ΤΕΙ ΔΥΤΙΚΗΣ ΕΛΛΑΔΑΣ

ΣΧΟΛΗ ΔΙΟΙΚΗΣΗΣ ΚΑΙ ΟΙΚΟΝΟΜΙΑΣ

ΤΜΗΜΑ ΔΙΟΙΚΗΣΗΣ ΕΠΙΧΕΙΡΗΣΕΩΝ / ΜΕΣΟΛΟΓΓΙ

Πτυχιακή εργασία

«ΜΕΛΕΤΗ ΤΕΧΝΙΚΩΝ ΑΝΑΓΝΩΡΙΣΗΣ ΣΥΜΠΕΡΙΦΟΡΑΣ ΧΡΗΣΤΩΝ ΣΕ ΚΟΙΝΩΝΙΚΑ ΔΙΚΤΥΑ»

Σωτηρία Κωτσάκη ΑΜ 15647

Επιβλέπουσα Καθηγήτρια : Δρ.Φωτεινή Γριβοκωστοπούλου

Μεσολόγγι 2017

Η έγκριση της πτυχιακής εργασίας από το Τμήμα Διοίκησης Επιχειρήσεων/Μεσολογγίου του ΤΕΙ Δυτικής Ελλάδας δεν υποδηλώνει απαραιτήτως και αποδοχή των απόψεων του συγγραφέα εκ μέρους του Τμήματος.

ΠΕΡΙΛΗΨΗ

Η ταχεία εξάπλωση των Μέσων Κοινωνικής Δικτύωσης τα τελευταία χρόνια, ως μέσο επικοινωνίας και ανταλλαγής πληροφοριών είχε σαν αποτέλεσμα τη δημιουργία συνόλου δεδομένων μεγάλου μεγέθους, τα οποία μπορούν να χρησιμοποιηθούν προκειμένου να αξιοποιηθούν πληροφορίες που σχετίζονται με το περιβάλλον του χρήστη. Ωστόσο, η ευκολία με την οποία τα δεδομένα μπορούν να συλλεχθούν και να αποθηκευτούν έχει δημιουργήσει μια νέα στάση σχετικά με την ανάλυση τους που οφείλεται στους περιορισμούς των υπαρχόντων μεθόδων ανάλυσης δεδομένων στην προσπάθεια αντιμετώπισης των προκλήσεων που έθεσαν οι νέοι τύποι δεδομένων. Κρίνοντας λοιπόν επιτακτική την ανάγκη για αξιοποίηση αυτής της κρυμμένης γνώσης σε αυτό το τεράστιο όγκο δεδομένων κυρίως υπό τη μορφή μη δομημένου κειμένου οδήγησε τους ερευνητές στη δημιουργία ενός νέου πεδίου ανάλυσης δεδομένων , το οποίο ονομάζεται Εξόρυξη Γνώσης από Κείμενο (Text Mining). Στόχος της παρούσα πτυχιακής εργασίας είναι να παρουσιάσει τα Μέσα Κοινωνικής Δικτύωσης και τις μεθόδους εξόρυξης δεδομένων καθώς και να παρουσιάσει βασικές τεχνικές που υπάρχουν για την εξόρυξη γνώσης από κείμενο.

Λέξεις-Κλειδιά: Κοινωνικά Δίκτυα, Εξόρυξη Κειμένου, Τεχνικές εξόρυξης κειμένου, Αλγόριθμοι Μηχανικής Μάθησης

ABSTRACT

The rapid deployment of Social Media Networks in recent years as a mean of communication and information exchange has as a result the creation of a large set of data that can be used in order to exploit information related to the user's environment. However, the ease with which data can be collected and stored has created a new attitude to their analysis due to the limitations of existing data analysis methods in addressing the challenges posed by new types of data. Therefore, judging the need to exploit this hidden knowledge in this vast amount of data that is mostly under the form of unstructured text, led researchers to create a new data analysis field, called Text Mining. The aim of this dissertation is to introduce the reader to the concept of Social Media Networking but also to this new method of data mining as well as to present all the available techniques available for extracting knowledge from text.

Key-Words: Social Media, Text Mining, Techniques of Text Mining, Machine Learning Algorithms'

Περιεχόμενα

ΕΙΣΑΓΩΓΗ

Η τελευταία δεκαετία λόγω της αλματώδης ανάπτυξης του Διαδικτύου επέφερε σημαντικές εξελίξεις στην ανταλλαγή πληροφοριών με αποτέλεσμα, τη δημιουργία πολλών δικτύων για το σκοπό αυτό, με πιο γνωστό το World Wide Web. Ωστόσο, πρόσφατα έκαναν την εμφάνιση τους μια νέα κατηγορία πληροφοριακών δικτύων, τα οποία χρήζουν ιδιαίτερης δημοτικότητας και είναι τα λεγόμενα Κοινωνικά Δίκτυα (Social Media). Η ανάλυση κοινωνικών δικτύων έχει προκαλέσει το έντονο ενδιαφέρον των ακαδημαϊκών σε ένα ευρύ φάσμα επιστημονικών κλάδων από τη κοινωνιολογία, την πολιτική έως το μάρκετινγκ ή την εγκληματολογία καθώς η μελέτη τους βοηθά στο να κατανοηθεί ο τρόπος με τον οποίο οι άνθρωποι επικοινωνούν και συνεργάζονται αλλά βοηθά επιπλέον τους ερευνητές να αναγνωρίσουν τη ροή της γνώσης σε επίπεδο τόσο μεταξύ διαφόρων οργανισμών όσο και μέσα σε κάθε οργανισμό. Καθίσταται, λοιπόν, σαφές πως η μελέτη και ανάλυση αυτών των δικτύων αποτελούν ιδιαίτερης σημασίας καθώς προσφέρουν πληροφορίες που είναι σημαντικές για θέματα δομής και χαρακτηριστικών του δικτύου όπως και θέματα εμπιστοσύνης και διάδοσης πληροφοριών.

Το βασικό πρόβλημα που παρόλα αυτά ανακύπτει είναι, ότι ενώ υπάρχει διαθεσιμότητα ενός τεράστιου όγκου πληροφοριών προς εκμετάλλευση ,το οποίο αυξάνεται συνεχώς, η επεξεργασία και εξαγωγής της χρήσιμης γνώσης από αυτά εξαιτίας του γεγονότος ότι αυτά βρίσκονται με τη μορφή κειμένου, καθιστούν τις παραδοσιακές τεχνικές εξόρυξης δεδομένων μη εφικτές Συνεπώς, είναι αναγκαία η δημιουργία νέων εργαλείων και τεχνικών που θα μπορούν να επεξεργαστούν αυτού του είδους των πληροφοριών και θα εξάγουν τη χρήσιμη γνώση. Στο πλαίσιο αυτό και με τη συνεργασία διαφόρων επιστημονικών πεδίων όπως της στατιστικής, της μηχανικής εκμάθησης, της θεωρίας της πληροφορίας και της εξόρυξης δεδομένων έχει δημιουργηθεί ένα νέο επιστημονικό πεδίο το οποίο καλείται Εξόρυξη Γνώσης από Κείμενο (Text Mining). Στόχος των εργαλείων και των αλγόριθμων της Εξόρυξης Γνώσης από Κείμενο είναι η εύρεση χρήσιμων και κατανοητών προτύπων σε μεγάλες συλλογές εγγράφων. Οι τεχνικές της Εξόρυξης Κειμένου είναι διαδεδομένες όχι μόνο για την εξαγωγή προτύπων από τα Κοινωνικά Δίκτυα αλλά χρήζουν ευρέος αποδοχής και από άλλους επιστημονικούς κλάδους όπως η Βιολογία, οι Επιχειρήσεις και η Εκπαίδευση.

Σκοπός της παρούσας πτυχιακής εργασίας είναι να εμβαθύνει στην έννοια των Κοινωνικών Δικτύων και να γίνει η σύνδεση τους με αυτό το νέο πεδίο εξαγωγής γνώσης της Εξόρυξης Κειμένου για το οποίο θα παρουσιαστούν λεπτομερώς οι τεχνικές και τα εργαλεία

.Αναλυτικότερα θα παρουσιαστούν τα στάδια της εξόρυξη γνώσης από κείμενο, όπου θα σταθούμε στα στάδια της προ-επεξεργασίας των εγγράφων, της μείωσης των διαστάσεων καθώς και στους αλγόριθμους μηχανικής μάθησης καθώς τα τελευταία είναι υψίστης σημασίας για την εξαγωγή της χρήσιμης γνώσης.

Η εργασία αυτή αποτελείται από τρία κύρια κεφάλαια.. Πιο συγκεκριμένα, στο πρώτο κεφαλαίο θα εισαχθούμε στο πεδίο της Εξόρυξης Δεδομένων που αποτελεί τη βάση για τη δημιουργία της Εξόρυξης Γνώσης από Κείμενο καθώς πολλές από τις τεχνικές και τα εργαλεία που χρησιμοποιούνται είναι κοινά για τα δύο πεδία.

Στο δεύτερο κεφάλαιο παρουσιάζεται το κύριο θέμα της εργασίας πού είναι η Εξόρυξη Κειμένου από τα Κοινωνικά Δίκτυα . Το κεφάλαιο αυτό χωρίζεται σε δύο μέρη όπου στο πρώτο μέρος γίνεται εκτενής αναφορά στα Κοινωνικά Δίκτυα, δίνοντας τον ακριβή ορισμό των Κοινωνικών Δικτύων, επιχειρείται μια κατηγοριοποίησης τους ενώ παρουσιάζονται και κάποια από τα δημοφιλέστερα μέσα Κοινωνικής Δικτύωσης. Στο δεύτερο μέρος τώρα, γίνεται οριοθέτηση της έννοιας του Text Mining όπως αυτό έχει παρουσιαστεί στη υπάρχουσα βιβλιογραφία ενώ ακολούθως παρατίθενται αναλυτικά τα βήματα πραγματοποίησης της διαδικασία Text Mining, όπου σε κάθε βήμα αναλύονται όλες οι υπάρχουσες τεχνικές που μπορούν να χρησιμοποιηθούν στην εν λόγω περίσταση όπως οι μέθοδοι προ-επεξεργασίας των εγγράφων ή κάποιοι από τους γνωστότερους αλγόριθμους εξόρυξης κειμένου.

Τέλος, στο τελευταίο κεφάλαιο γίνεται η παρουσίαση και εξαγωγή των αντίστοιχων συμπερασμάτων που έχουν προκύψει από όλη την εργασία..

ΚΕΦΑΛΑΙΟ 1
ΑΝΑΚΑΛΥΨΗ ΓΝΩΣΗΣ ΑΠΟ ΔΕΔΟΜΕΝΑ

1.1 Εισαγωγή

Οι ραγδαίες εξελίξεις στη συλλογή δεδομένων και στη τεχνολογία αποθήκευσης έχουν επιτρέψει στους οργανισμούς να αποθηκεύουν τεράστιες ποσότητες δεδομένων. Ωστόσο, η εξαγωγή χρήσιμων πληροφοριών έχει αποδειχτεί μια τεράστια πρόκληση. Συχνά, οι παραδοσιακές τεχνικές και τα εργαλεία δεν μπορούν να χρησιμοποιηθούν, εξαιτίας του τεράστιου όγκου ενός συνόλου δεδομένων. Μερικές φορές, η μη παραδοσιακή φύση των δεδομένων είναι εκείνη που υποδεικνύει ότι οι κλασσικές προσεγγίσεις δεν μπορούν να εφαρμοστούν, ακόμη και σε σχετικά μικρό σύνολο δεδομένων. Σε άλλες περιπτώσεις, αντίθετα, οι ερωτήσεις που πρέπει να απαντηθούν δεν μπορούν να αντιμετωπιστούν με τις υπάρχουσες τεχνικές ανάλυσης δεδομένων και επομένως, πρέπει να αναπτυχθούν νέες μέθοδοι.

Γενικότερα, οι παραδοσιακές μέθοδοι ανάλυσης δεδομένων συχνά αντιμετώπιζαν πρακτικές δυσκολίες στο να ανταποκριθούν στις προκλήσεις που δημιουργούνταν από τα νέα σύνολα δεδομένων. Οι κυριότερες προκλήσεις που συναντούσαν οι ερευνητές έως τώρα σύμφωνα με τους Tan et al (2005), ήταν οι εξής :

- *Κλιμάκωση (Scalability)*: Λόγω της προόδου στην παραγωγή και συλλογή δεδομένων, τα σύνολα δεδομένων μεγέθους gigabyte, terabyte ή ακόμη και peta-byte γίνονται όλο και πιο συνήθη. Για να είναι σε θέση οι παραδοσιακοί αλγόριθμοι να χειριστούν αυτά τα ογκώδη δεδομένα, θα πρέπει να είναι κλιμακωτοί. Δηλαδή να είναι σε θέση να χρησιμοποιούν ειδικές στρατηγικές αναζήτησης για να διαχειριστούν προβλήματα εκθετικής αναζήτησης.
- *Πολλές Διαστάσεις (High Dimensionality)*: Είναι πλέον σύνηθες οι συλλογές δεδομένων με εκατοντάδες ή χιλιάδες χαρακτηριστικά σε αντίθεση με τα λίγα χαρακτηριστικά που συνηθίζονταν μερικές δεκαετίες πριν. Κλασσικά παραδείγματα έρχονται κυρίως από τη βιο-πληροφορική όπου η πρόοδος στη τεχνολογία μικροσυστοιχιών έχει οδηγήσει στην παραγωγή δεδομένων για τα γονίδια, τα οποία περιέχουν χιλιάδες χαρακτηριστικά. Σε αυτή την περίπτωση οι παραδοσιακές τεχνικές ανάλυσης δεδομένων που αναπτύχθηκαν για λίγα δεδομένα συχνά δεν είναι κατάλληλες για πολυδιάστατα δεδομένα. Επίσης, για ορισμένους αλγόριθμους

ανάλυσης δεδομένων η υπολογιστική πολυπλοκότητα αυξάνεται ραγδαία καθώς αυξάνεται και το πλήθος των διαστάσεων των δεδομένων.

- *Ετερογενή και Πολύπλοκα Δεδομένα (Heterogeneous and Complex Data)*: Οι παραδοσιακές μέθοδοι ανάλυσης δεδομένων συχνά διαχειρίζονται σύνολα δεδομένων, τα οποία περιέχουν χαρακτηριστικά ιδίου τύπου, είτε συνεχή, είτε κατηγορικά. Παραδείγματα, τέτοιων μη παραδοσιακών τύπων δεδομένων αποτελούν τα τελευταία χρόνια, οι συλλογές ιστοσελίδων που περιέχουν ήμι-δομημένο κείμενο και υπερσυνδέσμους ή κλιματολογικά δεδομένα αποτελούμενα από μετρήσεις χρονικών σειρών. Οι τεχνικές που αναπτύσσονται για την εξόρυξη τέτοιων πολύπλοκων αντικειμένων πρέπει να λαμβάνουν υπόψη τις σχέσεις που υπάρχουν μέσα στα δεδομένα.
- *Κυριότητα και Διανομή Δεδομένων (Data Ownership and Distribution)*: Μερικές φορές τα δεδομένα που απαιτούνται για την ανάλυση δεν είναι αποθηκευμένα σε μια μόνο θέση ή δεν αποτελούν ιδιοκτησία κάποιου οργανισμού. Αντίθετα, κατανέμονται γεωγραφικά μεταξύ πηγών που ανήκουν σε διαφορετικές οντότητες. Αυτό απαιτεί την ανάπτυξη κατανεμημένων τεχνικών κάτι που οι μέχρι σήμερα τεχνικές ανάλυσης δεδομένων είναι δύσκολο να διαχειριστούν.
- *Μη παραδοσιακή Ανάλυση (Non-traditional analysis)*: Η παραδοσιακή στατιστική προσέγγιση βασίζεται σε ένα πρότυπο υπόθεσης και ελέγχου. Ουσιαστικά δηλαδή, προτείνεται μια υπόθεση, σχεδιάζεται ένα πείραμα για τη συλλογή δεδομένων, και έπειτα τα δεδομένα αναλύονται σε σχέση με την υπόθεση. Ωστόσο, μια τέτοια διαδικασία είναι χρονοβόρα και οι σύγχρονες εργασίες ανάλυσης δεδομένων συχνά απαιτούν τη δημιουργία και αξιολόγηση πολλών υποθέσεων ταυτόχρονα. Επιπλέον, τα σύνολα δεδομένων συχνά περιλαμβάνουν μη κλασικούς τύπους και κατανομές δεδομένων, με αποτέλεσμα το όλο εγχείρημα να δυσχεραίνει ακόμα περισσότερο.

Υπό αυτές τις συνθήκες είναι αναγκαίο η εύρεση νέων τεχνικών και εργαλείων που θα μετατρέπουν με ένα έξυπνο και αυτοματοποιημένο τρόπο τα δεδομένα σε χρήσιμες πληροφορίες και γνώση. Ο επιστημονικός κλάδος που διακυβεύεται το αντικείμενο αυτό ορίζεται ως Data Mining ή Εξόρυξης Γνώσης από Δεδομένα . Καθότι είναι δύσκολο να καθοριστεί η ακρίβεια του εύρους και των ορίων μελέτης αυτού του πεδίου, παραβλέποντας τις λεπτομέρειες, μπορούμε να ορίσουμε ως Data Mining (Hand et al, 2001):

«*Την ανάλυση των συχνά μεγάλων παρατηρούμενων όγκου δεδομένων με στόχο να ανακαλύψουμε κρυμμένες σχέσεις και να συλλέξουμε τα δεδομένα με νέους τρόπους, καινοτόμους , χρήσιμους και κατανοητούς για τον κάτοχο των δεδομένων*».

Η ιδέα στην οποία στηρίζεται η Εξόρυξη Γνώσης από Δεδομένα είναι η κατασκευή εργαλείων δηλαδή υπολογιστικών προγραμμάτων τα οποία θα χρησιμοποιηθούν με στόχο την εξαγωγή προτύπων και άλλων πληροφοριών. Ουσιαστικά, η εξόρυξη δεδομένων είναι η τεχνολογία που συνδυάζει τις παραδοσιακές μεθόδους ανάλυσης δεδομένων με τους συγχρόνους αλγόριθμους επεξεργασίας μεγάλου όγκου δεδομένων. Επιπλέον είναι εκείνη η οποία έχει ανοίξει ενδιαφέρουσες προοπτικές εξερεύνησης και ανάλυσης νέων τύπων δεδομένων αλλά και ανάλυσης παλαιών τύπων δεδομένων με νέες μεθόδους. Στο εισαγωγικό αυτό κεφάλαιο παρουσιάζεται μια γενική επισκόπηση της εξόρυξης δεδομένων ενώ σκιαγραφούνται θέματα-κλειδιά που θα αποτελέσουν τη βάση για τα επόμενα κεφάλαια της παρούσας πτυχιακής εργασίας.

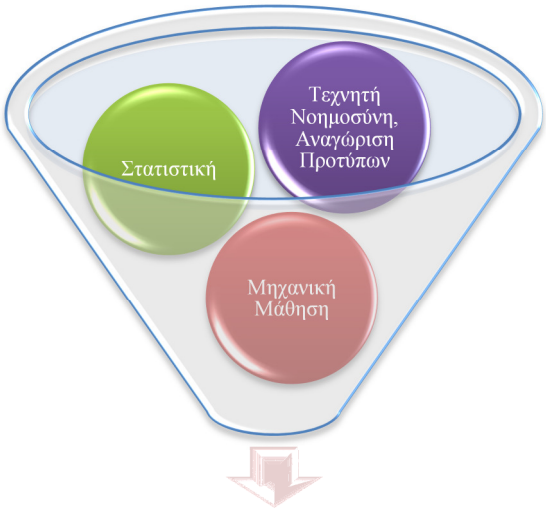
1.2 Η Εξόρυξη ως Στάδιο της Ανακάλυψης Γνώσης σε Βάσεις Δεδομένων

Οι όροι *ανακάλυψη γνώσης σε βάσεις δεδομένων* (Knowledge Discovery in Databases-KDD) και *εξόρυξη γνώσης από δεδομένα* (data mining) συχνά χρησιμοποιούνται εναλλακτικά για την ίδια έννοια. Στην πραγματικότητα έχουν δοθεί πολλές διαφορετικές ονομασίες σε αυτήν τη διαδικασία ανακάλυψης χρήσιμων προτύπων από δεδομένα όπως εξαγωγή γνώσης, ανακάλυψη πληροφοριών, ανάλυση δεδομένων ή συγκομιδή πληροφοριών. Τα τελευταία χρόνια, ο όρος KDD έχει χρησιμοποιηθεί ωστόσο για να περιγράψει μια διαδικασία που αποτελείται από πολλά βήματα , ένα από τα οποία είναι και η εξόρυξη γνώσης από δεδομένα. Σύμφωνα με τον επίσημο ορισμό των Frawley, Piatesky-Shapiro & Matheus (1991), με τον όρο ανακάλυψη γνώσης από βάσεις δεδομένων ορίζεται «*η ντετερμινιστική διαδικασία αναγνώρισης έγκυρων, καινοτόμων, ενδεχομένως χρήσιμων και εν τέλει κατανοητών προτύπων στα δεδομένα*.»

Η διαδικασία KDD συχνά θεωρείται πολύπλοκη καθώς όπως προαναφέρθηκε είναι μία διαδικασία που περιλαμβάνει πολλά και διαφορετικά βήματα. Η είσοδος σε αυτή τη διαδικασία είναι τα δεδομένα , και οι χρήσιμες πληροφορίες που επιθυμούν οι χρήστες είναι η έξοδος. Όμως, ο αντικειμενικός σκοπός δεν είναι πάντα από την αρχή ξεκάθαρος. Η

διαδικασία από μόνης της είναι διαδραστική και συνήθως απαιτείται πολύς χρόνος για την ολοκλήρωση της. Για να διασφαλιστεί η χρησιμότητα και η ακρίβεια των αποτελεσμάτων αυτής της διαδικασίας , συνήθως χρειάζεται η συνεργασία δικών του πεδίου εφαρμογής με ειδικούς της διαδικασίας KDD καθ' όλη τη διάρκεια της διαδικασίας αυτής.

