ΤΕΧΝΟΛΟΓΙΚΟ
ΕΚΠΑΙΔΕΥΤΙΚΟ
ΙΔΡΥΜΑ
ΔΥΤΙΚΗΣ ΕΛΛΑΔΑΣ

Τ.Ε.Ι. ΔΥΤΙΚΗΣ ΕΛΛΑΔΑΣ

ΤΜΗΜΑ ΔΙΟΙΚΗΣΗΣ ΕΠΙΧΕΙΡΗΣΕΩΝ ΠΑΤΡΑΣ

Πτυχιακή Εργασία

της

Μανιά Αγγελικής

" Έξυπνη Εξόρυξη Δεδομένων από το Διαδίκτυο "

Επιβλέπων Καθηγητής:

Κωνσταντίνος Γιωτόπουλος

ΠΑΤΡΑ

ΑΠΡΙΛΙΟΣ 2018

Πρόλογος

Η συνεχόμενη αύξηση του όγκου των δεδομένων στους τομείς των σύγχρονων επιχειρήσεων και της επιστήμης απαιτεί πιο πολύπλοκα και εξελιγμένα εργαλεία. Παρά το γεγονός ότι οι συνεχείς πρόοδοι της τεχνολογίας στον τομέα της εξόρυξης δεδομένων (data mining) έχουν κάνει την διαδικασία της εκτεταμένης συλλογής δεδομένων πολύ πιο εύκολη, εντούτοις υπάρχει συνεχής ανάγκη για νέες τεχνικές και εργαλεία που μπορούν να μας βοηθήσουν να μετατρέψουμε αυτά τα δεδομένα σε χρήσιμες πληροφορίες και γνώσεις.

Η πρόοδος της τεχνολογίας στο υλικό και το λογισμικό μας επιτρέπει να συλλέγουμε, αποθηκεύουμε και αναλύουμε μεγάλα ποσά δεδομένων (big data), από ποικίλες πηγές, για μεγάλες και μικρές επιχειρήσεις. Η επιστήμη δεδομένων (data science) εφαρμόζει αναλυτικές τεχνικές στα δεδομένα για την εύρεση υποκείμενων σχέσεων με στόχο τη βελτίωση της λειτουργίας του οργανισμού και την παραγωγή αξίας.

Ιδιαίτερα στον σύγχρονο κόσμο των επιχειρήσεων, η εκμετάλλευση αυτών των δεδομένων και η γνώση που μπορεί να εξαχθεί από αυτά, αποτελεί, μαζί με το ανθρώπινο δυναμικό της, τον πλέον πολύτιμο πόρο για την ανάπτυξη, την κερδοφορία και την επιβίωση της επιχείρησης μέσα σε ένα άκρως ανταγωνιστικό περιβάλλον.

Η επιτυχής διαχείριση και εκμετάλλευση των δεδομένων σε συνδυασμό με τα λεγόμενα πληροφοριακά συστήματα, συστήματα που μας επιτρέπουν να οργανώνουμε και να αναλύουμε τα δεδομένα, αποτελούν τον κύριο συντελεστή μεγέθυνσης της παραγωγικότητας μιας επιχείρησης και έχουν καίρια σημασία για την αποτελεσματικότητα των σύγχρονων επιχειρήσεων και των οργανισμών, μέσα σε μια ταχύτατα εξελισσόμενη παγκόσμια οικονομία.

Στην παρούσα πτυχιακή εργασία θα προσπαθήσουμε αρχικά να ορίσουμε και να περιγράψουμε ένα σχετικά πρόσφατο ερευνητικό πεδίο, που αποκαλείται “**Εξόρυξη Δεδομένων από τον**

Παγκόσμιο Ιστό”. (Web Data Mining) και εκμεταλλεύεται πλήθος μεθόδων του (π.χ. ομαδοποίηση).

Πιο αναλυτικά, η παρούσα πτυχιακή εργασία εστιάζει αρχικά στις βασικές αρχές και τις εφαρμογές εξόρυξης δεδομένων. Δίνεται ιδιαίτερη βάση στα είδη Βάσεων Δεδομένων που μπορούν να χρησιμοποιηθούν από τους διάφορους μηχανισμούς Εξόρυξης Δεδομένων και γίνεται παρουσίαση του συστήματος εξόρυξης γνώσης.

Στη συνέχεια μελετούνται οι αλγόριθμοι εξόρυξης δεδομένων. Οι αλγόριθμοι εξόρυξης δεδομένων που χρησιμοποιούνται στον Παγκόσμιο Ιστό και ειδικά στις μηχανές αναζήτησης είναι πολυάριθμοι. Στην παρούσα ενότητα θα παρουσιάσουμε τους πιο σημαντικοί από αυτούς, όπως τα δέντρα απόφασης (decision trees), τα τεχνητά νευρωνικά δίκτυα (artificial neural networks), η ανάλυση συσχετίσεων (Link Analysis), η συσταδοποίηση (Clustering) και η παλινδρόμηση (Regression) και τέλος πως γίνεται η προετοιμασία των δεδομένων για εξόρυξη.

Ακολούθως, στο τρίτο κεφάλαιο θα αναφερθούμε στις μηχανές αναζήτησης που υπάρχουν στον Παγκόσμιο Ιστό. Στην ενότητα αυτή παρουσιάζονται ορισμένες βασικές έννοιες σχετικά με την Ανάκτηση Πληροφορίας στον Παγκόσμιο Ιστό. Αρχικά δίνεται μία σύντομη περιγραφή του διαδικτύου και κάποια στοιχεία για την ανάπτυξή του. Μετά αναφέρονται οι ιδιαιτερότητες της ανάκτησης πληροφορίας στον Παγκόσμιο Ιστό σε σχέση με τις συμβατικές εφαρμογές ανάκτησης πληροφορίας. Στη συνέχεια περιγράφονται τα βασικά χαρακτηριστικά μηχανών αναζήτησης, περιγράφεται περιληπτικά η γλώσσα HTML, που χρησιμοποιείται για τη δημιουργία σελίδων του παγκόσμιου ιστού και τέλος θα γίνει μια σύντομη περιγραφή του τρόπου λειτουργίας των μηχανών αναζήτησης, του τρόπου με τον οποίο οι συγκεκριμένες μηχανές διαχειρίζονται ένα τόσο μεγάλο αριθμό δεδομένων από όλες τις ιστοσελίδες του διαδικτύου και γίνεται μια ιστορική αναδρομή σε αυτές, καθώς και παρουσίαση των πιο ευρέως χρησιμοποιουμένων μηχανών αναζήτησης (Search Engines) της σύγχρονης εποχής.

Τέλος, θα αναλύσουμε παραδείγματα εφαρμογών εξόρυξης δεδομένων στον παγκόσμιο ιστό. Θα παρουσιαστούν πρακτικές εφαρμογές των παραπάνω, σε τομείς των σύγχρονων Κοινωνικών Δικτύων, της σύγχρονης ψηφιακής αγοράς, των ψηφιακών μέσων επικοινωνίας καθώς και του τομέα των σύγχρονων γεωγραφικών συστημάτων πληροφορίας (G.I.S.), ως αποτελέσματα των μεθοδολογιών αυτών.

Abstract

The constant increase in the field of data in modern business and science requires more sophisticated tools. Although the constant advances in data mining technology have made the data collection process much easier, there is always a need for new techniques and tools that can help us turn this data into useful information and knowledge.

The advancement of technology in hardware and software allows us to collect, store and analyze large amounts of data from a variety of sources for all kind of businesses. Data science applies analytical techniques to data for finding underlying relationships to improve the organization's performance and value creation.

The exploitation of this data, as well as the knowledge it can derive from it, together with its human resources, is the most important resource for the company's growth, profitability and persistence in a highly competitive environment.

Successful data management and exploitation in conjunction with the so-called information systems, systems that allow data organization and analysis, are the main factor for increasing a company's productivity and are crucial for the efficiency of modern businesses and organizations within a rapidly evolving global economy.

In this diploma thesis we will initially attempt to define and describe a relatively recent research field, called "Data Mining from the World Wide Web" (Web Data Mining) and exploit its multitude of methods (e.g., clustering).

More specifically, this thesis focuses primarily on basic data mining and data mining applications. Particular importance is given to the kinds of databases that can be used by the various Data Mining mechanisms and the knowledge mining system is presented.

Then data mining algorithms are being studied. Data mining algorithms used on the Web, and especially on search engines, are numerous. In this section, the most important of these will be

presented, such as decision trees, artificial neural networks, relationship analysis, clustering and regression and finally how data is being prepared for extraction.

Then, in the third chapter, we will refer to the search engines on the Web. This section introduces some basic concepts of Web Information Recovery. Firstly, a brief description of the internet and some details about its development are given. Here are the peculiarities of information retrieval on the Web in relation to conventional information retrieval applications. The basic search engines are outlined, the HTML language, which is used to create web pages is briefly described, and finally we will give a short description of how search engines work, how these machines handle such a large number data from all web sites and a historical review of them, as well as a presentation of the most widely used search engines (Search Engines) of the modern season.

Finally, we will analyze examples of data mining applications on the web. Practical applications of the above, in the areas of modern social networks, the modern digital market, digital media and the modern field of Geographic Information Systems (G.I.S.), will be presented as a result of these methodologies.

Περιεχόμενα

Ευρετήριο Εικόνων

Ευρετήριο Πινάκων

Κεφάλαιο 1

1.1 Γενικά για την Εξόρυξη Δεδομένων Στον Παγκόσμιο Ιστό

Το Διαδίκτυο (Internet) σήμερα έχει εδραιωθεί ως ένα από τα μεγαλύτερα μέσα μαζικής επικοινωνίας. Σε αυτή την επιτυχία έχει συμβάλει ο τεράστιος όγκος πληροφορίας που αποθηκεύεται στους υπολογιστές που το συγκροτούν σε παγκόσμια κλίμακα, καθώς και η χρήση νέων τεχνολογιών που εξασφαλίζουν γρήγορη πρόσβαση σε αυτό (aDSL, T1, T3 κ.λπ.). Η ευκολία διακίνησης και ανάκτησης της πληροφορίας αυτής μέσω του Παγκοσμίου Ιστού(Π.Ι.) (World Wide Web ή WWW) παρέχει νέες δυνατότητες πληροφόρησης (web logs, news groups) μάθησης (e-learning), ανταλλαγής ιδεών (forums), κατανάλωσης αγαθών (e-commerce) κ.λπ. Ωστόσο, τα δεδομένα που διακινούνται μέσω του Παγκοσμίου Ιστού είναι συνήθως μη-δομημένα, ακατέργαστα και προσφέρονται με τυχαίο τρόπο. Ταυτόχρονα οι χρήστες /πελάτες / καταναλωτές συχνά απορρίπτουν ή αγνοούν μεγάλο μέρος της διαθέσιμης πληροφορίας. Αυτό οφείλεται αρχικά στο ότι οι περισσότεροι χρήστες δεν είναι συχνά καλοί γνώστες του Παγκοσμίου Ιστού και αντιμετωπίζουν δυσκολίες στον εντοπισμό της, στους μεγάλους χρόνους απόκρισης των διακομιστών, ακόμη και στη μη ελκυστική εμφάνιση του διαδικτυακού τόπου που την φιλοξενεί. Από την άλλη μεριά οι επιχειρήσεις / οργανισμοί / ανεξάρτητοι φορείς που παρέχουν αυτά τα δεδομένα είτε ως ηλεκτρονικές υπηρεσίες (ecommerce, e-learning, e-tourism) είτε ως αντικείμενα απλής πληροφόρησης (σελίδες, βίντεο, μουσική, εικόνες) ενδιαφέρονται κυρίως για την ικανοποίηση των χρηστών, για την ασφάλεια του διαδικτυακού τόπου καθώς και για την απόδοση του συστήματος. Για την αντιμετώπιση αυτών των προβλημάτων, την τελευταία δεκαετία έχουν πραγματοποιηθεί σημαντικές ερευνητικές εργασίες αντλούμενες από την Ανακάλυψη Γνώσης σε Βάσεις Δεδομένων (Α.Γ.) (KDD – Knowledge Discovery in Databases). Η Ανακάλυψη Γνώσης σε Βάσεις Δεδομένων είναι η μη τετριμμένη εξαγωγή συνεπαγόμενης, άγνωστης και ενδεχομένως χρήσιμης πληροφορίας από τα υπάρχοντα δεδομένα [FPM91]. Η εφαρμογή Α.Γ. σε δεδομένα που έχουν συλλεχθεί από τον Π.Ι. ονομάστηκε Εξόρυξη Δεδομένων στον Παγκόσμιο Ιστό

(Web Usage Mining). Ως δεδομένα ή αντικείμενα στον Π.Ι. θεωρείται η πληροφορία που παρέχεται με τη μορφή:

- Ημι-δομημένων εγγράφων (π.χ. HTML αρχεία)
- Δομημένων εγγράφων (π.χ. XML αρχεία)
- Πολυμεσικών εφαρμογών (π.χ. εικόνες, βίντεο, μουσική)
- Μικρό-εφαρμογές (π.χ. cgi-scripts)
- Άλλα είδη αρχείων (π.χ. pdf,ps,cookies)
- Αρχείων καταγραφής των συναλλαγών μεταξύ ενός εξυπηρέτη ιστοσελίδων (Web Server) ή πληρεξούσιου εξυπηρέτη (Proxy Server) και ενός πελάτη που ζητά αντικείμενα ή δεδομένα από αυτόν.

Σκοπός της Εξόρυξης Δεδομένων από τον Π.Ι. είναι η εξαγωγή (ανακάλυψη) μοτίβων από τα παραπάνω αντικείμενα, που ενώ υπάρχουν δεν διακρίνονται εύκολα. Ως μοτίβα θεωρούμε -μη φανερές χωρίς επεξεργασία-σχέσεις μεταξύ των αντικειμένων που θα μπορούσαν να φανούν χρήσιμες και να βελτιώσουν τον Π.Ι.
Η Εξόρυξη Δεδομένων στον Π.Ι. έχει χωριστεί σε τρεις διαφορετικές περιοχές:

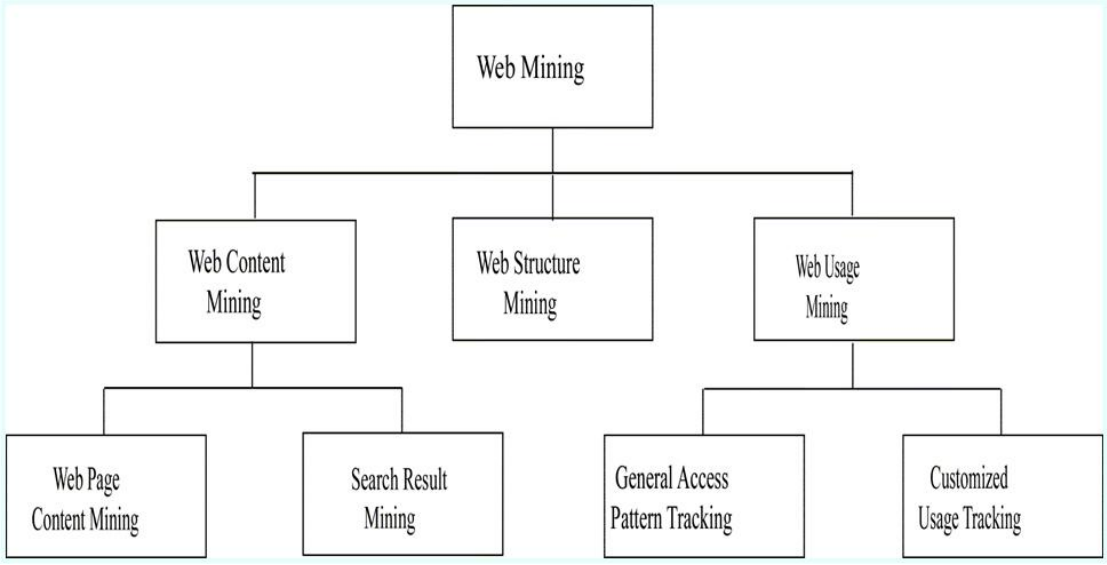
1. ***Εξόρυξη Δεδομένων βασισμένη στο Περιεχόμενο (Web Content Mining)*:** Η Ε.Δ.Π. είναι η διαδικασία εξαγωγής μοτίβων από το περιεχόμενο των ιστοσελίδων (π.χ. HTML έγγραφα). Ως περιεχόμενο των ιστοσελίδων θεωρούνται τα αντικείμενα του Π.Ι. με τα οποία αλληλεπιδρούν οι χρήστες, όπως είναι τα πολυμεσικά αρχεία και τα ημι-δομημένα έγγραφα. Ωστόσο η Ε.Δ.Π. ασχολείται κυρίως με την επεξεργασία του κειμένου των HTML σελίδων και για αυτό συχνά αποκαλείται Εξόρυξη Κειμένου στον Π.Ι. (Web Text Mining). Επειδή το περιεχόμενο στον Π.Ι. είναι συνήθως αδόμητο ή ημι-δομημένο χρειάζεται να εφαρμοστούν ειδικές τεχνικές ώστε τα υπάρχοντα δεδομένα να αναπαρασταθούν σε κατάλληλη για επεξεργασία μορφή. Παράδειγμα αποτελεί η εξαγωγή των λέξεων από τα HTML έγγραφα ενός ή περισσότερων διαδικτυακών τόπων και η αναπαράστασή τους σε κατάλληλη δομή δεδομένων (π.χ. πίνακας). Για την εξαγωγή μοτίβων από τα δομημένα μοντέλα που προκύπτουν χρησιμοποιούνται κυρίως τεχνικές Ανάκτησης Πληροφορίας (Information Retrieval) και Επεξεργασίας Φυσικής Γλώσσας (Natural Language Processing). Κύριος στόχος της Ε.Β.Π. είναι η δόμηση των αντικειμένων στον Π.Ι.

