3.17.**
- Υπολόγισε την prior πιθανότητα $P(t_a)$ ένα τυχαίο έγγραφο να ανήκει στην κατηγορία t_a, από τη σχέση 3.18.
- Εφαρμογή EM μέχρις ότου η μεταβολή στην λογαριθμική πιθανοφάνεια $L(\theta_{t_a}^{(t+1)}) - L(\theta_{t_a}^{(t)})$ να γίνει μικρότερη από μία καθορισμένη τιμή:
  * **E–Βήμα :** Υπολόγισε τις posterior $P(c_j \mid d)$ με βάση τη σχέση 3.13.
  * **M–Βήμα :** Υπολόγισε τις παραμέτρους θ_{t_a} της μικτής πολυωνυμικής κατανομής $p(d \mid t_a)$ από τις σχέσεις 3.16 και 3.17.

- *Έξοδος:* Ο kd–subtopic ταξινομητής ο οποίος ορίζεται από το σύνολο των παραμέτρων θ (σχέση 3.4).

Κεφάλαιο 5

Αυξητική Μέθοδος Εκπαίδευσης του Subtopic Ταξινομητή

5.1 Γενικά

Η μέθοδος ταξινόμησης subtopic, η οποία μελετήθηκε στο κεφάλαιο 3, εφαρμόζει τον αλγόριθμο ΕΜ για μικτές πολυωνυμικές κατανομές προκειμένου να υπολογίσει τις παραμέτρους θ του μοντέλου παραγωγής εγγράφων, όπως αυτό παρουσιάστηκε στην ενότητα 3.2 . Ο αλγόριθμος ΕΜ είναι ιδιαίτερα δημοφιλής γιατί είναι απλός στην υλοποίηση και εγγυάται την μονότονη αύξηση της πιθανοφάνειας του συνόλου εκπαίδευσης. Ωστόσο, η εκπαίδευση μίας μικτής πολυωνυμικής κατανομής με χρήση ενός τοπικού αλγορίθμου όπως είναι ο ΕΜ χαρακτηρίζεται από τα ακόλουθα μειονεκτήματα:

- Ο πραγματικός αριθμός των πυρήνων είναι συνήθως άγνωστος.
- Δεν υπάρχει γενικά αποδεκτή μέθοδος για την αρχικοποίηση των παραμέτρων του μοντέλου. Διαφορετικές αρχικοποιήσεις οδηγούν σε διαφορετικές εκτιμήσεις των παραμέτρων.
- Ο αλγόριθμος μπορεί να παγιδευτεί σε κάποιο από τα πολλά τοπικά μέγιστα της συνάρτησης της πιθανοφάνειας.

Για το λόγο αυτό, προτείνουμε μία αυξητική μέθοδο για την εκπαίδευση των μικτών πολυωνυμικών κατανομών $p(d \mid t_a), t_a \in \mathcal{T}$, η οποία προσπαθεί να υπερκεράσει αυτούς τους περιορισμούς.

Ας θεωρήσουμε τη γενική περίπτωση κατά την οποία διαθέτουμε ένα σύνολο εγγράφων $\mathcal{D}$ και επιθυμούμε να εκτιμήσουμε τη συνάρτηση πυκνότητας πιθανότητας $p(d)$, βάση της οποίας παράγονται τα έγγραφα του $\mathcal{D}$. Έστω ότι η $p(d)$ είναι μία μικτή πολυωνυμική κατανομή. Ο προτεινόμενος αλγόριθμος βασίζεται σε ανάλογες εργασίες που έχουν γίνει για την επίλυση του προβλήματος εκτίμησης κατανομών με Gaussian μικτά μοντέλα. Έχει αποδειχθεί θεωρητικά στο [49] ότι, υπό προϋποθέσεις, η εκπαίδευση ενός μικτού μοντέλου μεγιστοποιώντας την πιθανοφάνεια μπορεί να επιτευχθεί με έναν αυξητικό τρόπο, προσθέτοντας διαδοχικά πυρήνες στο μοντέλο. Συγκεκριμένα, υποθέτουμε ότι ένας νέος πυρήνας $\phi(d;\theta)$ προστίθεται σε ένα μικτό μοντέλο $f_k(d)$ με k πυρήνες για την δημιουργία του μοντέλου με $k+1$ πυρήνες:

$$f_{k+1}(d) = (1-\alpha)f_k(d) + \alpha\phi(d;\theta), \tag{5.1}$$

όπου $\alpha \in (0,1)$. Αν για κάθε k, δοθέντος του $f_k(d)$, το βάρος α και το σύνολο των παραμέτρων θ του $\phi(d;\theta)$ επιλέγονται βέλτιστα έτσι ώστε η νέα λογαριθμική πιθανοφάνεια

$$L_{k+1} = \sum_{i=1}^{|D|} \log f_{k+1}(d_i) = \sum_{i=1}^{|D|} \log\left[(1-\alpha)f_k(d_i) + \alpha\phi(d_i;\theta)\right] \tag{5.2}$$

να μεγιστοποιείται, τότε για μεγάλο k το τελικό μοντέλο έχει λογαριθμική πιθανοφάνεια σχεδόν τουλάχιστον τόσο μεγάλη όσο κάθε μικτή κατανομή του τύπου

$$f_k(d) = \sum_{j=1}^{k} P(c_j)\phi(d;\theta_j), \tag{5.3}$$

όπου με c_j συμβολίζουμε τον j–πυρήνα του μικτού μοντέλου, $P(c_j)$ είναι η εκ των προτέρων πιθανότητα ένα τυχαίο έγγραφο να έχει παραχθεί από τον πυρήνα c_j και $\phi(d;\theta_j)$ είναι η συστατική κατανομή του πυρήνα c_j με παραμέτρους θ_j. Δηλαδή για κάθε μικτή κατανομή και σύνολο δεδομένων, υπάρχει ένας αριθμός C τέτοιος ώστε η λογαριθμική πιθανοφάνεια που επιτυγχάνεται με τον αυξητικό αλγόριθμο είναι το πολύ C/k μικρότερη από την λογαριθμική πιθανοφάνεια της μικτής κατανομής, όπως αποδεικνύεται στο [49]. Επιπλέον μια αξιοσημείωτη ιδιότητα αυτής της τεχνικής μεγιστοποίησης είναι ότι οι παράμετροι του $f_k(d)$ παραμένουν σταθερές κατά την μεγιστοποίηση της L_{k+1}.