Η εξόρυξη δεδομένων από την άλλη, ως βήμα της διαδικασίας της Ανακάλυψης Γνώσης στρέφει το ενδιαφέρον της κυρίως στις μεθοδολογίες και τις τεχνικές εξόρυξης προτύπων δεδομένων ή τις περιγραφές των δεδομένων από τις μεγάλες αποθήκες δεδομένων. Συνεπώς, περιλαμβάνει μοντέλα συναρμολογήσεων των υπό εξέταση δεδομένων ή εναλλακτικά την εξαγωγή προτύπων από αυτά (Χαλκίδη, Βαζιργιάννης, 2008). Υπάρχει μια μεγάλη συλλογή αλγορίθμων εξόρυξης δεδομένων, πολλοί από τους οποίους χρησιμοποιούν έννοιες και τεχνικές από διαφορετικούς τομείς όπως η στατιστική, η αναγνώριση προτύπων, η μηχανική μάθηση, οι αλγόριθμοι και οι βάσεις δεδομένων.



Σχήμα 1: *Η Εξόρυξη Δεδομένων στη Συμβολή Άλλων Επιστημονικών Πεδίων*

Μια θεμελιώδης ιδιότητα των αλγορίθμων εξόρυξης δεδομένων, και αυτή που διαφοροποιεί τους περισσότερους από αυτούς από άλλες παρόμοιες τεχνικές που υιοθετούνται στη μηχανική μάθηση και τη στατιστική, είναι ότι οι αλγόριθμοι εξόρυξης δεδομένων έχουν σχεδιαστεί με έμφαση στην εξελιξιμότητα όσον αφορά το μέγεθος του

συνόλου δεδομένων εισαγωγής. Πιο συγκεκριμένα, ένας αλγόριθμος εξόρυξης δεδομένων μπορεί να αντιμετωπισθεί ως σύνθεση τριών βασικών συστατικών, το πρώτο εκ των οποίων είναι η *περιγραφή του μοντέλου*. Η περιγραφή του μοντέλου επικεντρώνεται κατά κύριο λόγο στη λειτουργία του μοντέλου όπου καθορίζονται οι βασικοί στόχοι κατά τη διάρκεια της διαδικασίας εξόρυξης δεδομένων καθώς και στην παραστατική μορφή του μοντέλου μέσω της οποίας γίνεται προσπάθεια απεικόνισης του μοντέλου και ερμηνείας του με κατανοητούς όρους. Χαρακτηριστικά, πιο περίπλοκα μοντέλα ταιριάζουν καλύτερα στα δεδομένα αλλά μπορεί να είναι δυσκολότερο να γίνουν κατανοητά και να ανταποκριθούν σε πραγματικές συνθήκες.

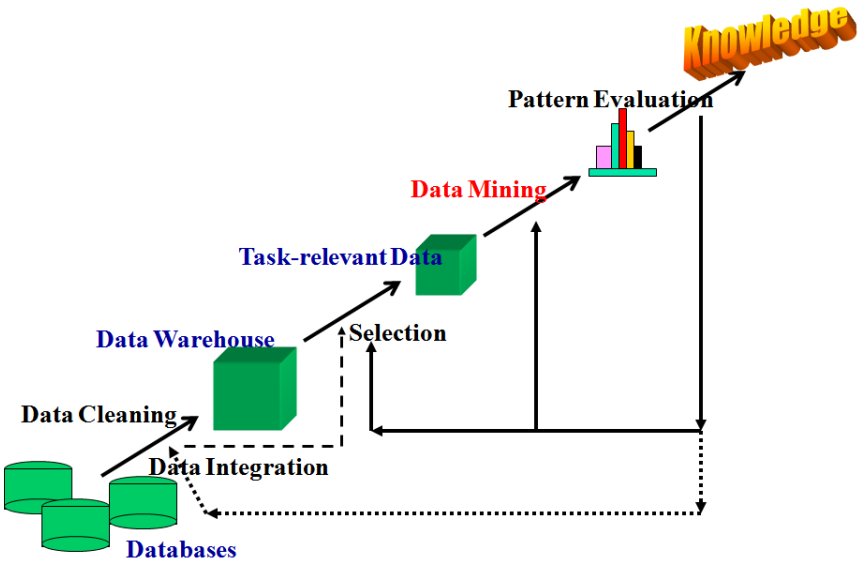
Ως δεύτερο βασικό συστατικό είναι *η αξιολόγηση του μοντέλου*. Με βάση κάποια κριτήρια αξιολόγησης όπως η μέγιστη πιθανότητα καθορίζεται πόσο καλά ένα συγκεκριμένο μοντέλο ταιριάζει με τα κριτήρια της KDD διαδικασίας. Γενικά, η αξιολόγηση του μοντέλου αναφέρεται και στην εγκυρότητα των προτύπων και στην αξιολόγηση της ακρίβειας, της χρησιμότητας και της δυνατότητας κατανόησης του μοντέλου.

Ένας βασικός παράγοντας είναι οι *αλγόριθμοι αναζήτησης*. Αναφέρεται στην προδιαγραφή ενός αλγορίθμου να βρίσκει συγκεκριμένα μοντέλα και παραμέτρους, δοσμένου ενός συνόλου δεδομένων, μιας οικογένειας μοντέλων και ενός κριτηρίου αξιολόγησης. Γενικά υπάρχουν δύο τύποι αλγορίθμων αναζήτησης: 1) Αυτοί που αναζητούν παραμέτρους δηλαδή ψάχνουν για παραμέτρους, οι οποίες βελτιστοποιούν ένα κριτήριο αξιολόγησης για το μοντέλο και 2) αυτοί που αναζητούν μοντέλα και οι οποίοι εκτελούν μια επαναληπτική διαδικασία αναζήτησης για την αντιπροσώπευση των δεδομένων.

Έτσι, προκειμένου να υπάρχει ξεκάθαρη διαφοροποίησης μεταξύ της διαδικασίας και των εργαλείων , θα χρησιμοποιούμε τον όρο της Ανακάλυψης Γνώσης για την περιγραφή του συνόλου της διαδικασίας ανάλυσης και ως Data Mining τις μεθόδους και τις τεχνικές που χρησιμοποιούνται για το σκοπό αυτό. Η εξόρυξη δεδομένων βέβαια ως όρος είναι αυτός που έχει επικρατήσει τελευταία και αναφέρεται στη διαδικασία της εύρεσης δομών γνώσης οι οποίες περιγράφουν με ακρίβεια μεγάλα σύνολα πρωτογενών δεδομένων.

1.3 Η Διαδικασία της Ανακάλυψης Γνώσης

Η διαδικασία της ανακάλυψη γνώσης αποτελείται κυρίως από πέντε βήματα τα οποία και είναι *η επιλογή, η προ-επεξεργασία, ο μετασχηματισμός, η εξόρυξη γνώσης από δεδομένα, η ερμηνεία* και *η αξιολόγηση* , τα οποία και αναλύονται με λεπτομέρεια ακολούθως.



Εικόνα 1: *Στάδια Ανακάλυψης Γνώσης*

Πηγή slidewiki.org

1.3.1 Επιλογή

Σε αυτό το πρώτο στάδιο συλλέγονται τα δεδομένα από διάφορες βάσεις δεδομένων , αρχεία και μη ηλεκτρονικές πηγές. Το σύνολο των δεδομένων αυτών θα αποτελέσουν τη βάση προκειμένου να ανακαλύψουμε την κρυμμένη γνώση. Τα δεδομένα που χρειάζονται για τη διαδικασία αυτή μπορούν να προέλθουν από πολλές και διαφορετικές και ετερογενείς πηγές δεδομένων.

1.3.2 Προ-επεξεργασία

Τα δεδομένα που πρόκειται να χρησιμοποιηθούν κατά τη διαδικασία ίσως να είναι λανθασμένα ή ελλιπή. Ίσως υπάρχουν ανώμαλα δεδομένα από πολλαπλές πηγές που περιλαμβάνουν διαφορετικούς τύπους δεδομένων και διαφορετικές μονάδες μέτρησης. Σε αυτό το βήμα μπορούν να πραγματοποιηθούν πολλές και διαφορετικές δραστηριότητες. Τα

λανθασμένα δεδομένα μπορούν να διορθωθούν ή να αφαιρεθούν, ενώ τα ελλιπή δεδομένα πρέπει να συλλεχθούν ή να εκτιμηθούν.

1.3.3 Μετασχηματισμός

Τα δεδομένα που προέρχονται από διαφορετικές πηγές χρειάζεται να μετατραπούν σε ένα κοινό σχήμα για την περαιτέρω επεξεργασία. Μερικά δεδομένα ίσως απαιτείται να κωδικοποιηθούν ή να μετασχηματιστούν σε πιο χρήσιμα σχήματα. Μπορεί παράλληλα να μειωθεί και ο όγκος των δεδομένων προκειμένου να ελαττωθεί ο αριθμός των πιθανών τιμών των δεδομένων που θα ληφθούν υπόψη.

1.3.4 Εξόρυξη γνώσης από δεδομένα

Ένα από τα σημαντικότερα στάδια στη διαδικασίας ανακάλυψης γνώσης, αποτελεί η εξόρυξη γνώσης από δεδομένα (Data Mining). Με βάση το είδος της εξόρυξης που είναι να εκτελεστεί, σε αυτό το βήμα εφαρμόζονται αλγόριθμοι μηχανικής μάθησης στα δεδομένα που έχουν μετασχηματιστεί από το προηγούμενο στάδιο προκειμένου να προκύψουν τα επιθυμητά αποτελέσματα.

1.3.5 Ερμηνεία/Αξιολόγηση

Σε αυτό το τελευταίο στάδιο της διαδικασίας KDD απαιτείται να γίνει ερμηνεία και αξιολόγηση του μοντέλου που χρησιμοποιήθηκε. Ένα πολύ σημαντικό βήμα σε αυτό, είναι ο τρόπος παρουσίασης των αποτελεσμάτων με τρόπο κατανοητό στο τελικό χρήστη καθώς από αυτό θα κριθεί η χρησιμότητα τους ή μη. Για το σκοπό της παρουσίασης συνήθως χρησιμοποιούνται στρατηγικές οπτικοποίησης των εξαγόμενων προτύπων όπως και γραφικές διεπαφές χρήστη (GUI).

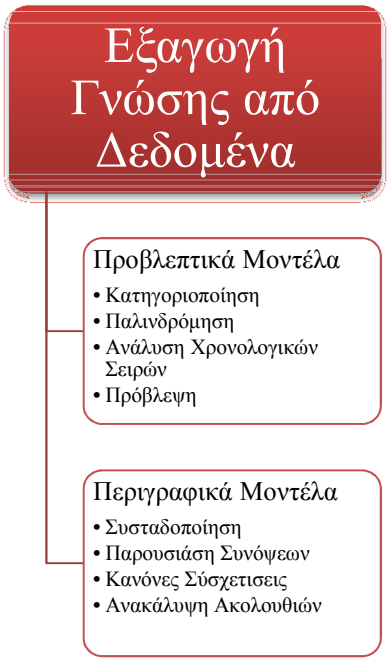
1.4 Βασικές Τεχνικές της Εξόρυξης Δεδομένων

Οι εργασίες της εξόρυξης δεδομένων χωρίζονται γενικά σε δύο βασικές κατηγορίες-μοντέλα, στις λεγόμενες *Προγνωστικές Εργασίες* (Predictive tasks) και στις *Περιγραφικές Εργασίες* (Descriptive tasks) (Tan et al, 2005).

Οι **Προγνωστικές Εργασίες (Predictive tasks)** έχουν ως στόχο να προβλέπουν τη τιμή ενός συγκεκριμένου χαρακτηριστικού βασιζόμενες στις τιμές άλλων χαρακτηριστικών. Το υπό επίβλεψη χαρακτηριστικό είναι γνωστό και ως **στόχο (target)** ή **εξαρτημένη μεταβλητή (dependent variable),** ενώ τα χαρακτηριστικά που χρησιμοποιούνται για να γίνει η πρόβλεψη είναι γνωστά ως **επεξηγηματικές (explanatory)** ή **ανεξάρτητες μεταβλητές (independent variables).**

Οι Περιγραφικές Εργασίες (Descriptive tasks) από την άλλη, αποσκοπούν στο να εξάγουν υποδείγματα (συσχετίσεις, τάσεις, τροχιές) που συνοψίζουν τις βασικές σχέσεις που υπάρχουν στα δεδομένα. Οι περιγραφικές εργασίες της εξόρυξης δεδομένων είναι πολλές φορές από τη φύση τους διερευνητικές και συχνά απαιτούνται τεχνικές μετεπεξεργασίας ώστε να επικυρωθούν και ερμηνευτούν τα αποτελέσματα.

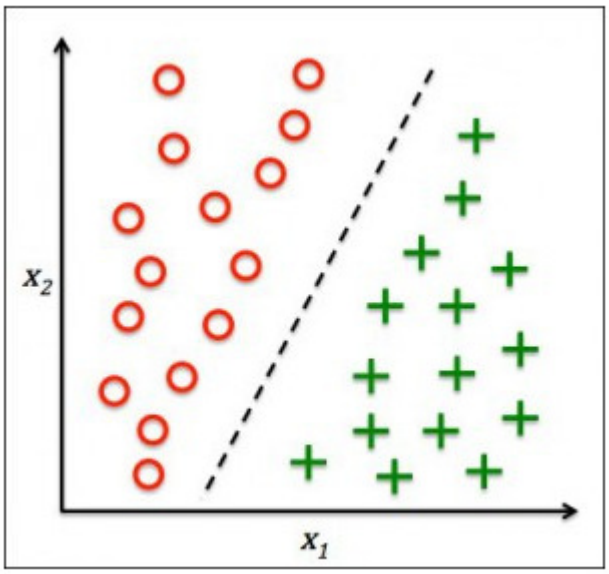
Για κάθε μία από τις βασικές αυτές εργασίες της εξόρυξης δεδομένων υπάρχουν διάφορες τεχνικές οι οποίες χρησιμοποιούνται προκείμενου να επιτευχθεί ο στόχος αυτός, ανάλογα φυσικά και με το βαθμό απαίτησης των εφαρμογών που πρόκειται να εκτελεστούν. Οι τεχνικές αυτές αρχικά εμφανίζονται συνοπτικά στην εικόνα που ακολουθεί ενώ στη συνέχεια αναλύονται.



Σχήμα 2: *Μοντέλα και τεχνικές της εξόρυξη γνώσης από δεδομένα*

- **Κατηγοριοποίηση**

Η κατηγοριοποίηση [1](classification) είναι μία τεχνική της εξόρυξης δεδομένων, κατά την οποία ένα στοιχείο ανατίθεται σε ένα προκαθορισμένο σύνολο κατηγοριών. Ο όρος κατηγοριοποίηση συναντάται στην βιβλιογραφία και ως ταξινόμηση. Γενικότερα, ο στόχος της διαδικασίας αυτής είναι η ανάπτυξη ενός μοντέλου, το οποίο αργότερα θα μπορεί να χρησιμοποιηθεί για την κατηγοριοποίηση μελλοντικών δεδομένων.



Εικόνα 2: *Κατηγοριοποίηση*

Πηγή jaxenter.com

Η κατηγοριοποίηση μπορεί να περιγραφεί ως μία διαδικασία δύο βημάτων:

1. **Εκμάθηση (Learning):**Στο πρώτο βήμα της διαδικασίας δημιουργείται/προσδιορίζεται το μοντέλο με βάση ένα σύνολο προκατηγοριοποιημένων παραδειγμάτων, που ονομάζεται δεδομένα εκπαίδευσης (training data).Τα δεδομένα εκπαίδευσης αναλύονται από ένα αλγόριθμο κατηγοριοποίησης, προκειμένου να σχηματιστεί το μοντέλο. Λόγω του ότι τα δεδομένα εκπαίδευσης ανήκουν σε μία προκαθορισμένη κατηγορία, η οποία είναι γνωστή, η κατηγοριοποίηση αποτελεί μέθοδος εποπτευομένης μάθησης (supervised learning). Το μοντέλο, που λέγεται και αλλιώς κατηγοριοποιητής (classifier), αναπαρίσταται με τη μορφή κανόνων

[1] https://el.wikipedia.org/- Κατηγοριοποίηση

κατηγοριοποίησης (classification rules), δέντρων απόφασης (decision trees) ή μαθηματικών τύπων.

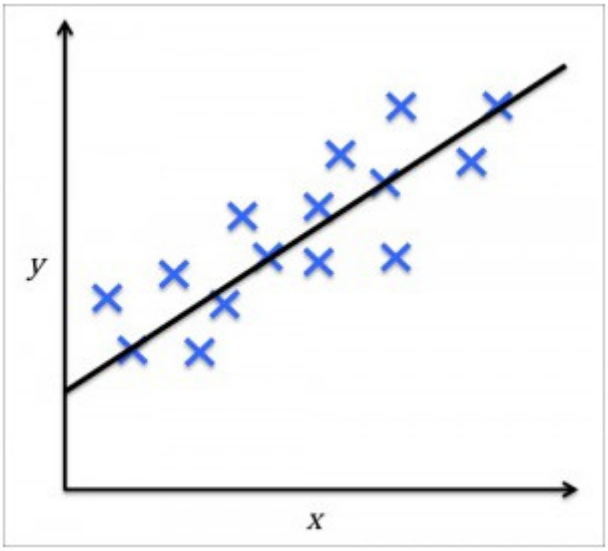
2. **Κατηγοριοποίηση (Classification)**: Μετά την δημιουργία του μοντέλου, το επόμενο βήμα είναι η αξιολόγησή του. Για να επιτευχθεί αυτό, χρησιμοποιούμε τα δοκιμαστικά δεδομένα (test data) για να υπολογίσουν την ακρίβεια του μοντέλου. Το μοντέλο κατηγοριοποιεί τα δοκιμαστικά δεδομένα. Έπειτα, η κατηγορία που σχηματίστηκε με βάση τα δοκιμαστικά δεδομένα συγκρίνεται με την πρόβλεψη που έγινε για τα δεδομένα εκπαίδευσης, τα οποία είναι ανεξάρτητα από αυτά της δοκιμής. Η ακρίβεια του μοντέλου υπολογίζεται από το ποσοστό των δειγμάτων δοκιμής που κατηγοριοποιήθηκαν σωστά σε σχέση με το υπό εκπαίδευση μοντέλο.

Στην περίπτωση που το μοντέλο κριθεί αποδεκτό, τότε μπορεί να χρησιμοποιηθεί για την κατηγοριοποίηση μελλοντικών δειγμάτων δεδομένων, των οποίων η κατηγοριοποίηση είναι άγνωστη. Παραδείγματα μιας τέτοιας τεχνικής αποτελούν η ανίχνευση ανεπιθύμητων μηνυμάτων με βάση την επικεφαλίδα τους ή το περιεχόμενό τους, η πρόβλεψη καρκινικών κυττάρων χαρακτηρίζοντας τα ως καλοήθη ή κακοήθη ή η κατηγοριοποίηση πελατών μιας τράπεζας ανάλογα με την πιστωτική τους ικανότητα.

- **Παλινδρόμηση**

Η παλινδρόμηση (regression) χρησιμοποιείται για να απεικονιστεί ένα στοιχειώδες δεδομένο σε μια πραγματική μεταβλητή πρόβλεψης. Στην πραγματικότητα, η παλινδρόμηση περιλαμβάνει την εκμάθηση της συνάρτησης που κάνει αυτή την απεικόνιση. Η παλινδρόμηση προϋποθέτει ότι τα σχετικά δεδομένα ταιριάζουν με μερικά γνωστά είδη συνάρτησης όπως η γραμμική ή λογιστική παλινδρόμηση και μετά καθορίζει την καλύτερη συνάρτηση αυτού του είδους και μοντελοποιεί τα δεδομένα που έχουν δοθεί. Προκειμένου να γίνει περισσότερο κατανοητό τα όσα προαναφέρθηκαν θα δώσουμε ένα απλό παράδειγμα παλινδρόμησης που αναφέρεται στην περίπτωση της τυπικής γραμμικής παλινδρόμησης. Μία καθηγήτρια πανεπιστημίου επιθυμεί οι αποταμιεύσεις της να φτάσουν σε ένα ορισμένο επίπεδο πριν από τη συνταξιοδότηση της. Περιοδικά, προβλέπει ποιες θα είναι οι αποταμιεύσεις της κατά τη συνταξιοδότηση της βασιζόμενη στην τρέχουσα τιμή τους και σε προηγούμενες τιμές. Χρησιμοποιεί έναν απλό γραμμικό τύπο παλινδρόμησης για να προβλέψει αυτήν την τιμή ταιριάζοντας προηγούμενες συμπεριφορές σε μία γραμμική

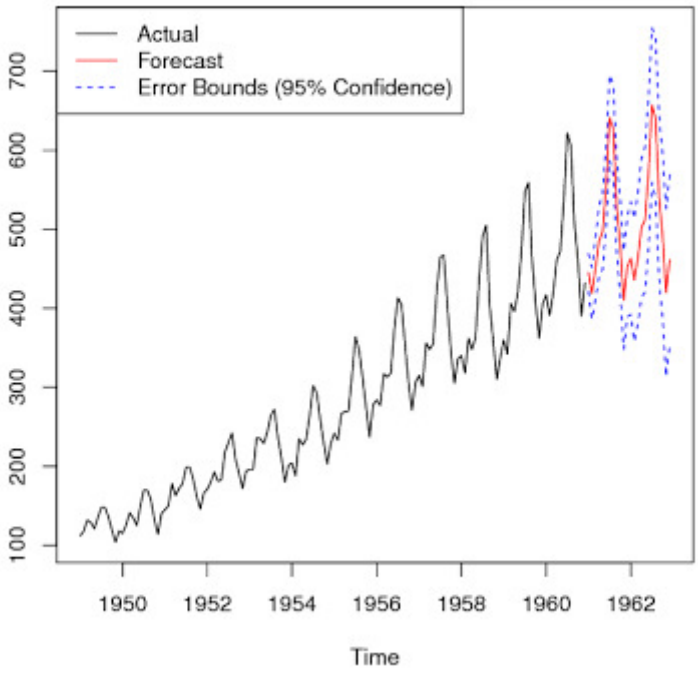
συνάρτηση και στη συνέχεια χρησιμοποιεί αυτή τη συνάρτηση για να προβλέψει τις τιμές σε κάποιες στιγμές στο μέλλον. Βασιζόμενη σε αυτές τις τιμές, στη συνέχεια τροποποιεί το χαρτοφυλάκιο των επενδύσεων της (Dunham,2004) .