για ευκολότερη διαχείριση της πληροφορίας που θα παρέχει αποτελεσματικότερες αναζητήσεις (queries) περιεχομένου στον Π.Ι. Η εννοιολογική τιμαριθμοποίηση (concept indexing) των αντικειμένων του Π.Ι. που χρησιμοποιείται από πολλές μηχανές αναζήτησης (search engines) αποτελεί εφαρμογή της Ε.Β.Π..

2. ***Εξόρυξη Δεδομένων βασισμένη στη Δομή (Web Structure Mining)*:** Η Ε.Δ.Δ. είναι η διαδικασία εξαγωγής μοτίβων από τη δομή του Π.Ι. Ως δομή του Π.Ι. θεωρείται η οργάνωση που διαμορφώνεται από τους (από και προς) συνδέσμους (hyperlinks, links, υπερ-σύνδεσμοι) που περιλαμβάνουν οι HTML σελίδες. Ο Π.Ι. μπορεί να αναπαρασταθεί ως ένας γράφος με κόμβους (κορυφές) τις ιστοσελίδες και με ακμές τους συνδέσμους μεταξύ των σελίδων. Με μελέτη αυτών των δομών (γράφων) είναι δυνατόν να ανακαλυφθούν μοτίβα όπως κοινότητες ιστοσελίδων ή διαδικτυακών τόπων που είναι στενά συνδεδεμένες μεταξύ τους καθώς επίσης και αρκετά δημοφιλείς διαδικτυακοί τόποι.

3. ***Εξόρυξη Δεδομένων βασισμένη στη Χρήση (Web Usage Mining)*:** Η Ε.Δ.Χ. είναι η διαδικασία εξαγωγής μοτίβων (μη φανερή αλλά χρήσιμη πληροφορία) από τα δεδομένα χρήσης (usage data) που αποθηκεύονται σε έναν πηγαίο1 ή πληρεξούσιο εξυπηρέτη ή ακόμη και στην πλευρά του χρήστη (π.χ. cookies). Τα δεδομένα χρήσης που χρησιμοποιεί κυρίως η Ε.Δ.Χ. είναι τα αρχεία καταγραφής κίνησης2 των χρηστών (Web Log Files) ενός διαδικτυακού τόπου και για αυτό η Ε.Δ.Χ. συχνά αποκαλείται Εξόρυξη Δεδομένων από τα Αρχεία Καταγραφής (Web Log Mining). Αυτά τα αρχεία αποθηκεύουν αρκετή πληροφορία, όπως είναι η ώρα και το αντικείμενο που ζητήθηκαν από κάποιον χρήστη. Μοντελοποιώντας και φιλτράροντας κατάλληλα αυτά τα αρχεία είναι δυνατόν να μελετήσουμε την συμπεριφορά των χρηστών στον Π.Ι. εφαρμόζοντας διάφορες μεθόδους από την Α.Γ. Η μελέτη της συμπεριφοράς των χρηστών μπορεί να αποκαλύψει γνώση που προηγουμένως δεν ήταν ορατή. Παραδείγματος χάριν, είναι πιθανόν μετά από μελέτη να διαπιστωθεί ότι οι χρήστες από Ελλάδα που επισκέπτονται το διαδικτυακό τόπο κοιτούν κάποιες συγκεκριμένες ιστοσελίδες. Αυτή η γνώση μπορεί να βοηθήσει στη βελτίωση του Π.Ι. δημιουργώντας νέους αλλά

και τροποποιώντας τους υπάρχοντες διαδικτυακούς τόπους με βάση τις συμπεριφορές των χρηστών. Σε αυτήν την πτυχιακή εργασία εξετάστηκε αναλυτικά η τρίτη κατηγορία της Εξόρυξης Δεδομένων στον Π.Ι. η οποία παρουσιάζεται εν συντομία στη συνέχεια και αναλυτικότερα σε επόμενο κεφάλαιο. Οι περιοχές της εξόρυξης δεδομένων του Π.Ι απεικονίζονται στην εικόνα 1.1.



Εικόνα 1.1: Περιοχές data mining στον Παγκόσμιο Ιστό

Παρά το γεγονός ότι υπάρχει μια γενικότερη συμφωνία ότι ο στόχος της εξόρυξης δεδομένων είναι η ανακάλυψη νέας και χρήσιμης πληροφορίας σε βάσεις δεδομένων, τα μέσα για την επίτευξη του στόχου αυτού ποικίλουν σε πολύ υψηλό βαθμό. Η εξόρυξη γνώσης περιλαμβάνει ένα ευρύ πεδίο υπολογιστικών μεθόδων που μεταξύ άλλων περιλαμβάνουν, την στατιστική ανάλυση (statistical analysis), τα δένδρα αποφάσεων (decision trees), τα νευρωνικά δίκτυα (neural networks), την εξαγωγή κανόνων (rule induction) και την γραφική οπτικοποίηση (graphic visualization).

Τέτοιες μέθοδοι χρησιμοποιούνται για την εύρεση συσχετίσεων, προτύπων και δομών σε μεγάλες και διαρκώς αυξανόμενες βάσεις δεδομένων. Ειδικά η εύρεση εργαλείων είναι ένα ιδιαίτερα σημαντικό εξαγόμενο της εξόρυξης δεδομένων μέσω σχέσεων μεταξύ των χαρακτηριστικών των βάσεων δεδομένων.

1.2 Εξόρυξη Δεδομένων και Ανακάλυψη Γνώσης

Ο όρος εξόρυξη δεδομένων αναφέρεται στην εξόρυξη ή την ανακάλυψη νέων πληροφοριών με την μορφή κανόνων ή προτύπων από πηγές δεδομένων. Για να είναι πρακτικά χρήσιμες αυτές οι πληροφορίες πρέπει να έχουν εξαχθεί από μεγάλες βάσεις δεδομένων και αρχεία. Η εξόρυξη δεδομένων χρησιμοποιεί τεχνικές από την μηχανική μάθηση, την στατιστική, τα νευρωνικά δίκτυα κοκ.

Εικόνα 1.2: Η Εξόρυξη Δεδομένων ως συμβολή πολλαπλών αρχών

Οι αποθήκες δεδομένων μπορούν να χρησιμοποιηθούν για να υποστηρίξουν την εξόρυξη δεδομένων. Στο σύγχρονο κόσμο του Διαδικτύου και των επιχειρήσεων δεδομένα δημιουργούνται διαρκώς από τράπεζες, τηλεπικοινωνίες, εμπορικές συναλλαγές, το διαδίκτυο, το ηλεκτρονικό εμπόριο ακόμη και από επιστημονικά άρθρα, πειράματα και αποτελέσματα. Το σύνολο αυτών των δεδομένων αποθηκεύονται στις λεγόμενες αποθήκες δεδομένων.

Ο ορισμός της αποθήκης δεδομένων εστιάζει στην αποθήκευση δεδομένων. Η κύρια πηγή των δεδομένων ξεδιαλύνεται, μεταμορφώνεται, κατηγοριοποιείται και διατίθεται με σκοπό τη χρήση της για την εξόρυξη δεδομένων, διαδικτυακής αναλυτικής επεξεργασίας, έρευνα αγοράς και υποστήριξης αποφάσεων.

Ωστόσο, τα μέσα για την ανάκτηση και την ανάλυση δεδομένων, την εξαγωγή, μετατροπή και φόρτωση των δεδομένων, καθώς και η διαχείριση του λεξικού δεδομένων θεωρούνται επίσης ουσιώδεις συνιστώσες ενός συστήματος αποθήκευσης δεδομένων.

Σημαντικό ρόλο στην ανακάλυψη των συγκεκριμένων μέσων και τεχνικών για τη σωστή αξιοποίηση των δεδομένων διαδραματίζει η ανακάλυψη γνώσης από βάσεις δεδομένων, η οποία και αποτελεί τη διαδικασία εντοπισμού έγκυρων, εν δυνάμει χρήσιμων και κατανοητών πρότυπων (patterns) σε δεδομένα.

Η ανακάλυψη γνώσης από βάσεις δεδομένων (KDD - Knowledge Discovery in Databases) αφορά την παραγωγή λειτουργικής γνώσης, μέσω της ανάλυσης των δεδομένων από μεγάλες αποθήκες, καθώς και την εύρεση των απαραίτητων δομών γνώσης που αναδεικνύουν συσχετίσεις ή κανόνες που είναι κρυμμένοι εντός των δεδομένων και δεν μπορούν να εξαχθούν από τον ίδιο τον άνθρωπο χωρίς τη χρήση μηχανής. Στην ουσία αναφέρεται στην όλη διαδικασία που πρέπει να ακολουθηθεί ώστε να προκύψει η απαραίτητη γνώση από την απλή αποθήκευση των δεδομένων.

Ένας ορισμός του τι είναι ανακάλυψη γνώσης από βάσεις δεδομένων είναι ο παρακάτω:

Ορισμός - "KDD είναι η ντετερμινιστική διαδικασία αναγνώρισης έγκυρων, καινοτόμων, ενδεχομένως χρήσιμων και εν τέλει κατανοητών προτύπων στα δεδομένα" (Frawley, Piatesky-Shaphiro and Matheus, 1992)[1]

Η εξόρυξη των δεδομένων αυτών αποτελεί τμήμα της διαδικασίας ανακάλυψης γνώσης από βάσεις δεδομένων (KDD - Knowledge Discovery in Databases) και αναφέρεται κυρίως στις τεχνικές που μπορεί κάποιος να ακολουθήσει ώστε να φθάσει στην απαιτούμενη γνώση.

Η βασική διαδικασία της ανακάλυψης γνώσης αποτελείται από 6 φάσεις: καθορισμό και καθαρισμό των δεδομένων που θα χρησιμοποιηθούν, εμπλουτισμό και ενσωμάτωση, επιλογή των κατάλληλων δεδομένων προς επεξεργασία, κατάλληλη

[1] **William J. Frawley, Gregory Piatetsky-Shapiro, and Christopher J. Matheus**, Knowledge Discovery in Databases: An Overview, 1992

τροποποίησή τους ώστε να εφαρμοστούν οι τεχνικές εξόρυξης, εξόρυξη, και τέλος η δημιουργία αναφορών και η αξιολόγησή τους.

Παρακάτω ακολουθούν επιγραμματικά τα βήματα που ακολουθούνται για την ανακάλυψη γνώσης, καθώς θα αναλυθούν εκτενέστερα σε επόμενα κεφάλαιο.

- **Επιλογή δεδομένων (Data selection):** Απόκτηση δεδομένων προς επεξεργασία από ετερογενείς πηγές. Επιλέγονται προσεκτικά εκείνα που είναι σχετικά και χρήσιμα για την ανάλυση που θα ακολουθήσει.

- **Προεπεξεργασία Δεδομένων** (**Data preprocessing**): Στο βήμα αυτό, αφαιρούνται τα δεδομένα που παράγουν θόρυβο, δηλαδή όλα εκείνα τα στοιχεία που μπορούν να επηρεάσουν ή και να διαστρεβλώσουν το αποτέλεσμα, γίνεται τακτοποίηση και επεξεργασία των δεδομένων και εσφαλμένα, προβληματικά ή ελλείποντα δεδομένα τροποποιούνται. Το συγκεκριμένο στάδιο καταναλώνει το 50% - 60% της γενικής προσπάθειας για ανακάλυψη γνώσης από βάσεις δεδομένων (KDD)

- **Τροποποίηση δεδομένων (Data transformation):** Γίνεται μετατροπή των ετερογενών δεδομένων σε κοινή τυποποίηση κατάλληλη προς επεξεργασία. Το σύνολο των δεδομένων λαμβάνει την απαραίτητη μορφή, ώστε η μορφή τους να είναι κατάλληλη για την διαδικασία της εξόρυξης.

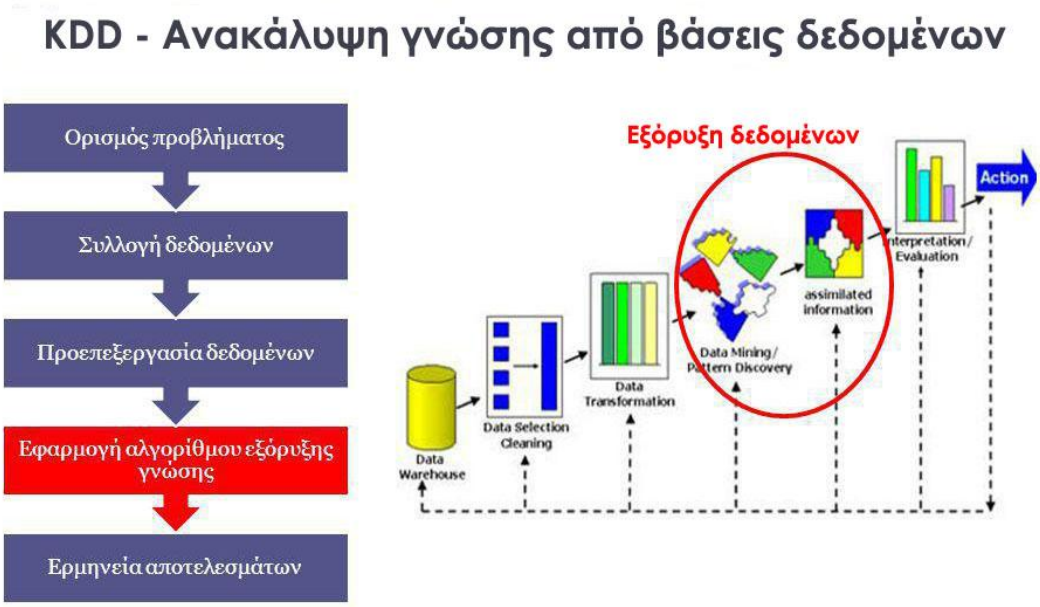
- **Εξόρυξη δεδομένων (Data mining):** Είναι το σημαντικότερο από τα βήματα της διαδικασίας και αυτό γιατί στο συγκεκριμένο στάδιο γίνεται η εφαρμογή των αλγορίθμων ώστε να καταλήξουμε στην παραγωγή του απαραίτητου μοντέλου-προτύπου για την ανακάλυψη της γνώσης.