Η σπουδαιότητα αυτού του αποτελέσματος είναι ότι η μεγιστοποίηση της πιθανοφάνειας ενός μικτού πολυωνυμικού μοντέλου μπορεί να αντικατασταθεί από την επαναληπτική εκπαίδευση ενός μικτού μοντέλου $f_{k+1}(d)$ δύο στοιχείων, όπου το πρώτο στοιχείο είναι το παλιό μοντέλο $f_k(d)$ και το δεύτερο

είναι ένας πολυωνυμικός πυρήνας $\phi(d;\theta)$ όπου $\theta = \{\theta_{w_t|c_{k+1}} : w_t \in V\}$ το σύνολο των παραμέτρων του. Αυτό αποτελεί πλεονέκτημα από πρακτική άποψη, αφού ένα μικτό μοντέλο με δύο στοιχεία είναι πιο εύκολο να εκπαιδευτεί από ότι ένα πιο γενικό-μοντέλο. Παρ' όλα αυτά χρειάζονται κατάλληλες τεχνικές αναζήτησης προκειμένου να προσδιοριστούν οι βέλτιστες παράμετροι α και θ, που μεγιστοποιούν την L_{k+1}.

Μια αποδοτική τεχνική που αντιμετωπίζει αυτό το πρόβλημα έχει προταθεί στο [48]. Η μέθοδος χρησιμοποιεί ένα συνδυασμό τοπικής και καθολικής αναζήτησης, κάθε φορά που προστίθεται ένας καινούριος πυρήνας στο μικτό μοντέλο. Εφόσον πρέπει να εκπαιδευτεί ένα μικτό μοντέλο με δύο στοιχεία, γίνεται τοπική αναζήτηση με τον αλγόριθμο EM για να βρεθεί ένα μέγιστο της L_{k+1} ως προς τα α και θ, ενώ οι παράμετροι του $f_k(d)$ παραμένουν σταθερές. Προκειμένου να εφαρμοστεί ο EM, οι παράμετροι α και θ αρχικοποιούνται υλοποιώντας μια καθολική αναζήτηση στον χώρο παραμέτρων. Από την στιγμή που εκτιμηθούν οι παράμετροι του νέου πυρήνα και το βάρος α, εφαρμόζεται και πάλι ο EM για να μεγιστοποιηθεί η L_{k+1} ως προς όλες τις παραμέτρους του μοντέλου. Ο αλγόριθμος που προτείνουμε στην συνέχεια για την άπληστη εκπαίδευση του subtopic ταξινομητή αποτελεί μια τροποποίηση αυτής της μεθόδου για την επίλυση του προβλήματος ταξινόμησης κειμένου.

5.2 Τοπική Αναζήτηση

Θεωρούμε ένα πρόβλημα ταξινόμησης κειμένου με $|\mathcal{T}|$ κατηγορίες και ένα σύνολο εκπαίδευσης $D = \{(d_i, t_{d_i}), i = 1, \ldots, |D|\}$, όπου d_i είναι το i–έγγραφο του συνόλου και $t_{d_i} \in \mathcal{T} = \{\mathrm{t}_1, \ldots, \mathrm{t}_{|\mathrm{T}|}\}$ είναι η κατηγορία του εγγράφου d_i. Το αρχικό σύνολο D μπορεί εύκολα να διαιρεθεί σε $|T|$ ανεξάρτητα υποσύνολα D_{t_a}, έτσι ώστε κάθε υποσύνολο να περιέχει έγγραφα μόνο της αντίστοιχης κατηγορίας.

Προκειμένου να υπολογίσουμε τις παραμέτρους $\theta = \{P(t_1), \theta_{t_1}, \ldots, P(t_{|T|}), \theta_{t_{|T|}}\}$ του ταξινομητή κειμένου, αρκεί να υπολογίσουμε για κάθε κατηγορία $t_a \in T$ την εκ των προτέρων πιθανότητα ένα τυχαίο έγγραφο να ανήκει στην κατηγορία t_a, $P(t_a)$, καθώς επίσης και το σύνολο των παραμέτρων θ_{t_a} της υπό συνθήκη κατανομής $p(d \mid t_a)$ χρησιμοποιώντας το σύνολο εκπαίδευσης της κατηγορίας t_a, D_{t_a}. Κατά συνέπεια, θα μελετήσουμε τον αυξητικό αλγόριθμο EM (greedy EM) για την εκτίμηση της μικτής πολυωνυμικής κατανομής $p(d \mid t_a)$.

Υποθέτουμε ότι η υπό συνθήκη κατανομή $p(d \mid t_a)$ περιγράφεται από ένα μικτό πολυωνυμικό μοντέλο $p_k(d \mid t_a)$ με k πυρήνες και σύνολο παραμέτρων

$\theta^k_{t_a}$. Η $p_k(d \mid t_a)$ ορίζεται κατά τα γνωστά ως:

$$p_k(d \mid t_a; \theta^k_{t_a}) = \sum_{j=1}^{k} P(c_j) p(d \mid c_j). \tag{5.4}$$

Αν προστεθεί ένας νέος πυρήνας $p(d \mid c_{k+1})$ στο υπάρχον μικτό μοντέλο με τους k πυρήνες, τότε η υπό συνθήκη κατανομή της κατηγορίας t_a γίνεται:

$$p_{k+1}(d \mid t_a; \theta^{k+1}_{t_a}) = (1-\alpha) p_k(d \mid t_a; \theta^k_{t_a}) + \alpha p(d \mid c_{k+1}), \tag{5.5}$$

όπου $a \in (0,1)$. Η λογαριθμική πιθανοφάνεια των εγγράφων του συνόλου εκπαίδευσης D_{t_a} μετά την προσθήκη του νέου πυρήνα γίνεται:

$$L_{k+1} = \sum_{i=1}^{|D_{t_a}|} \log p_{k+1}(d_i \mid t_a) = \sum_{i=1}^{|D_{t_a}|} \log \left[(1-\alpha) p_k(d_i \mid t_a) + \alpha p(d \mid c_{k+1})\right] \tag{5.6}$$

Κατά τη φάση της τοπικής αναζήτησης, επιθυμούμε να μεγιστοποιήσουμε τη συνάρτηση της λογαριθμικής πιθανοφάνειας 5.6. Εφόσον οι παράμετροι της $p_k(d \mid t_a)$ παραμένουν σταθερές, μπορούμε να εφαρμόσουμε τον αλγόριθμο EM για την εκτίμηση του βάρους α και των παραμέτρων $\theta = \{\theta_{w_t|c_{k+1}} : w_t \in V\}$ του νέου πυρήνα, χρησιμοποιώντας ως αρχικές τιμές (όπως θα δούμε στην επόμενη ενότητα) αυτές που προέκυψαν από την καθολική αναζήτηση. Τα βήματα του *τμηματικού EM* (partial EM) είναι τα ακόλουθα:

1. Ε–βήμα: Για κάθε έγγραφο d_i του συνόλου εκπαίδευσης D_{t_a} υπολόγισε τις εκ των υστέρων πιθανότητες $P(c_{k+1} \mid d_i)$

$$P(c_{k+1} \mid d_i) = \frac{\alpha p(d_i \mid c_{k+1})}{(1-\alpha) p_k(d_i \mid t_a) + \alpha p(d_i \mid c_{k+1})} \tag{5.7}$$