Εικόνα 3: *Τεχνική Παλινδρόμησης*
Πηγή jaxenter.com

- **Ανάλυση Χρονοσειρών**

Με την ανάλυση χρονολογικών σειρών ή χρονοσειρών (time series analysis), μελετάται η τιμή ενός γνωρίσματος καθώς μεταβάλλεται στο χρόνο. Οι μέθοδοι ανάλυσης χρονοσειρών αναλύουν τα δεδομένα διαφορετικών χρονικών περιόδων και εξάγουν χρήσιμα συμπεράσματα για το φαινόμενο. Έτσι για παράδειγμα, εάν οι τιμές παρουσιάζουν κανονικότητες στις διακυμάνσεις τους στη διάρκεια του χρόνου, ο εντοπισμός αυτών των διακυμάνσεων μπορεί να χρησιμοποιηθεί για τη διατύπωση προβλέψεων. Το συνηθέστερο παράδειγμα χρονοσειρών είναι ο δείκτης τιμών του χρηματιστήριου (Dunham,2004).



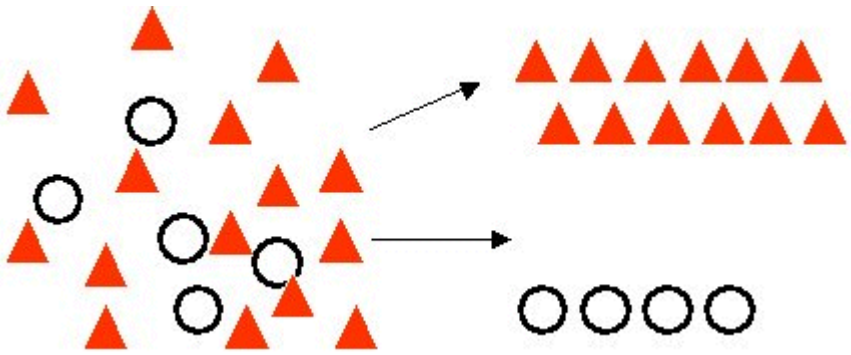
Εικόνα 3: *Ανάλυση Χρονολογικών Σειρών*

Πηγή RDataMining.com

- **Συσταδοποίηση**

Η συσταδοποίηση ή ανάλυση συστάδων (clustering) ομαδοποιεί τα αντικείμενα δεδομένων με βάση μόνο τις πληροφορίες που βρίσκονται στα δεδομένα και που περιγράφουν τα αντικείμενα και τις σχέσεις τους. Ο στόχος είναι τα αντικείμενα μιας ομάδας να είναι όμοια μεταξύ τους και διαφορετικά με αντικείμενα άλλων ομάδων. Όσο πιο μεγάλη η ομοιότητα εντός μια ομάδας και όσο πιο μεγάλη είναι η διαφορά μεταξύ των ομάδων, τόσο πιο καλή ή πιο διακριτή είναι η συσταδοποίηση . Η ανάλυση συστάδων σχετίζεται με άλλες τεχνικές που χρησιμοποιούνται για το διαχωρισμό των αντικειμένων δεδομένων σε ομάδες. Για παράδειγμα, η συσταδοποίηση πολλές φορές συγχέεται με την έννοια της κατηγοριοποίησης που προαναφερθήκαμε υπό την έννοια ότι δημιουργεί έναν προσδιορισμό αντικειμένων με ετικέτες κατηγοριών. Ωστόσο, λαμβάνει αυτές τις ετικέτες μόνο από τα δεδομένα. Η έννοια της κατηγοριοποίησης όπως ορίστηκε προηγουμένως καλείται ως εποπτευόμενη κατηγοριοποίηση καθώς στα νέα δεδομένα, χωρίς ετικέτα, αντικείμενα αποδίδεται μια ετικέτα κατηγορίας χρησιμοποιώντας ένα μοντέλο που αναπτύχθηκε από

αντικείμενα με γνωστές ετικέτες κατηγορίας. Αντίθετα, στην ανάλυση συστάδων οι κατηγορίες δεν είναι προκαθορισμένες για αυτό και η συσταδοποίηση εναλλακτικά καλείται και ως μη εποπτευόμενη κατηγοριοποίηση.

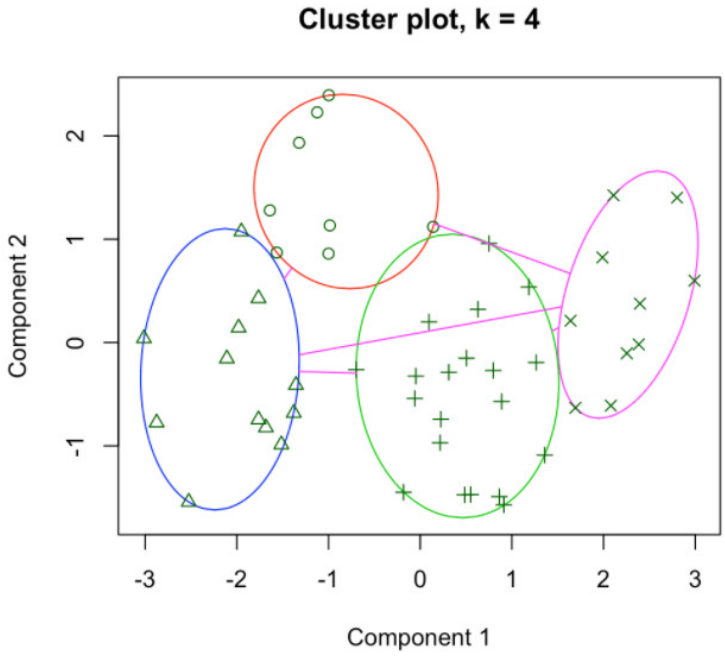


Εικόνα 4: *Τεχνική Συσταδοποίησης*
Πηγή http://ehindistudy.com

Σύμφωνα με τους Hand et al, (2001) διακρίνονται τρεις βασικές κατηγορίες μεθόδων ανάλυσης συστάδων:

1. ***Μέθοδοι Διαχωρισμού (partioning methods)***: από ένα αρχικό σύνολο n δεδομένων δημιουργούνται k ομάδες όπου η κάθε ομάδα αντιπροσωπεύει μία συστάδα για την οποία θα πρέπει να ισχύει ότι κάθε συστάδα περιέχει τουλάχιστον ένα αντικείμενο και κάθε αντικείμενο ανήκει σε μία μόνο συστάδα.

Οι αλγόριθμοι αυτής της κατηγορίας ονομάζονται διαιρετικοί και λειτουργούν κατασκευάζοντας σε μια βάση δεδομένων D που αποτελείται από n αντικείμενα, ένα σύνολο k συστάδων. Ο αλγόριθμος συνήθως ξεκινάει με μια αρχική διάσπαση της βάσης δεδομένων και εν συνεχεία κάνει χρήση μιας στρατηγικής για τη βελτιστοποίηση της αντικειμενικής λειτουργίας. Συνηθίζεται, κάθε συστάδα να αντιπροσωπεύεται από το κέντρο της ή από ένα από τα n αντικείμενα που βρίσκονται κοντά στο κέντρο της. Οι αλγόριθμοι αυτού του είδους εκτελούνται κατά κύριο λόγο σε δύο στάδια. Στο πρώτο στάδιο γίνεται ο καθορισμός των k αντιπροσώπων ενώ στο δεύτερο κάθε αντικείμενο ανατίθεται στη συστάδα που είναι σε πιο κοντινή απόσταση με αυτό.

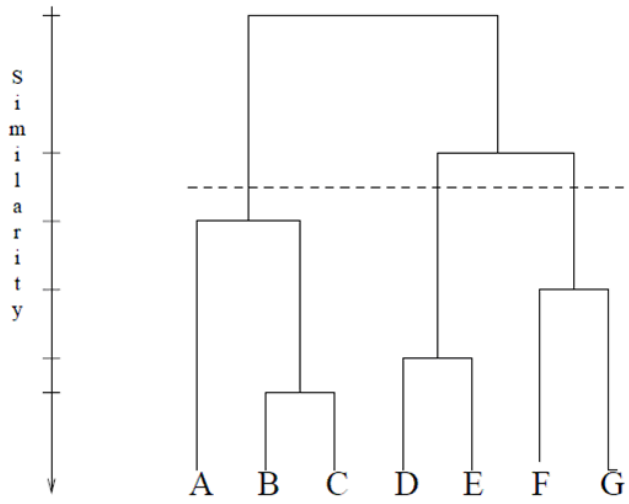


Εικόνα 5: *Μέθοδος Διαχωρισμού*

Πηγή STHDA.com

2. ***Ιεραρχικές Μέθοδοι (Hierarchical clustering):*** Στις μεθόδους αυτές το αρχικό σύνολο δεδομένων διασπάται, δημιουργώντας μια ιεραρχική δομή από συστάδες. Ανάλογα με τη μέθοδο διάσπασης οι ιεραρχικοί μέθοδοι διακρίνονται σε agglomerative και divisive.

Οι ιεραρχικοί αλγόριθμοι αποσυνθέτουν μια βάση δεδομένων σε ένα σύνολο από φωλιασμένες συστάδες που είναι οργανωμένες σαν δέντρο. Κάθε κόμβος (συστάδα) στο δέντρο είναι μια ένωση των παιδιών του (υποσυστάδες) και η ρίζα του δέντρου αποτελεί τη συστάδα που περιέχει όλα τα αντικείμενα. Το δενδρογράφημα αυτό να δημιουργηθεί είτε από τα δέντρα στη ρίζα (agglomerative) είτε από τη ρίζα στα φύλλα (divisive).



Εικόνα 6: *Δενδογράφημα*

Πηγή Wikibooks

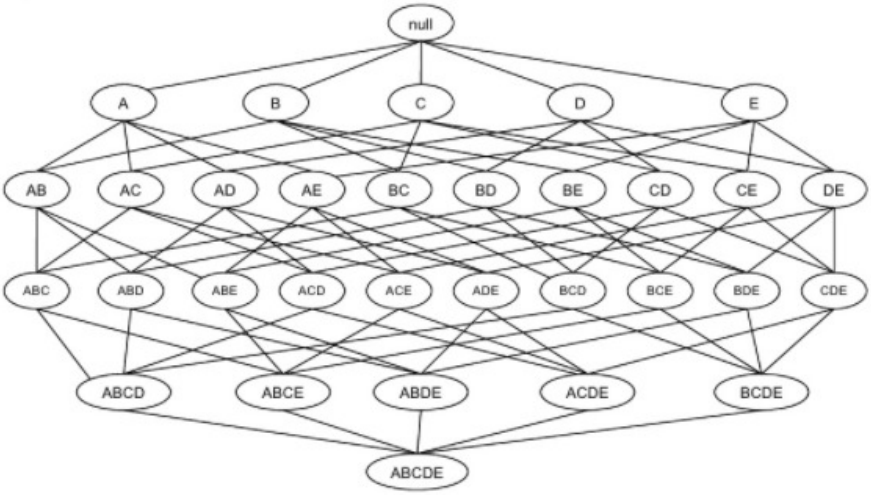
3. ***Μέθοδοι βασισμένες σε Μοντέλα (model-based methods):*** Στην περίπτωση αυτή υποθέτεται ότι κάθε συστάδα μπορεί να αντιπροσωπευθεί από ένα μαθηματικό μοντέλο για αυτό και στις μεθόδους αυτούς προσπαθούν να εντοπιστούν τα αντικείμενα που ανήκουν σε κάθε συστάδα ώστε να υπάρχει αντιστοίχιση με το σχετικό μαθηματικό μοντέλο. Οι τεχνικές που χρησιμοποιούνται σε αυτή τη κατηγορία ανάλυσης συστάδων συνήθως χρησιμοποιούν για τη λειτουργία τους μεθόδους που βασίζονται στη στατιστική και στα νευρωνικά δίκτυα.

- **Παρουσίαση Συνόψεων**

Η παρουσίαση συνόψεων (summarization) απεικονίζει τα δεδομένα σε υποσύνολα τους με συνοδευτικές απλές περιγραφές. Η σύνοψη των δεδομένων ονομάζεται επίσης και *χαρακτηρισμός* ή *γενίκευση*. Εξάγει ή παράγει αντιπροσωπευτικές πληροφορίες σχετικά με τις βάσεις δεδομένων. Αυτό γίνεται ανακτώντας, στην πραγματικότητα, τμήματα από τα δεδομένα. Εναλλακτικά, μπορούν να εξαχθούν από τα δεδομένα συνοπτικές πληροφορίες. Εν ολίγοις, η παρουσίαση συνόψεων χαρακτηρίζει τα περιεχόμενα της βάσης δεδομένων.

- **Κανόνες Συσχέτισης**

Η ανάλυση συνδέσμων που εναλλακτικά αναφέρεται και ως *ανάλυση συγγένειας* ή *συσχέτιση*, αναφέρεται στη διαδικασία εκείνη της εξόρυξη γνώσης που αποκαλύπτει συσχετίσεις μεταξύ των δεδομένων. Το καλύτερο παράδειγμα αυτού του είδους της εφαρμογής είναι ο προσδιορισμός κανόνων συσχετίσεων. Ένας κανόνας συσχέτισης (association rule) είναι ένα μοντέλο που αναγνωρίζει ειδικούς τύπους συσχέτισης μεταξύ των δεδομένων. Η χρήση των κανόνων συσχετίσεων για τις όποιες αποφάσεις πρέπει να γίνεται με πολύ προσοχή καθώς υπάρχει κίνδυνος σε αυτές τις συσχετίσεις να είναι τυχαίες και να μην αντιπροσωπεύουν καμία έμφυτη σχέση ανάμεσα στα δεδομένα.



Εικόνα 6: *Κανόνες Συσχέτισης*

Πηγή Orriols A.

- **Ανακάλυψη Ακολουθιών**

Η ανακάλυψη ακολουθιών ή ακολουθιακή ανακάλυψη χρησιμοποιείται ως μία τεχνική καθορισμού σειριακών προτύπων στα δεδομένα. Συνήθως πρόκειται για πρότυπα τα οποία βασίζονται σε μια χρονική ακολουθία ενεργειών, τα οποία αν και παρουσιάζουν συσχετίσεις εξαιτίας του γεγονότος ότι συσχετίζονται τα δεδομένα, η συσχέτιση τους αυτή βασίζεται στο χρόνο.

ΚΕΦΑΛΑΙΟ 2
ΑΝΑΚΑΛΥΨΗ ΓΝΩΣΗΣ ΑΠΟ ΤΑ ΚΟΙΝΩΝΙΚΑ ΔΙΚΤΥΑ

2.1 Κοινωνικά Δίκτυα

Η εξέλιξη του Παγκόσμιου Ιστού τη τελευταία δεκαετία και συγκεκριμένα η νέα εποχή του Web 2.0, χαρακτηρίζεται από την εξάπλωση του διαδικτύου, την πληθώρα των διαδικτυακών εφαρμογών, κυρίως όμως από την απίστευτη ευκολία στη δημιουργία περιεχομένου και την αξιοποίηση του Παγκόσμιου Ιστού ως μία πλατφόρμα συνεργασίας και συμμετοχής μεταξύ των χρηστών. Σε αντίθεση με το Web 1.0, την εποχή του Web 2.0 καταργούνται οι ρόλοι «συγγραφέα» και «αναγνώστη». Ο Παγκόσμιος Ιστός πέρασε στην εποχή του Read-Write Web. Οι χρήστες έχουν τη δυνατότητα να παράγουν περιεχόμενο (User Generated Content – UGC) και να συμμετέχουν στη συγγραφή ιστοσελίδων, με ποικίλους τρόπους. Ένας από αυτούς τους τρόπους αποτελούν και τα κοινωνικά δίκτυα. Από την είσοδο στην αγορά του πρώτου αναγνωρίσιμου δικτύου, του Six-Degrees το 1997 (Boyd & Ellison, 2007), ένας μεγάλος αριθμός κοινωνικών δικτύων (Online Social Networks) όπως το Facebook, το Twitter, το Instagram και το LinkedIn, έχουν μετατραπεί σε δημοφιλή διαδικτυακές πλατφόρμες όπου οι άνθρωποι επικεντρώνονται και έρχονται σε επαφή, με αποτέλεσμα να χρήζουν ιδιαίτερης δημοτικότητας.

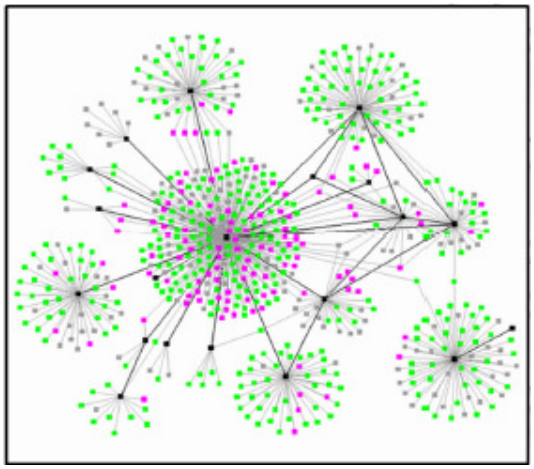
Ως κοινωνικό δίκτυο σύμφωνα με τον ορισμό των Walker, MacBride & Vachon (1977), θεωρείται το άθροισμα των προσωπικών επαφών μέσω των οποίων το άτομο διατηρεί την κοινωνική του ταυτότητα, λαμβάνει συναισθηματική υποστήριξη, υλική ενίσχυση και συμμετοχή στις υπηρεσίες, έχει πρόσβαση στις πληροφορίες και δημιουργεί νέες κοινωνικές επαφές. Τα κοινωνικά δίκτυα αναφέρονται ουσιαστικά στις κοινωνικές σχέσεις του ατόμου, στον τρόπο με τον οποίο αυτά αντιλαμβάνονται και αξιολογούν τις εν λόγω σχέσεις (Παπάνης, Γιαβρίμης, Βίκη & Παπάνης, 2011). Κοινωνική δικτύωση είναι η σύσταση και αξιοποίηση κοινοτήτων ανθρώπων με κοινά ενδιαφέροντα. Γενικότερα, ως κοινωνικό δίκτυο (social network) ορίζεται κάθε δίκτυο σχέσεων και αλληλεπιδράσεων. Οι κόμβοι απαρτίζονται από δρώντες (actors) ή μέλη και οι ακμές απαρτίζονται από τις σχέσεις ή τις αλληλεπιδράσεις μεταξύ των μελών (Σωτηριάδου & Παπαδάκης, 2012).

Εναλλακτική σημασία των κοινωνικών δικτύων δίνεται μέσα από τον όρο «ιστοσελίδες κοινωνικής δικτύωσης», όπου ο όρος «δικτύωση» σημαίνει, κυρίως, την έναρξη σχέσεων μεταξύ άγνωστων ή γνωστών ατόμων. Αφενός, η έναρξη σχέσεων με άγνωστα άτομα δεν

θεωρείται αδύνατη, αφετέρου, δεν είναι η πρωτεύουσα τακτική για την πλειοψηφία των χρηστών (Σωτηριάδου & Παπαδάκης, 2012). Τα κοινωνικά δίκτυα είναι σχεδιασμένα έτσι, ώστε να παρέχουν εύκολη πρόσβαση σε όλους, ανεξαρτήτου ηλικίας, φύλου, εθνικότητας και μορφωτικού επιπέδου. Επιπλέον, είναι διαθέσιμα ανά πάσα στιγμή και από διαφορετικά μέσα (laptop, netbook, smartphone, iPad κτλ.). Υπό την έννοια αυτή και ο όρος «κοινωνική δικτύωση» (social networking), που χρησιμοποιείται συχνά συγχέεται λανθασμένα με τον όρο «social media». Ο όρος «κοινωνικά μέσα» (social media) αναφέρεται στα μέσα διαμοιρασμού πληροφορίας, ενημέρωσης και κοινωνικής δικτύωσης και αξιοποιούν τεχνολογίες Web 2.0 των οποίων η φιλοσοφία βασίζεται στη δημιουργία και ανταλλαγή περιεχομένου από τους χρήστες και στη μεταξύ τους αλληλεπίδραση και υλοποιούν πτυχές της κοινωνικής δικτύωσης. Αυτό φαίνεται και στη συνέχεια μέσα από παραδείγματα συγκεκριμένων μέσων κοινωνικής δικτύωσης.

Τα «κοινωνικά μέσα» αποτελούν λοιπόν διαδικτυακές υπηρεσίες, οι οποίες επιτρέπουν στους χρήστες να:

- δημιουργούν ιδιωτικό ή δημόσιο προφίλ, το οποίο οριοθετείται από το κάθε σύστημα,
- να δημιουργούν λίστες από άλλους χρήστες, οι οποίοι να διαμοιράζονται τη σύνδεση,
- να έχουν τη δυνατότητα να περιηγούνται και να μεταφέρουν τις λίστες των ιδίων αλλά και εκείνων που δημιουργήθηκαν από άλλους χρήστες του ιδίου συστήματος και να διαμοιράζουν περιεχόμενο με τους άλλους χρήστες.