- **Αξιολόγηση προτύπων (Pattern evaluation):** Στο βήμα αυτό αναγνωρίζονται χρήσιμα πρότυπα που αναπαριστούν γνώση, βάσει συγκεκριμένων μέτρων αξιολόγησης (evaluation measures).

- **Αναπαράσταση γνώσης (Knowledge representation):** Στο τελικό αυτό στάδιο, η γνώση που έχει ανακαλυφθεί

παρουσιάζεται στον χρήστη, γίνεται παρουσίαση των αποτελεσμάτων της εξόρυξης δεδομένων στους χρήστες. Τα αποτελέσματα τίθενται προς αξιολόγηση και στο στάδιο αυτό γίνεται ευρεία χρήση μεθόδων οπτικοποίησης και GUI.

Σε αρκετές βιβλιογραφίες τα δύο τελευταία βήματα, της Αξιολόγησης και της Αναπαράστασης, θεωρούνται ως ένα ενιαίο βήμα, αυτό της **Ερμηνείας (interpretation) – Αξιολόγησης (evaluation),** στο οποίο ο ίδιος ο χρήστης με τη βοήθεια γραφικών απεικονίσεων των προτύπων ή/και των δεδομένων προχωράει στην ερμηνεία και αξιολόγησή τους.



Εικόνα 1.3: Στάδιο της εξόρυξης δεδομένων στη διαδικασία της KDD - Ανακάλυψης γνώσης

Από τα παραπάνω γίνεται εύκολα αντιληπτό ότι η εξόρυξη δεδομένων, αν και αποτελεί ένα στάδιο στην όλη διαδικασία της ανακάλυψης γνώσης, εν τούτοις αποτελεί μια διαδικασία-κλειδί και μια από τις πλέον κρίσιμες. Η εξόρυξη δεδομένων ως βήμα της διαδικασίας ανακάλυψης γνώσης ενδιαφέρεται κυρίως για τις μεθοδολογίες και τις τεχνικές εξαγωγής προτύπων δεδομένων ή τις περιγραφές δεδομένων από τις μεγάλες αποθήκες δεδομένων.

Αξίζει να αναφέρουμε, ότι η διαδικασία της ανακάλυψης γνώσης είναι επαναληπτική και θα μπορούσε να περιέχει βρόχους μεταξύ οποιωνδήποτε από τα ανωτέρω βήματα. Αν και η κύρια εργασία στη διαδικασία ανακάλυψης γνώσης εστιάζεται στη διαδικασία

εξόρυξης δεδομένων, τα άλλα βήματα είναι εξίσου σημαντικά για την επιτυχή εφαρμογή της τεχνικής KDD.

1.3 Στόχοι της Εξόρυξης Δεδομένων

Υπάρχει μια μεγάλη συλλογή αλγορίθμων εξόρυξης δεδομένων, πολλοί από τους οποίους χρησιμοποιούν έννοιες και τεχνικές από διαφορετικούς τομείς όπως η μηχανική μάθηση, η τεχνητή νοημοσύνη, η αναγνώριση προτύπων, η στατιστική και οι βάσεις δεδομένων. 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.

Στην ουσία η εξόρυξη δεδομένων αποτελεί μια διαδικασία ημί-αυτόματης ανάλυσης μεγάλων βάσεων δεδομένων για την εύρεση προτύπων (patterns) που ήταν μη-προφανή για το σύστημα μέχρι εκείνη τη στιγμή, που θα ισχύουν για νέα δεδομένα με κάποια βεβαιότητα, τα οποία πιθανόν να εφαρμόζονται σε αντικείμενα, επιχειρήσεις και επιστήμες και τέλος να είναι απόλυτα κατανοητά και προς ερμηνεία από τον άνθρωπο.

Βασικός στόχος της εξόρυξης δεδομένων είναι η υποστήριξη στρατηγικών αποφάσεων. Οι μέθοδοι εξόρυξης γνώσης στοχεύουν στην ανακάλυψη στοιχείων που θα είναι χρήσιμα για τους οργανισμούς και τις επιχειρήσεις. Για παράδειγμα, οι Fayyad, Piatetsky-Shapiro & Smyth[2] , αναφέρουν ως εφαρμογές της KDD (ανακάλυψη γνώσης από βάσεις δεδομένων) στον χώρο των επιχειρήσεων σε δραστηριότητες όπως το Marketing, οι Επενδύσεις, ο Προσδιορισμό απειλών (fraud detection), οι Αγορές σε πολυκαταστήματα/αλυσίδες/eshops και οι Συναλλαγές με τράπεζες/πιστωτικές κάρτες.

Η εξόρυξη δεδομένων έχει λοιπόν σαν βασικούς της στόχους την εφαρμογή τεχνικών πρόβλεψης και συμπεριφοράς τάσεων (prediction), την αναγνώριση, την περιγραφή (description) σε μεγάλες βάσεις δεδομένων, καθώς επίσης την ταξινόμηση και την

[2] **U. Fayyad, G. Piatetsky-Shapiro, P. Smyth**, Knowledge Discovery and Data Mining: Towards a Unifying Framework, 1996

βελτιστοποίηση των πόρων της. Παρακάτω περιγράφονται και αναλύονται οι στόχοι της εξόρυξης δεδομένων.

- **Πρόβλεψη:** Στο προβλεπτικό πρότυπο υπάρχουν πρόσθετα ζητήματα σχετικά με τη χρονική σχέση των μεταβλητών. Έχοντας ως δεδομένα την αξία της μεταβλητής που προβλέπεται και τα ιστορικά στοιχεία μπορούμε να χτίσουμε ένα μοντέλο θα μπορέσει να εξηγήσει την προς επεξεργασία συμπεριφορά. Η **Προγνωστική Ανάλυση** στοχεύει στη διατύπωση προβλέψεων για το μέλλον, συνήθως με την οικοδόμηση κάποιου μοντέλου. Όταν αυτό το μοντέλο εφαρμόζεται στις τρέχουσες εισαγωγές, το αποτέλεσμα είναι μια πρόβλεψη της μελλοντικής συμπεριφοράς. Με άλλα λόγια, οι διαδικασίες πρόβλεψης της εξόρυξης δεδομένων (predictive data mining tasks), προσπαθούν να κάνουν εκτιμήσεις βγάζοντας συμπεράσματα από τα διαθέσιμα δεδομένα και μεταβλητές. Η προσπάθεια πρόβλεψης μελλοντικών συμπεριφορών έχει ως στόχο να ληφθούν αποφάσεις που να μεγιστοποιούν το κέρδος και να προλαμβάνουν δυσάρεστες καταστάσεις. Για παράδειγμα, ένα ηλεκτρονικό κατάστημα μπορεί να χρησιμοποιήσει τα συμπεράσματα του μοντέλου που προέκυψε για να προβλέψει ποιοι πελάτες θα εμφανίσουν τάσεις φυγής μέσα στους επόμενους 6 μήνες. Εφαρμόζοντας τεχνικές πρόβλεψης ο ιδιοκτήτης του συγκεκριμένου e-shop θα μπορέσει να εντοπίσει αυτούς τους πελάτες, να εφαρμόσει τεχνικές marketing και προσέλκυσης ενδιαφέροντος με στόχο να μειώσει τη ζημία και τον αριθμό μη ευχαριστημένων πελατών.

- **Αναγνώριση:** Κατ' αυτόν τον τρόπο, ενδεικτικά, τα δεδομένα αυτόματα ταξινομούνται σε κατηγορίες ή διαχωρίζονται σε ομάδες με βάση κάποια κριτήρια, ακόμα και υπό την παρουσία θορύβου ο οποίος δυσκολεύει την αναγνώριση, ωθώντας συνήθως τα δεδομένα να μοιάζουν περισσότερο τυχαία απ' όσο πραγματικά είναι. Οι τυποποιημένες αυτές μορφές των δεδομένων χρησιμοποιούνται για να δείξουν την ύπαρξη μιας δραστηριότητας ή ενός γεγονότος. Για παράδειγμα, η εξόρυξη βιοϊατρικών δεδομένων και η ανάλυση δεδομένων DNA έχει γνωρίσει τεράστια ανάπτυξη από τα μέσα της δεκαετίας του 1990. Όλες οι ακολουθίες DNA, αποτελούνται από τέσσερα βασικά δομικά στοιχεία, τα νουκλεοτίδια: Αδενίνη (A), Κυτοσίνη (C), Γουανίνη (G) και

Θυμίνη (Τ). Αυτά τα τέσσερα στοιχεία συνδυάζονται και σχηματίζουν μακριές ακολουθίες με τη μορφή συνεστραμμένης έλικας. Τα γονίδια αποτελούνται συνήθως από εκατοντάδες νουκλεοτιδίων που οργανώνονται σε συγκεκριμένη διάταξη. Συγκεκριμένα πρότυπα ακολουθιών γονιδίων σχετίζονται με ορισμένες ασθένειες και παίζουν σημαντικό ρόλο στην ιατρική. Προς την κατεύθυνση αυτή, η αναγνώριση προτύπων είναι μια περιοχή-κλειδί που προσφέρει πληθώρα εργαλείων για την εύρεση ομοιότητας και τη σύγκριση μεταξύ ακολουθιών DNA.

- **Περιγραφή:** Στόχος είναι να βρεθούν κατανοητά πρότυπα που περιγράφουν τα δεδομένα και τις ιδιότητες τους. Στην ουσία αναφερόμαστε στην έννοια της **Περιγραφικής Ανάλυσης**, η οποία και στοχεύει στην κατάδειξη ομαδοποιήσεων και ιδιοτήτων των δεδομένων, χωρίς να επιδιώκει τη διατύπωση προβλέψεων. Είναι η διαδικασία η οποία επικεντρώνεται στην ανακάλυψη προτύπων και αναπαριστά τα δεδομένα μιας πολύπλοκης βάσης δεδομένων με όσο το δυνατό πιο κατανοητό και αξιοποιήσιμο τρόπο.

- **Ταξινόμηση:** Σε αυτό το στάδιο έχουμε διαχωρισμό των στοιχείων, με αποτέλεσμα να προκύπτουν διαφορετικές κατηγορίες ή κλάσεις. Επιλύουμε στην ουσία μεταβλητές στόχευσης σαν συνάρτηση των υπολοίπων μεταβλητών εισόδων. Για παράδειγμα, μπορούμε να επιλύσουμε ένα πρόβλημα στόχευσης αγοραστικού κοινού για ένα ηλεκτρονικό κατάστημα, όπου οι δυνητικοί πελάτες του κατατάσσονται σε διαφορετικές κατηγορίες (μεταβλητές) για την προσέλκυση των οποίων θα πρέπει να εφαρμόσουμε συγκεκριμένες τεχνικές marketing. Ο στόχος είναι να χτιστεί ένα πρότυπο κάποιου είδους που μπορεί να εφαρμοστεί στα αταξινόμητα στοιχεία προκειμένου να τα ταξινομήσει.

- **Βελτιστοποίηση:** Γίνεται στην ουσία επιλογή της κατάλληλης αλγοριθμικής διαδικασίας για να βελτιστοποιήσουμε το αποτέλεσμα λειτουργίας. Η βέλτιστη χρήση κάποιων πόρων, όπως ο χρόνος, ο χώρος, το χρήμα και η μεγιστοποίηση κάποιων μεγεθών, όπως είναι τα κέρδη είτε οι πωλήσεις. Στις σύγχρονες επιχειρήσεις, η βελτιστοποίηση των διαδικασιών αποτελεί μόνιμη επιδίωξη,

ώστε να επιτευχθεί η μεγιστοποίηση της αποτελεσματικότητας και της αποδοτικότητας

Εικόνα 1.4: Βασικοί Στόχοι Εξόρυξης Δεδομένων (Data Mining)

1.4 Είδη Βάσεων Δεδομένων που χρησιμοποιούνται για την Εξόρυξη Δεδομένων

Η έννοια του Data Mining (Εξόρυξη Δεδομένων) είναι άρρηκτα συνδεδεμένη με την αξιοποίηση των δεδομένων που μας προσφέρουν οι Βάσεις Δεδομένων. Εξ' ορισμού η έννοια της εξόρυξης δεδομένων εστιάζει στην περιγραφή των δεδομένων μίας μεγάλης βάσης και στην πρόβλεψη και εξήγηση νέων δεδομένων που μπορεί να εισαχθούν.

Επομένως, η εξόρυξη γνώσης από μία βάση δεδομένων αναφέρεται σε ολόκληρη τη διαδικασία ανακάλυψης χρήσιμης πληροφορίας από μεγάλα σύνολα δεδομένων.

Ένα σύστημα εξόρυξης δεδομένων θα μπορούσε να ταξινομηθεί σύμφωνα με τα είδη βάσεων δεδομένων στις οποίες εφαρμόζεται η εξόρυξη δεδομένων. Παραδείγματος χάριν, ένα σύστημα που χρησιμοποιείται για την εξαγωγή γνώσης από σχεσιακά δεδομένα

καλείται σχεσιακό σύστημα γνώσης. Εάν εξάγει τη γνώση από αντικειμενοστραφείς βάσεις δεδομένων καλείται αντικειμενοστραφές σύστημα εξόρυξης δεδομένων. Γενικά, ένα σύστημα εξόρυξης δεδομένων θα μπορούσε να ταξινομηθεί βασισμένο στους διάφορους τύπους συστημάτων βάσεων δεδομένων, όπως τα σχεσιακά συστήματα βάσεων δεδομένων, τα αντικειμενοστραφή συστήματα βάσεων δεδομένων, οι χωροχρονικές βάσεις δεδομένων, τα συστήματα βάσεων δεδομένων πολυμέσων, κ.λπ.

1.4.1 Σχεσιακές Βάσεις Δεδομένων

Ένα Σύστημα διαχείρισης βάσεων δεδομένων (ΣΔΒΔ) (database management system-DBMS) είναι μια συλλογή προγραμμάτων που επιτρέπουν στους χρήστες να δημιουργούν και να συντηρούν μια βάση δεδομένων. Τα ΣΔΒΔ συχνά απεικονίζουν τα δεδομένα σε μια δομή τύπου πίνακα. Αυτή είναι η αφορμή για την εισαγωγή του σχεσιακού μοντέλου (relational model), όπου τα δεδομένα απεικονίζονται να αποτελούνται από σχέσεις. (Ξένος & Χριστοδουλάκης, 2000)[3]

Μια σχεσιακή βάση δεδομένων είναι προσβάσιμη μέσω της γλώσσας SQL (Structured Query Language), κώδικας της οποίας συχνά ενσωματώνεται σε κώδικα κάποιας άλλης γλώσσας όπως η C++ ή η Java. Με την εξαίρεση μερικών συστημάτων βάσεων δεδομένων βασιζόμενων σε αντικείμενα, που κυκλοφόρησαν στα τέλη της δεκαετίας 1980 και κατά τη διάρκεια της δεκαετίας του 1990, η τεχνολογία σχεσιακών βάσεων δεδομένων είναι η κυρίαρχη τεχνολογία βάσεων δεδομένων τα τελευταία 20 χρόνια.