2. Μ–βήμα: Υπολόγισε τις νέες τιμές των παραμέτρων του k+1 πυρήνα από τις εξισώσεις:

$$\alpha = \frac{1 + \sum_{i=1}^{|D_{t_a}|} P(c_{k+1} \mid d_i)}{(k+1) + |D_{t_a}|} \tag{5.8}$$

$$\theta_{w_t|c_{k+1}} \equiv P(w_t \mid c_{k+1}) = \frac{1 + \sum_{i=1}^{|D_{t_a}|} N(w_t, d_i) P(c_{k+1} \mid d_i)}{|V| + \sum_{s=1}^{|V|} \sum_{i=1}^{|D_{t_a}|} N(w_s, d_i) P(c_{k+1} \mid d_i)} \tag{5.9}$$

Επειδή υπολογίζονται μόνο οι παράμετροι του νέου πυρήνα, η εφαρμογή του *τμηματικού* ΕΜ αποτελεί μία απλή και γρήγορη μέθοδο για την τοπική αναζήτηση ενός μεγίστου της L_{k+1}, χωρίς να χρειαστεί να καταφύγουμε σε άλλες μη γραμμικές μεθόδους βελτιστοποίησης με μεγάλο υπολογιστικό κόστος. Ωστόσο, ο τμηματικός ΕΜ παραμένει ένας τοπικός αλγόριθμος μεγιστοποίησης της πιθανοφάνειας, γεγονός που τονκαθιστά ευαίσθητο στις αρχικές τιμές των παραμέτρων α και θ του νέου πυρήνα. Στην συνέχεια, προτείνουμε μία στρατηγική καθολικής αναζήτησης για την αρχικοποίηση των παραμέτρων του τμηματικού ΕΜ.

5.3 Καθολική Αναζήτηση

Ο στόχος της καθολικής αναζήτησης είναι να βρεθούν κατάλληλες αρχικές τιμές για τις παραμέτρους του νέου πυρήνα $\theta = \{\theta_{w_t|c_{k+1}} : w_t \in V\}$ καθώς και για το βάρος α. Η στρατηγική που ακολουθούμε αποτελείται από τα ακόλουθα βήματα:

- Εφάρμοσε τον τμηματικό ΕΜ για χάθε έγγραφο $d \in D_{t_a}$, αρχικοποιώντας τον νέο πυρήνα στο έγγραφο d, δηλαδή $\theta_{w_t|c_{k+1}} = d_t, t = 1, \ldots, |V|$.
- Βρες το έγγραφο d^* από όλα τα έγγραφα του συνόλου εκπαίδευσης, το οποίο οδηγεί στη μέγιστη λογαριθμική πιθανοφάνεια του συνόλου εκπαίδευσης μετά τον τερματισμό του αντίστοιχου τμηματικού ΕΜ.
- Αρχικοποίησε τον νέο πυρήνα στο έγγραφο d^*, δηλαδή $\theta_{w_t|c_{k+1}} = d^*_t, t = 1, \ldots, |V|$ και θέσε το α ίσο με την τιμή που προέκυψε μετά το τέλος του τμηματικού ΕΜ για το έγγραφο d^*.

Μετά το πέρας της καθολικής αναζήτησης, οι εκτιμήσεις για τις παραμέτρους $\theta_{w_t|c_{k+1}, t \in V}$ του νέου πυρήνα αντιστοιχούν στο έγγραφο d^*, το οποίο οδηγεί στη μέγιστη λογαριθμική πιθανοφάνεια του συνόλου εκπαίδευσης μετά τον τερματισμό του αντίστοιχου τμηματικού ΕΜ.

5.4 Εκπαίδευση του Subtopic Ταξινομητή Κειμένου Με Χρήση του Αυξητικού Αλγορίθμου ΕΜ

Συνοψίζοντας όσα αναφέρθηκαν μέχρι τώρα, προκύπτει η αυξητική μέθοδος για την κατασκευή ενός ταξινομητή κειμένου, την οποία ονομάζουμε ως μέθοδο greedy subtopic. Η μέθοδος greedy subtopic υπολογίζει τις παραμέτρους θ, όπως εκείνες ορίζονται στη σχέση 3.4, εφαρμόζοντας κατάλληλα τον

αυξητικό αλγόριθμο ΕΜ για μικτές πολυωνυμικές κατανομές. Στη συνέχεια παρουσιάζουμε τα βήματα της μεθόδου.

Μέθοδος Greedy Subtopic

- *Είσοδος:* το σύνολο εκπαίδευσης $\mathcal{D} = \{(\mathrm{d_i}, \mathrm{t_{d_i}}), \mathrm{i} = 1, \ldots, |\mathrm{D}|\}$, το οποίο χωρίζεται στα ανεξάρτητα υποσύνολα $D_{t_a}, a = 1, \ldots, |T|$, όπου T το σύνολο των κατηγοριών του προβλήματος ταξινόμησης.
- Καθορισμός του πλήθους $|C_{t_a}|$ των πυρήνων ανά κατηγορία.
- Για κάθε κατηγορία $t_a \in T$ κάνε τα εξής:
  - Υπολόγισε την prior πιθανότητα $P(t_a)$ ένα τυχαίο έγγραφο να ανήκει στην κατηγορία t_a, από τη σχέση:

$$P(t_a) = \frac{1 + |D_{t_a}|}{|T| + |D|} \tag{5.10}$$

  - Αρχικοποίησε το μικτό πολυωνυμικό μοντέλο της υπό συνθήκη κατανομής $p(d \mid t_a)$ χρησιμοποιώντας έναν πυρήνα, τοποθετώντας τον στο κέντρο των εγγράφων του συνόλου εκπαίδευσης D_{t_a}. Δηλαδή: $\theta = E[d]$.
  - Για $k = 2, \ldots, |C_{t_a}|$ κάνε τα εξής:
    * Εκτέλεσε καθολική αναζήτηση μεταξύ όλων των εγγράφων $d \in D_{t_a}$, για την τοποθέτηση του νέου πυρήνα. Αρχικοποίησε τις παραμέτρους θ του νέου πυρήνα στο έγγραφο d^* που μεγιστοποιεί την 5.6 έπειτα από εφαρμογή τμηματικού ΕΜ.
    * Εκτέλεσε τοπική αναζήτηση αρχικοποιώντας τον τμηματικό ΕΜ με την εκτίμηση των παραμέτρων του νέου πυρήνα καθώς και του α, που βρέθηκε κατά την καθολική αναζήτηση.
    * Εφάρμοσε επαναληπτικά τα βήματα του τμηματικού ΕΜ μέχρι να συγκλίνει.
    * Εφάρμοσε τον αλγόριθμο ΕΜ (όπως παρουσιάστηκε στο κεφάλαιο 3) χρησιμοποιώντας όλους τους πυρήνες του μικτού μοντέλου μέχρι να συγκλίνει.
- *Έξοδος:* Ο greedy subtopic ταξινομητής ο οποίος ορίζεται από το σύνολο των παραμέτρων θ, όπως καθορίζεται από τη σχέση 3.4.