Εικόνα 7: *Απεικόνιση ενός Κοινωνικού Δικτύου*

Πηγή Google

2.2 Ταξινόμηση Κοινωνικών Δικτύων

Οι χρήσεις, όσο και τα μέσα κοινωνικής δικτύωσης αυξάνονται συνεχώς και ίσως αυτός είναι ο λόγος που πολλοί ερευνητές έχουν προσπαθήσει να τα κατηγοριοποιήσουν χρησιμοποιώντας διαφορετικά κριτήρια κάθε φορά. Στη συνέχεια ακολουθεί ο πιο συνηθισμένος τρόπος κατηγοριοποίησης των μέσων κοινωνικής δικτύωσης όπως αυτός παρουσιάστηκε από την Mirma Bard (2010).

Σύμφωνα με αυτήν την κατηγοριοποίηση, τα μέσα κοινωνικής δικτύωσης χωρίζονται στις εξής κατηγορίες όπως φαίνεται και στον πίνακα που ακολουθεί:

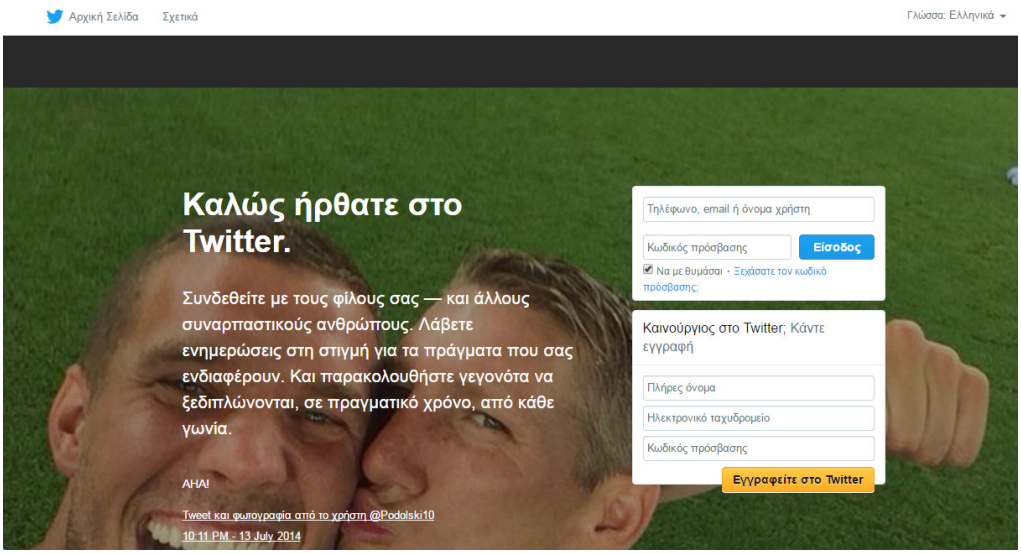
Κατηγορία	• Παραδείγματα
Ήχος – Audio	• i-Tunes
Διαμοίραση φωτογραφιών – Photo Sharing	• Instagram
Εκδόσεις – Publishing	• Wordpress
Μικροϊστολόγια – Microblogging	• Twitter
Απευθείας μετάδοση – Live-casting	• Justin-Tv
Βίντεο – Video	• Youtube
Συγκέντρωση πληροφορίας – Aggregation	• Digg
Wikis	• Wikipedia
Κοινωνική δικτύωση – Social Networking	• Facebook
Αναζήτηση	• Google
RSS	• FeedBurner

Πίνακας 1: *Κατηγοριοποίηση Κοινωνικών Δικτύων*

Στη συνέχεια παρουσιάζουμε λίγο πιο αναλυτικά κάποια από τα πιο αντιπροσωπευτικά παραδείγματα social media όπως είναι το Twitter ,το Facebook, το LinkedIn και το Youtube.

2.2.1 Twitter

Το Twitter (Τουίτερ) όπως παρουσιάστηκε και στον Πίνακα 1 αποτελεί έναν ιστόχωρο κοινωνικής δικτύωσης που εμπίπτει στην κατηγορία του micro-blogging και επιτρέπει στους χρήστες του να στέλνουν και να διαβάζουν σύντομα μηνύματα (μέχρι 140 χαρακτήρες), τα οποία ονομάζονται τουίτς (tweets). Τα μηνύματα μπορούν να αναγνωστούν και από μη συνδεδεμένους χρήστες, αλλά μόνο οι συνδεδεμένοι μπορούν να δημοσιεύσουν κείμενα (Wikipedia). Ο συγκεκριμένος χώρος αποτέλεσε δημιουργία του Τζακ Ντόρσεϊ και «ανέβηκε» για πρώτη φορά τον Ιούλιο του 2006. Από τότε και έπειτα η υπηρεσία έγινε γρήγορα δημοφιλής και σήμερα έχει πάνω από 305 εκατομμύρια ενεργούς χρήστες (2015). Είναι ένας από τους δέκα πιο δημοφιλείς ιστότοπους του διαδικτύου ενώ έρχεται δεύτερος σε ότι αφορά τα social media.



Εικόνα 8: *Διαδικτυακός Χώρος Twitter*

Πηγή twitter.com

2.2.2 Facebook

Το Facebook είναι ένας ιστοχώρος κοινωνικής δικτύωσης που ξεκίνησε τη λειτουργία του στις 4 Φεβρουαρίου του 2004, δημιουργός του οποίου είναι ο Μαρκ Ζάρκεμπεργκ. Οι χρήστες μπορούν να επικοινωνούν μέσω μηνυμάτων με τις επαφές τους αλλά και να αλληλεπιδρούν μέσω των like με αυτούς και να τους ειδοποιούν όταν ανανεώνουν τις προσωπικές πληροφορίες τους. Το site ξεκίνησε αρχικά σαν ένα μέσο διασύνδεσης μεταξύ των φοιτητών του Harvard και μετά άρχισε να αυξάνεται συνεχώς, προσθέτοντας χρήστες από επιλεγμένα πανεπιστήμια, μέχρι που το 2006 η υπηρεσία έγινε προσβάσιμη σε κάθε άνθρωπο του πλανήτη που η ηλικία του ξεπερνούσε τα 13 χρόνια. Το όνομα της ιστοσελίδας προέρχεται από τα έγγραφα παρουσίασης των μελών πανεπιστημιακών κοινοτήτων μερικών Αμερικάνικων κολεγίων και προπαρασκευαστικών σχολείων που χρησιμοποιούσαν οι νεοεισερχόμενοι σπουδαστές για να γνωριστούν μεταξύ τους.

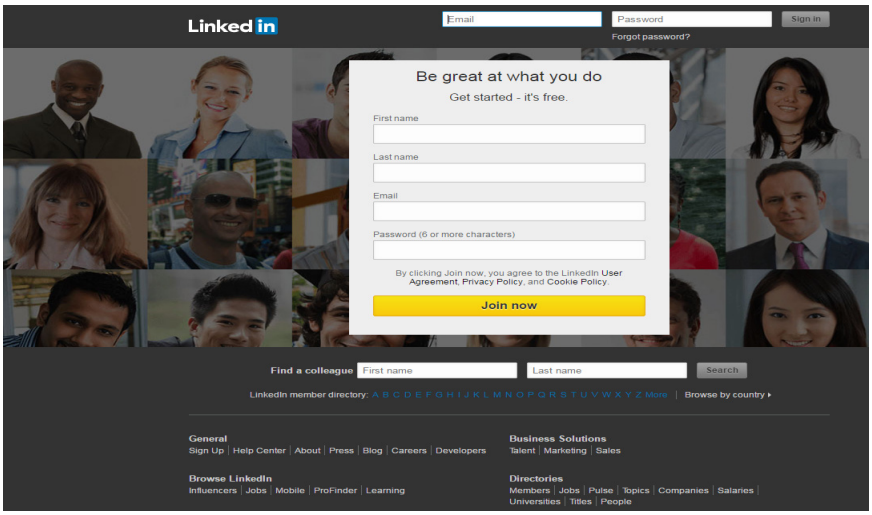
Εικόνα 9: *Διαδικτυακός Χώρος Facebook*

Πηγή facebook.com

Σήμερα, σύμφωνα με την comScore, το Facebook είναι το κορυφαίο σε επισκεψιμότητα μέσο κοινωνικής δικτύωσης, καθώς πέρασε το Myspace τον Απρίλιο του 2008. Μάλιστα σε έρευνα των ανεξάρτητων παρόχων στοιχείων επισκεψιμότητας Alexa και SimilarWeb, το Facebook είναι δεύτερο και πρώτο σε επισκεψιμότητα αντίστοιχα, έχει την υψηλότερη επισκεψιμότητα ανάμεσα στα μέσα κοινωνικής δικτύωσης, έχοντας περισσότερους από 20 δισεκατομμύρια επισκέπτες ανά μήνα .

2.2.3 LinkedIn

Ήταν 5 Μαΐου του 2003 όταν ξεκίνησε το LinkedIn από μια ομάδα ανθρώπων σε ένα γραφείο με όραμα τη σύνδεση των επαγγελματιών ανά το κόσμο ώστε να γίνουν πιο παραγωγικοί και επιτυχημένοι. Το σύνθημα με το οποίο ξεκίνησαν ήταν «Relationships matter», δηλαδή οι σχέσεις έχουν σημασία και αυτό είναι αλήθεια και ήταν εμπνευσμένο από τον ίδιο τον ιδρυτή της πλατφόρμας Reid Hoffman ο οποίος δηλώνει ότι αυτές ήταν που αποτελούσαν τη κινητήριο δύναμη και για τη δική του σταδιοδρομία . Έτσι το LinkedIn αποτέλεσε μία σελίδα κοινωνικής δικτύωσης που αφορά κατά κύριο λόγο επαγγελματίες και δίνει την δυνατότητα στους χρήστες να δημιουργήσουν το δικό τους προφίλ, να δικτυωθούν με συνεργάτες και δίκτυα συνεργατών επιχειρήσεων, να επικοινωνήσουν και να συνεργαστούν με καταρτισμένους επαγγελματίες. Η βασική του χρήση παρέχεται δωρεάν, υπάρχουν όμως χρεώσεις σε πρόσθετες παροχές που απευθύνονται σε εργοδότες οι οποίοι κάνουν δημοσιεύσεις για διαθέσιμες θέσεις εργασίας (linkedin.com).



Εικόνα 9: *Διαδικτυακός Χώρος LinkedIn*

Πηγή linkedin.com

2.2.4 Youtube

Το YouTube αποτελεί έναν ισχυρό ιστοχώρο το οποίο προσφέρει στους χρήστες τη δυνατότητα δημιουργίας και διαμοιρασμού βίντεο και μουσικής (Cayari, 2011). Πρόκειται πιο συγκεκριμένα για μια ιστοσελίδα διαμοιρασμού αρχείων βίντεο, η οποία οφείλει τη δημιουργία τους στους τρεις ιδρυτές της και πρώην υπάλληλους της Pay Pal, τους Chad Hurley, Steve Chen και Jawed Karim. Το YouTube λειτούργησε για πρώτη φορά το

Φεβρουάριο του 2005 ενώ τον Νοέμβριο του επόμενου έτους εξαγοράστηκε από την Google για 1.65 δισεκατομμύρια δολάρια (Kotler, Keller, Koshy & Jha, 2008). Οι χρήστες εκτός από τη δυνατότητα που έχουν για να παρακολουθούν και να ανεβάζουν βίντεο στην σελίδα μπορούν να προβούν και σε σχολιασμό των τελευταίων και να δείξουν αν τους αρέσουν ή όχι, απλά πατώντας ένα κουμπί. Το YouTube εκμεταλλευόμενο της ιδιαίτερης δημοτικότητας του τα τελευταία χρόνια , προωθεί και δράσεις μάρκετινγκ προσφέροντας την δυνατότητα στις επιχειρήσεις να δημιουργούν το δικό τους κανάλι και να διαφημίζονται μέσα από αυτό. Συγκεκριμένα το 94% των 100 κορυφαίων διαφημιστών, χρησιμοποιούν το YouTube για να προωθήσουν τις καμπάνιες τους (Wikipedia.org). Σύμφωνα με τους ανεξάρτητους παρόχους στοιχείων επισκεψιμότητας Alexa και SimilarWeb, το YouTube ήταν η τρίτη σε επισκεψιμότητα ιστοσελίδα στον κόσμο τον Ιούνιο του 2015.

Εικόνα 10: *Λογότυπο Youtube*

Πηγή youtube.com

2.3 Βασικά Είδη Ανάλυσης Κοινωνικών Δικτύων

Εξαιτίας της δημοτικότητας των social media, η ποσότητα των διαθέσιμων online δεδομένων που μπορούν να προκύψουν από αυτά έχει αυξηθεί ραγδαία συμπεριλαμβανομένου κειμένου, εικόνων, ήχου ή βίντεο. Πιο συγκεκριμένα, σύμφωνα με έρευνες κάθε λεπτό της ημέρας μπορούν να παραχθούν κατά μέσο όρο 350.000 tweets ενώ πάνω από 3000 φωτογραφίες ανεβαίνουν στον διαδικτυακό χώρο flickr. Γίνεται, λοιπόν εύκολα αντιληπτό πως αυτά τα δεδομένα παρέχουν πρωτοφανείς ευκαιρίες για έρευνα ανάλυσης των δεδομένων και ως εκ τούτου, η ανάλυση των κοινωνικών δικτύων έχει σημαντική αξία για πολλούς τομείς εφαρμογής όπως η χάραξη πολιτικής, τη διαφήμιση, και την εσωτερική ασφάλεια.

Κατά κύριο λόγο δύο είναι τα είδη δεδομένων τα οποία συχνά αναλύονται στο περιβάλλον των κοινωνικών δικτύων :

- **Δομική ανάλυση η οποία βασίζεται στους συνδέσμους (Linkage-based & Structural Analysis)** : Αυτό το είδος ανάλυσης χρησιμοποιείται για να αποκαλύψει τις δομικές ιδιότητες και τα πρότυπα της εξέλιξης των κοινωνικών δικτύων, τον εντοπισμό των κοινοτήτων, ή για μελλοντικές συνδέσεις, κ.λπ. Αυτό το είδος της ανάλυσης είναι ιδιαίτερα χρήσιμη για διάφορα πεδία εφαρμογής, όπως η κοινωνική ψυχολογία, το μάρκετινγκ, και την άμυνα κατά της τρομοκρατίας

- **Ανάλυση η οποία βασίζεται στο περιεχόμενο (Content-based Analysis) :** Η συγκεκριμένη μελετά το ετερογενές και αδόμητο περιεχόμενο που δημιουργείται από τους χρήστες κοινωνικών δικτύων, όπως τα blogs, εικόνες, βίντεο και ετικέτες (tags). Μερικές δημοφιλής πρακτικές ανάλυσης αυτού του τύπου περιλαμβάνουν την εξόρυξη γνώμης (opinion mining), τη διερεύνηση των τάσεων (trend detection) ή τη συνεργατική σύσταση (collaborative recommendation). Τα πεδία εφαρμογής της Content-based Analysis βρίσκει απήχηση σε επιχειρήσεις, την πολιτική, και την έρευνα των μέσων ενημέρωσης των καταναλωτών.

Έχει παρατηρηθεί, ωστόσο, ότι ο συνδυασμός της Linkage-based analysis με την Content-based analysis έχει πολύ ικανοποιητικά αποτελέσματα στις περισσότερες εφαρμογές.

2.4 Συσχέτιση Εξόρυξης Δεδομένων και Κοινωνικά Δίκτυα (Social Media Mining)

Όπως αναφέρθηκε και στο πρώτο κεφάλαιο της παρούσας πτυχιακής εργασίας η εξόρυξη γνώσης από τα δεδομένα αποτελεί ένα νέο ερευνητικό πεδίο που σαν σκοπό έχει την ανακάλυψη της κρυμμένης γνώσης από τεράστιες αποθήκες δεδομένων για την επίλυση προβλημάτων του πραγματικού κόσμου. Από την άλλη πλευρά, η ραγδαία εξάπλωση και χρήση των κοινωνικών δικτύων μέσω του παγκόσμιου ιστού, έχουν καταστήσει διαθέσιμο ένα πρωτοφανές ποσό δεδομένων το οποίο είναι διαθέσιμο προς επεξεργασίας και το οποίο αποτελεί αντικείμενο μελέτης πολλών και διαφορετικών πεδίων της επιστήμης, όπως η πολιτική, η ψυχολογία, οι επιχειρήσεις κ.α.

Συνδυάζοντας λοιπόν τεχνικές από την εξόρυξη γνώσης με τα κοινωνικά δίκτυα μπορούμε να ανακαλύψουμε νέες ενδιαφέρουσες πλευρές της ανθρώπινης συμπεριφοράς και της ανθρώπινης αλληλεπίδρασης, να βελτιώσουμε την αντίληψη που έχουν οι άνθρωποι σχετικά με ένα θέμα, να προσδιορίσουμε ομάδες ανθρώπων ανάμεσα στις μάζες του πληθυσμού, να μελετήσουμε ομάδες που αλλάζουν με το χρόνο, να βρεθούν άνθρωποι με επιρροή, ή ακόμα και να γίνει η σύσταση ενός προϊόντος ή μιας δραστηριότητας σε ένα άτομο (Μπιρμπίλη, Πασχάλης, Κωτσιαντής,2013).

. Η εφαρμογή της εξόρυξης γνώσης στα δεδομένα των social media, οδήγησε σε σημαντική άνοδο των online social media τα τελευταία χρόνια και στη δημιουργία επίσης ενός νέου ερευνητικού πεδίου το οποίο αποτελεί υπο-πεδίο της εξόρυξης δεδομένων, το λεγόμενο *social media mining*. Η εξόρυξη δεδομένων ωστόσο από τα κοινωνικά δίκτυα δεν είναι εύκολη υπόθεση. Τα δεδομένα των social media έχουν τρία χαρακτηριστικά τα οποία δημιουργούν προκλήσεις στους ερευνητές : Τα δεδομένα είναι μεγάλα, θορυβώδη, και δυναμικά. Επιπλέον, τίθεται και το ζήτημα και του εμπορικού περιορισμού καθώς σε ότι αναφορά τα κοινωνικά μέσα δικτύωσης όπως το facebook δεν συναντάται εύκολη πρόσβαση στα δεδομένα. Μόνο το για να ξεπεράσουν αυτές τις προκλήσεις, κυρίως σε ότι αναφορά τη δομή των δεδομένων αναπτύσσονται οι τεχνικές του data mining και ιδιαίτερα του text mining, για το οποίο θα αναφερθούμε εκτενώς στη συνέχεια, που χρησιμοποιούνται από τους ερευνητές για να δώσουν μία βαθύτερη ματιά στα δεδομένα των κοινωνικών δικτύων που διαφορετικά δε θα ήταν δυνατόν.

2.5 Εξόρυξη Γνώσης από Κείμενο (Text Mining)

Η μορφή των δεδομένων που συναντά κανείς τις περισσότερες φορές στα δεδομένα των κοινωνικών δικτύων είναι είτε δομημένο είτε ημι-δομημένο κείμενο. Συνεπώς, για να μπορέσει να πραγματοποιηθεί η διαδικασία της ανακάλυψης γνώσης δεν αρκούν οι τεχνικές του data mining αλλά και οι τεχνικές της εξόρυξης γνώσης από κείμενο (Text Mining) ή της Ανακάλυψης Γνώσης από Κείμενο (Knowledge-Discovery in Text). Τι ορίζεται όμως Text Mining;

Η ανακάλυψη γνώσης σε κείμενο (Knowledge Discovery in Text - KDT) καθώς και η εξόρυξη κειμένου (Text Mining – TM) περιλαμβάνουν αυτοματοποιημένες τεχνικές για την ανάλυση πολύ μεγάλων συλλογών από δεδομένα αλλά και την εξαγωγή χρήσιμων

πληροφοριών από αυτά, οι οποίες βρίσκονται σήμερα στο επίκεντρο του ενδιαφέροντος τόσο από εμπορική όσο και από επιστημονική πλευρά. Χρησιμοποιώντας τεχνικές από την εξόρυξη δεδομένων (text mining), την μηχανική μάθηση (machine learning), τη στατιστική (statistics) την επεξεργασία φυσικής γλώσσας (natural language processing), την ανάκτηση πληροφορίας (information retrieval), την εξαγωγή πληροφορίας (information extraction) και τη διαχείριση γνώσης (knowledge management), οι τεχνικές αυτές προσπαθούν να επιλύσουν το πρόβλημα της μετατροπής των τεραστίων ποσοτήτων από δεδομένα, σε χρήσιμη γνώση. Καθώς δεν υπάρχει καθιερωμένο λεξιλόγιο για αυτό τον αναπτυσσόμενο ερευνητικό τομέα, συχνά απαντώνται διαφορετικοί όροι για να δηλώσουν το ίδιο πράγμα: Ανακάλυψη γνώσης σε κείμενο (Knowledge Discovery in Text), Κειμενική Εξόρυξη Δεδομένων (Text Data Mining), Εξόρυξη Κειμένου ή Εξόρυξη Γνώσης από Κείμενα (Text Mining).