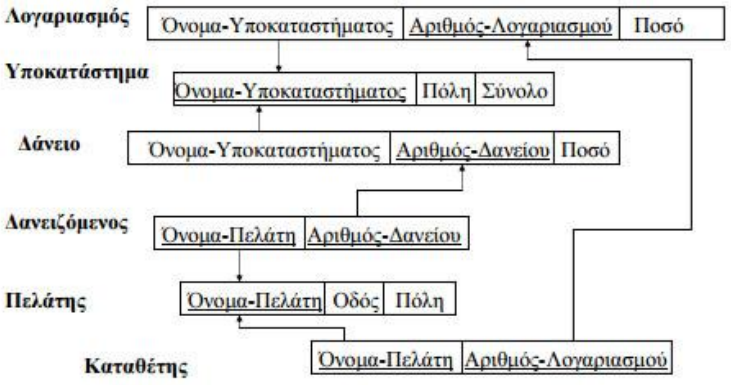
Τα δεδομένα που χρησιμοποιούμε για την εξόρυξη δεδομένων συνήθως βρίσκονται (ή τα φέρνουμε) στη μορφή ενός πίνακα–σχέσης. Κάθε πίνακας αποτελείται από ένα σύνολο πεδίων (συνήθως στήλες) και σε αυτόν βρίσκονται αποθηκευμένα ένας μεγάλος αριθμός δεδομένων-εγγραφών (συνήθως γραμμές). Κάθε εγγραφή σε έναν σχεσιακό πίνακα αναπαριστά ένα αντικείμενο και χαρακτηρίζεται από ένα μοναδικό “κλειδί”. Παράδειγμα πίνακα δεδομένων για εργαζομένους:

[3] **Μ. Ξένος, Δ. Χριστοδουλάκης**, Βάσεις Δεδομένων, 2000

Εικόνα 1.5: Παράδειγμα απεικόνισης βάσης δεδομένων[4]

Μια βάση δεδομένων είναι μια συλλογή από, σχετικά μεταξύ τους, δεδομένα. Μια σχεσιακή βάση δεδομένων (ΣΒΔ) είναι μια συλλογή δεδομένων οργανωμένα σε ένα σύνολο πινάκων-σχέσεων. Στους πίνακες και μεταξύ των πινάκων μπορούν να ισχύουν περιορισμοί που υποστηρίζει το σχεσιακό μοντέλο δεδομένων. Ένα Σύστημα διαχείρισης ΒΔ (ΣΔΒΔ) είναι το σύνολο των εργαλείων για τη δημιουργία και διαχείριση ΒΔ. Όλα τα γνωστά εμπορικά ΣΔΒΔ βασίζονται στο σχεσιακό μοντέλο δεδομένων και σταδιακά έχουν επεκταθεί σε αντικειμενο-σχεσιακά ΣΔΒΔ. Ακολουθεί παράδειγμα μιας σχεσιακής ΒΔ στο οποίο και παρουσιάζονται το σχήμα της ΒΔ, ενδεικτικά δεδομένα, καθώς και η σχέση μεταξύ τους:

[4] Β ΕΠΑΛ - ΣΔΒΔ και Εφαρμογές τους στο WEB, https://epal-b-sdbd-web.wikispaces.com/%CE%A3%CF%85%CF%83%CF%84%CE%AE%CE%BC%CE%B1%CF%84%CE%B1%20%CE%94%CE%B9%CE%B1%CF%87%CE%B5%CE%AF%CF%81%CE%B9%CF%83%CE%B7%CF%82%20%CE%92%CE%AC%CF%83%CE%B5%CF%89%CE%BD%20%CE%94%CE%B5%CE%B4%CE%BF%CE%BC%CE%AD%CE%BD%CF%89%CE%BD%20%CE%BA%CE%B1%CE%B9%20%CE%95%CF%86%CE%B1%CF%81%CE%BC%CE%BF%CE%B3%CE%AD%CF%82%20%CF%84%CE%BF%CF%85%CF%82%20%CF%83%CF%84%CE%BF%20WEB



Εικόνα 1.6: Παράδειγμα του σχεσιακού μοντέλου

1.4.2. Βάσεις Δεδομένων Συναλλαγών

Μια συναλλαγή βάσεων δεδομένων είναι μια μονάδα εργασιών που εκτελούνται από ένα σύστημα διαχείρισης βάσεων δεδομένων (DBMS). Μια συναλλαγή βάσεων δεδομένων, εξ ορισμού, πρέπει να χαρακτηρίζεται από Ατομικότητα, Συνέπεια, Απομόνωση και Μονιμότητα. Αυτές οι ιδιότητες των συναλλαγών βάσεων δεδομένων που αναφέρονται συχνά με το ακρωνύμιο ACID.

Μια συναλλαγή είναι μια ομάδα από SQL εντολές που εκτελούνται μαζί ως ενιαίο τμήμα των εργασιών που πρέπει να υλοποιηθούν. Αν όλες αυτές οι εντολές εκτελεστούν με επιτυχία, τότε η πράξη έχει γίνει commit, δηλαδή οι αλλαγές που γίνονται στα δεδομένα είναι οριστικές. Εάν οποιαδήποτε από τις εντολές μέσα σε συναλλαγή αποτύχει, τότε ολόκληρη η συναλλαγή ακυρώνεται: συγκεκριμένα στην SQL, η συναλλαγή θα πρέπει να αναιρεθεί όλη.

Εφαρμογές δυο επιπέδων, όπου το μεγαλύτερο μέρος της επιχειρηματικής λογικής είναι επικεντρωμένο σε κώδικα βάσεων δεδομένων, κάνουν έντονη χρήση των συναλλαγών. Ακόμη και σε πολυεπίπεδες εφαρμογές μπορούν να ωφεληθούν σε μεγάλο βαθμό με τη χρησιμοποίηση των συναλλαγών - κώδικας βάσεων δεδομένων μπορεί να γίνει μέρος μιας μεγαλύτερης συναλλαγής. Ακόμη και οι πολυεπίπεδες εφαρμογές πρέπει να χρησιμοποιούν τις συναλλαγές, δεδομένου ότι η βάση δεδομένων ερμηνεύει όλο

το σύνολο των δραστηριοτήτων της σε μορφή συναλλαγών. Ωστόσο, η εφαρμογή ερμηνεύει συναλλαγές μόνο με insert, update και delete δηλώσεις.

Ο τρόπος που γράφονται τις συναλλαγές επηρεάζει τις επιδόσεις του SQL Server σε μεγάλο βαθμό. Αυτό οφείλεται στον τρόπο υλοποίησης των συναλλαγών. Όταν γίνεται μια συναλλαγή, το κομμάτι των δεδομένων που πρόκειται να τροποποιηθεί είναι δεσμευμένο για αποκλειστική χρήση από το άτομο ή την επεξεργασία που ξεκίνησε τη συναλλαγή: στον SQL Server αυτή αναφέρεται ως κλείδωμα (locking). Δεδομένου ότι ο πίνακας και τα στοιχεία του ευρετηρίου είναι αποθηκευμένα σε "σελίδες", ο SQL Server μπορεί να κλειδώσει μια μοναδική γραμμή, σελίδα, ομάδα σελίδων ή ένα ολόκληρο πίνακα, ανάλογα με το ποσό των διαθέσιμων πόρων και των τροποποιήσεων των δεδομένων που πρέπει να γίνουν. Μία συναλλαγή είναι συνήθως το αποτέλεσμα της εκτέλεσης ενός προγράμματος που είναι γραμμένο σε μία γλώσσα προγραμματισμού υψηλού επιπέδου. Οι εντολές που προσδιορίζουν μια συναλλαγή περικλείονται μεταξύ των εκφράσεων «begin transaction» και «end transaction». Μια συναλλαγή μπορεί να έχει τις εξής καταστάσεις:

Ενεργή Κατάσταση(Active State)
Χωρίζεται σε δύο φάσεις.

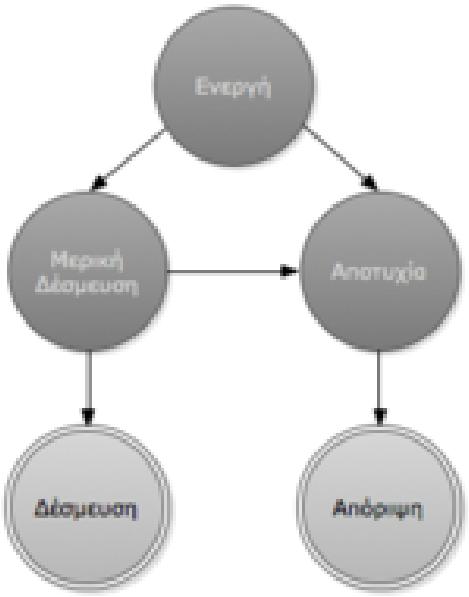
- Αρχική φάση(Initial Phase): μια συναλλαγή βάσης δεδομένων είναι στη φάση που οι σχέσεις της έχουν αρχίσει να εκτελούνται.
- Μερικώς Δεσμευμένη Φάση (Partially Committed Phase): μια συναλλαγή βάσης δεδομένων μπαίνει σε αυτή τη φάση όταν έχει εκτελεστεί η τελευταία δήλωση. Στην παρούσα φάση, έχει ολοκληρώσει την εκτέλεσή της, αλλά εξακολουθεί να είναι δυνατή η ματαίωση της συναλλαγής, επειδή η έξοδος από την εκτέλεση μπορεί να παραμείνει προσωρινά στην κύρια μνήμη - μια βλάβη υλικού μπορεί να διαγράψει την έξοδο.

Αποτυχής Κατάσταση (Failed State)
Μια συναλλαγή βάσης δεδομένων μπαίνει σε Αποτυχή Φάση, όταν η κανονική εκτέλεση του δεν μπορεί πλέον να προχωρήσει λόγω του βλάβης υλικού ή λαθών προγράμματος.

Ματαιωμένη Κατάσταση (Aborted State)

Μια συναλλαγή βάσης δεδομένων, εφόσον το DBMS την έχει καθορίσει αποτυχημένη, μπαίνει στη Ματαιωμένη Φάση. Μια ματαιωμένη συναλλαγή δεν πρέπει να έχει συνέπειες για τη βάση δεδομένων, επομένως τυχόν αλλαγές που έχουν γίνει στη βάση δεδομένων πρέπει να αναιρεθούν, είτε σε τεχνικούς όρους κάνει roll back. Η βάση δεδομένων θα επιστρέψει στη συνεπή μορφή της όταν η ανεπιτυχής συναλλαγή έχει αναιρεθεί. Το DBMS είναι υπεύθυνο για τη διαχείριση της ματαίωσης των συναλλαγών (abort).

Επιβεβαιωμένη Κατάσταση(Committed State)
Μια συναλλαγή βάσης δεδομένων μπαίνει στην επιβεβαιωμένη κατάσταση, όταν αρκετές πληροφορίες έχουν γραφτεί στο δίσκο, μετά την ολοκλήρωση της εκτέλεσής του με επιτυχία.



Εικόνα 1.7: Καταστάσεις συναλλαγής

Κατηγορίες Συναλλαγών

Υπάρχουν δύο είδη συναλλαγών στον SQL Server οι έμμεσες (implicit) και άμεσες (explicit). Οι έμμεσες συναλλαγές διαχειρίζονται από τον SQL Server - για κάθε σχέση insert, update και delete, καθώς για όλα τα στοιχεία σχεσιακής γλώσσας (DDL) εκτελούνται ως έμμεσες (implicit) συναλλαγές. Αυτό σημαίνει ότι ακόμη και αν δεν βλέπουμε τις άμεσες (explicit) εντολές των

συναλλαγών, ο SQL Server θα αναιρέσει το σύνολο της δήλωσης ή της ομάδας δηλώσεων αν αντιμετωπίσει ένα σφάλμα.

Οι άμεσες(explicit) συναλλαγές αποτελούνται από begin, commit και rollback συναλλαγές και συντάσσονται από τους προγραμματιστές.

Συνήθως, οι δηλώσεις που περικλείονται στο εσωτερικό μιας συναλλαγής θα πρέπει να περιλαμβάνουν το εντολών DML (π.χ. select, insert, update, merge - upsert, call - call a PL/SQL, explain plan, lock table) που πρέπει να εκτελεστούν ως μονάδα. Οι δηλώσεις select δεν τροποποιούν τα δεδομένα και γενικά δεν χρειάζεται να περιληφθούν σε μια συναλλαγή από τις ίδιες.

Ιδιότητες Συναλλαγών

Κάθε συναλλαγή σε SQL Server θα πρέπει να περάσει το τεστ ACID για να είναι έγκυρη. Οι ιδιότητες ACID μιας συναλλαγής είναι:

A - Ατομικότητα (Atomicity)
Αυτό σημαίνει ότι ο SQL Server απαιτεί συναλλαγές που είτε όλες οι πράξεις της συναλλαγής επιτυγχάνουν, είτε όλες αποτυγχάνουν.

C - Συνέπεια (Consistency)
Στο τέλος μιας συναλλαγής, η βάση του SQL Server πρέπει να είναι σε συνεπή μορφή.

I - Απομόνωση (Isolated)
Ακόμα κι αν τρέχουν πολλές συναλλαγές ταυτόχρονα, κάθε συναλλαγή πρέπει να νομίζει ότι τρέχει μόνη της.

D - Μονιμότητα (Durable)
Αν η συναλλαγή επιτύχει, πρέπει το αποτέλεσμά της να επιβιώνει, ακόμα και σε περίπτωση αποτυχίας του συστήματος.

Ο SQL Server πραγματοποιεί συναλλαγές σταθερά μέσω της διαδικασίας ανάκτησης. Κάθε φορά που ο διακομιστής ξεκινά κάθε σύστημα και κάθε χρήστη η βάση δεδομένων κάνει ανάκτηση. Κατά τη διάρκεια της διαδικασίας ανάκαμψης όλες οι δεσμευμένες συναλλαγές πραγματοποιούνται μόνιμα - γράφονται στο δίσκο, σε

περίπτωση που παραμείνουν στο αρχείο καταγραφής συναλλαγών (log) κατά την αποτυχία του συστήματος. Το σύνολο των μη επιβεβαιωμένων συναλλαγών αναιρούνται.

1.4.3 Βάσεις Κειμένου

Οι βάσεις κειμένων είναι βάσεις οι οποίες περιέχουν λέξεις ή ολόκληρα κείμενα, ή εναλλακτικά που περιέχουν λεκτικές περιγραφές αντικειμένων. Αυτές οι περιγραφές μπορούν να κυμαίνονται από απλές λέξεις κλειδιά, μέχρι ολόκληρες προτάσεις, όπως για παράδειγμα περιγραφές προϊόντων, απαντήσεις σε ερωτήματα – παράπονα χρηστών σε ένα call-center κ.α. Η πληροφορία που μπορεί να ανακαλύψει κάποιος από τέτοιες βάσεις είναι ανεξάντλητη. Παράδειγμα από μια βάση κειμένων μπορεί να δημιουργήσει έναν θησαυρό λέξεων, ή μια λίστα συνωνύμων. Από μια βάση παραπόνων – απαντήσεων χρηστών σε ένα call – center μπορεί πάλι να δημιουργήσει μια λίστα σχετικών αυτοματοποιημένων απαντήσεων σε αντίστοιχα ερωτήματα.

1.4.4. Χωροχρονικές Βάσεις Δεδομένων

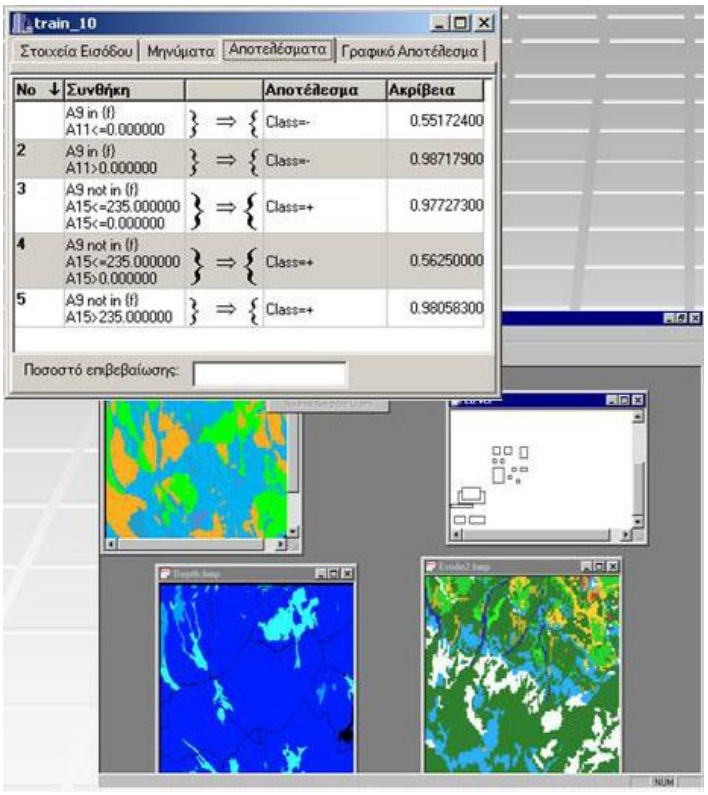
Τα χώρο-χρονικά αποτελούν μια ειδική κατηγορία δεδομένων που αποτελούνται από ένα σύνολο παρατηρούμενων μεταβλητών που σχετίζονται με ένα σύνολο σταθερών ή μεταβλητών χωρικών συντεταγμένων. Σε αντίθεση με τα αμιγώς χρονικά (π.χ. χρονοσειρές) ή χωρικά (π.χ. εικόνες) τα δεδομένα αυτά χρήζουν ιδιαίτερης μεταχείρισης καθώς οι αλγόριθμου εξόρυξης θα πρέπει να αξιοποιούν παράλληλα τόσο την χωρική όσο και την χρονική συνιστώσα.