Κεφάλαιο 6

Πειραματική Μελέτη

6.1 Προεπεξεργασία ενός Συνόλου Εγγράφων

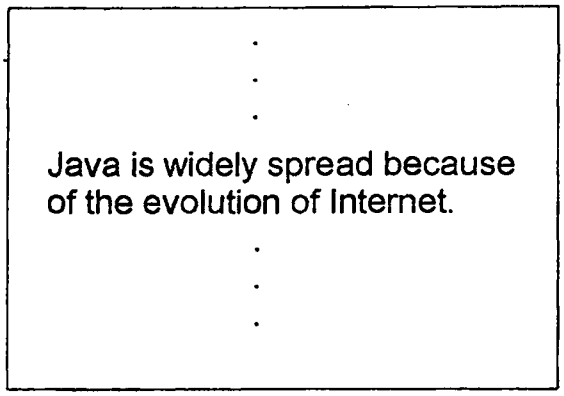
Υποθέτουμε ένα σύνολο εγγράφων (Dataset) $D = \{d_1, \ldots, d_{|D|}\}$. Τα έγγραφα που απαρτίζουν το σύνολο D έχουν μορφή που επιτρέπει στον άνθρωπο να τα αναγνώσει και να επεξεργαστεί τα περιεχόμενά τους. Παράδειγμα της μορφής αυτής παρουσιάζεται στο σχήμα 6.1 . Ωστόσο, η μορφή αυτή δεν είναι κατάλληλη για έναν ταξινομητή κειμένου καθώς, όπως περιγράψαμε αναλυτικά στην ενότητα 1.5, ένα κείμενο $d \in D$ αναπαρίσταται συνήθως ως ένα διάνυσμα από βάρη. Στην εργασία μας, κάθε έγγραφο αναπαρίσταται ως ένα διάνυσμα από εμφανίσεις λέξεων:

$$d = \langle N(w_1, d), \ldots, N(w_{|V|}, d) \rangle, \tag{6.1}$$

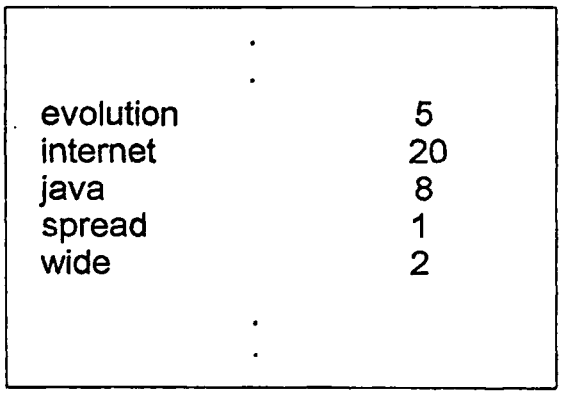
όπου V είναι ένα λεξικό που αποτελείται από τις λέξεις που εμφανίζονται τουλάχιστον μία φορά σε τουλάχιστον ένα έγγραφο του συνόλου εγγράφων D και $N(w_i, d)$ είναι το πλήθος των εμφανίσεων της λέξης w_i στο έγγραφο d. Η επιθυμητή μορφή του εγγράφου d παρουσιάζεται στο σχήμα 6.2, όπου στην πρώτη στήλη του εγγράφου εμφανίζονται οι λέξεις του εγγράφου d, ενώ στη δεύτερη στήλη τα αντίστοιχα πλήθη των εμφανίσεων των λέξεων.

Προκειμένου να φέρουμε τα έγγραφα του συνόλου δεδομένων D από τη μορφή του σχήματος 6.1 σε αυτήν του σχήματος 6.2, πραγματοποιούμε ένα στάδιο *προεπεξεργασίας* των εγγράφων. Το στάδιο αυτό χωρίζεται στις ακόλουθες φάσεις :

1. *Εύρεση λεξικού V:* βρίσκουμε τις λέξεις που εμφανίζονται τουλάχιστον μία φορά σε τουλάχιστον ένα έγγραφο του συνόλου εγγράφων D. Το σύνολο αυτών των λέξεων αποτελεί αρχική μας εκτίμηση για το λεξικό V, το οποίο πρόκειται να χρησιμοποιηθεί για την αναπαράσταση



Σχήμα 6.1: παρουσιάζεται η μορφή ενός συνηθισμένου εγγράφου, όπως αυτό διατίθεται ηλεκτρονικά στο διαδίκτυο. Οι λέξεις που το απαρτίζουν είναι τοποθετημένες κατά τέτοιο τρόπο ώστε ο άνθρωπος που το αναγνώσκει να είναι σε θέση να επεξεργαστεί τα περιεχόμενά του.



Σχήμα 6.2: παρουσιάζεται η επιθυμητή μορφή του εγγράφου του σχήματος 6.1 . Ένας στατιστικός ταξινομητής κειμένου πρέπει να αναπαριστά ένα έγγραφο ως διάνυσμα συχνοτήτων εμφανίσεων λέξεων, κάτι το οποίο είναι εφικτό στο σχήμα αυτό. Στην πρώτη στήλη περιέχονται οι λέξεις του εγγράφου και στη δεύτερη οι αντίστοιχες εμφανίσεις τους.

Σχήμα 6.3: το στάδιο της προεπεξεργασίας (preprocessing) για ένα σύνολο εγγράφων.

των εγγράφων. Χαρακτηριστικό αυτής της φάσης αποτελεί το μεγάλο μέγεθος του λεξικού, γεγονός το οποίο μας οδηγεί στην σταδιακή και επιτακτική μείωση του μέσω των επομένων φάσεων. Πρέπει να σημειωθεί ότι κάθε φάση δέχεται ως είσοδο το λεξικό το οποίο προκύπτει από την προηγούμενή της.

2. *Αφαίρεση συνηθισμένων λέξεων από το λεξικό:* αφαιρούμε από το λεξικό λέξεις οι οποίες δεν παρέχουν κάποια ιδιαίτερη πληροφορία για το περιέχόμενο των εγγράφων όπως άρθρα, αντωνυμίες, προθέσεις, σύνδεσμοι κτλ. Η φάση αυτή λαμβάνει χώρα σχεδόν πάντα.

3. *Δυνατότητα επιλογής της μορφολογικής ρίζας για κάθε λέξη (stemming):* για κάθε λέξη του εγγράφου κρατάμε μόνο τη ρίζα, κάτι το οποίο έχει ως αποτέλεσμα τη συρρίκνωση του λεξικού. Για παράδειγμα ας υποθέσουμε ότι το λεξικό μας περιέχει τις λέξεις "test", "tests", "testing", "tested" οι οποίες έχουν ως ρίζα τη λέξη "test". Τότε, με εφαρμογή αυτής της φάσης διατηρούμε έναν όρο αντί για τέσσερις. Ο αλγόριθμος για την επιλογή της ρίζας που χρησιμοποιήθηκε στην εργασία μας είναι αυτός του Porter [13]. Η φάση αυτή είναι προαιρετική.