Διαχωρίζοντας τον όρο knowledge discovery in text από τον όρο text mining, μπορούμε να πούμε ότι η εξόρυξη κειμένου αποτελεί ένα στάδιο της ανακάλυψης γνώσης σε κείμενο, η οποία είναι μια διαδικασία που περιλαμβάνει πολλά βήματα για την ανεύρεση χρήσιμης πληροφορίας από κείμενα, από την συλλογή των εγγράφων, την προ-επεξεργασία τους (ώστε να μετατραπούν σε κάποια επιθυμητή αναπαράσταση όπως XML, SGML κλπ), την εξαγωγή λεκτικών πληροφοριών σχετικών με το περιεχόμενο κάθε εγγράφου, την εξόρυξη κειμένου μέσω της δημιουργίας μεταδεδομένων (metadata creation) και της αναγνώρισης προτύπων και συσχετίσεων μεταξύ των δεδομένων, μέχρι και την απεικόνιση (οπτικοποίηση- visualization) της γνώσης που προκύπτει.

Σύμφωνα με τους Choudhary et al, 2009 η διαδικασία της ανακάλυψης γνώσης σε κείμενο περιλαμβάνει τρία στάδια τα οποία και είναι:

- ***Συλλογή των Δεδομένων (Document Collection)***

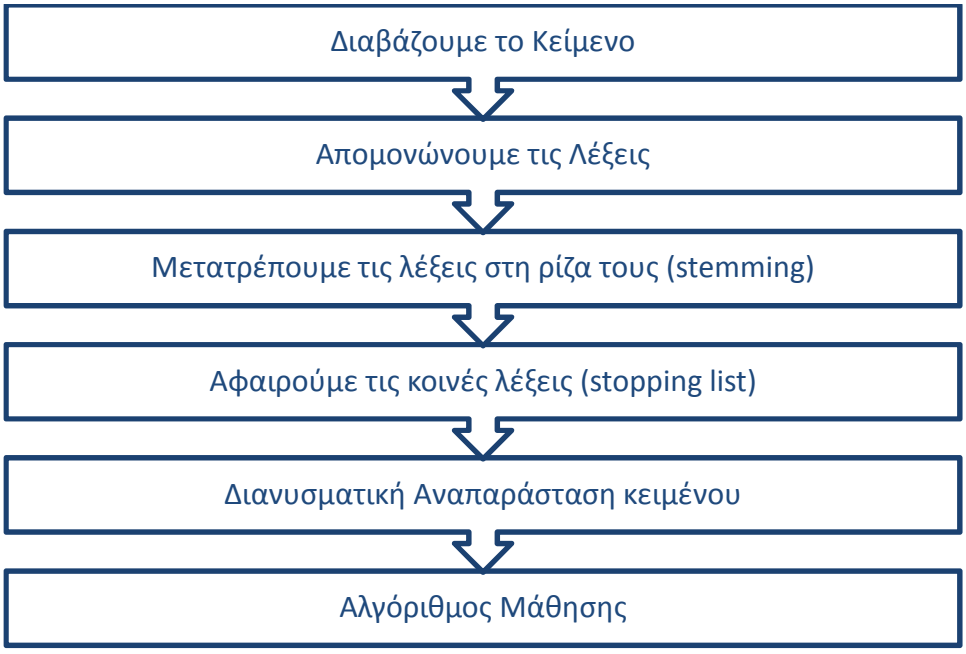
Αποτελεί το πρώτο στάδιο στη διαδικασία και σαν βασικό στόχο έχει τον εντοπισμό της πηγής από την οποία θα γίνει η συλλογή των αναγκαίων εγγράφων που είναι απαραίτητα για να ανακαλυφθεί η κρυμμένη γνώση. Εφόσον, καταλήξουμε στη τελική επιλογή αυτών ακολουθεί η ανάκτηση τους.

- ***Προ-επεξεργασία των Εγγράφων (Pre-processing)***

Το επόμενο στάδιο μετά τη συλλογή των κειμένων, είναι αυτό της προ-επεξεργασίας των κειμένων ώστε τα έγγραφα να μετατραπούν σε μια μορφή που να μπορούν να εφαρμοστούν οι τεχνικές του Text Mining. Συνήθως η προ-επεξεργασία αφορά το καθάρισμα των κειμένων από τετριμμένες λέξεις (stopwords removal), όπως είναι τα άρθρα, οι σύνδεσμοι, οι αντωνυμίες, αλλά και συχνά χρησιμοποιούμενες λέξεις που δεν ανήκουν στις ανωτέρω κατηγορίες και οι οποίες δεν φέρουν ιδιαίτερη σημασιολογική πληροφορία, όπως τα επιρρήματα. Στόχος μέσα από το συγκεκριμένο στάδιο είναι αφού τα έγγραφα επεξεργαστούν, να λάβουν τη θέση για την τελική ανακάλυψη της νέας γνώσης.

- ***Εξόρυξη Κειμένου (Text Mining)***

Η εξόρυξη κειμένου χρησιμοποιεί διάφορους αλγόριθμους από τις τεχνικές της μηχανικής μάθησης όπως και διάφορα άλλα εργαλεία προκειμένου να εξάγει μεταδεδομένα ή υψηλής πληροφορίας περιεχόμενο και/ή να ανακαλύψει πρότυπα και σχέσεις μέσα από την εξαγόμενη πληροφορία.



Σχήμα 3: *Στάδια Ανακάλυψης Γνώσης από Κείμενο*

2.6 Προ-επεξεργασία Κειμένου

Για την αποτελεσματική εξαγωγή των χαρακτηριστικών γνωρισμάτων είναι απαραίτητο να προηγηθεί η προεπεξεργασία των κειμένων. Η προεπεξεργασία των κειμένων χωρίζεται σε δύο στάδια:

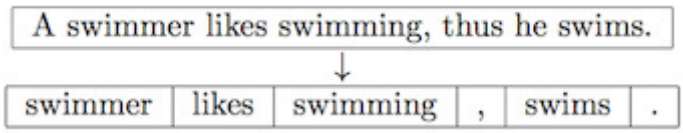
- ✓ ***1ο Στάδιο***

Το πρώτο στάδιο αφορά την αφαίρεση των τετριμμένων που δεν βοηθούν στο χαρακτηρισμό των κειμένων. Η διαδικασία αυτή αναφέρεται ως *stop-words removal*.

- ✓ ***2ο Στάδιο***

Το δεύτερο στάδιο αφορά στην αναγνώριση των ριζών των λέξεων έτσι ώστε να μην παίζει ρόλο η πτώση ή ο χρόνος στον οποίο βρίσκονται. Η διαδικασία αυτή αναφέρεται ως stemming.

Όσον αφορά στη διαδικασία της απομάκρυνσης τετριμμένων λέξεων (stop-word removal), μπορεί εύκολα κάποιος να παρατηρήσει ότι υπάρχουν πολλές λέξεις σε ένα έγγραφο που δεν προσφέρουν καμία βοήθεια στην ανάλυση ή στο χαρακτηρισμό του. Οι λέξεις αυτές είναι τα άρθρα (ο, το, κλπ), οι σύνδεσμοι (και, όμως, κλπ) καθώς και διάφορες λέξεις που συναντώνται σε πολλά κείμενα. Εάν αυτές οι λέξεις συμπεριληφθούν στα χαρακτηριστικά γνωρίσματα ενός κειμένου τότε θα λειτουργήσουν σαν θόρυβος και μπορεί να μειώσουν σημαντικά ενός αλγορίθμου εξόρυξης γνώσης. Για τη βέλτιστη αφαίρεση των τετριμμένων λέξεων , συνήθως επικαλείται ένας μηχανισμός αυτόματης αναγνώρισης των συντακτικών μερών μιας πρότασης (part of speech tagger), ο οποίος κατ' επέκταση εφαρμόζεται διαδοχικά σε όλο το κείμενο. Ορισμένοι από τους πιο συχνά χρησιμοποιούμενους τέτοιους μηχανισμούς είναι αυτός του Brill, ή αυτός που έχει αναπτυχθεί από την ομάδα επεξεργασίας φυσικής γλώσσας του Πανεπιστημίου Stanford. Ένας τέτοιος μηχανισμός βοηθάει στην αναγνώριση των ουσιαστικών, των ρημάτων, άρθρων, και γενικότερα όλων των μερών του λόγου, διευκολύνοντας την απαλοιφή των τετριμμένων λέξεων (Βαρζιγιάννης, Χαλκίδη, 2005).

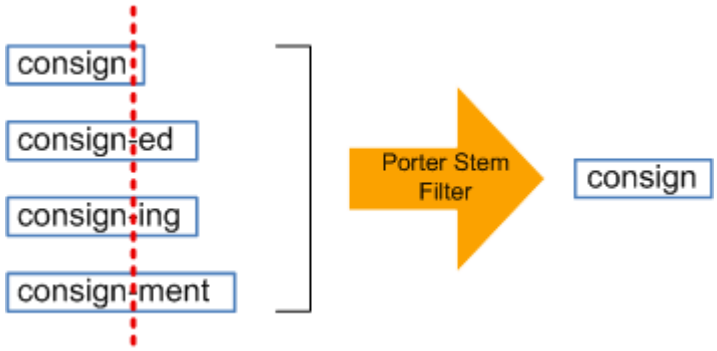


Εικόνα 11: *Απαλοιφή Τετριμμένων Λέξεων*

Πηγή Gladwin Analytics

Η διαδικασία του stemming, δηλαδή της αναγνώρισης των ριζών των λέξεων ανεξάρτητα από το χρόνο ή την πτώση στην οποία βρίσκονται, μπορεί να βοηθήσει σημαντικά στην εξαγωγή χαρακτηριστικών καθώς και στη βελτίωση της απόδοσης των αλγόριθμων εξόρυξης γνώσης. Η επιτυχής εφαρμογή ενός αλγόριθμου stemming θα αντιστοιχούσε τις λέξεις «σταματώ», «σταμάτημα» στην ίδια ρίζα και θα μπορούσε να βοηθήσει στην αναγνώριση της λέξης ως χαρακτηριστικό γνώρισμα. Επιπλέον, η ύπαρξη λέξεων σε διαφορετικές πτώσεις ή χρόνους μπορεί να εμποδίσει τον σωστό υπολογισμό ομοιότητας μεταξύ δύο κειμένων ή τον σωστό υπολογισμό των πιθανοτήτων εμφάνισης των λέξεων στα κείμενα και άρα μπορεί να έχει σημαντικές επιπτώσεις στην απόδοση των αλγόριθμων που θα εφαρμοστούν. Για τη διαδικασία του stemming έχουν προταθεί αρκετοί αλγόριθμοι στη βιβλιογραφία, με πιο γνωστούς να είναι οι Porter Stemmer και Lovins Stemmer (Βαρζιγιάννης, Χαλκίδη, 2005).

Μετά το πέρας της διαδικασίας προεπεξεργασίας των εγγράφων, οι λέξεις-όροι που έχουν μείνει στο κείμενο θεωρούνται υποψήφιες για την επιλογή ως χαρακτηριστικά γνωρίσματα.



Εικόνα 12: *Διαδικασία Stemming*

Πηγή Lean Java Engineering

2.7 Αναπαράσταση Κειμένου

Οι τεχνικές εξόρυξης γνώσης από κείμενα μπορούν να εφαρμοστούν αφού αυτά έχουν αναπαρασταθεί σε κάποια επεξεργάσιμη μορφή. Συχνά, η αναπαράσταση των κειμένων μεταφράζεται στη δημιουργία ενός διανυσματικού χώρου, όπου κάθε κείμενο αποτελεί και ένα διάνυσμα στο χώρο αυτό. Ο λόγος που η συνήθης τακτική αναπαράστασης κειμένων για εξόρυξη γνώσης αποτελείται από τη κατασκευή ενός διανυσματικού χώρου, είναι το γεγονός ότι με αυτόν τον τρόπο μπορούμε να αναπαραστήσουμε τα κείμενα σαν ένα σύνολο όρων (bag of words), που στο καθένα μπορεί να αποδοθεί διαφορετικό βάρος. Κατά συνέπεια, η αναπαράσταση αυτή ενσωματώνει την ενστικτώδη αντίληψη που έχει ο χρήστης για τα κείμενα, ότι δηλαδή το νόημα ενός κειμένου μπορεί να εξαχθεί από τους όρους που το αποτελούν.

Η αναπαράσταση ενός κειμένου σε κάποιον διανυσματικό χώρο υπονοεί ότι ο χώρος αυτός θα έχει τόσες διαφορετικές αναπαραστάσεις όσες και οι διαφορετικοί όροι του κειμένου. Είναι σαφές ότι στο στάδιο αυτό έχει προηγηθεί, το στάδιο της προεπεξεργασίας , ώστε να αφαιρεθούν εκείνες οι διαστάσεις-όροι, οι οποίες δεν προσφέρουν καμία πληροφορία.

Οι δύο βασικοί τρόποι αναπαράστασης των κειμένων σε κάποιο διανυσματικό χώρο είναι:

- *Το μοντέλο Boolean*
- *Μοντέλο Διανυσματικού Χώρου (Vector Space)*

Οι δυο τρόποι αναπαράστασης έχουν ένα κοινός στοιχείο, που είναι το γεγονός ότι το κείμενο διάνυσμα και στις δύο περιπτώσεις θα έχει τόσες διαστάσεις, όσοι και οι διαφορετικοί όροι του κειμένου.

2.7.1 Μοντέλο Boolean

Η Boolean αναπαράσταση ή αλλιώς λογικό μοντέλο, η τιμή που μπορεί να πάρει κάθε διάσταση ανήκει στο σύνολο {0,1}. Εάν για μια διάσταση του διανύσματος η τιμή είναι 1, αυτό υποδηλώνει πως ο όρος που αντιστοιχεί στη διάσταση αυτή υπάρχει στο κείμενο. Σε αντίθετη περίπτωση, δηλαδή στην περίπτωση όπου η τιμή είναι 0, αυτό σημαίνει ότι ο αντίστοιχος όρος δεν περιέχεται στο συγκεκριμένο κείμενο.

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	1	1	0	0	0	1
Brutus	1	1	0	1	0	0
Caesar	1	1	0	1	1	1
Calpurnia	0	1	0	0	0	0
Cleopatra	1	0	0	0	0	0
mercy	1	0	1	1	1	1
worser	1	0	1	1	1	0
...						

Εικόνα 13: *Παράδειγμα Booelan Αναπαράστασης*

Πηγή Stanford NLP Group

2.7.2 Μοντέλο Vector Space

Στο μοντέλου διανυσματικού χώρου, κάθε έγγραφο αναπαρίσταται σαν ένα διάνυσμα χαρακτηριστικών, του οποίου το μήκος ισούται με τον αριθμό των μοναδικών γνωρισμάτων των εγγράφων σε μια συλλογή. Κάθε στοιχείο του διανύσματος έχει ένα βάρος που υποδεικνύει τη σημαντικότητα κάθε γνωρίσματος στον χαρακτηρισμό του εγγράφου. Συνήθως αυτά τα γνωρίσματα είναι όροι που εξάγονται από το έγγραφο χρησιμοποιώντας τεχνικές ανάκτησης πληροφορίας.

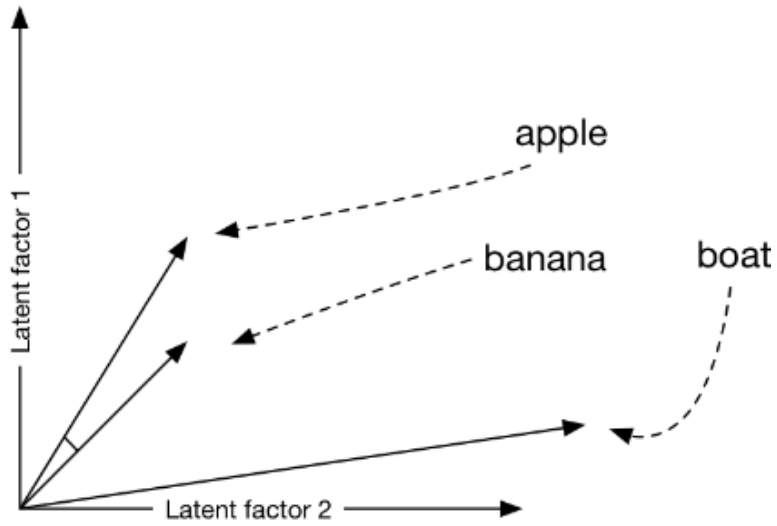
Η φάση της εξαγωγής των όρων που χαρακτηρίζουν ένα έγγραφο καλείται *ευρετηρίαση εγγράφου*. Στη συνέχεια, στη φάση *ορισμού βαρών όρων* , στου όρους αυτούς ανατίθεται βάρη υποδεικνύοντας τη σημαντικότητα τους στο χαρακτηρισμό εγγράφων. Τα βάρη μπορεί να είναι δυαδικά, υποδεικνύοντας την ύπαρξη (1) ή όχι (0) του όρου του κειμένου. Ωστόσο ως πιο σύνηθες τακτική χρησιμοποιείται η συχνότητα με την οποία ένας όρος εμφανίζεται στο κείμενο (*Συχνότητα μη επεξεργασμένων όρων*) ή έναν αλγόριθμο που ανήκει στην οικογένεια *Tf*Idf*. Η συχνότητα μη επεξεργασμένων όρων βασίζεται στη στατιστική του όρου μέσα στο έγγραφο και είναι ο απλούστερος τρόπος να ανατεθούν βάρη στους όρους. Το *Tf*Idf* είναι ένα μέτρο που χρησιμοποιείται στις συλλογές των εγγράφων , που ευνοεί όρους που είναι συχνοί σε σχετικά έγγραφα, αλλά μη συχνοί στη συλλογή σαν

ολότητα. Ο όρος *Tf* αφορά στη συχνότητα εμφάνισης του όρου στο έγγραφο και ο όρος *Idf* είναι η αντίστροφη συχνότητα συμβάντων του όρου σε όλη τη συλλογή η οποία ορίζεται ως

$$Idf = log\ (n_k / N)$$

Όπου n_k είναι ο αριθμός των εγγράφων που περιλαμβάνουν τον όρο και N είναι ο συνολικός αριθμός των εγγράφων.

Μετά τον ορισμό των βαρών, επιλέγεται ένα μέτρο ομοιότητας για τον υπολογισμό μεταξύ δύο εγγράφων. Υποθέτοντας ότι κάθε έγγραφο αναπαρίσταται από ένα διάνυσμα με βάρη, η ομοιότητα μπορεί να βρεθεί με τον απλό τρόπο υπολογισμού του εσωτερικού γινομένου. Ωστόσο, αυτό το μέτρο ομοιότητας δεν χρησιμοποιείται ποτέ και συνήθως το πιο δημοφιλές μέτρο αποτελεί ο συντελεστής Cosine, ο οποίος μετρά το συνημίτονο της γωνίας μεταξύ δύο διανυσμάτων χαρακτηριστικών (Βαρζιγιάννης, Χαλκίδη, 2005).



Εικόνα 14: *Παράδειγμα Vector Space Μοντέλου*

Πηγή Stanford NLP Group

2.8 Μείωση Διαστάσεων Χαρακτηριστικών

Έχοντας αναπαραστήσει τα κείμενα σε κάποιον διανυσματικό χώρο το τελευταίο στάδιο πριν την εφαρμογή του κατάλληλου αλγόριθμου εξόρυξης γνώσης είναι η μείωση των διαστάσεων των χαρακτηριστικών γνωρισμάτων των κειμένων. Όπως ειπώθηκε στην προηγούμενη ενότητα , τα κείμενα αναπαρίστανται σε διανυσματικό χώρο, έτσι ώστε οι διαστάσεις να είναι οι διαφορετικοί όροι των κειμένων. Είναι προφανές ότι για μεγάλες συλλογές κειμένων, τα διανύσματα αυτά, πέραν του γεγονότος ότι θα είναι αραιά, θα καταλαμβάνουν και τεράστιο χώρο στη μνήμη ή τον δίσκο του υπολογιστή με αποτέλεσμα η όλη διαδικασία να είναι αρκετά αργή και όχι ύψιστης ποιότητας.

Καθίσταται λοιπόν αναγκαίο να πραγματοποιηθεί μια μείωση της διάστασης αυτών προκειμένου να αφαιρεθούν περιττά ή μη σχετικά γνωρίσματα. Διάφοροι μέθοδοι έχουν προταθεί για τη μείωση της διαστατικότητας είτε από τη θεωρία πληροφορίας είτε μέσα από τη βιβλιογραφία της Υπολογιστικής Γραμμικής Άλγεβρας. Γενικά όμως η μείωση των διαστάσεων των χαρακτηριστικών γνωρισμάτων ενός κειμένου μπορεί να πραγματοποιηθεί μέσω δύο μεθόδων

1. *της επιλογής των γνωρισμάτων (feature selection) ή*
2. *της εξαγωγής-μετατροπής γνωρισμάτων (feature extraction/feature transformation).*

2.8.1 Μέθοδοι Επιλογής Γνωρισμάτων

Οι μέθοδοι επιλογής των χαρακτηριστικών γνωρισμάτων ταξινομούν τους όρους βάσει ενός αριθμητικού μέτρου που υπολογίζεται από τη συλλογή εγγράφων για να επιλέγει στη συνέχεια ένα υποσύνολο των όρων βάσει αυτού του μέτρου. Στη συνέχεια αναλύουμε τέσσερα από τα πιο δημοφιλή μέτρα ποιότητας που χρησιμοποιούνται για την επιλογή γνωρισμάτων.

- **Συχνότητα όρου και αντίστροφη συχνότητα (TF-IDF)**

Το TF-IDF είναι ο πιο δημοφιλής τρόπος να αποδώσουμε βάρος στην κάθε διάσταση ενός διανύσματος κειμένου. Όπως αναφέρθηκε στο μοντέλο διανυσματικού χώρου αποτελείται από δύο ποσότητες, την ποσότητα TF και την ποσότητα IDF. Η ποσότητα TF είναι απλώς η

συχνότητα εμφάνισης ενός όρου t_i στο κείμενο d_k και συμβολίζεται με $TF(d_k, t_i)$. Η ποσότητα IDF πολλαπλασιάζεται με την TF για να δώσει το τελικό βάρος στη διάσταση t_i του διανύσματος d_k. Η IDF λειτουργεί σαν βάρος ένα βάρος σημαντικότητας ενός όρου ως προς ένα κείμενο , σε σχέση όμως με ολόκληρη τη συλλογή των κειμένων που ανήκει.