Χαρακτηριστικά παραδείγματα χώρο-χρονικών δεδομένων ειναι:

- Δεδομένα δικτύων αισθητήρων, στα οποία ένα σύνολο αισθητήρων καταγράφει διάφορες μεταβλητές όπως θερμοκρασία, φωτεινότητα, κ.λ.π. σε συγκεκριμένες τοποθεσίες. Στην κατηγορία αυτή ανήκουν και δεδομένα από μετρήσεις σεισμικής δραστηριότητας καθώς και από μετεωρολογικούς σταθμούς.
- Δεδομένα συστημάτων πλοήγησης (π.χ. Global Positioning System - GPS), στα οποία καταγράφονται συνεχώς οι θέσεις διάφορων αντικειμένων.

- Δεδομένα Video, στα οποία δειγματοληπτούνται με μεγάλους ρυθμούς διαδοχικά στιγμιότυπα εικόνων. Στην επεξεργασία video η χωρική πληροφορία σχετίζεται με τη θέση των εικονοστοιχείων (pixels) σε κάθε στιγμιότυπο/εικόνα.
- Δεδομένα ηλεκτροεγκεφαλογραφήματος, στα οποία ένα σύνολο αισθητήρων τοποθετημένων σε συγκεκριμένες θέσεις στην επιφάνεια του κρανίου καταγράφει την ηλεκτρική εγκεφαλική δραστηριότητα.

Στα παραπάνω δεδομένα η εξόρυξη γνώσης καλείται να δώσει λύσεις σε προβλήματα όπως η ανακάλυψη γεγονότων ενδιαφέροντος (τα οποία μπορεί να σχετίζονται με τον εντοπισμό επικίνδυνων φαινομένων στα κλιματολογικά δεδομένα ή με την ανίχνευση επιληπτικών κρίσεων στα δεδομένα ηλεκτροεγκεφαλογραφήματος), η αναζήτηση αντικειμένων με παρόμοια χώρο-χρονική συμπεριφορά και η πρόβλεψη της χώρο-χρονικής εξέλιξης ενός γεγονότος ενδιαφέροντος. Μολονότι τα χώρο-χρονικά δεδομένα εμφανίζονται σε πληθώρα εφαρμογών που ανήκουν σε διαφορετικούς τομείς, η ανάλυσή τους μπορεί να τμηματοποιηθεί ιεραρχικά έχοντας στα χαμηλότερα επίπεδα κοινές μεθόδους και αλγόριθμους εξόρυξης δεδομένων, προσαρμοσμένες φυσικά στις εκάστοτε ιδιαιτερότητες και απαιτήσεις.



Εικόνα 1.8: Χωροχρονικά Δεδομένα

1.4.5 Πολυμεσικές Βάσεις

Οι βάσεις δεδομένων πολυμέσων παρέχουν χαρακτηριστικά που επιτρέπουν στους χρήστες να αποθηκεύουν και να διατυπώνουν ερωτήματα ή επερωτήσεις (queries) σε διαφορετικούς τύπους δεδομένων πολυμέσων (πολυμεσικά δεδομένα), που περιλαμβάνουν εικόνα (image), όπως φωτογραφίες ή σχέδια, κινούμενη εικόνα (video), ταινίες, ειδήσεις κλπ, ήχο (audio), όπως τραγούδια τηλεφωνικά μηνύματα, ή διαλέξεις, και κείμενο (text), όπως βιβλία και άρθρα, καθώς και σε παραδοσιακούς τύπους δεδομένων (όπως αριθμούς και σειρές χαρακτήρων). Οι βασικοί τύποι ερωτημάτων που απαιτούνται για τη βάση δεδομένων περιλαμβάνουν τον εντοπισμό των πηγών πολυμέσων που περιέχουν κάποια αντικείμενα που ενδιαφέρουν. Για παράδειγμα, μπορεί κάποιος να θέλει να εντοπίσει, από μια βάση δεδομένων video, όλες τις ακολουθίες video, όπου εμφανίζεται ένα συγκεκριμένο πρόσωπο. Μπορεί επίσης να θέλει να ανακτήσει ακολουθίες video που να περιέχουν κάποιες δραστηριότητες, όπως video στα οποία επιτυγχάνεται γκολ σε ένα ποδοσφαιρικό παιγνίδι από συγκεκριμένο παίκτη ή ομάδα.

Τα χαρακτηριστικά των πολυμεσικών τύπων δεδομένων οδηγούν σε ορισμένες απαιτήσεις που πρέπει να ικανοποιούνται από ένα Σύστημα Διαχείρισης Βάσεων Δεδομένων (ΣΔΒΔ) για την αποτελεσματική και αποδοτική υποστήριξή τους. Μερικά από αυτά τα χαρακτηριστικά είναι: η μεταβολή των δεδομένων σε σχέση με το χρόνο, ο μεγάλος όγκος των δεδομένων και οι εξειδικευμένες απαιτούμενες λειτουργίες. Τα παραδοσιακά σχεσιακά ΣΔΒΔ δεν είναι σε θέση να καλύψουν τις ανάγκες των πολυμεσικών εφαρμογών και για το λόγο αυτό έχουν αναπτυχθεί νέα μοντέλα Βάσεων Δεδομένων (ΒΔ), όπως οι Αντικειμενοστραφείς και οι Αντικειμενοσχεσιακές ΒΔ (Silberschatz, Korth, Sudarshan, 2011)[5].

Τα πολυμεσικά δεδομένα αποτελούνται από την περιγραφική πληροφορία (π.χ. τίτλος ταινίας) και την πληροφορία περιεχομένου (content). Ένας τρόπος διαχείρισης των πολυμεσικών τύπων δεδομένων χρησιμοποιώντας ένα παραδοσιακό ΣΔΒΔ είναι να αποθηκεύσουμε την περιγραφική πληροφορία στη βάση δεδομένων του ΣΔΒΔ και να χρησιμοποιήσουμε εξωτερικά αρχεία για την αποθήκευση του περιεχομένου. Το βασικό μειονέκτημα αυτής της προσέγγισης είναι ότι δεν μπορούμε να

[5] **A. Silberschatz, H. F. Korth, S. Sudarshan**, «Συστήματα Βάσεων Δεδομένων, η Πλήρης Θεωρία των Βάσεων Δεδομένων», 6η έκδοση, Εκδόσεις Μ. Γκιούρδας, Αθήνα 2011.

χρησιμοποιήσουμε τη λειτουργικότητα του ΣΔΒΔ για το περιεχόμενο των τύπων δεδομένων (π.χ. την κατασκευή δομών καταλόγων / ευρετηρίων - indexes). Μπορεί επίσης να καταλήξει σε ασυνέπειες, όπως ένα αρχείο που είναι σημειωμένο στην βάση δεδομένων, αλλά του οποίου τα περιεχόμενα λείπουν ή το αντίστροφο. Συνεπώς είναι επιθυμητό να αποθηκεύονται στην βάση δεδομένων τα ίδια τα δεδομένα (Μανωλόπουλος, Παπαδόπουλος, 2009)[6].

Οι περισσότερο γνωστοί τύποι δεδομένων πολυμέσων που είναι διαθέσιμοι στις Βάσεις Δεδομένων Πολυμέσων είναι οι παρακάτω.

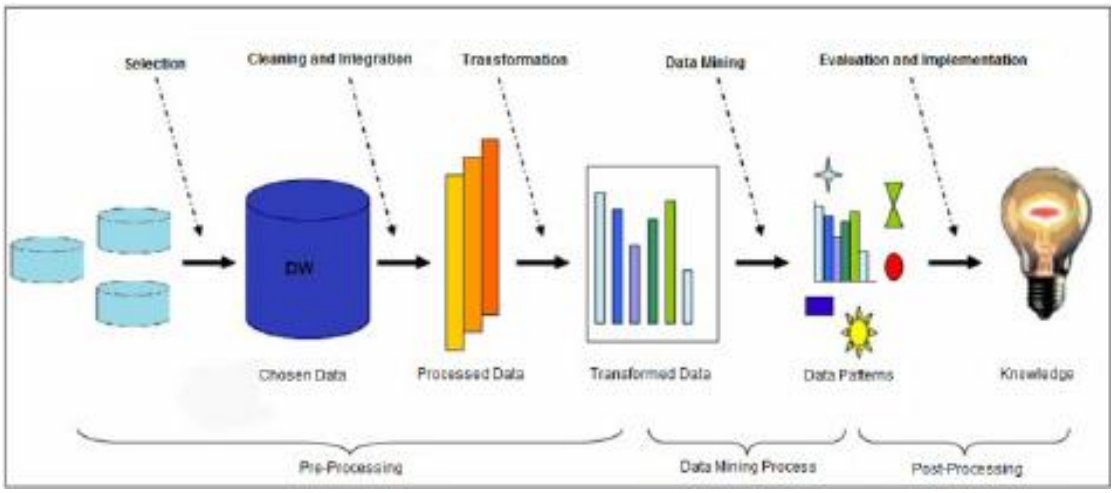
- Κείμενο: Μπορεί να είναι ή να μην είναι μορφοποιημένο. Για ευκολία ανάλυσης δομημένων εγγράφων, χρησιμοποιούνται πρότυπα όπως η SGML και παραλλαγές όπως η HTML.
- Γραφικά: Παραδείγματα περιλαμβάνουν σχέδια και εικονογραφήσεις που κωδικοποιούνται με χρήση κάποιου πρότυπου (πχ., CGM, PICT, postscript).
- Εικόνες: Περιλαμβάνουν σχέδια, φωτογραφίες, κοκ., κωδικοποιημένα σε τυπικές μορφοποιήσεις όπως bitmap, JPEG, και MPEG. Στα JPEG, και MPEG υπάρχει συμπίεση. Οι εικόνες αυτές δεν διαιρούνται σε επί μέρους στοιχεία. Επομένως τα ερωτήματα με βάση το περιεχόμενο (πχ., βρες όλες τις εικόνες που περιέχουν κύκλους) δεν είναι εύκολες.
- Κινούμενες Εικόνες: Χρονικές ακολουθίες από δεδομένα εικόνων ή γραφικών.
- Βίντεο: Ένα σύνολο από φωτογραφικά δεδομένα σε χρονική ακολουθία με καθορισμένο ρυθμό - για παράδειγμα 30 καρέ το δευτερόλεπτο.
- Δομημένος Ήχος: Μια ακολουθία από στοιχεία ήχου που περιλαμβάνουν νότες, τόνο, διάρκεια, κοκ.
- Ήχος: Δειγματοληπτικά δεδομένα από ηχητικές ηχογραφήσεις σαν συμβολοσειρές από bits σε ψηφιακή μορφή. Τυπικά οι αναλογικές ηχογραφήσεις μετατρέπονται σε ψηφιακή μορφή πριν την αποθήκευση.
- Σύνθετα Δεδομένα Πολυμέσων: Ένας συνδυασμός από τύπους δεδομένων πολυμέσων όπως ήχος και βίντεο που μπορεί να αναμειγνύονται φυσικά για να δώσουν ένα νέο τύπο μορφοποίησης αποθήκευσης ή λογική ανάμειξη ενώ διατηρούν τους αρχικούς τύπους και τις μορφοποιήσεις.

[6] **Μανωλόπουλος, Παπαδόπουλος**, «Συστήματα Βάσεων Δεδομένων – Θεωρία και Πρακτική Εφαρμογή», Αθήνα, 2006.

1.5 Διαδικασία Εξόρυξης Γνώσης

Η διαδικασία ανακάλυψης γνώσης από βάσεις δεδομένων (KDD) συνήθως ορίζεται από τα εξής στάδια:

1. **Επιλογή δεδομένων (Data selection)**
2. **Προεπεξεργασία Δεδομένων (Data preprocessing)**
3. **Τροποποίηση δεδομένων (Data transformation)**
4. **Εξόρυξη δεδομένων (Data mining)**
5. **Ερμηνείας (interpretation) – Αξιολόγησης (evaluation)**



Εικόνα 1.9: Στάδια εξόρυξης γνώσης από Βάση Δεδομένων

Υπάρχουν και παραλλαγές για τον ορισμό των σταδίων αυτών σύμφωνα και με το Cross Industry Standard Process for Data Mining (CRISP-DM) όπου τα στάδια έχουν ως εξής:

1. Κατανόηση Θέματος
2. Κατανόηση δεδομένων
3. Προετοιμασία δεδομένων
4. Μοντελοποίηση
5. Αξιολόγηση
6. Ανάπτυξη ή απλοποιημένη διαδικασία όπως
   - Προ-επεξεργασία
   - Εξόρυξη δεδομένων
   - Επικύρωση αποτελέσματος

Σε αυτό το σημείο αξίζει να κάνουμε μια αναφορά σε δύο βασικά στάδια που περιλαμβάνει η διαδικασία εξόρυξης γνώσης, αυτό της προ-επεξεργασίας και της μοντελοποίησης.

1.5.1 Προεπεξεργασία Δεδομένων (Data preprocessing)

Η προεπεξεργασία των δεδομένων αποτελεί ένα αναγκαίο στάδιο στη διαδικασία εξόρυξης γνώσης. Πριν την εφαρμογή των αλγορίθμων εξόρυξης δεδομένων, το ερευνώμενο σύνολο αυτών πρέπει να συναρμολογείται. Καθώς η εξόρυξη δεδομένων μπορεί να αποκαλύψει μόνο τα πρότυπα που πράγματι εμφανίζονται στα δεδομένα, το σύνολο δεδομένων που ερευνούμε, πρέπει να είναι αρκετά μεγάλο για να περιέχει αυτά τα πρότυπα παραμένοντας να εξορυχτεί σε ένα αποδεκτό χρονικό διάστημα. Η προ επεξεργασία είναι απαραίτητη για την ανάλυση πολλών παραγόντων-συνόλων πριν την εξόρυξη δεδομένων.

Ένα πολύ συνηθισμένο πρόβλημα είναι η ύπαρξη χαμένων τιμών, η έλλειψη δηλαδή τιμών σε ορισμένα πεδία καταχωρημένων εγγραφών ή να λείπουν σημαντικά χαρακτηριστικά ή να περιέχουν συναθροιστικά δεδομένα. Ένα άλλο πρόβλημα των δεδομένων είναι ο λεγόμενος «θόρυβος». Τα δεδομένα μπορεί να περιέχουν λανθασμένες τιμές. Ένας όρος που έχει επικρατήσει στη βιβλιογραφία της εξόρυξης δεδομένων και περιγράφει δεδομένα με χαμένες τιμές, θόρυβο και άλλα προβλήματα, είναι ο όρος «ακάθαρτα δεδομένα» (dirty data). Επομένως, για να έχουμε ποιοτικά αποτελέσματα από την εξόρυξη γνώσης χρειαζόμαστε ποιοτικά δεδομένα. Η διαδικασία αντιμετώπισης των χαμένων τιμών, του θορύβου, των ασυνεπειών και άλλων προβλημάτων των δεδομένων ονομάζεται **«καθαρισμός δεδομένων»** (data cleansing) και αποτελεί μέρος των εργασιών της προεπεξεργασίας τους.