4. *Δυνατότητα αφαίρεσης των λέξεων από το λεξικό, οι οποίες περιέχονται σε επικεφαλίδες (headers):* ανάλογα με τη μορφή των εγγράφων κάποιος μπορεί να επιλέξει κάποιο τμήμα τους από το οποίο θα προκύψουν οι λέξεις του λεξικού. Για παράδειγμα, στο σύνολο εγγράφων "20 Newsgroups" [12], κάθε έγγραφο ξεκινάει με επικεφαλίδες (headers) και στη συνέχεια περιέχει το κείμενο που αποτελεί ουσιαστικά το περιεχόμενό του. Επομένως είναι επιθυμητό να επιλεχθούν οι λέξεις από το τμήμα του εγγράφου που προκύπτει με αφαίρεση των επικεφαλίδων που βρίσκονται στην αρχή του. Η φάση αυτή είναι συνίσταται να λαμβάνει χώρα για τα σύνολα των εγγράφων που περιέχουν επικεφαλίδες (headers) (π.χ. έγρραφα σε html μορφή).

5. *Δυνατότητα αφαίρεσης των λέξεων από το λεξικό με βάση τη συχνότητα εμφάνισης τους στα έγγραφα του συνόλου D:* στόχος αυτής της φάσης είναι να αφαιρέσει από το λεξικό λέξεις οι οποίες εμφανίζονται σπάνια στα έγγραφα του συνόλου D. Για το λόγο αυτό ορίζουμε ένα κατώφλι (threshold) το οποίο και συγκρίνουμε με το πλήθος των εμφανίσεων κάθε λέξης του λεξικού σε όλα τα έγγραφα του συνόλου εγγράφων D. Οι λέξεις οι οποίες συνολικά εμφανίζονται στο D λιγότερο από *threshold* φορές αφαιρούνται από το λεξικό.

6. *Κατασκευή ενός νέου συνόλου εγγράφων $\hat{D}$ με βάση το λεξικό που προκύπτει ύστερα από την εφαρμογή των φάσεων 2, 3, 4 και 5:* βασιζόμενοι στο λεξικό, το οποίο τελικά προέκυψε με το πέρας των προηγούμενων φάσεων, παράγουμε ένα νέο σύνολο εγγράφων $\hat{D}$ το οποίο αποτελείται από έγγραφα που έχουν την επιθυμητή μορφή του σχήματος 6.2 .

Στο σχήμα 6.3 παρουσιάζεται το στάδιο της προεπεξεργασίας ενός συνόλου εγγράφων, όπως αυτό υλοποιήθηκε στα πλαίσια της εργασίας.

Στο σημείο αυτό πρέπει να επισημάνουμε, ότι η φάση της προεπεξεργασίας υλοποιήθηκε σε C με τη συνεπικούρηση του πακέτου λογισμικού *Rainbow* [11], το οποίο αναπτύχθηκε από τον Andrew McCallum. Το Rainbow χρησιμοποιήθηκε για την ανάγνωση του συνόλου εγγράφων D και τον εντοπισμό του λεξικού V βάση του οποίου μπορούν να αναπαρασταθούν τα έγγραφα του D.

6.2 Πειράματα

Σε αυτήν την ενότητα συγκρίνουμε τις στατιστικές μεθόδους ταξινόμησης κειμένων Naive Bayes, Subtopic, Kd-Subtopic (με τις δύο εκδόσεις που προτείνουμε) και Greedy Subtopic. Εξετάζουμε τρία σύνολα δεδομένων, το 20 Newsgroups [12] καθώς και τα C_2, C_4 τα οποία συνθέσαμε βασιζόμενοι στο σύνολο 20 Newsgroups. Ακολουθεί μία σύντομη περιγραφή των παραπάνω συνόλων.

20 Newsgroups: Πρόκειται για ένα σύνολο εγγράφων το οποίο είναι ευρέως γνωστό στον κλάδο της ταξινόμησης κειμένου. Αποτελείται από 2000 άρθρα τα οποία συλλέχθησαν από ομάδες συζήτησης (discussion groups) και τα οποία χωρίζονται σε 20 κατηγορίες (πίνακας 6.1). Πολλές από αυτές είναι συγχεόμενες, για παράδειγμα πέντε από αυτές προέρχονται από την ομάδα συζητήσεων comp.* , ενώ τρεις αναφέρονται σε θέματα θρησκείας (alt.atheism, soc.religion.christian και talk.religion.misc). Μετά το τέλος της προεπεξεργασίας (preprocessing), το λεξικό το οποίο προκύπτει και το οποίο χρησιμοποιείται για την αναπαράσταση των εγγράφων του συνόλου 20 Newsgroups, αποτελείται από 3191 λέξεις.

C_2 : Είναι ένα συνθετικό σύνολο εγγράφων το οποίο δημιουργήσαμε με βάση το σύνολο 20 Newsgroups. Αποτελείται από 1600 έγγραφα τα οποία ανήκουν σε δύο κατηγορίες. Η πρώτη κατηγορία περιέχει όλα τα έγγραφα των κατηγοριών comp.graphics, comp.os.ms-windows.misc, comp.sys.ibm.pc.hardware, comp.sys.mac.hardware, sci.crypt, sci.electronics, sci.med και sci.space του συνόλου 20 Newsgroups, ενώ η δεύτερη τα έγγραφα των κατηγοριών rec.autos, rec.motorcycles, rec.sport.baseball, rec.sport.hockey, talk.politics.guns, talk.politics.mideast,

Κατηγορίες Συνόλου Εγγράφων 20 NewsGroups	
alt.atheism	comp.graphics
comp.os.ms-windows.misc	comp.sys.ibm.pc.hardware
comp.sys.mac.hardware	comp.windows.x
misc.forsale	rec.autos
rec.motorcycles	rec.sport.baseball
rec.sport.hockey	sci.crypt
sci.electronics	sci.med
sci.space	soc.religion.christian
talk.politics.guns	talk.politics.mideast
talk.politics.misc	talk.religion.misc

Πίνακας 6.1: Οι κατηγορίες που απαρτίζουν το σύνολο εγγράφων 20 Newsgroups

talk.politics.misc και talk.religion.misc . Μετά το τέλος της προεπεξεργασίας (preprocessing), το λεξικό το οποίο προκύπτει και το οποίο χρησιμοποιείται για την αναπαράσταση των εγγράφων του συνόλου C_2, αποτελείται από 2719 λέξεις.