Η ποσότητα TF-IDF θα δώσει μεγάλο βάρος σε έναν όρο που εμφανίζεται συχνά στο κείμενο και που συνολικά στη συλλογή είναι σπάνιος. Αυτό τον μετατρέπει αυτόματα σε χαρακτηριστικό όρο του κειμένου και συνεπώς το κριτήριο για την εξαγωγή καλών όρων είναι να λαμβάνουν μεγάλη τιμή για την ποσότητα TF-IDF.

- **Πληροφοριακό Κέρδος (Information Gain-IG)**

Στην περίπτωση επεξεργασίας μιας συλλογής κειμένων, γνωρίζουμε εκ των προτέρων συνήθως για κάθε κείμενο την ή τις κατηγόριες στις οποίες ανήκει. Η γνώση μας αυτή μας επιτρέπει να την μεταφέρουμε και να την αξιοποιήσουμε μέσω ενός ακόμη ενός μέτρου ποιότητας όπως είναι το Πληροφοριακό Κέρδος.

Το IG ποσοτικοποιεί το πληροφοριακό κέρδος που θα έχουμε αν ενσωματώσουμε στη διαδικασία πρόβλεψης της κατηγοριοποίησης κειμένων έναν συγκεκριμένο όρο. Μια διαφορετική οπτική αυτού είναι ότι το μέτρο αυτό στην ουσία μετράει τον αριθμό των bites που μας προσδίδει η ενσωμάτωση ή μη ενός όρου κάποιου κειμένου στη διαδικασία κατηγοριοποίησης κειμένων. Ο υπολογισμός του πληροφοριακού κέρδους μπορεί να διακρίνει τους ποιοτικότερους όρους της συλλογής, καθώς όσο μεγαλύτερη είναι η τελική τιμή του IG τόσο πιο ποιοτικός χαρακτηρίζεται και ο αντίστοιχός όρος.

- **Αμοιβαία Πληροφορία (Mutual Information-MI)**

Το μέτρο της αμοιβαίας πληροφόρησης είναι ένα κριτήριο που συχνά χρησιμοποιείται στη στατιστική μοντελοποίηση συσχετίσεων μιας γλώσσας, και σε σχετικές εφαρμογές. Αυτό το οποίο πραγματοποιεί είναι η σύγκριση της από κοινού πιθανότητα ενός όρου t και μια κλάσης c με την πιθανότητα του όρου t και της κλάσης c ξεχωριστά. Η μαθηματική μορφή του κριτηρίου είναι

$$I(t,c) = log\frac{P(t,c)}{P(t) * P(c)} = log\frac{P(t\hat{}c)}{P(t) * P(c)}$$

Εάν το συγκεκριμένο κριτήριο εμφανίσει μια μεγάλη τιμή αυτό θα σήμαινε και ένα ποιοτικότερο όρο που αξίζει να συμπεριληφθεί ως χαρακτηριστικό γνώρισμα.

- **Στατιστική X^2**

Η στατιστική X^2 είναι ένας από τους απλούστερους μη παραμετρικούς ελέγχους. Βασίζεται στην ιδέα ελέγχου όλων των δεδομένων και όχι μόνο για παράδειγμα μόνο του μέσου. Σε πολλές έρευνες οι συγκεντρωμένες παρατηρήσεις χωρίζονται σε κατηγορίες και τα δεδομένα έχουν τη μορφή συχνοτήτων για κάθε κατηγορία. Ο χωρισμός σε κατηγορίες γίνεται τόσο με βάση ποσοτικά χαρακτηριστικά δηλαδή το βάρος όσο και με ποιοτικά χαρακτηριστικά δηλαδή την ικανοποίηση από την εξυπηρέτηση σε ένα εστιατόριο. Ο έλεγχος αυτό συνήθως εφαρμόζεται στις περιπτώσεις με σκοπό:

- Τον έλεγχο καλής προσαρμογής (goodness of fit)
- Τον έλεγχο ανεξαρτησίας
- Τον έλεγχο ομογένειας

Ο πρώτος έλεγχος συγκρίνει μια δειγματική κατανομή σε σχέση με κάποια εξειδικευμένη πληθυσμιακή κατανομή. Δηλαδή σύμφωνα με την πρώτη εφαρμογή ελέγχουμε αν κάποια δεδομένα προσαρμόζονται σε κάποια θεωρητική κατανομή που θα αναμένονταν σύμφωνα με τις υποθέσεις και τη θεωρία. Ο δεύτερος έλεγχος πραγματοποιείται όταν κάθε τμήμα του δείγματος μπορεί να ταξινομηθεί σε τουλάχιστον δύο κατηγορίες. Όταν τα δεδομένα μπορούν να ταξινομηθούν σε περισσότερα από ένα χαρακτηριστικά τότε αναπαριστώνται μέσα από πίνακες που ονομάζονται πίνακες συνάφειας. Οι πίνακες αυτοί χρησιμοποιούνται για τον έλεγχο της ανεξαρτησίας ή όχι των χαρακτηριστικών που υπάρχουν στις διαστάσεις του πίνακα. Επίσης μπορεί να ελεγχθεί αν τα υπό εξέταση χαρακτηριστικά προέρχονται από τον ίδιο πληθυσμό. Τέλος στον έλεγχο ομογένειας έχουμε ταξινόμηση που περιλαμβάνει τη διαίρεση κάποιου συνόλου σε υποσύνολα σύμφωνα με κάποιο χαρακτηριστικό.

Σε όλες τις περιπτώσεις εφαρμογής της στατιστική X^2 συγκρίνουμε τις παρατηρημένες τιμές του δείγματος με τις αναμενόμενες τιμές ώστε να διαπιστωθεί εάν υπάρχει σημαντική

στατιστική διαφορά μεταξύ των δύο συνόλων. Όσο μεγαλύτερη είναι η εν λόγω διαφορά, τόσο πιθανότερο είναι να προκύψει στατιστικώς σημαντικό αποτέλεσμα.

Η σύνδεση του μέτρου X^2 με τις τεχνικές του Text Mining έγκειται στο γεγονός ότι το πρώτο μέτρα την έλλειψη ανεξαρτησίας ανάμεσα σε ένα όρο t και μιας κατηγορίας c της συλλογής κειμένων. Η αλγεβρική μορφή του κριτηρίου είναι

$$x^2(t,c) = \frac{N \times (AD - CB)^2}{(A + C) \times (B + D) \times (A + B) \times (C + D)}$$

Όπου A ο αριθμός των φορών κατά τις οποίες ο t και η c συνυπάρχουν, B ο αριθμός των φορών κατά τις οποίες ο t εμφανίζεται, χωρίς να εμφανίζεται η c, C ο αριθμός των φορών κατά τις οποίες εμφανίζεται η c αλλά δεν εμφανίζεται ο t και D ο αριθμός των φορών κατά τις οποίες ταυτόχρονα δεν εμφανίζεται ούτε η c, ούτε και ο t, και N ο συνολικός αριθμός όλων των κειμένων.

Το στατιστικό αποτέλεσμα του συγκεκριμένου ελέγχου προσπαθεί να εντοπίσει τους καλύτερους όρους μέσα σε μια κλάση c και είναι αυτοί που διανέμονται πιο διαφορετικά στα σύνολα των θετικών και αρνητικών παραδειγμάτων της κλάσης c.

2.8.2 Μέθοδοι Εξαγωγής Χαρακτηριστικών

Οι μέθοδοι εξαγωγής ή μετασχηματισμού των χαρακτηριστικών εκτελούν ένα μετασχηματισμό της διανυσματικής αντιπροσώπευσης της συλλογής των εγγράφων σε ένα χαμηλότερο διανυσματικό χώρο, όπου οι νέες διαστάσεις μπορούν να αντιμετωπισθούν ως γραμμικοί συνδυασμοί των αρχικών διαστάσεων.

Η **Ανάλυση Κυρίων Συνιστωσών** (ΑΚΣ) (Principal Components Analysis (PCA)) είναι μια μέθοδος εξαγωγής χαρακτηριστικών, η οποία επιτρέπει τη μείωση του πλήθος των διαστάσεων ενός συνόλου δεδομένων. Θεωρούμε ότι έχουμε ένα σύνολο δεδομένων με K γραμμές και N διαστάσεις (στήλες). Με την ΑΚΣ βρίσκουμε ένα σύστημα M κάθετων διανυσμάτων, όπου M < N, και προβάλλουμε τα δεδομένα στον νέο χώρο M διαστάσεων. Με τον τρόπο αυτό δημιουργούμε γραμμικούς συνδυασμούς των αρχικών μεταβλητών οι οποίοι:

● είναι ασυσχέτιστοι μεταξύ τους,

● περιέχουν το μεγαλύτερο μέρος της διακύμανσης των αρχικών μεταβλητών.

Τέλος, η **Διάσπαση Μοναδιαίων Τιμών** (ΔΜΤ) (Singular Value Decomposition-SVD) είναι ακόμη μία τεχνική μετασχηματισμού η οποία είναι ισοδύναμη με την ανάλυση σε κύριες συνιστώσες αν αφαιρεθεί ο μέσος κάθε μεταβλητής και χρησιμοποιείται ως εναλλακτική μέθοδος για τη μείωση των διαστάσεων.

2.9 Τεχνικές Εξόρυξης Κειμένου

Η διαχείριση ενός τεράστιου όγκου δεδομένων όπως είναι αυτή των κειμένων απαιτεί την ύπαρξη διαφορετικών τεχνικών που να μπορούν να εφαρμοστούν στα πλαίσια του Text Mining προκειμένου να μπορέσει να εξαχθεί και η ανάλογη γνώση. Συνήθως αυτές οι τεχνικές αντλούνται από τα πλαίσια της εξόρυξης δεδομένων αλλά με κάποιες διαφοροποιήσεις. Κάποιες από αυτές τις τεχνικές εξόρυξη κειμένου είναι:

- Εξαγωγή Πληροφοριών (Information Extraction)
- Κατηγοριοποίηση (Categorization)
- Ομαδοποίηση (Clustering)
- Συνόψιση (Summarization)
- Απεικόνιση Πληροφορίας (Information Visualization)
- Διασύνδεση Εννοιών (Concept Linkage)
- Εξαγωγή Οντολογιών (Ontology Extraction)

Στη συνέχεια θα παρουσιάσουμε αναλυτικά κάθε μία από τις παραπάνω τεχνολογίες.

2.9.1 Εξαγωγή Πληροφοριών

Η εξαγωγή πληροφοριών είναι ένα είδος ανάκτησης πληροφοριών από μη δομημένα δεδομένα ή κείμενα γραμμένα σε φυσική γλώσσα και αποτελεί κλάδο της σύγχρονης επιστήμης των υπολογιστών. Είναι βασισμένη στην τεχνολογία της επεξεργασίας του φυσικού λόγου. Πρακτικά, συνίσταται στην αναγνώριση συγκεκριμένου είδους πληροφοριών, όπως κύρια ονόματα (ονόματα ανθρώπων, τοπωνύμια, ονόματα εταιρειών, ημερών, μηνών, κτλ.), χρονικές πληροφορίες (ημερομηνίες), σχέσεις και γεγονότα από (συνήθως) ηλεκτρονικά κείμενα. Απώτερος στόχος της εξαγωγής πληροφοριών είναι η «κατανόηση» των βασικών συστατικών του υπό ανάλυση κειμένου, τα οποία αργότερα μπορούν να χρησιμοποιηθούν από εφαρμογές όπως αυτόματη εξαγωγή περιλήψεων κειμένων, αυτόματη απάντηση ερωτήσεων, αυτόματη μετάφραση κ.α.

Δεδομένου του μεγάλου όγκου πληροφοριών που παράγονται και διακινούνται σήμερα, όπου είναι σχεδόν και το κύριο χαρακτηριστικό του διαδικτύου, το ζητούμενο στις μέρες μας είναι όχι απλώς η κατοχή της πληροφορίας, αλλά η διαχείριση της πληροφορίας και ο εντοπισμός της «σχετικής» πληροφορίας (Wikipedia).

2.9.2 Κατηγοριοποίηση

Η κατηγοριοποίηση ή διαφορετικά ταξινόμηση κειμένου είναι η πιο δημοφιλής από τις τεχνικές της εξόρυξης κειμένου. Στόχος της είναι εναποθέσει τα δεδομένα-έγγραφα σε προκαθορισμένες κατηγορίες / κλάσεις. Είναι μία μέθοδος η οποία εντάσσεται στην επιβλεπόμενη μηχανική μάθηση, καθώς οι κλάσεις καθορίζονται εκ των προτέρων, πριν ακόμη εξεταστούν τα δεδομένα. Στην κατηγοριοποίηση κειμένου όπως και στη κατηγοριοποίηση δεδομένων, πραγματοποιείται η χρήση ενός συνόλου εκπαίδευσης (training set) το οποίο χρησιμοποιείται για να εκπαιδεύσει το μοντέλο κατηγοριοποίησης, μέσω μιας στατιστικής ανάλυσης ενός συνόλου λεκτικών προτύπων. Έπειτα, γίνεται εφαρμογή του μοντέλο που αναπτύχθηκε, στην ταξινόμηση του συνόλου ελέγχου (test set) και αξιολογείται η απόδοσή του.

2.9.3 Ομαδοποίηση

Μία ομάδα (cluster) είναι μια συλλογή από σχετικά έγγραφα, και η ομαδοποίηση (clustering) είναι η διαδικασία της δημιουργίας ομάδων εγγράφων βάσει κάποιου κριτηρίου

ομοιότητας. Η ομαδοποίηση κειμένων είναι χρήσιμη για τον προσδιορισμό κρυμμένων ομοιοτήτων, για να διευκολύνει τη διαδικασία του να βρούμε παρόμοιες ή σχετικές πληροφορίες, ενώ επιπλέον μπορούμε όταν εξερευνούμε μια καινούρια συλλογή δεδομένων ώστε να έχουμε μια γενική επισκόπηση της συλλογής.

Η κατηγοριοποίηση διαφέρει από την ομαδοποίηση στο γεγονός ότι τα κείμενα ομαδοποιούνται εκείνη τη στιγμή με βάση την ομοιότητά τους, χωρίς να υπάρχει η ανάγκη χρησιμοποίησης προκαθορισμένων θεμάτων.

Οι πιο γνωστοί αλγόριθμοι που χρησιμοποιούνται είναι ιεραρχικοί (hierarchical), διαχωριστικοί (partitional), δυαδικοί σχεσιακοί (binary relational) και ασαφείς (fuzzy). Επίσης ο πιο σημαντικός παράγοντας στη λειτουργία της ομαδοποίησης είναι το μέτρο ομοιότητας που χρησιμοποιεί ο εκάστοτε αλγόριθμος.

2.9.4 Συνόψιση

Η συνόψιση ενός κειμένου είναι σημαντική στη προσπάθεια κάποιος να κατανοήσει εάν ένα μεγάλο σε μέγεθος κείμενο μπορεί να καλύψει τις ανάγκες του, προχωρώντας στην λεπτομερή ανάγνωσή του. Ο στόχος στη συνόψιση είναι η ελάττωση της έκτασης και της λεπτομέρειας ενός κειμένου, διατηρώντας όμως τα βασικά του σημεία και το συνολικό του νόημα. Στο σημείο αυτό αξίζει να σημειωθεί ότι ενώ οι ηλεκτρονικοί υπολογιστές μπορούν να αν γνωρίζουν πρόσωπα, τοποθεσίες και χρονικές αναφορές, καθίσταται ακόμα δύσκολη η σημασιολογική ανάλυση και ερμηνεία του κειμένου, διότι δεν κατέχουν τις αντίστοιχες γλωσσικές δεξιότητες με τον άνθρωπο.

Μία από τις πιο ευρέως χρησιμοποιούμενες στρατηγικές είναι η εξαγωγή προτάσεων. Για παράδειγμα θα μπορούσε ένα λογισμικό συνόψισης να εξάγει φράσεις που ακολουθούν εκφράσεις όπως «συνοψίζοντας», «εν κατακλείδι» κ.α. οι οποίες γενικά περιλαμβάνουν τα πιο βασικά στοιχεία ενός κειμένου. Τέλος, η συνόψιση μπορεί να λειτουργήσει με την κατηγοριοποίηση για τη δημιουργία περιλήψεων σε κείμενα που ανακτώνται σε ένα συγκεκριμένο θέμα. Εάν σε ένα ιατρικό προσωπικό δίνονταν εκατοντάδες κείμενα για ένα συγκεκριμένο τομέα, με τη βοήθεια εργαλείων αυτόματης δημιουργίας περιλήψεων, θα μπορούσε να μειωθεί σημαντικά ο χρόνος για την ταξινόμηση αφενός του υλικού αυτού και αφετέρου ο χρόνος απόκτησης της σχετικής με ένα συγκεκριμένο θέμα πληροφορίας.

2.9.5 Απεικόνιση Πληροφορίας

Στόχος της απεικόνισης πληροφοριών είναι να οργανώσει μεγάλες πηγές κειμένου σε μία οπτική ιεραρχία. Αυτό δίνει τη δυνατότητα της περιήγησης, η οποία είναι πιο σημαντική σε σχέση από μία απλή αναζήτηση και καθιστά πιο εύκολη τη διαδικασία περιορισμού μεγάλου όγκου κειμένων.

Επίσης, η απεικόνιση πληροφοριών χρησιμοποιεί την εξαγωγή χαρακτηριστικών γνωρισμάτων προκειμένου να δημιουργηθεί μια γραφική αντιπροσώπευση της συλλογής κειμένων. Αυτή η προσέγγιση βοηθάει το χρήστη στον προσδιορισμό των κύριων θεμάτων ή των πιο σημαντικών εννοιών. Σκοπός είναι, με τη χρήση υπολογιστικών μετασχηματισμών, να μειωθεί η γνωστική προσπάθεια εξέτασης μεγάλων συλλογών από κείμενα γεγονός το οποίο θα βοηθήσει την ανακάλυψη νέας γνώσης. Μία εφαρμογή απεικόνισης πληροφοριών αποτελεί το SPIRE (Spatial Paradigm for Information Retrieval and Exploration) το οποίο αναπτύχθηκε από τον Wise το 1999.

2.9.6 Διασύνδεση Εννοιών

Η διαδικασία διασύνδεσης εννοιών χρησιμοποιεί εργαλεία που συνδέουν σχετικά κείμενα, αναγνωρίζοντας τις κοινές μεταξύ τους έννοιες και βοηθώντας το χρήστη να ανακαλύψει πληροφορίες οι οποίες πιθανόν να μην ήταν δυνατό να βρεθούν με τη χρήση παραδοσιακών μεθόδων.

Πρόκειται δηλαδή, για μία ιδιαίτερα σημαντική διαδικασία στον τομέα του Text Mining. Για παράδειγμα στον ιατρικό κλάδο όπου η έρευνα είναι αρκετά εκτεταμένη, είναι δύσκολο για έναν ερευνητή να διαβάσει όλες τις επιστημονικές δημοσιεύσεις και να κάνει τις απαραίτητες διασυνδέσεις με άλλη έρευνα, ώστε να αποκτήσει τη σχετική πληροφορία. Για το λόγο αυτό υπάρχουν λογισμικά Text Mining τα οποία μπορούν εύκολα να αναγνωρίσουν τη σύνδεση μεταξύ δύο ή περισσότερων θεμάτων.

2.9.7 Εξαγωγή Οντολογιών

Οι οντολογίες έχουν καθιερωθεί ως δομημένα πλαίσια για την οργάνωση πληροφορίας και χρησιμοποιούνται κυρίως στην Τεχνητή Νοημοσύνη, στον Σημασιολογικό Ιστό, στη Βιοπληροφορική, στην επιστήμη Βιβλιοθηκονομίας, και σε άλλες επιστήμες -

κλάδους ως μια μορφή αναπαράστασης γνώσης για τον κόσμο. Αξίζει να σημειωθεί πως η κύρια ώθηση στις οντολογίες δόθηκε από την ανάπτυξη του Σημασιολογικού ιστού (Semantic Web), ο οποίος εφευρέθηκε από τον Tim Berners Lee.

Ο κύριος σκοπός του σημασιολογικού ιστού είναι να εξελίξει τον τωρινό ιστό, ο οποίος αποτελείται από απλές σελίδες που μπορούν να αναγνωστούν μόνο από ανθρώπους, σε σελίδες που περιέχουν πληροφορίες ανάγνωσης για τις μηχανές (μεταδεδομένα) και στο πως συνδέονται μεταξύ τους οι σελίδες, δημιουργώντας έτσι αυτόματες υπηρεσίες που χρησιμοποιούν τον ιστό πιο έξυπνα και πραγματοποιούν εργασίες για τους χρήστες. Επομένως, ο ρόλος των οντολογιών σε αυτό το σημείο είναι να παρέχουν εννοιολογική υποστήριξη για να καταστήσουν τη σημασιολογία μιας μηχανής μεταδεδομένων ερμηνεύσιμη.

2.10 Αλγόριθμοι Text Mining

2.10.1 Μηχανική Μάθηση

Η Τεχνητή Νοημοσύνη περιλαμβάνει πολλές τεχνικές εξόρυξης γνώσης, όπως τα νευρωνικά δίκτυα και τη κατηγοριοποίηση. Ωστόσο, η τεχνική νοημοσύνη είναι πιο γενική και περιλαμβάνει περιοχές εκτός της παραδοσιακής εξόρυξη γνώσης. Οι εφαρμογές της τεχνικής νοημοσύνης μπορεί να μην απασχολούνται με το θέμα αντιμετώπισης μεγάλων όγκων δεδομένων καθώς χειρίζονται συνήθως μικρά σύνολα δεδομένων.