Η προεπεξεργασία των δεδομένων αποτελεί ένα βασικό στάδιο της διαδικασίας ανακάλυψης γνώσης. Στα πλαίσια της εκτελούνται εργασίες καθαρισμού των δεδομένων, ανίχνευσης ανωμαλιών, μετασχηματισμού τους καθώς και σύνοψής τους. Παρακάτω ακολουθούν ορισμένες τεχνικές, οι οποίες εφαρμόζονται για τη διεξαγωγή αυτών των εργασιών.

- ✓ **Ανίχνευση ανωμαλιών (Anomaly detection):** Καλείται η αναγνώριση προτύπων από ένα σύνολο δεδομένων που εμφανίζουν διαφορετική συμπεριφορά από την

προσδοκώμενη. Στόχος είναι η υψηλού επιπέδου ανίχνευση πιθανών ανωμαλιών, διατηρώντας όμως χαμηλά ποσοστά λανθασμένης προειδοποίησης. Ως εφαρμογή μπορούμε να αναφέρουμε τον προσδιορισμό απειλής στην έγκριση δανείων ή πιστωτικών καρτών από μια τράπεζα.

- ✓ **Κανόνες συσχέτισης (Μοντέλο αλληλεξάρτησης)**: Αποτελούν μία από τις σημαντικότερες και νεότερες τεχνικές εξόρυξης γνώσης από μεγάλες βάσεις δεδομένων. Οι πληροφορίες που συγκεντρώνονται παράγουν ενδιαφέρουσες συσχετίσεις και πρότυπα, που βρίσκουν εφαρμογή από τους τομείς της ζωής και της ενασχόλησης του ανθρώπου μέχρι τα τηλεπικοινωνιακά δίκτυα, την αγορά και διαχείριση ρίσκου. Η ανάγκη κατανόησης και ανάλυσης του καλαθιού αγοράς(market basket analysis) των σύγχρονων καταναλωτών οδήγησε στην χρησιμοποίηση των κανόνων συσχέτισης. Για παράδειγμα, ένα ηλεκτρονικό κατάστημα, μπορεί να υπολογίσει ποια προϊόντα συνδυάζονται και αγοράζονται συνήθως μαζί και να χρησιμοποιήσει αυτή την πληροφορία για σκοπούς marketing και προώθησης σχετικών προϊόντων.

- ✓ **Συσταδοποίηση (Clustering)**: Αποτελεί τη διαδικασία εύρεσης συστάδων(ομάδων) αντικειμένων έτσι ώστε τα αντικείμενα σε κάθε ομάδα να είναι όμοια (ή να σχετίζονται) και διαφορετικά (ή μη σχετιζόμενα) από τα αντικείμενα των άλλων ομάδων.

- ✓ **Κατηγοριοποίηση (Classification)**: Είναι η διαδικασία κατά την οποία ένα στοιχείο ανατίθεται σε ένα προκαθορισμένο σύνολο κατηγοριών. Ο στόχος της διαδικασίας αυτής είναι η ανάπτυξη ενός μοντέλου, το οποίο αργότερα θα μπορεί να χρησιμοποιηθεί για την κατηγοριοποίηση μελλοντικών δεδομένων. Ενδιαφέρουσες εφαρμογές της τεχνικής αυτής αποτελούν η πρόβλεψη καρκινικών κυττάρων χαρακτηρίζοντας τα ως καλοήθη ή κακοήθη, η κατηγοριοποίηση πελατών μιας τράπεζας ανάλογα με την πιστωτική τους ικανότητα κ.ά.

- ✓ **Παλινδρόμηση (στατιστική) (Regression):** Γίνεται έρευνα της συσχέτισης μεταξύ μίας εξαρτώμενης μεταβλητής και μιας ή περισσότερων ανεξάρτητων μεταβλητών. Η παλινδρόμηση προϋποθέτει ότι τα σχετικά δεδομένα ταιριάζουν με μερικά γνωστά είδη συνάρτησης και μετά καθορίζει την καλύτερη συνάρτηση αυτού του είδους που μοντελοποιεί τα δεδομένα που έχουν δοθεί. Αποτέλεσμα της παλινδρόμησης όταν χρησιμοποιείται ως τεχνική εξόρυξης δεδομένων, αποτελεί ένα μοντέλο που χρησιμοποιείται αργότερα για να προβλέψει τις τιμές της κατηγορίας για τα νέα δεδομένα. Στην τεχνική αυτή επιδιώκεται να εξακριβωθεί η αιτιώδης επίδραση μιας μεταβλητής επάνω σε άλλη. Για παράδειγμα, η επίδραση της αύξησης των τιμών των προϊόντων με την προσφορά/ζήτηση.

- ✓ **Σκιαγράφηση:** Μερικές φορές ο σκοπός της εξόρυξης δεδομένων είναι απλά να περιγραφεί τι συμβαίνει σε μια περίπλοκη βάση δεδομένων με έναν τρόπο που να αυξάνει την κατανόησή μας όσον αφορά τους ανθρώπους, τα προϊόντα, ή τις διαδικασίες που παρήγαγαν τα στοιχεία αρχικά. Μια καλή περιγραφή μιας συμπεριφοράς συχνά θα προτείνει και μια εξήγηση για αυτήν. Στο ελάχιστο, μια καλή περιγραφή μπορεί να προτείνει από πού να αρχίσει η έρευνα για την εξήγηση.

1.5.2 Μοντελοποίηση δεδομένων (Data Modeling)

Βασική διαδικασία και στόχος του τομέα της εξόρυξης δεδομένων είναι η μοντελοποίηση των σχέσεων ανάμεσα σε ένα εξαρτημένο γνώρισμα-στόχο και σε άλλα ανεξάρτητα γνωρίσματα. Τα δεδομένα από μόνα τους, δεν επιτρέπουν στον άνθρωπο την εξαγωγή συμπερασμάτων. Οπότε περνάνε από την διαδικασία της ανάλυσης και αναγνώρισης προτύπων ώστε να διαπιστωθεί/εξαχθεί η απαραίτητη πληροφορία. Η ανάλυση συνίσταται στην τυποποίηση των σχέσεων ανάμεσα στην εξαρτημένη και στις ανεξάρτητες μεταβλητές, συνήθως με τη δημιουργία ενός μοντέλου, που επιτρέπει τον υπολογισμό της εξαρτημένης μεταβλητής από τις ανεξάρτητες. Με αυτό τον τρόπο παρέχει τη δυνατότητα της αυτόματης δημιουργίας προγνωστικών μοντέλων ακριβείας σχετικά με το μέλλον. Υπάρχει επίσης η

επιλογή να επιλεχθεί η καλύτερη λύση από την αξιολόγηση πολλών διαφορετικών μοντέλων.

Στη συγκεκριμένη διαδικασία γίνεται χρησιμοποίηση εργαλείων λογισμικού εξόρυξης, όπως η οπτικοποίηση (εκτυπώνοντας δεδομένα και τη θέσπιση σχέσεων) και ανάλυση διασποράς (για τον εντοπισμό μεταβλητών που πηγαίνουν καλά μαζί) είναι χρήσιμη για την αρχική ανάλυση. Εργαλεία όπως η γενικευμένη επαγωγή κανόνα μπορεί να αναπτύξει αρχικούς κανόνες συσχέτισης. Μόλις επιτευχθεί μεγαλύτερη κατανόηση των δεδομένων (συχνά μέσω μοτίβου αναγνώρισης, που προκλήθηκε από την προβολή της παραγωγής μοντέλο), μπορούν να εφαρμοστούν πιο λεπτομερή μοντέλα κατάλληλα για το είδος των δεδομένων. Η διαίρεση των δεδομένων σε σύνολα εκπαίδευσης (training) και test sets είναι επίσης απαραίτητη για τη μοντελοποίηση.

Κεφάλαιο 2

2.1 Μέθοδοι εξόρυξης Γνώσης και Δεδομένων

Οι βασικότερες από τις μεθόδους της εξόρυξης δεδομένων (Fayyad, Berry & Linoff, 2004)[7], μέσω των οποίων επιτυγχάνονται οι στόχοι που αναφέραμε προηγουμένως, είναι οι εξής:

- Κατηγοριοποίηση
  Η Κατηγοριοποίηση αποτελεί μία από τις πιο βασικές τεχνικές εξόρυξης δεδομένων. Η μέθοδος της κατηγοριοποίησης βασίζεται στην εξέταση των χαρακτηριστικών ενός αντικειμένου και στην αντιστοίχηση του σε ένα προκαθορισμένο σύνολο κλάσεων με βάση αυτά τα χαρακτηριστικά.

- Συσταδοποίηση
  Η συσταδοποίηση είναι η μέθοδος διαχωρισμού ενός συνόλου δεδομένων σε ένα σύνολο συστάδων (clusters) που περιέχουν όμοια στοιχεία. Η συσταδοποίηση διαφοροποιείται από την κατηγοριοποίηση διότι η συσταδοποίηση δε διαθέτει

[7] **Michael J.A. Berry, Gordon S. Linoff,** Data Mining Techniques For Marketing, Sales, and Customer Relationship Management Second Edition, 2004

προκαθορισμένες κατηγορίες. Τα δεδομένα οργανώνονται σε συστάδες με βάση την ομοιότητα που έχουν μεταξύ τους.

- Ανάλυση Συσχέτισης
  Η ανάλυση συσχέτισης αποτελεί επίσης μία από τις πιο σημαντικές τεχνικές εξόρυξης δεδομένων. Αυτό που καθιστά ιδιαίτερα ενδιαφέρουσα τη διαδικασία ανάλυσης συσχέτισης είναι ο συνοπτικός τρόπος με τον οποίο παρουσιάζουν χρήσιμες πληροφορίες που γίνονται εύκολα κατανοητές από τους χρήστες. Η διαδικασία της συσχέτισης ανακαλύπτει «κρυμμένες» συσχετίσεις των γνωρισμάτων ενός συνόλου δεδομένων. Οι συσχετισμοί αυτοί παρουσιάζονται στην εξής μορφή: $A \rightarrow B$, όπου τα A και B αναφέρονται στα σύνολα γνωρισμάτων που υπάρχουν στα υπό ανάλυση δεδομένα.

- Παλινδρόμηση
  Η παλινδρόμηση αναφέρεται στην εκμάθηση μιας λειτουργίας που εκχωρεί δεδομένα σε μια μεταβλητή πρόβλεψης που παίρνει πραγματικές τιμές. Η παλινδρόμηση μπορεί να χρησιμοποιηθεί παραδείγματος χάριν, για τον υπολογισμό της πιθανότητας με την οποία ένας ασθενής θα αναρρώσει με βάση τα αποτελέσματα της διάγνωσης.

2.2 Κατηγοριοποίηση (Classification)

Η διαδικασία της κατηγοριοποίησης, ή αλλιώς ταξινόμησης (classification) περιλαμβάνει την οργάνωση ενός συνόλου αντικειμένων (objects) που περιγράφονται από ένα σύνολο χαρακτηριστικών (attributes), σε μια σειρά από προκαθορισμένες κλάσεις (classes), χρησιμοποιώντας μεθόδους μάθησης με επίβλεψη (supervised learning methods). Στην ουσία δηλαδή η κατηγοριοποίηση (classification) είναι η διαδικασία η οποία απεικονίζει ένα σύνολο δεδομένων σε προκαθορισμένες ομάδες. Τις ομάδες αυτές συχνά τις καλούμε κατηγορίες ή κλάσεις.

Η κατηγοριοποίηση αποτελεί μια από τις βασικές εργασίες στο στάδιο της Εξόρυξης Δεδομένων. Βασίζεται στην εξέταση των χαρακτηριστικών ενός αντικειμένου, το οποίο με βάση τα χαρακτηριστικά αυτά αντιστοιχίζεται σε ένα προκαθορισμένο σύνολο κλάσεων. Η βασική ιδέα είναι η εξής: έχοντας ένα σύνολο

από κατηγορίες (κλάσεις) και ένα σύνολο δεδομένων με δείγματα, για τα οποία ξέρουμε σε ποια κλάση ανήκουν, στόχος της κατηγοριοποίησης είναι η δημιουργία ενός μοντέλου, το οποίο θα μπορεί να κατηγοριοποιήσει αυτόματα σε αυτές τις κατηγορίες νέα, άγνωστα, μη-κατηγοριοποιημένα δείγματα. Συνήθως η συγκεκριμένη τεχνική επιλύει προβλήματα, όπως ανάλυσης αποχωρήσεων (churn analysis), διαχείρισης κινδύνων (risk management) και στόχευσης.

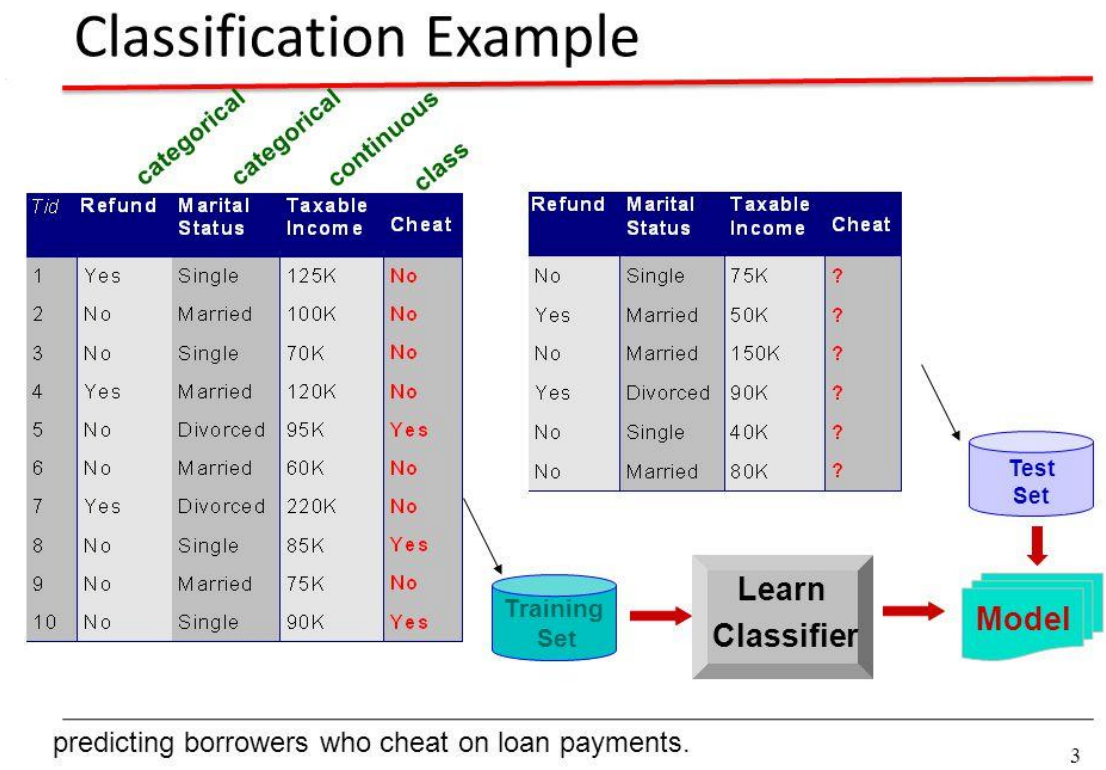
Στην πράξη μια διαδικασία κατηγοριοποίησης μπορεί να οριστεί ως η εκτέλεση δύο συγκεκριμένων βημάτων:

1. Δημιουργία μοντέλου βασιζόμενου σε δεδομένα εκπαίδευσης
2. Εφαρμογή του μοντέλου στο σύνολο των δεδομένων

Ένα παράδειγμα συνόλου δεδομένων, κατάλληλο για κατηγοριοποίηση, είναι τα στοιχεία αιτήσεων χορήγησης τραπεζικών δανείων, παράδειγμα που φαίνεται στην εικόνα 2.1. Στα γνωρίσματα καταγράφονται τα στοιχεία των πελατών, όπως οικογενειακή κατάσταση, ύψος εισοδήματος κλπ. και σε ένα γνώρισμα αναφέρεται το εάν εγκρίνεται ή απορρίπτεται το δάνειο. Η έγκριση ή η απόρριψη εξαρτάται από τα στοιχεία του κάθε πελάτη. Με την κατηγοριοποίηση δημιουργείται ένας μηχανισμός υπολογισμού της κατηγορίας του κάθε αντικειμένου από τα υπόλοιπα γνωρίσματα του. Στο παράδειγμα με τα τραπεζικά δάνεια, ένα σύνολο κανόνων, οι οποίοι ορίζουν για ποια εισοδήματα, για ποια οικογενειακή κατάσταση και άλλα στοιχεία εγκρίνεται το δάνειο, ενώ για ποιες όχι, είναι ένα μοντέλο κατηγοριοποίησης. Το μοντέλο δεν είναι υποχρεωτικά κανόνες, αλλά μπορεί να έχει άλλες μορφές, όπως πχ να συνίσταται σε ένα πλέγμα κόμβων και συνδέσεων ενός νευρωνικού δικτύου. Αφού δημιουργηθεί το μοντέλο, μπορεί να χρησιμοποιηθεί για τη διατύπωση προβλέψεων. Για μια νέα αίτηση δανείου μπορούν να εισαχθούν τα στοιχεία του πελάτη στο μοντέλο και αυτό να προβλέψει εάν θα εγκριθεί ή θα απορριφθεί η αίτηση. Σημειωτέον ότι αυτό που θα προβλεφθεί είναι η κατηγορία του κάθε αντικειμένου, δηλαδή μια ονομαστική τιμή.