C_4 : Πρόκειται επίσης για ένα συνθετικό σύνολο εγγράφων το οποίο δημιουργήσαμε βασιζόμενοι στο σύνολο 20 Newsgroups. Αποτελείται από 1600 έγγραφα τα οποία ανήκουν σε τέσσερις κατηγορίες. Η πρώτη κατηγορία περιέχει όλα τα έγγραφα των κατηγοριών comp.graphics, comp.os.ms-windows.misc, comp.sys.ibm.pc.hardware και comp.sys.mac.hardware του συνόλου 20 Newsgroups, η δεύτερη τα έγγραφα των κατηγοριών sci.crypt, sci.electronics, sci.med και sci.space, η τρίτη τα έγγραφα των κατηγοριών rec.autos, rec.motorcycles, rec.sport.baseball και rec.sport.hockey, ενώ η τέταρτη περιέχει τα έγγραφα που ανήκουν στις κατηγορίες talk.politics.guns, talk.politics.mideast, talk.politics.misc και talk.religion.misc . Μετά το τέλος της προεπεξεργασίας (preprocessing), το λεξικό το οποίο προκύπτει και το οποίο χρησιμοποιείται για την αναπαράσταση των εγγράφων του συνόλου C_2, αποτελείται από 2719 λέξεις.

Όσον αφορά την αυξητική μέθοδο εκπαίδευσης του subtopic ταξινομητή (greedy subtopic), εφαρμόσαμε μία τροποποίηση στο στάδιο της καθολικής αναζήτησης. Κατασκευάσαμε ένα kd–δέντρο (kd–tree) για τα έγγραφα κάθε κατηγορίας. Στη συνέχεια προκειμένου να αρχικοποιήσουμε τον νέο πυρήνα που προστίθεται σε κάθε επανάληψη της αυξητικής μεθόδου εφαρμόζουμε τον τμηματικό EM (partial EM) για τα έγγραφα d που αντιστοιχούν στο κέντρα των κόμβων του αντίστοιχου kd–δέντρου, εκτός της ρίζας. Ο νέος πυρήνας αρχικοποιείται στο έγγραφο d^*, το οποίο οδηγεί στη μέγιστη λογαριθμική πιθανοφάνεια του συνόλου εκπαίδευσης μετά τον τερματισμό του αντίστοιχου

τμηματικού ΕΜ. Ο λόγος για τον οποίο δοκιμάζουμε ως υποψήφια έγγραφα για την αρχικοποίηση του νέου πυρήνα τα κέντρα των κόμβων του kd–δέντρου είναι για να εξασφαλίσουμε ότι ο νέος πυρήνας θα τοποθετηθεί αρχικά "κοντά" σε όσο το δυνατόν περισσότερα έγγραφα του συνόλου εκπαίδευσης.

Στη συνέχεια της ενότητας θα αναλύσουμε τα πειραματικά αποτελέσματα των μεθόδων που μελετήθηκαν. Οι μέθοδοι δοκιμάστηκαν στα σύνολα δεδομένων 20 Newsgroups, C_2 και C_4. Για κάθε σύνολο δεδομένων, προκειμένου να έχουμε μία εκτίμηση της ακρίβειας ταξινόμησης (classification accuracy) δηλαδή του ποσοστού των εγγράφων που ταξινομούνται επιτυχώς, χρησιμοποιήσαμε τη μέθοδο 5–fold cross validation [10]. Οι πίνακες 6.2, 6.3 και 6.4 εμφανίζουν τα αποτελέσματα για όλες τις στατιστικές μεθόδους ταξινόμησης κειμένων που μελετήθηκαν για διάφορα πλήθη πυρήνων ανά κατηγορία.

Τα πειραματικά αποτελέσματα πιστοποιούν ότι:

- Η μέθοδος *Naive Bayes* προσφέρει μία αρχική εκτίμηση της ακρίβειας ταξινόμησης στο πρόβλημα στο οποίο εφαρμόζεται. Επιτυγχάνει τη χαμηλότερη απόδοση και στα 3 σύνολα δεδομένων στα οποία τη μελετήσαμε, γεγονός το οποίο οφείλεται στην παραβίαση των υποθέσεων του μοντέλου παραγωγής των εγγράφων το οποίο υϊοθετεί. Ωστόσο, βασικό της προτέρημα αποτελεί ο χαμηλός χρόνος εκτέλεσης, διότι πρόκειται για μη επαναληπτική μέθοδο.

- Η μέθοδος *Subtopic* χρησιμοποιεί το βελτιωμένο μοντέλο παραγωγής εγγράφων της ενότητας 3.2 κι επομένως αναμένουμε καλύτερα αποτελέσματα από τα αντίστοιχα της μεθόδου Naive Bayes. Μελετώντας τους πίνακες 6.2, 6.3 και 6.4 παρατηρούμε ότι ο Subtopic ταξινομητής επιτυγχάνει σημαντική βελτίωση στην ακρίβεια ταξινόμησης της τάξης του 5-10% σε σχέση με τον Naive Bayes ταξινομητή.

- Η μέθοδος *Kd–Subtopic* (και με τις 2 εκδόσεις της Kd–Subtopic Version 1, Kd–Subtopic Version 2) επιχειρεί να επιλύσει το πρόβλημα της αρχικοποίησης που χαρακτηρίζει τον Subtopic ταξινομητή μέσω της κατασκευής και εκμετάλλευσης kd–δέντρων. Τα αποτελέσματα δείχνουν ότι αποφέρει ελαφρώς καλύτερες αποδόσεις σε σχέση με τον Subtopic ταξινομητή ενώ παράλληλα επιτυγχάνει την ανεξαρτησία του επαγόμενου ταξινομητή από την αρχικοποίηση των παραμέτρων του. Όσον αφορά τη σύγκριση των δύο εκδόσεων της Kd–Subtopic μεθόδου, και οι δύο δίνουν εξίσου καλά αποτελέσματα, και το ποια από τις δύο υπερισχύει εξαρτάται από τα χαρακτηριστικά του προβλήματος ταξινόμησης που επιλύουν.

Σύνολο Εγγράφων 20 NewsGroups		
	Αριθμός Πυρήνων	
Μέθοδος Ταξινόμησης	2	4
Subtopic	63.799 %	60.819 %
Kd-Subtopic Version 1	64.366 %	62.599 %
Kd-Subtopic Version 2	64.433 %	63.233 %
Greedy Subtopic	64.766 %	64.766 %
Naive Bayes	55.166 %	

Πίνακας 6.2: Ακρίβεια ταξινόμησης (classification accuracy) στο σύνολο 20 Newsgroups

Σύνολο Εγγράφων C2				
	Αριθμός Πυρήνων			
Μέθοδος Ταξινόμησης	2	4	6	8
Subtopic	90.807 %	91.166 %	90.924 %	90.957 %
Kd-Subtopic Version 1	90.907 %	91.456 %	91.146 %	90.994 %
Kd-Subtopic Version 2	91.066 %	91.097 %	91.160 %	91.124 %
Greedy Subtopic	91.499 %	91.549 %	91.258 %	91.141 %
Naive Bayes	84.499 %			

Πίνακας 6.3: Ακρίβεια ταξινόμησης (classification accuracy) στο σύνολο C_2

- Η αυξητική μέθοδος εκπαίδευσης του Subtopic ταξινομητή, *Greedy Subtopic*, επιτυγχάνει την υψηλότερη απόδοση συγκριτικά με τις προηγούμενες μεθόδους ταξινόμησης κειμένου, γεγονός που επαληθεύεται σε όλες τις εκτελέσεις των πειραμάτων που πραγματοποιήθηκαν. Επιπλέον, όπως και στην περίπτωση της μεθόδου Kd–Subtopic, ο επαγόμενος ταξινομητής είναι ανεξάρτητος από την αρχικοποίηση των παραμέτρων του.