Μηχανική Μάθηση είναι η περιοχή της τεχνητής νοημοσύνης η οποία εξετάζει πώς να γράφονται προγράμματα που να μαθαίνουν. Για τους σκοπούς της εξόρυξης γνώσης, η μηχανική μάθηση χρησιμοποιείται συχνά για την πρόβλεψη και κατηγοριοποίηση. Με τη μηχανική μάθηση, ο υπολογιστής κάνει μία πρόβλεψη και κατόπιν, βασιζόμενος σε ανάδραση περί της ορθότητας της πρόβλεψης «μαθαίνει» από την ανάδραση αυτή. Όταν η μηχανική μάθηση εφαρμόζεται σε εργασίες εξόρυξης γνώσης, χρησιμοποιείται ένα μοντέλο για να αναπαραστήσει τα δεδομένα. Κατά τη διάρκεια της διαδικασίας μάθησης χρησιμοποιείται ένα δείγμα από τη βάση δεδομένων για να εκπαιδεύσει το σύστημα να εκτελέσει σωστά την επιθυμητή εργασία. Κατόπιν, το σύστημα εφαρμόζεται στη γενική βάση δεδομένων για να εκτελέσει στην πραγματικότητα την εργασία.

Ολοκληρώνοντας, η μηχανική μάθηση έχει δύο γενικές προσεγγίσεις: τη *εποπτευόμενη μάθηση* (supervised learning) και τη *μη εποπτευόμενη μάθηση* (unsupervised learning). Η εποπτευομένη προσέγγιση μαθαίνει από παραδείγματα. Δοθέντος ενός συνόλου εκπαίδευσης συν των σωστών απαντήσεων, το υπολογιστικό μοντέλο εφαρμόζει επιτυχώς κάθε καταχώρηση στο σύνολο εκπαίδευσης. Βασιζόμενο στην ικανότητα του να χειρίζεται μία κάθε μία από αυτές τις καταχωρήσεις, το μοντέλο αλλάζει για να εξασφαλισθεί ότι δουλεύει καλύτερα με αυτή τη καταχώρηση, αν εφαρμοζόταν ξανά. Αφού δοθούν αρκετές τιμές εισόδου, το μοντέλο θα μάθει τη σωστή συμπεριφορά για κάθε δυνητική καταχώρηση. Στη μη εποπτευόμενη μάθηση τα δεδομένα υπάρχουν αλλά δεν υπάρχει γνώση της σωστής απάντησης εφαρμογής του μοντέλου στα δεδομένα.

2.10.2 Αλγόριθμοι εξόρυξης κειμένου

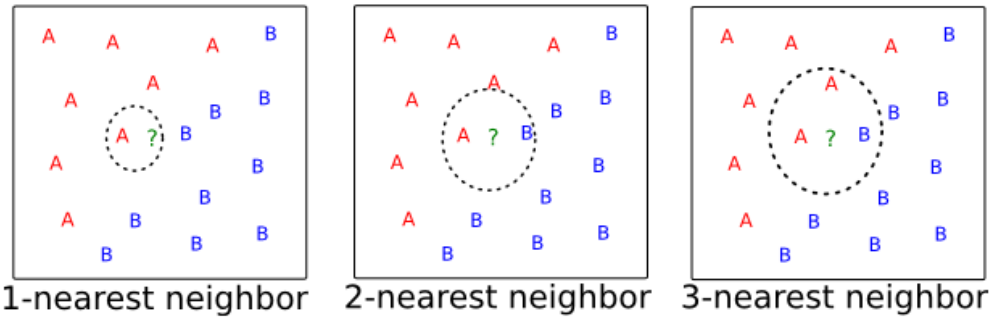
Προκειμένου να μπορέσει να πραγματοποιηθεί το τελικό στάδιο της ανακάλυψης γνώσης από κείμενο είναι απαραίτητο να εφαρμοστούν οι αντίστοιχες τεχνικές του Text Mining σε συνδυασμό με κάποιο αλγόριθμο της μηχανικής μάθησης. Ως αλγόριθμος ορίζεται μια πεπερασμένη σειρά ενεργειών, αυστηρά καθορισμένων και εκτελέσιμων σε πεπερασμένο χρόνο, που στοχεύουν στην επίλυση ενός προβλήματος. Πιο απλά (αλγόριθμο) ονομάζουμε μία σειρά από εντολές που έχουν αρχή και τέλος, είναι σαφείς ως σκοπό έχουν την επίλυση κάποιου προβλήματος. Οι πιο γνωστοί αλγόριθμοι είναι τα νευρωνικά δίκτυα, τα δέντρα απόφασης, οι ταξινομητές κοντινότερου γείτονα, οι μηχανές διανυσμάτων υποστήριξης, τα δέντρα απόφασης και οι ταξινομητές Bayes. Στη συνέχεια γίνεται ανάλυση των μεθόδων αυτών της ταξινόμησης.

- ***K-Nearest Neighbor (KNN)***

Η αρχή αυτής της μεθόδου είναι να κατηγοριοποιήσει ένα νέο έγγραφο, βρίσκοντας τα πιο όμοια με αυτό έγγραφα στο σύνολο εκπαίδευσης. Μέθοδοι που χρησιμοποιούν αυτή την αρχή καλούνται *μέθοδοι μάθησης βασισμένες στη μνήμη*. Η λογική η οποία ακολουθεί ο αλγόριθμος είναι η εξής: Στην αρχή επιλέγονται Κ αρχικά κέντρα βάρους, όπου Κ είναι μία παράμετρος ορισμένη από το χρήστη, συγκεκριμένα το πλήθος των συστάδων. Κάθε σημείο στη συνέχεια αποδίδεται στο πιο κοντινό κέντρο βάρους, και κάθε σύνολο σημείων που αποδίδεται σε ένα κέντρο βάρους συνιστά μια κατηγορία. Το κέντρο βάρους κάθε συστάδας

στη συνέχεια ενημερώνεται με βάση τα σημεία που αποδίδονται στη συστάδα. Τα βήματα της εκχώρησης επαναλαμβάνονται μέχρι να μην υπάρχει σημείο που να αλλάζει συστάδα ή ισοδύναμα, μέχρι τα κέντρα βάρους να παραμένουν σταθερά.

Για την ανάθεση των βαρών γίνεται χρήση της συχνότητας-αντίστροφης συχνότητας όρου (*Tf*idf*), υπολογίζοντας την ομοιότητα μεταξύ των παραδειγμάτων ελέγχου και κέντρων κατηγοριών. Το βάρος που ανατίθεται σε ένα όρο είναι ένας συνδυασμός του βάρους του σε μια ερώτηση εκτίμησης της σχετικότητας και μη σχετικότητας των εγγράφων. Η ομοιότητα μεταξύ δύο εγγράφων είναι εύκολο να μετρηθεί χρησιμοποιώντας την ομοιότητα συνημίτονων. Γενικά, ο αλγόριθμος των κ-κοντινότερων γειτόνων ερμηνεύεται εύκολα και αποδίδει καλά. Από την άλλη πλευρά όμως δεν μειώνει τη διάσταση του χώρου και δεν μπορεί να χρησιμοποιήσει offline επεξεργασία.

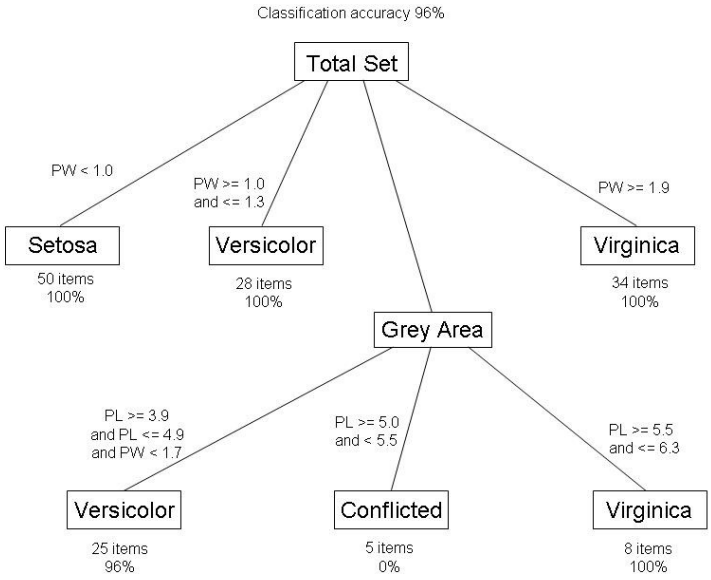


Εικόνα 15: *Αναπαράσταση K- Nearest Neighbor*
Πηγή trevorwhitney.com

- ***Δέντρα Αποφάσεων***

Το μοντέλο που βασίζεται σε δέντρα αποφάσεων αποτελείται από σειρά απλών κανόνων αποφάσεων, οι οποίοι συχνά παρουσιάζονται με τη μορφή γράφου. Είναι μια από τις πιο δημοφιλής τεχνικές που χρησιμοποιήθηκαν πρόσφατα και χρησιμοποιείται συνήθως στην περίπτωση που θέλουμε να εφαρμόσουμε τη τεχνική της κατηγοριοποίησης. Ένα δένδρο απόφασης κατασκευάζεται με βάση ένα σύνολο εκπαίδευσης προ-κατηγοριοποιημένων δεδομένων. Κάθε ένας από τους εσωτερικούς κόμβους απόφασης προσδιορίζει τον έλεγχο

ενός γνωρίσματος και κάθε κλαδί που «κατεβαίνει» από εκείνον τον κόμβο αντιστοιχεί σε μια από τις πιθανές τιμές για το συγκεκριμένο γνώρισμα. Επίσης, κάθε φύλλο αντιστοιχεί σε μια από τις κατηγορίες που έχουν οριστεί. Η διαδικασία για την κατηγοριοποίηση ενός νέου δείγματος με βάση ένα δένδρο απόφασης είναι η ακόλουθη: ξεκινώντας από τη ρίζα του δέντρου και εξετάζοντας τα γνωρίσματα που καθορίζονται από το κόμβο αυτό προσδιορίζονται διαδοχικά οι εσωτερικοί κόμβοι που θα αναμένονται προς επίσκεψη έως ότου γίνει κατάληξη σε ένα φύλλο. Σε κάθε εσωτερικό κόμβο ελέγχεται εάν το δείγμα ικανοποιεί το συγκεκριμένο κόμβο. Η έκβαση αυτής της δοκιμής καθορίζει εν συνεχεία το κλαδί που θα ακολουθηθεί καθώς και τον επόμενο κόμβο. Η κατηγορία του υπό μελέτη δείγματος είναι η κατηγορία του τελικού κόμβου ο οποίος αντιστοιχεί σε φύλλο του δέντρου.



Εικόνα 16: *Αναπαράσταση Δέντρων Απόφασης*
Πηγή Webster University

Διάφοροι μέθοδοι έχουν προταθεί για την εξαγωγή δέντρων αποφάσεων όπως ο αλγόριθμος C4.5. Τα δέντρα αποφάσεων είναι ένας κατηγοριοποιητής βασισμένος στις πιθανότητες όπου η συνάρτηση confidence (κλάση) αναπαριστά μια κατανομή πιθανότητας. Είναι εύκολο να ερμηνευτούν, ωστόσο απαιτούν έναν αριθμό παραμέτρων μοντέλου που είναι δύσκολο να βρεθούν και η εκτίμηση του λάθους είναι δύσκολη (Βαρζιγιάννης, Χαλκίδη, 2005).

- ***Αφελής Bayes (Naïve Bayes)***

Ένα ακόμη δημοφιλής ταξινομητής για την ανάκτηση πληροφορίας σε κείμενο, είναι το «απλοϊκό» μοντέλο Bayes (Naive Bayes model) ή αφελής Bayes, ο οποίος επίσης είναι ένας ταξινομητής που βασίζεται στις πιθανότητες. Ένας Αφελής κατά Bayes αλγόριθμος , εκτιμά την εξαρτώμενη από την κατηγορία πιθανότητα υποθέτοντας, ότι τα χαρακτηριστικά είναι υπό συνθήκη ανεξάρτητα δεδομένης μιας κατηγορίας *y*. Η υπόθεση της υπό συνθήκη ανεξαρτησίας μπορεί να εκφραστεί τυπικά ακολούθως:

$$P(X|Y=y) = \prod_{i=1}^{d} P(X_i|Y=y)$$

Όπου κάθε σύνολο χαρακτηριστικών $X=\{X_1, X_2, \ldots\ldots\ldots X_d\}$ αποτελείται από *d* ανεξάρτητα χαρακτηριστικά. Με την υπόθεση της υπό συνθήκη ανεξαρτησίας, αντί να υπολογίζεται η εξαρτώμενη από τη κατηγορία πιθανότητα για κάθε συνδυασμό του X, αρκεί να εκτιμηθεί η υπό συνθήκη πιθανότητα για κάθε X_i δοθέντος του Y. Η τελευταία προσέγγιση είναι πιο πρακτική επειδή δεν απαιτεί ένα πολύ μεγάλο σύνολο εκπαίδευσης για μια καλή εκτίμηση της πιθανότητας.

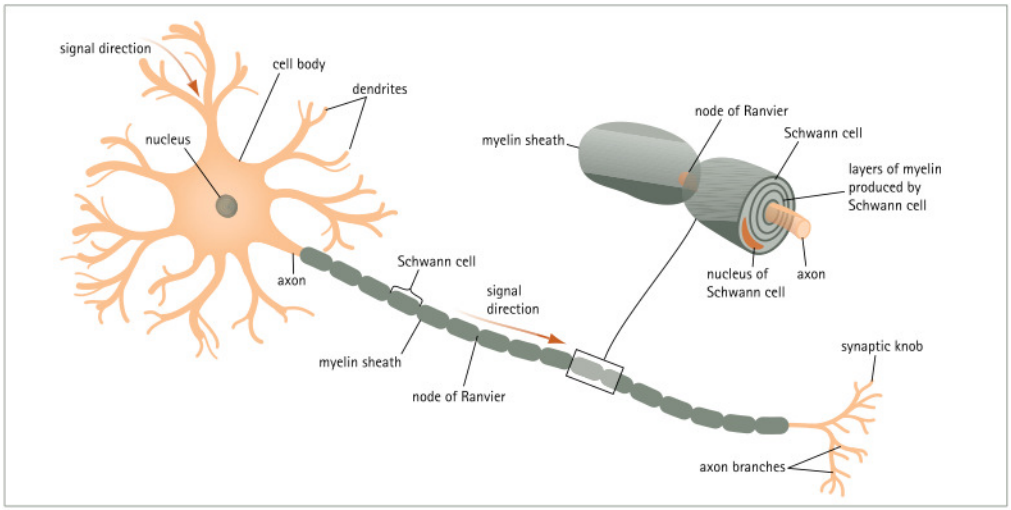
Για να κατηγοριοποιήσει μια εγγραφή ελέγχου, ο αφελής Bayes υπολογίζει την εκ των υστέρων πιθανότητα για κάθε κατηγορία Υμε βάση τον ακόλουθο τύπο του θεωρήματος Bayes:

$$P(X|Y) = \frac{P(Y)\prod_{i=1}^{d} P(X_i|Y)}{P(X)}$$

Δεδομένου ότι η τιμή P(X) είναι σταθερή για κάθε Y, αρκεί να επιλεγεί η κατηγορία που μεγιστοποιεί τον αριθμητή. Συνεπώς, κατασκευάζεται από ένα σύνολο εκπαίδευσης για να εκτιμήσει την πιθανότητα κάθε κατηγορίας, δεδομένων των τιμών των χαρακτηριστικών λέξεων ενός κειμένου. Το «απλοϊκό» μοντέλο Bayes (Naive Bayes model) ή αφελής Bayes είναι ένας αλγόριθμος που έχει καλή ανταπόκριση ακόμα και όταν τα χαρακτηριστικά ανεξαρτησίας που υποδεικνύει η θεωρία Naïve Bayes σαν υπόθεση, δεν ισχύουν. Ωστόσο, βασίζεται σε απλουστευμένες υποθέσεις.

- ***Νευρωνικά Δίκτυα***

Τα νευρωνικά δίκτυα αποτελούν μία νέα εξελικτική μέθοδο ταξινόμησης κειμένου όπου λαμβάνει χώρα τα τελευταία χρόνια. Η μελέτη των τεχνητών νευρωνικών δικτύων, όπως είναι πιο γνωστή η ονομασία τους, ξεκίνησε από τις προσπάθειες για προσομοίωση των βιολογικών νευρωνικών συστημάτων. Ο ανθρώπινος εγκέφαλος αποτελείται κυρίως από νευρικά κύτταρα, που ονομάζονται *νευρώνες* (neurons), τα οποία διασυνδέονται με άλλους νευρώνες μέσω *νηματοειδών ινών* (axons) που ονομάζονται *νευρίτες*. Οι νευρίτες χρησιμοποιούνται για τη μετάδοση νευρικών διεγέρσεων από έναν νευρώνα σε έναν άλλο, όταν αυτοί διεγείρονται. Ένας νευρώνας συνδέεται με τους άξονες άλλων νευρώνων μέσω *δενδριτών* (dendrites), οι οποίοι είναι προεκτάσεις από το σώμα των κυττάρων του νευρώνα. Το σημείο επαφής ανάμεσα σε έναν δενδρίτη και έναν νευρίτη ονομάζεται *νευρική σύναψη* (synapse). Οι νευρολόγοι έχουν ανακαλύψει ότι ο ανθρώπινος εγκέφαλος μαθαίνει, αλλάζοντας την ισχύ της συναπτικής σύνδεσης μεταξύ των νευρώνων κατά την επαναλαμβανόμενη διέγερση από το ίδιο ερέθισμα.



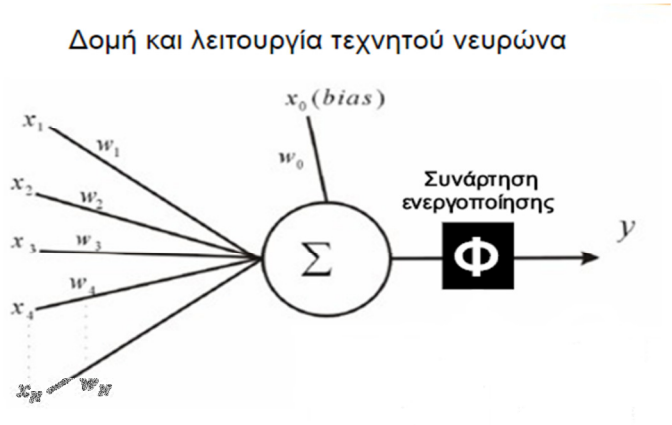
Εικόνα 17: *Δομή νευρικού συστήματος*

Πηγή forbes.com

Με τρόπο ανάλογο του ανθρώπινου εγκεφάλου, ένα τεχνητό νευρωνικό δίκτυο αποτελείται, από ένα σύνολο διασυνδεδεμένων κόμβων και προσανατολισμένων βελών. Οι ομοιότητες με τον ανθρώπινο εγκέφαλο έγκειται αρχικά στο γεγονός ότι ένα νευρωνικό

δίκτυο προσλαμβάνει τη γνώση από τον περιβάλλον του, μέσω μιας διαδικασίας μάθησης και έπειτα ότι η ισχύς των συνδέσεων μεταξύ των νευρώνων, που αποκαλείται συνοπτικό βάρος, χρησιμοποιείται για την αποθήκευση της γνώσης που αποκτιέται. Αναλυτικότερα, η δομή ενός μοντέλο νευρωνικού δικτύου αναπαρίστανται ως ένα μείγμα τριών βασικών στοιχείων:

- Ένα σύνολο *συνάψεων*, κάθε μία εκ των οποίων χαρακτηρίζεται από ένα δικό της βάρος ή δύναμη. Συγκεκριμένα ένα σήμα x_j στην είσοδο της σύναψης j που συνδέεται με ένα νευρώνα k πολλαπλασιάζεται με ένα συνοπτικό βάρος $w_{kj.}$
- Ένας *αθροιστή* για την άθροιση των σημάτων εισόδου, σταθμισμένων από τα αντίστοιχα συνοπτικά βάρη του νευρώνα και
- Μια *συνάρτηση ενεργοποίησης* για τον περιορισμό του πλάτους του σήματος εξόδου ενός νευρώνα. Η συνάρτηση ενεργοποίησης αναφέρεται επίσης και ως *συνάρτηση περιορισμού*, καθώς περιορίζει το επιτρεπτό εύρος πλάτους του σήματος εξόδου σε κάποια πεπερασμένη τιμή.



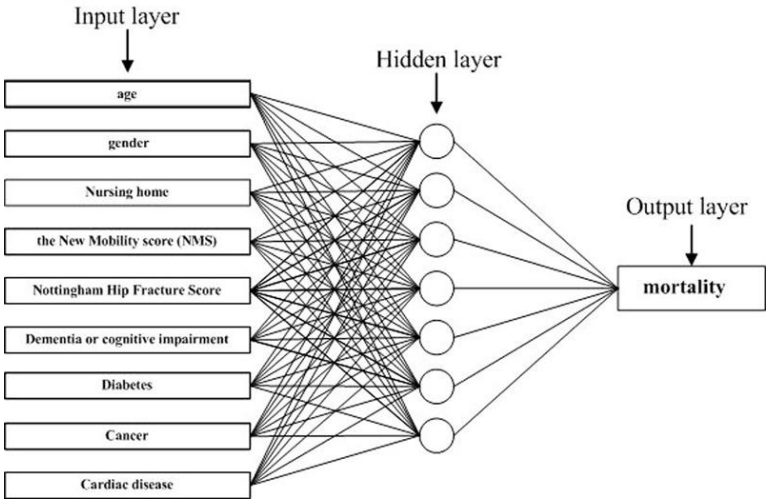
Εικόνα 17: *Δομή και λειτουργεία τεχνητού νευρώνα*
Πηγή Tan et al

Όπως είναι φανερό, οι αριθμοί οι οποίοι συναποτελούν το διάνυσμα εισόδου, αλλά και οι αριθμοί οι οποίοι συναποτελούν το διάνυσμα εξόδου , περιγράφουν χαρακτηριστικά του προς επίλυση προβλήματος. Συνήθως αυτό που μας ενδιαφέρει είναι το δίκτυο να απεικονίζει με ορθό τρόπο διανύσματα εισόδου σε κατάλληλα διανύσματα εξόδου, το πρόβλημα δηλαδή είναι η υλοποίηση μίας συνάρτησης πολλαπλών μεταβλητών, κατά κανόνα περίπλοκης και με άγνωστο ακριβή τύπο. Τέτοιες απεικονίσεις έχουν εφαρμογή σε ποικιλία τομέων της

επιστήμης και της τεχνολογίας, αφού λειτουργούν ως αριθμητικά μοντέλα για πολλά διαφορετικά ζητήματα. Το ίδιο δίκτυο μπορεί να υλοποιήσει άπειρες διαφορετικές απεικονίσεις, μία για κάθε διαφορετική επιλογή συνόλου συναπτικών βαρών[2].