Υπάρχουν πολλές δυνατότητες εφαρμογής τεχνικών κατηγοριοποίησης στον κόσμο των επιχειρήσεων. Η πρόβλεψη

χρεοκοπίας επιχειρήσεων και άλλων οργανισμών είναι ίσως το πιο γνωστό παράδειγμα. Επίσης, στον χρηματοπιστωτικό τομέα, η εκτίμηση της πιστοληπτικής ικανότητας και η διαχείριση του ρίσκου είναι δύο τυπικά πεδία εφαρμογής. Σημαντικές εφαρμογές βρίσκει η κατηγοριοποίηση και στον χώρο της διαφήμισης και των πωλήσεων. Η ένταξη πελατών σε προκαθορισμένες κατηγορίες χρησιμοποιείται για την προσέλκυση πελατών με άμεσο διαπροσωπικό μάρκετινγκ, καθώς και για την κατάρτιση προγραμμάτων επιβράβευσης πελατών και εκχώρησης πελατών σε συγκεκριμένα προγράμματα



Εικόνα 2.1: Παράδειγμα Κατηγοριοποίησης[8]

Υπάρχουν τρεις βασικές μέθοδοι που χρησιμοποιούνται για να λύσουν το πρόβλημα της κατηγοριοποίησης:

[8] **Xiangliang Zhang**, Classification I: Decision Tree

- **Καθορισμός των ορίων:** Η κατηγοριοποίηση εκτελείται με διαίρεση του χώρου της εισόδου των εν δυνάμει πλειάδων της Βάσης Δεδομένων σε περιοχές όπου κάθε περιοχή συνδέεται με μια κατηγορία
- **Χρήση κατανομών πιθανότητας:** Για κάθε κατηγορία που δίνεται Cj P(ti | Cj) είναι η συνάρτηση κατανομής πιθανότητας (probability distribution function) για την κατηγορία υπολογισμένη σε ένα σημείο, ti. Αν η πιθανότητα εμφάνισης κάθε κατηγορίας P(Cj), είναι γνωστή (ίσως να έχει οριστεί από κάποιον ειδικό του πεδίου εφαρμογής – domain expert), τότε P(Cj) P(ti | Cj) είναι η εκτίμηση της πιθανότητας ότι η ti ανήκει στην κατηγορία Cj
- **Χρήση εκ των υστέρων πιθανοτήτων:** Με δεδομένη μια τιμή δεδομένων ti, θέλουμε να καθορίσουμε την πιθανότητα για την οποία η ti ανήκει στην κατηγορία Cj. Αυτό υποδηλώνεται με το P(Cj | ti) που ονομάζεται εκ των υστέρων πιθανότητα (posterior probability).

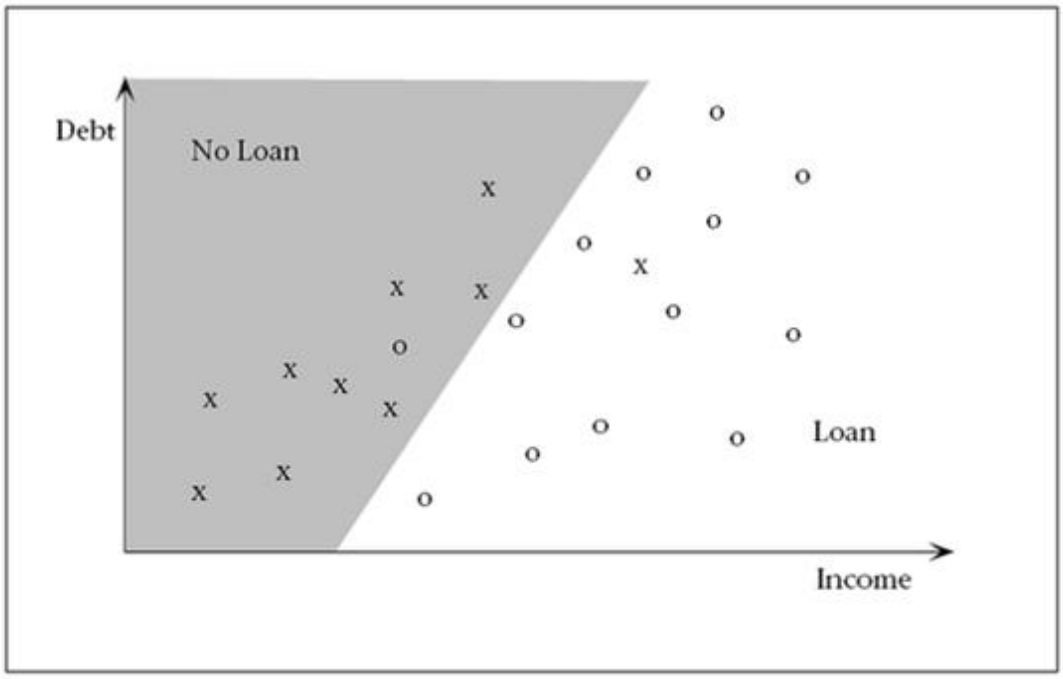
Η επίδοση των αλγορίθμων εξετάζεται με την εκτίμηση της ακρίβειας (accuracy) της κατηγοριοποίησης, δηλαδή την ικανότητα του μοντέλου να προβλέπει την κατηγορία μιας νέας περίπτωσης. Η εκτίμηση της ακρίβειας είναι ένα πολύ σημαντικό ζήτημα στο χώρο της κατηγοριοποίησης αφού κάτι τέτοιο μας δείχνει το πόσο καλά ανταποκρίνεται ο αλγόριθμος μας για δεδομένα, καθώς και μας παρέχει ένα μέτρο σύγκρισης των διαφόρων αλγορίθμων κατηγοριοποίησης.

$$\text{Ακρίβεια} = \frac{\text{Σωστές προβλέψεις}}{\text{Σύνολο προβλέψεων}}$$

$$\text{Αποτίμηση σφάλματος} = \frac{\text{Λάθος προβλέψεις}}{\text{Σύνολο προβλέψεων}}$$

Έτσι τελικά επιλέγεται και ο αλγόριθμος με τη μεγαλύτερη ακρίβεια και τη μικρότερη αποτίμηση σφάλματος, ο οποίος και θα έχει οριστεί ως αυτός με τις καλύτερες προβλέψεις.

Στην παρακάτω εικόνα έχουμε έναν απλό διαχωρισμό των στοιχείων δανείου σε δύο περιοχές κατηγοριών. Η τράπεζα πιθανώς να θελήσει να χρησιμοποιήσει τις περιοχές ταξινόμησης για να αποφασίσει, εάν θα δοθεί δάνειο ή όχι στους μελλοντικούς υποψηφίους.

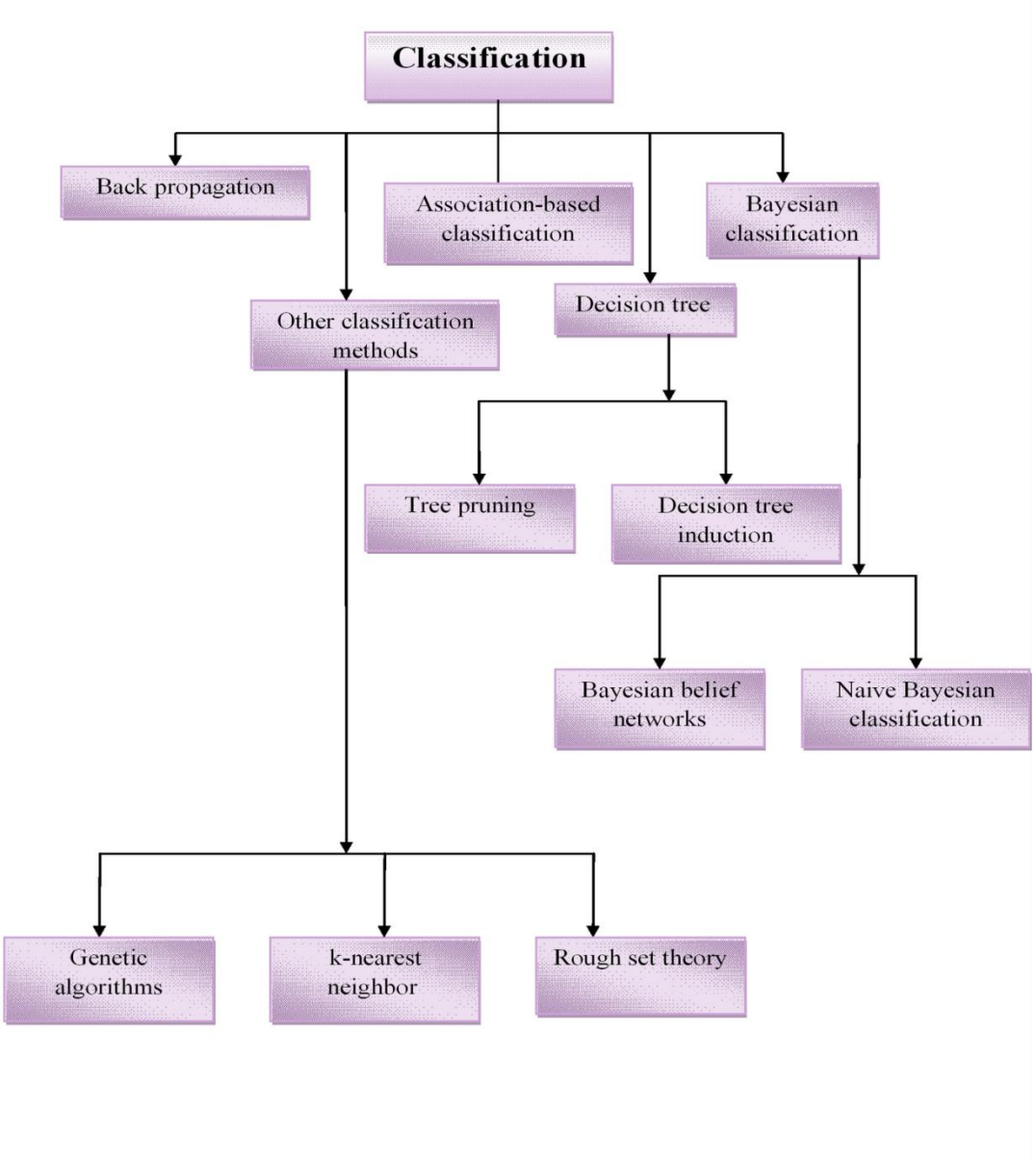


Εικόνα 2.2: Παράδειγμα Κατηγοριοποίησης Υποψηφίων Δανειοληπτών

Ένα απλό γραμμικό όριο κατηγοριοποίησης για το σύνολο των στοιχείων δανείου. Η διαμορφωμένη περιοχή δείχνει την κατηγορία (απόρριψη-έγκριση) και όχι το δάνειο.

Από τα παραπάνω γίνεται αντιληπτό ότι η τεχνικής της κατηγοριοποίησης χρησιμοποιείται για την πρόβλεψη της κλάσης που ανήκει ένα αντικείμενο. Τυπικοί αλγόριθμοι ταξινόμησης είναι τα δέντρα αποφάσεων (decison trees), τα νευρωνικά δίκτυα (neural networks) και Naive Bayes μοντέλα πιθανοτήτων. Γενικά μπορούμε να πούμε ότι οι αλγόριθμοι κατηγοριοποίησης μπορούν

να διαχωριστούν στις ακόλουθες κατηγορίες, κάποιες από τις οποίες θα αναλύσουμε στη συνέχεια:



Εικόνα 2.3: Διαχωρισμός Αλγορίθμων Κατηγοριοποίησης

2.2.1 Bayesian Κατηγοριοποίηση

Η Bayesian κατηγοριοποίηση αποτελεί μία κατηγορία μεθόδων της κατηγοριοποίησης και βασίζεται στη στατιστική θεωρία κατηγοριοποίησης του Bayes. Αυτό σημαίνει ότι πραγματοποιείται μια πιθανοτική πρόβλεψη, δηλαδή προβλέπει την πιθανότητα ένα δείγμα Χ να ανήκει σε κάποια κατηγορία. Η απόδοση αυτού του

είδους κατηγοριοποίησης είναι αρκετά υψηλή και χαρακτηρίζεται από την μεγάλη ταχύτητα της διαδικασίας κατηγοριοποίησης σε μεγάλες Βάσεις Δεδομένων.

Θεωρώντας ότι η συνεισφορά όλων των χαρακτηριστικών του συνόλου εκπαίδευσης είναι ανεξάρτητη και ότι κάθε ένα συνεισφέρει εξίσου στο πρόβλημα της κατηγοριοποίησης, έχει προταθεί μια απλή μέθοδος κατηγοριοποίησης η οποία είναι γνωστή ως απλοϊκή κατηγοριοποίηση κατά Bayes και βασίζεται στον κανόνα του Bayes για την υπό συνθήκη πιθανότητα. Ο κανόνας Bayes, είναι μια τεχνική που εκτιμά την πιθανοφάνεια μιας ιδιότητας παίρνοντας το σύνολο των δεδομένων σαν απόδειξη ή σαν είσοδο.

Ο κανόνας Bayes μας επιτρέπει να προσδιορίζουμε τις πιθανότητες των υποθέσεων, με δεδομένη την τιμή κάποιου δεδομένου, P(h|D).

Δοσμένου ενός σύνολου εκπαίδευσης D, η εκ των υστέρων πιθανότητα (posteriori probability) της υπόθεσης h, P(h|D) ακολουθεί το θεώρημα του Bayes:

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

Ο απλούστερος Bayesian κατηγοριοποιητής είναι ο Naïve Bayesian. Αυτός υποθέτει ότι η επίδραση ενός γνωρίσματος σε μία κατηγορία είναι ανεξάρτητη από τις τιμές των υπόλοιπων γνωρισμάτων. Ο λόγος που γίνεται αυτό είναι για να αποφεύγονται οι πολύπλοκοι υπολογισμοί κατά τη συνθήκη ανεξαρτησίας της κατηγορίας.