6.3 Μελλοντική Έρευνα

Το πεδίο της στατιστικής ταξινόμησης κειμένων παρουσιάζει πολλά ζητήματα που προσφέρονται για μελλοντική έρευνα. Ενδεικτικά αναφέρουμε τα ακόλουθα:

- *Περαιτέρω διερεύνηση της αυξητικής μεθόδου εκπαίδευσης του Subtopic ταξινομητή (Greedy Subtopic):* Κατά τη διάρκεια των πειραμάτων παρατηρήσαμε ότι η βελτίωση στην επίδοση που αποφέρει η μέθοδος σε σχέση

Σύνολο Εγγράφων C4				
	Αριθμός Πυρήνων			
Μέθοδος Ταξινόμησης	2	4	6	8
Subtopic	83.582 %	83.632 %	83.307 %	83.074 %
Kd-Subtopic Version 1	83.958 %	83.916 %	83.458 %	83.583 %
Kd-Subtopic Version 2	83.957 %	83.849 %	83.516 %	83.499 %
Greedy Subtopic	84.624 %	84.416 %	84.416 %	84.374 %
Naive Bayes	74.833 %			

Πίνακας 6.4: Ακρίβεια ταξινόμησης (classification accuracy) στο σύνολο C_4

με την επίδοση της Subtopic μεθόδου δεν είναι αντίστοιχη σε μέγεθος με τη βελτίωση που σημειώθηκε στην περίπτωση της μοντελοποίησης με χρήση μικτών κανονικών (Gaussian) κατανομών σε συνεχή χώρο δεδομένων. Αυτό οφείλεται αφενός στο γεγονός ότι το πρόβλημα της ταξινόμησης κειμένων χαρακτηρίζεται από τη μεγάλη διάσταση των προτύπων κειμένου κι επομένως στην αναγκαία ύπαρξη μεγάλου πλήθους προτύπων κειμένου για την εκπαίδευση των ταξινομητών κι αφετέρου στο γεγονός ότι η χρήση μικτών πολυωνυμικών κατανομών για τη μοντελοποίηση των εγγράφων των κατηγοριών παρουσιάζει σοβαρά αριθμητικά προβλήματα. Πιο συγκεκριμένα πολλές φορές η εκτίμηση μίας πολυωνυμικής κατανομής για ένα συγκεκριμένο έγγραφο d (σχέση 3.1) προκαλεί αριθμητικά προβλήματα λόγω υπερχείλισης (overflow), δεδομένου ότι προκύπτουν τιμές που δεν μπορούν να αναπαρασταθούν μέσα στα όρια ακρίβειας της υπολογιστικής μηχανής.

- *Εκτενέστερη πειραματική μελέτη:* Οι στατιστικές μέθοδοι που μελετήσαμε συνίσταται να εφαρμοστούν σε περισσότερα σύνολα δεδομένων (datasets) καθώς και σε μεγαλύτερα πλήθη εγγράφων ανά κατηγορία, ώστε να επαχθούν περισσότερα συμπεράσματα σχετικά με την απόδοση των μεθόδων.

- *Βελτίωση του τρόπου κατασκευής των kd-δέντρων:* Ενδιαφέρον παρουσιάζει ο τρόπος κατασκευής των kd-δέντρων. Στην υλοποίηση μας υπολογίζουμε τον ορθογώνιο άξονα στον οποίο σημειώνεται η μέγιστη διακύμανση προκειμένου να διασπάσουμε έναν κόμβο του kd-δέντρου σε δύο παιδιά. Αντί αυτού θα μπορούσαμε να υπολογίσουμε τον άξονα της κύριας κατεύθυνσης (principal direction) με χρήση μίας αποδοτικής μεθόδου ανάλυσης ιδιοτιμών (SVD - Singular Value Decomposition), η οποία ωστόσο θα επιβαρύνει σημαντικά το χρόνο εκτέλεσης της με-

θόδου.

- *Μελέτη για αποδοτικότερη μείωση της διάστασης του προβλήματος ταξινόμησης:* Ένα επιπλέον ζήτημα προς μελλοντική έρευνα αποτελεί η μελέτη για περεταίρω μείωση του μεγέθους του λεξικού V το οποίο χρησιμοποιείται για την αναπαράσταση των διανυσμάτων κειμένου, κατά τέτοιον τρόπο ώστε να μην χάνεται πολύτιμη πληροφορία για τον επαγόμενο ταξινομητή κειμένου.

Βιβλιογραφία

[1] F. Sebastiani, *Machine Learning in Automated Text Categorization*, ACM Computing Surveys, Vol. 34, No. 1, March 2002, pp. 1-47

[2] K. P. Nigam, *Using Unlabeled Data to Improve Text Classification*, PhD Thesis, Carnegie Mellon University, Pittsburgh, May 2001

[3] K. Nigam, A. K. McCallum, S. Thrun and T. Mitchell, *Text Classification from Labeled and Unlabeled Documents using EM*, Machine Learning, Vol. 39, pp. 103-134, 2000

[4] D. Boley, *Hierarchical Taxonomies using Divisive Partitioning*, Technical Report TR-98-012, Department of Computer Science, University of Minnesota, Minneapolis, 1998

[5] D. Boley, *Principal Direction Divisive Partitioning*, Data Mining and Knowledge Discovery, Vol. 2, No. 4, pp. 325-344, 1998

[6] D. Boley, M. Gini, R. Gross, E. Han, K. Hastings, G. Karypis, V. Kumar, B. Mobasher and J. Moore, *Partitioning-Based Clustering for Web Document Categorization*, Decision Support Systems, 1999

[7] A. Likas and N. Vlassis, *A Greedy EM Algorithm for Gaussian Mixture Learning*, Neural Processing Letters 15: pp. 77-87, 2002

[8] K. Blekas, D. I. Fotiadis and A. Likas, *Greedy Mixture Learning for Multiple Motif Discovery in Biological Sequences*, Bioinformatics, Vol. 19, No. 0 2003, pp. 1-11

[9] T. M. Mitchell, *Machine Learning*, McGraw-Hill International Editions, 1996

[10] C. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, 1995.