Το κύριο χαρακτηριστικό των νευρωνικών δικτύων είναι η εγγενής ικανότητα μάθησης. Ως μάθηση μπορεί να οριστεί η σταδιακή βελτίωση της ικανότητας του δικτύου να επιλύει κάποιο πρόβλημα Η μάθηση επιτυγχάνεται μέσω της εκπαίδευσης, μίας επαναληπτικής διαδικασίας σταδιακής προσαρμογής των παραμέτρων του δικτύου (συνήθως των βαρών και της πόλωσής του) σε τιμές κατάλληλες ώστε να επιλύεται με επαρκή επιτυχία το προς εξέταση πρόβλημα. Αφού ένα δίκτυο εκπαιδευτεί, οι παράμετροί του συνήθως «παγώνουν» στις κατάλληλες τιμές και από εκεί κι έπειτα είναι σε λειτουργική κατάσταση. Το ζητούμενο είναι το λειτουργικό δίκτυο να χαρακτηρίζεται από μία ικανότητα γενίκευσης, αυτό σημαίνει πως δίνει ορθές εξόδους για εισόδους καινοφανείς και διαφορετικές από αυτές με τις οποίες εκπαιδεύτηκε.

Χρησιμοποιώντας αυτή τη μέθοδο, ως ένας αλγόριθμος Text Mining, κατασκευάζεται ένα ξεχωριστό νευρωνικό δίκτυο ανά κατηγορία, εκπαιδεύοντας μια μη γραμμική αντιστοίχηση από τις λέξεις εισόδου σε μια κατηγορία. Ο σχεδιασμός του είναι εύκολος για να τροποποιηθεί και διάφορα μοντέλα μπορούν να κατασκευαστούν εύκολα και γρήγορα. Το μοντέλου εξόδου ωστόσο δεν περιέχει καμία σαφή εξήγηση. Επιπρόσθετα το κόστος εκπαίδευσης είναι ιδιαίτερα υψηλό.



Εικόνα 18: *Νευρωνικό Δίκτυο*

Πηγή Braz J Med Biol

[2]https://el.wikipedia.org/wiki- Νευρωνικό Δίκτυο

- ***Support Vector Machine***

Η μέθοδος SVM (Support Vector Machine) ή μέθοδος Μηχανών Διανυσμάτων Υποστήριξης παρουσιάστηκε αρχικά από τον Vapnik και στη συνέχεια μελετήθηκε περισσότερο από τον Joachims. Η μέθοδος SVMs έχουν αποδειχθεί γρήγοροι και αποδοτικοί κατηγοριοποιητές για έγγραφα κειμένου και λύνουν το πρόβλημα των διαστάσεων, καθώς αντί του περιορισμένου αριθμού των χαρακτηριστικών, χρησιμοποιούν μια εκλεπτυσμένη δομή, η οποία δεν εξαρτάται απαραίτητα από τις διαστάσεις του χώρου εισόδου. Η δομή Support Vector Machine είναι ένα υπερεπίπεδο που διαχωρίζει ένα σύνολο θετικών παραδειγμάτων από ένα σύνολο αρνητικών παραδειγμάτων με το μέγιστο περιθώριο (margin). Το περιθώριο ορίζεται από την απόσταση του υπερεπιπέδου στα κοντινότερα θετικά και αρνητικά παραδείγματα. Το πρόβλημα βελτιστοποίησης των Μηχανών Διανυσμάτων Υποστήριξης είναι να βρεθεί μια επιφάνεια απόφασης που να μεγιστοποιεί το περιθώριο ανάμεσα στα σημεία δεδομένων σε ένα σύνολο εκπαίδευσης. Εξαιτίας, αυτής της απεικόνισης πυρήνα (kernel) ,οι υπολογισμοί περιλαμβάνουν μόνο εσωτερικά γινόμενα, οι οποίοι είναι υπολογιστικά αποδοτικοί. Έτσι, καθώς οι SVMs δεν χάνουν την αποδοτικότητα τους ή την ικανότητα τους να γενικεύουν καθώς ο αριθμός των χαρακτηριστικών εισόδου μεγαλώνει, τους κάνει ένα από τα πιο ιδανικά μοντέλα για τη ταξινόμηση εγγράφων χρησιμοποιώντας όλες τις λέξεις σε ένα κείμενο απευθείας σαν γνωρίσματα-χαρακτηριστικά.

ΚΕΦΑΛΑΙΟ 3

ΣΥΜΠΕΡΑΣΜΑΤΑ

Στην πτυχιακή αυτή εργασία εμβαθύναμε σε ένα νέο επιστημονικό πεδίο το οποίο έχει κερδίσει το αυξανόμενο ενδιαφέρον των ερευνητών τα τελευταία χρόνια και ονομάζεται Social Media Mining ή Εξόρυξη Γνώσης από Κοινωνικά Δίκτυα. Ο τομέας αυτός, της Εξόρυξης Γνώσης από τα Κοινωνικά Δίκτυα αποτελεί αλληλένδετο κομμάτι της Εξόρυξης Γνώσης από Δεδομένα καθώς με τη βοήθεια των τεχνικών του τελευταίου γίνεται προσπάθεια ανακάλυψης νέων προτύπων ή σχέσεων σε ενδιαφέρουσες πλευρές της ανθρώπινης συμπεριφοράς και της ανθρώπινης αλληλεπίδρασης.

Ο τομέας αυτός προήλθε σαν αποτέλεσμα της αλματώδης εξέλιξης του Παγκόσμιου Ιστού και συγκεκριμένα της εποχής του Web 2.0 όπου οι χρήστες έχουν τη δυνατότητα να παράγουν περιεχόμενο και να συμμετέχουν στη συγγραφή ιστοσελίδων με ποικίλους τρόπους. Έτσι τα κοινωνικά δίκτυα ως ένα μέρος αυτής της εξέλιξης έχουν μετατραπεί και αναδειχτεί στις πιο δημοφιλείς διαδικτυακές πλατφόρμες με κάποια από αυτά όπως το Facebook ή το Youtube να απαριθμούν εκατομμύρια χρήστες. Εξαιτίας αυτής λοιπόν της δημοτικότητας συμπεραίνεται ότι η ποσότητα των διαθέσιμων online δεδομένων που μπορούν να προκύψουν από αυτά έχει αυξηθεί σε εκθετικό ρυθμό συμπεριλαμβανομένου κειμένου, εικόνων, ήχου ή βίντεο. Γίνεται, λοιπόν εύκολα αντιληπτό πως αυτά τα δεδομένα παρέχουν πρωτοφανείς ευκαιρίες για έρευνα ανάλυσης των δεδομένων και ως εκ τούτου, η ανάλυση των κοινωνικών δικτύων έχει σημαντική αξία για πολλούς τομείς εφαρμογής.

Μέχρι τώρα η ανάλυση των κοινωνικών δικτύων κατηγοριοποιείται σε δύο βασικά είδη, τη Δομική Ανάλυση η οποία βασίζεται στους συνδέσμους (Linkage-based & Structural Analysis) και την Ανάλυση η οποία η οποία βασίζεται στο περιεχόμενο (Content-based Analysis). Η Δομική Ανάλυση χρησιμοποιείται για να αποκαλύψει τις δομικές ιδιότητες και τα πρότυπα της εξέλιξης των κοινωνικών δικτύων, τον εντοπισμό των κοινοτήτων, ή για μελλοντικές συνδέσεις ενώ η Ανάλυση η οποία βασίζεται στο περιεχόμενο (Content-based Analysis) αποσκοπά στη μελετή του ετερογενούς και αδόμητου περιεχόμενου που δημιουργείται από τους χρήστες κοινωνικών δικτύων, όπως τα blogs, εικόνες, βίντεο και ετικέτες (tags). Στις σημαντικότερες πρακτικές ανάλυσης για την τελευταία μέθοδο ανάλυσης περιλαμβάνονται η εξόρυξη γνώμης (opinion mining), η διερεύνηση τάσεων (trend

detection) και η συνεργατική σύσταση (collaborative recommendation). Ωστόσο και στις δύο αναλύσεις απαιτούνται τεχνικές υψηλού επιπέδου για την πραγματοποίηση τους εξαιτίας του νέου τύπου δεδομένων και εδώ είναι που η Εξόρυξη Γνώσης από Δεδομένα έρχεται να δώσει τη λύση.

Η Εξόρυξη Γνώσης από Δεδομένα αποτελεί εκείνο τον επιστημονικό κλάδο ο οποίος περιλαμβάνει την ανάπτυξη βασικών εργαλείων και μεθόδων για την αποδοτική συλλογή, αποθήκευση και αναζήτηση των συνόλων δεδομένων μέσα από μεγάλες βάσεις δεδομένων. Είναι ένας τομέας που συνδυάζει μέσω των τεχνικών του πολλές επιστήμες καθώς βασίζεται σε ιδέες όπως η δειγματοληψία ή η εκτίμηση και ο έλεγχος υποθέσεων από τον τομέα της στατιστικής, σε αλγόριθμους αναζήτησης, τεχνικές μοντελοποίησης και θεωρίες μάθησης από τη τεχνητή νοημοσύνη καθώς και αναγνώριση προτύπων και μηχανική μάθηση. Η εξόρυξη δεδομένων ή γνώσης αναφέρεται στη διαδικασία της αυτόματης ανακάλυψης χρήσιμων πληροφοριών μέσα από μεγάλες δεξαμενές δεδομένων. Οι τεχνικές εξόρυξης δεδομένων εφαρμόζονται για να ερευνήσουν σε βάθος μεγάλες βάσεις δεδομένων με σκοπό να βρεθούν νέα πρότυπα, τα οποία σε διαφορετική περίπτωση θα παρέμεναν άγνωστα. Η εξόρυξη δεδομένων αποτελεί άρρηκτο κομμάτι της ανακάλυψης γνώσης από τις βάσεις δεδομένων (KDD) η οποία αποτελεί τη συνολική διεργασία της μετατροπής ακατέργαστων δεδομένων σε γνώση για αυτό δεν είναι λίγοι αυτοί οι οποίοι συγχέουν τις δύο έννοιες. Ωστόσο, η διαδικασία ανακάλυψης γνώσης (KDD) συνηθίζεται να αναφέρεται σε ολόκληρη τη διαδικασία ανακάλυψης χρήσιμης πληροφορίας από τα μεγάλα σύνολα δεδομένων ενώ η εξόρυξη αποτελεί μόνο ένα στάδιο αυτής.

Οι εργασίες της εξόρυξης δεδομένων χωρίζονται γενικά σε δύο βασικές κατηγορίες-μοντέλα, στις λεγόμενες *Προγνωστικές Εργασίες* (Predictive tasks) και στις *Περιγραφικές Εργασίες* (Descriptive tasks).Οι προγνωστικές εργασίες έχουν ως στόχο να προβλέπουν τη τιμή ενός συγκεκριμένου χαρακτηριστικού βασιζόμενες στις τιμές άλλων χαρακτηριστικών ενώ οι περιγραφικές εργασίες από την άλλη, αποσκοπούν στο να εξάγουν υποδείγματα που συνοψίζουν τις βασικές σχέσεις που υπάρχουν στα δεδομένα.

Για κάθε μία από τις βασικές αυτές εργασίες της εξόρυξης δεδομένων υπάρχουν διάφορες τεχνικές οι οποίες χρησιμοποιούνται προκείμενου να επιτευχθεί ο στόχος αυτός, ανάλογα φυσικά και με το βαθμό απαίτησης των εφαρμογών που πρόκειται να εκτελεστούν. Οι κυριότερες τεχνικές της εξόρυξης δεδομένων είναι η παλινδρόμηση, η κατηγοριοποίηση, η

συσταδοποίηση, η ανάλυση χρονολογικών σειρών, η παρουσίαση συνόψεων, οι κανόνες συσχέτισης και η ανακάλυψη ακολουθιών.

Η Εξόρυξη Γνώσης από Κείμενο, έρχεται να αντικαταστήσει την Εξόρυξη Γνώσης από Δεδομένα στην περίπτωση που η μορφή των δεδομένων βρίσκονται υπό την μορφή κειμένων. Κατά κανόνα προσεγγίζεται από μεθόδους και αλγορίθμους της μηχανικής μάθησης και της εξόρυξης δεδομένων και αποσκοπεί να ανακαλύψει ενδιαφέροντα στοιχεία όπως ομάδες ή συσχετίσεις μέσα από μεγάλες συλλογές εγγράφων. Η σύνδεση της Εξόρυξης Γνώσης από Κείμενο με τη διαδικασία της Εξόρυξης Γνώσης από τα Κοινωνικά Δίκτυα έγκειται στο γεγονός ότι τα δεδομένα των κοινωνικών δικτύων είναι σε δομημένο ή μη δομημένο κείμενο. Η διαδικασία της ανακάλυψης από κείμενο πραγματοποιείται σε τρία στάδια, τα οποία και είναι η Συλλογή των Δεδομένων, η Προεπεξεργασία των εγγράφων και η Εξόρυξη Γνώσης από κείμενο. Το στάδιο της προ-επεξεργασίας των κειμένων είναι ένα ιδιαίτερα σημαντικό στάδιο της διαδικασίας της ανακάλυψης γνώσης από κείμενο , καθώς τα δεδομένα των κοινωνικών δικτύων είναι μεγάλα, θορυβώδη και δυναμικά. Μεταξύ των σημαντικότερων μεθόδων προ-επεξεργασίας είναι η αφαίρεση των τετριμμένων λέξεων καθώς και η διαδικασία του stemming.

Ωστόσο εκτός από την αφαίρεση των τετριμμένων λέξεων και τη διαδικασία του stemming, το ίδιο επίπεδο σημαντικότητας χαρακτηρίζει και τη διαδικασία της μείωσης των διαστάσεων των χαρακτηριστικών εξαιτίας του υπολογιστικού κόστους και του επιπέδου λειτουργίας των αλγόριθμων μηχανικής μάθησης. Η μείωση των διαστάσεων μπορεί να προσεγγιστεί είτε με τη μέθοδο της επιλογής των χαρακτηριστικών όπου επιλέγονται εκείνα τα χαρακτηριστικά τα οποία θεωρούνται πιο σημαντικά με βάση συγκεκριμένες στατιστικές είτε με τη μέθοδο εξαγωγής χαρακτηριστικών όπου ο χώρος των διαστάσεων των χαρακτηριστικών μέσω μιας διαδικασίας μετασχηματισμού μετατρέπονται σε ένα χώρο με λιγότερες διαστάσεις. Τα κριτήρια με τα οποία επιλέγονται τα χαρακτηριστικά στην μέθοδο μείωσης των διαστάσεων είναι η Συχνότητα του Έγγραφου, το Πληροφοριακό Κέρδος, η Αμοιβαία Πληροφόρηση, η Στατιστική X^2 και ο Λόγος Πιθανοτήτων. Τα χαρακτηριστικά τα οποία λαμβάνουν μια υψηλή τιμή των κριτηρίων αυτομάτως θεωρούνται και πιο σημαντικά.

Όπως και στην Εξόρυξη Γνώσης από Δεδομένα έτσι και στην Εξόρυξη Γνώσης από Κείμενο έχουν αναπτυχθεί τεχνικές που χρησιμοποιούνται για τον σκοπό αυτό. Οι κυριότερες εκ των οποίων είναι η Εξαγωγή Πληροφοριών (Information Extraction), η Κατηγοριοποίηση

(Categorization), η Ομαδοποίηση (Clustering), η Απεικόνιση Πληροφορίας (Information Visualization) και η Σύνοψη (Summarization) .

Ολοκληρώνοντας, προκειμένου να μπορέσει να πραγματοποιηθεί το τελικό στάδιο της ανακάλυψης γνώσης από κείμενο είναι απαραίτητο να εφαρμοστούν οι αντίστοιχες τεχνικές του Text Mining σε συνδυασμό με κάποιο αλγόριθμο της μηχανικής μάθησης. Οι πιο διαδομένοι αλγόριθμοι εξόρυξης κειμένου είναι τα νευρωνικά δίκτυα, τα δέντρα απόφασης, ο K-nearest Neighbor, ο Support Vector Machine, τα δέντρα απόφασης και ο αφελής Bayes. Ο K-nearest Neighbor προσπαθεί να κατηγοριοποιήσει ένα νέο έγγραφο, βρίσκοντας τα πιο όμοια με αυτό έγγραφα στο σύνολο εκπαίδευσης. Τα δέντρα αποφάσεων είναι ένας κατηγοριοποιητής βασισμένος στις πιθανότητες όπου η συνάρτηση confidence (κλάση) αναπαριστά μια κατανομή πιθανότητας. Είναι εύκολο να ερμηνευτούν, ωστόσο απαιτούν έναν αριθμό παραμέτρων μοντέλου που είναι δύσκολο να βρεθούν και η εκτίμηση του λάθους είναι δύσκολη. Ο Αφελής κατά Bayes αλγόριθμος , εκτιμά την εξαρτώμενη από την κατηγορία πιθανότητα υποθέτοντας, ότι τα χαρακτηριστικά είναι υπό συνθήκη ανεξάρτητα δεδομένης μιας κατηγορίας *y* ενώ τέλος η δομή Support Vector Machine είναι ένα υπερεπίπεδο που διαχωρίζει ένα σύνολο θετικών παραδειγμάτων από ένα σύνολο αρνητικών παραδειγμάτων με το μέγιστο περιθώριο (margin).

ΕΛΛΗΝΙΚΗ ΒΙΒΛΙΟΓΡΑΦΙΑ

Μπιρμπίλη Μ, Πασχάλης Γ, Κωτσιαντής Σ,(2013), *Εφαρμογή αυτόματης κατάταξης γνώμης σε δεδομένα του Twitter*

Παπάνης, Ε., Γιαβρίμης, Π., Βίκη, Α. & Παπάνης Α. (2011). *Τα κοινωνικά δίκτυα μαθητών με ειδικές εκπαιδευτικές ανάγκες και η επίδρασή τους στη σχολική επίδοση, την αυτοεκτίμηση και την κοινωνική προσαρμογή τους σύμφωνα με τις απόψεις των εκπαιδευτικών*

Σωτηριάδου, Α. & Παπαδάκης, Σ. (2012), *Αξιοποίηση των Κοινωνικών Δικτύων για τη Διδακτική της Πληροφορικής σε Ενηλίκους – Βιβλιογραφική ανασκόπηση*, 6ο Πανελλήνιο Συνέδριο Διδακτική της Πληροφορικής, 417-426, Φλώρινα

Χαλκίδη Μ., Βαζιργιάννης Μ. (2005), *Εξόρυξη Γνώσης από τον Παγκόσμιο Ιστό*, Εκδόσεις Τυπωθύτω, Αθήνα

Χάλκος Γ. (2011), *Στατιστική- Θεωρία, Εφαρμογές και Χρήση Στατιστικών Προγραμμάτων σε Η/Υ*, Εκδόσεις Τυπωθύτω, Αθήνα

ΞΕΝΗ ΒΙΒΛΙΟΓΡΑΦΙΑ

Bogdan B , Treleaven P.(2015), *Social media analytics: a survey of techniques, tools and platforms, Open Forum*, AI & Soc (2015) 30:89–116

Boyd, D. M., & Ellison, N. B. (2010). *Social network sites: definition, history, and scholarship,* Engineering Management Review, IEEE, 38(3), 16-31.

Cayari, C. (2011), *The YouTube effect: How YouTube has provided new ways to consume, create, and share music*., International Journal of Education & the Arts, 12(6), 1-28.

Choudhary et al (2009), *The needs and benefits of Text Mining applications on Post-Project Reviews, Journal Computer in Industry*, 60 (9), pp. 728-740

Dunham M. (2004), *Εισαγωγικά και Προηγμένα Θέματα Εξόρυξης Γνώσεις από Δεδομένα, μετάφραση Εκδόσεις Νέων Τεχνολογιών*, Αθήνα

Fayyad U. et al (1991), *Advances in Knowledge Discovery and Data Mining*, Massachusetts Institute of Technology, London

Kotler, P et al (2008), Marketing Management, 13th ed., *Chapter 17: Designing and managing integrated marketing communications*, United States: Prentice Hall pp 459, 477.

Hand et al (2001), *Principles of Data Mining*, MIT Press, Cambridge

Hansen,D. et al (2011), *Analysing Social Media Networks with NodeXL*, *Insights from a Connected World*., Massachusetts, MA: Elsevier Inc.

Haykin S (2009), Νευρωνικά Δίκτυα και Μηχανική Μηχανική Μάθηση, μετάφραση Εκδόσεις Παπασωτηρίου, Αθήνα

Tan et al (2005), *Εισαγωγή στην Εξόρυξη Δεδομένων*, μετάφραση Εκδόσεις Τζιόλα, Αθήνα