[11] A. K. McCallum, *Bow: A Toolkit for Statistical Language Modeling, Text Retrieval, Classification and Clustering*, http://www-2.cs.cmu.edu/ mccallum/bow/, 1996

[12] 20 Newsgroups Dataset available via : *http://www-2.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/news20.html*

[13] M. F. Porter, *An algorithm for suffix stripping*, Program, Vol.14 No. 3, pp. 130-137, 1980

[14] Y. H. Li and A. K. Jain, *Classification of Text Documents*, The Computer Journal, Vol. 41, No. 8, 1998

[15] A. McCallum and K. Nigam, A Comparison of Event Models for Naive Bayes Text Classification.

[16] I. Androutsopoulos, J. koutsias, K. V. Chandrinos and C. D. Spyropoulos, *An Experimental Comparison of Naive Bayesian and Keyword-based Anti-spam Filtering with Personal E-mail Messages*, Proceedings of SIGIR-00, 23rd ACM International Conference on Research and Development in Information Retrieval, pp. 160-167, 2000

[17] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz , *A Bayesian approach to filtering junk e-mail*, AAAI'98 Workshop on Learning for Text Categorization, 1998

[18] C. Apte, F. Damerau and S. M. Weiss, *Automated Learning of Decision Rules for Text Categorization*, ACM Transactions on Information Systems, Vol. 12, No. 3, pp. 233-251

[19] L. D. Baker and A. K. McCallum, *Distributional Clustering of Words for Text Classification*, Proceedings of SIGIR-98, 21st ACM International Conference on Research an Development in Information Retrieval, pp. 96-103

[20] S. Chakrabarti, B. Dom and P. Indyk, *Enhanced Hypertext Categorization Using Hyperlinks*, Proceedings of SIGMOD-98, ACM International Conference in Management of Data, pp. 307-318, 1998

[21] C. Clack, J. Farringdon, P. Lidwell and T. Yu, *Autonomous Document Classification for Business*, Proceedings of the 1st International Conference on Autonomous Agents, pp. 201-208, 1997

[22] C. Cleverdon, *Optimizing Convenient Online Access to Bibliographic Databases*, Information Services and Use, Vol. 4, No. 1, pp. 37-47, 1984

[23] W. W. Cohen and Y. Singer, *Context-sensitive Learning Methods for Text Categorization*, ACM Transactions on Information Systems, Vol. 17, No. 2, pp. 141-173, 1999

[24] I. Dagan, Y. Karov and D. Roth, Mistake-Driven Learning in Text Categorization, Proceedings of EMNLP-97, 2nd Conference on Empirical Methods in Natural Language Processing, pp. 55-63, 1997

[25] H. Drucker, V. Vapnik and D. Wu, *Automatic Text Categorization and its Applications to Text Retrieval*, IEEE Transactions on Neural Networks, Vol. 10, No. 5, pp. 1048-1054, 1999

[26] S. T. Dumais and H. Chen, *Hierarchical Classification of Web Content*, Proceedings of SIGIR-00, 23rd ACM International Conference on Research and Development in Information Retrieval, pp. 256-263, 2000

[27] S. T. Dumais, J. Platt, D. Heckerman and M. Sahami, *Inductive Learning Algorithms and Representations for Text Categorization*, Proceedings of CIKM-98, 7th ACM International Conference on Information and Knowledge Management, pp. 148-155, 1998

[28] R. S. Forsyth, *New Directions in Text Categorization*, A. Gammerman Editions, Causal Models and Intelligent Data Manegement, pp. 151-185, 1999

[29] L. Galavotti, F. Sebastiani and M. Simi, *Experiments on the Use of Feature Selection and Negative Evidence in Automated Text Categorization*, Proceedings of ECDL-00, 4th European Conference on Research and Advanced Technology for Digital Libraries, pp. 59-68, 2000

[30] P. J. Hayes, S. P. Weinstein, *CONSTRUE/TIS: A System for Content-Based Indexing of a Database of News Stories*, Innovative Applications of Artificial Intelligence (IAAI), pp.49-64, 1990

[31] T. Joachims, *A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization*, Proceedings of ICML-97, 14th International Conference on Machine Learning, pp. 143-151, 1997

[32] T. Joachims, *Text Categorization with Support Vector Machines: Learning with Many Relevant Features*, Proceedings of ECML-98, 10th European Conference on Machine Learning, pp. 137-142, 1998

[33] W. Lam, M. E. Ruiz and P. Shrinivasan, *Automatic Text Categorization and its Applications to Text Retrieval*, IEEE Transactions on Knowledge and Data Engineering, Vol. 11, No. 6, pp. 865-879, 1999

[34] Ruiz, M. E. and Srinivasan, *P. Hierarchical text categorization using neural networks*, Information Retrieval, Vol. 5, No .1, pp. 87-118, 2002

[35] D. D. Lewis, *Naive Bayes at Forty: The Independence Assumption in Information Retrieval*, Proceedings of ECML-98, 10th European Conference on Machine Learning, pp. 4-15, 1998

[36] R. Lierre and P. Tadepalli, *Active Learning with Committees for Text Categorization*, Proceedings of AAAI-97, 14th Conference of The American Association for Artificial Intelligence, pp. 591-596, 1997

[37] A. K. McCallum and K. P. Nigam, *Employing EM in pool-based active learning for Text Classification*, Proceedings of ICML-98, 15th International Conference on Machine Learning, pp. 350-358, 1998

[38] Y. Yang, *An Evaluation of Statistical Approaches to Text Categorization*, Information Retrieval, Vol. 1, No. 1-2, pp. 69-90, 1999

[39] Y. Yang and X. Liu , *A Re-examination of Text Categorization Methods*, Proceedings of SIGIR-99, 22nd ACM International Conference on Research and Development in Information Retrieval, pp. 42-49, 1999

[40] Y. Yang and J. O. Pedersen, *A Comparative Study on Feature Selection in Text Categorization*, Proceedings of ICML-97, 14th International Conference on Machine Learning, pp. 412-420, 1997

[41] Y. Yang and C. G. Chute, *An Example-Based Mapping Method for Text Categorization and Retrieval*, ACM Transactions on Information Systems, Vol. 12, No, 3, pp. 252-277, 1994

[42] G. McLachlan, D. Peel, *Finite Mixture Models*, Wiley, 2000

[43] K. Fukunaga, *Introduction to Statistical Pattern recognition*, Academic Press, New York, 1990

[44] A. P. Dempster, N. M. Laird and D. B. Rubin, *Maximum Likelihood Estimation from Incomplete Data via the EM Algorithm*, Journal of the Royal Statistical Society B, vol. 39, pp. 1-38, 1977.

[45] R. Redner and H. Walker, *Mixture densities, Maximum Likelihood and the EM Algorithm*, SIAM Review, vol. 26, no. 2, pp. 195-239, 1984.