Η προσέγγιση της απλής κατηγοριοποίησης κατά Bayes έχει αρκετά πλεονεκτήματα. Πρώτον, είναι πολύ εύκολο να χρησιμοποιηθεί. Δεύτερον, αντίθετα με άλλες τεχνικές κατηγοριοποίησης χρειάζεται μόνο ένα πέρασμα των δεδομένων εκπαίδευσης. Επίσης, η προσέγγιση αυτή μπορεί εύκολα να χειριστεί ελλιπή δεδομένα, απλά παραλείποντας τις αντίστοιχες πιθανότητες. Σε περιπτώσεις όπου υπάρχουν απλές συσχετίσεις

στα δεδομένα, η τεχνική συνήθως δίνει καλά αποτελέσματα κατηγοριοποίησης σε σύντομο χρονικό διάστημα.

Από την άλλη πλευρά, υπάρχουν πολλές περιπτώσεις όπου ο αλγόριθμος κατηγοριοποίησης κατά Bayes δεν δίνει καλά αποτελέσματα. Πρώτον, σπάνιες είναι οι περιπτώσεις όπου τα χαρακτηριστικά δεν είναι ανεξάρτητα. Μια προσέγγιση είναι να αγνοήσουμε τα χαρακτηριστικά τα οποία εξαρτώνται από άλλα. Επιπρόσθετα, η τεχνική αυτή δεν μπορεί να χειριστεί συνεχή δεδομένα. Το μειονέκτημα αυτό λύνεται με το να χωρίσουμε τα συνεχή χαρακτηριστικά σε διαστήματα, ωστόσο αυτό δεν είναι κάτι απλό και ο τρόπος με το οποίον θα γίνει είναι πολύ πιθανό να επηρεάσει τα αποτελέσματα.

Παράδειγμα 2.1: Καλός καιρός για Tennis. Υπολογισμός P(xi|C)[9]

Τα δεδομένα εκπαίδευσης D που θα χρησιμοποιήσουμε για αυτό το παράδειγμα παρουσιάζονται στον πίνακα 2.1. Ας δούμε σε ποια κατηγορία (ναι / όχι) θα κατηγοριοποιηθεί το άγνωστο δείγμα X = <rain, hot, high, false>

Weather	Temperature	Humidity	Windy	Play Tennis
Sunny	Hot	High	Fasle	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

[9] Machine Learning – Naive Bayes Classifier, https://computersciencesource.wordpress.com/2010/01/28/year-2-machine-learning-naive-bayes-classifier/

Πίνακας 2.1: Δεδομένα εκπαίδευσης παραδείγματος "Play-Tennis"

Weather	
P(sunny|p) = 2/9	**P(sunny|n) = 3/5**
P(overcast|p) = 4/9	**P(overcast|n) = 0**
P(rain|p) = 3/9	**P(rain|n) = 2/5**
Temperature	
P(hot|p) = 2/9	**P(hot|n) = 2/5**
P(mild|p) = 4/9	**P(mild|n) = 2/5**
P(cool|p) = 3/9	**P(cool|n) = 1/5**
Humidity	
P(high|p) = 3/9	**P(high|n) = 4/5**
P(normal|p) = 6/9	**P(normal|n) = 1/5**
Windy	
P(true|p) = 3/9	**P(true|n) = 3/5**
P(false|p) = 6/9	**P(false|n) = 2/5**

Για το άγνωστο δείγμα μας X = <rain, hot, high, false>

$P(X|p)\cdot P(p) = P(rain|p)\cdot P(hot|p)\cdot P(high|p)\cdot P(false|p)\cdot P(p) = 3/9\cdot 2/9\cdot 3/9\cdot 6/9\cdot 9/14 =$ **0.010582**

$P(X|n)\cdot P(n) = P(rain|n)\cdot P(hot|n)\cdot P(high|n)\cdot P(false|n)\cdot P(n) = 2/5\cdot 2/5\cdot 4/5\cdot 2/5\cdot 5/14 =$ **0.018286**

Άρα κατηγοριοποιούμε την νέα πλειάδα X στην κατηγορία "NO - Play Tennis"

2.2.2 Naive Bayesian

Η Naive Bayesian κατηγοριοποίηση βασίζεται στην ανεξαρτησία των χαρακτηριστικών

$$P(x1,\ldots,xk|C) = P(x1|C)\cdot\ldots\cdot P(xk|C)$$

Αν το i-οστό χαρακτηριστικό είναι κατηγορικό, τότε η πιθανότητα P(xi|C) υπολογίζεται ως η σχετική συχνότητα των δειγμάτων που έχουν την τιμή xi ως το i-οστό χαρακτηριστικό στην κλάση C

Αν το i-οστό χαρακτηριστικό είναι συνεχές, τότε η πιθανότητα P(xi|C) υπολογίζεται μέσω μιας Γκαουσιανής συνάρτησης πυκνότητας πιθανότητας.

Υποθέτουμε ότι έχουμε ένα σύνολο δεδομένων S και έστω ότι κάθε δείγμα δεδομένων X=(x1,x2,...,xn) με m κατηγορίες C1,C2,..,Cm. Δεδομένου ενός αγνώστου δείγματος δεδομένων X, ο κατηγοριοποιητής θα προβλέψει ότι το X ανήκει στην κατηγορία C που έχει την μέγιστη εκ των υστέρων(posterior) πιθανότητα με βάση το X. Αυτό σημαίνει ότι το X κατηγοριοποιείται στην Ci αν και μόνο αν:

$$p(Ci|X)>p(Cj|X) \text{ για κάθε } 1\leq j\leq m \text{ και } j\neq i$$

Ο στόχος, λοιπόν, είναι να βρούμε την μέγιστη posterior πιθανότητα, δηλαδή το μέγιστο p(Ci|X) για κάθε κλάση, με αποτέλεσμα ο Naive Bayesian κατηγοριοποιητής να έχει υψηλή απόδοση. Η απόδοση του συγκρίνεται με αυτή των δέντρων απόφασης και κάποιους κατηγοριοποιητές που στηρίζονται σε νευρωνικά δίκτυα σε ορισμένες εφαρμογές.

2.2.3 Δένδρα Απόφασης

Τα δέντρα απόφασης ή δέντρα ταξινόμησης (decision trees ή classification trees) είναι μια από τις πιο συχνά χρησιμοποιούμενες τεχνικές κατηγοριοποίησης, γιατί προσφέρει σαφή και κατανοητά αποτελέσματα μέσα σε λίγο χρόνο. Η συγκεκριμένη τεχνική μπορεί να χρησιμοποιηθεί για την ταξινόμηση και την πρόβλεψη τόσο ονομαστικών, όσο και αριθμητικών ποσοτήτων. Σύμφωνα με τους

Quinlan (1986, 1987, 1993) και Murphy (1998)[10], τα δέντρα απόφασης είναι δομές που ταξινομούν τα αντικείμενα μιας βάσης δεδομένων βάσει των τιμών των χαρακτηριστικών αυτών. Τα δέντρα απόφασης κατασκευάζονται χρησιμοποιώντας μόνο εκείνα τα γνωρίσματα που είναι σε θέση να διακρίνουν τις έννοιες προς εκμάθηση.

Για να χτίσουμε ένα δέντρο απόφασης, πρέπει αρχικά να επιλέξουμε ένα υποσύνολο περιπτώσεων από το σύνολο των δεδομένων που θα χρησιμοποιηθούν στην εκπαίδευση (υποσύνολο δεδομένων εκπαίδευσης - training set). Αυτό το υποσύνολο (δεδομένα ελέγχου - test set) χρησιμοποιείται έπειτα από τον αλγόριθμο για να κατασκευάσει το δέντρο απόφασης. Τα υπόλοιπα δεδομένα, τα δεδομένα training set, χρησιμοποιούνται στην εξέταση της ακρίβειας του κατασκευασμένου δέντρου. Εάν το δέντρο απόφασης ταξινομεί τις περιπτώσεις σωστά, η διαδικασία ολοκληρώνεται. Εάν μια περίπτωση είναι ανακριβώς ταξινομημένη, η περίπτωση προστίθεται στο επιλεγμένο υποσύνολο των training set και ένα νέο δέντρο κατασκευάζεται.

Όλα τα δέντρα απόφασης χαρακτηρίζονται και περιγράφονται από συγκεκριμένους όρους για την κατασκευή τους και διέπονται από τις ακόλουθες ιδιότητες:

- Κάθε εσωτερικός κόμβος ονοματίζεται με το όνομα ενός χαρακτηριστικού Xi.
- Κάθε κλαδί/σύνδεση ονοματίζεται με ένα κατηγόρημα που μπορεί να εφαρμοστεί στο χαρακτηριστικό που αποτελεί το όνομα του κόμβου-πατέρα.
- Κάθε φύλλο ονοματίζεται με το όνομα μιας κλάσης

Τα χαρακτηριστικά βήματα που πρέπει να ακολουθεί ένας αλγόριθμος κατασκευής δέντρων απόφασης:

1. Έστω Χ είναι το υποσύνολο δεδομένων εκπαίδευσης (training set)

2. Επιλέγουμε ένα χαρακτηριστικό που διαφοροποιεί καλύτερα τις περιπτώσεις που περιλαμβάνονται στο Χ.

3. Δημιουργούμε έναν κόμβο στο δέντρο του οποίου η αξία είναι το επιλεγμένο χαρακτηριστικό. Δημιουργούμε

[10] **J.R. QUINLAN**, Induction of Decision Trees, 2007

θυγατρικούς δεσμούς από αυτόν τον κόμβο, όπου κάθε σύνδεση αντιπροσωπεύει μια μοναδική αξία για τα επιλεγμένα χαρακτηριστικά. Χρησιμοποιούμε τις τιμές των θυγατρικών δεσμών για να υποδιαιρέσουμε περαιτέρω τα στιγμιότυπα σε δευτερεύουσες κλάσεις.

4. Για κάθε δευτερεύουσα κλάση που δημιουργήθηκε στο βήμα 3:

    - Εάν τα στιγμιότυπα στη δευτερεύουσα κλάση ικανοποιούν προκαθορισμένα κριτήρια ή εάν το σύνολο των υπολοίπων επιλογών γνωρισμάτων γι' αυτή τη διαδρομή του δέντρου είναι μηδέν, καθορίζουμε την κατηγοριοποίηση των καινούργιων στιγμιότυπων που ακολουθούν αυτή τη διαδρομή αποφάσεων.
    - Εάν η δευτερεύουσα κλάση δεν ικανοποιεί τα προκαθορισμένα κριτήρια, και υπάρχει τουλάχιστον ένα γνώρισμα για να υποδιαιρέσει περαιτέρω τη διαδρομή του δέντρου, αφήστε το Χ να είναι το τρέχον σύνολο των στιγμιότυπων της δευτερεύουσας κλάσης και επιστρέφουμε στο βήμα 2.

Μερικά από τα κρίσιμα ζητήματα που αφορούν τους αλγόριθμους δημιουργίας δένδρων απόφασης ή κατηγοριοποίησης είναι τα ακόλουθα:

- ✓ **Χαρακτηριστικά διάσπασης (splitting attributes):** Είναι Τα χαρακτηριστικά των παραδειγμάτων στη βάση Χ που χρησιμοποιούνται σαν ονόματα κόμβων του δέντρου, δηλ. επιλέχτηκαν ως καλύτερα χαρακτηριστικά. Διαφορετικά σύνολα χαρακτηριστικών διάσπασης έχουν σαν αποτέλεσμα διαφορετικά Δέντρα Απόφασης με διαφορετική απόδοση. Η επιλογή τους στηρίζεται όχι μόνο στο σύνολο εκπαίδευσης, αλλά και στη γνώμη του εμπειρογνώμονα.

- ✓ **Κριτήρια διάσπασης (splitting criterion):** Τα κριτήρια με βάση τα οποία επιλέγεται το καλύτερο χαρακτηριστικό διάσπασης κάθε φορά. Η σειρά επιλογής των χαρακτηριστικών διάσπασης παίζει σημαντικό ρόλο στην απόδοση ενός Δέντρου Απόφασης. Ο αριθμός διασπάσεων συνδέεται με τη διάταξη των χαρακτηριστικών διάσπασης. Ο αριθμός διασπάσεων μπορεί εύκολα να προσδιοριστεί όταν

το πεδίο είναι μικρό (λίγα χαρακτηριστικά, λίγες και διακριτές τιμές), αλλιώς (πολλά χαρακτηριστικά ή πολλές/συνεχείς τιμές) η πολυπλοκότητα επιλογής και κατ' επέκταση δημιουργίας του Δέντρου Απόφασης ανεβαίνει.

- ✓ **Η δομή του δένδρου:** Επιθυμητό είναι να δημιουργούνται δέντρα που είναι ισορροπημένα και με τα λιγότερα επίπεδα (μικρότερο βάθος). Ισοζυγισμένα δένδρα λίγων επιπέδων ασφαλώς και βοηθούν στην αποδοτικότερη κατηγοριοποίηση. Αυτό όμως δεν είναι πάντα εφικτό ούτε το υπολογιστικά φτηνότερο. Μερικοί αλγόριθμοι δημιουργούν μόνο δυαδικά δέντρα.

- ✓ **Κριτήριο τερματισμού (stopping criterion):** Η δημιουργία ενός δέντρου σταματά οπωσδήποτε όταν όλα τα δεδομένα του (εναπομείναντος) συνόλου εκπαίδευσης κατηγοριοποιούνται πλήρως. Μπορεί όμως να είναι απαραίτητο να σταματήσει νωρίτερα για να αποφευχθούν π.χ. μεγάλα δέντρα. Το πότε ή πού θα σταματήσει είναι θέμα συναλλαγής (trade-off) μεταξύ ακρίβειας (accuracy) και απόδοσης (performance) του αλγορίθμου. Επίσης, πρώιμος τερματισμός μπορεί να γίνει για αποφυγή του φαινομένου της υπερπροσαρμογής (overfitting). Τέλος, μπορεί να προχωρήσει σε μεγαλύτερα δέντρα αν είναι γνωστό ότι υπάρχουν κατηγορίες δεδομένων που δεν αντιπροσωπεύονται στο σύνολο εκπαίδευσης.

- ✓ **Τα δεδομένα εκπαίδευσης:** Η δημιουργία ενός δένδρου απόφασης βασίζεται αποκλειστικά στα δεδομένα εκπαίδευσης. Αν το σύνολο δεδομένων είναι πολύ μικρό, τότε το δέντρο μπορεί να μην είναι τόσο λεπτομερές, ώστε να ταξινομεί γενικότερα δεδομένα. Αν είναι πολύ μεγάλο, το δέντρο πιθανόν να προκαλέσει υπερπροσαρμογή (overfitting).

- ✓ **Κλάδεμα (Pruning):** Μετά τη δημιουργία ενός Δέντρου Απόφασης μπορεί να χρειάζονται τροποποιήσεις για να βελτιώσουν την απόδοσή του, όπως π.χ. το κλάδεμα πλεοναζόντων συγκρίσεων ή υποδέντρων.

Οι αλγόριθμοι ταξινόμησης που βασίζονται στα δέντρα απόφασης, περιλαμβάνουν δύο διακριτές φάσεις:

1. **Τη φάση οικοδόμησης (building phase):** Σε αυτή την πρώτη φάση, η οποία χρίζει μεγαλύτερης έρευνας και προσπάθειας, το σύνολο των δεδομένων εκπαίδευσης χωρίζεται πολλές φορές, έως ότου όλα τα αντικείμενα σε ένα τμήμα του ανωτέρω συνόλου να ανήκουν στην ίδια κλάση.

2. **Τη φάση κλαδέματος (pruning phase):** Έπειτα, αφού έχει ήδη δημιουργηθεί το δέντρο απόφασης, οι περισσότεροι αλγόριθμοι εκτελούν τη φάση του κλαδέματος, περικόπτοντας κάποιους από τους κόμβους, προκειμένου αφενός να αποτραπούν επικαλύψεις, και αφετέρου το δέντρο να έχει υψηλότερη ακρίβεια ταξινόμησης.

Τα πλεονεκτήματα από τη χρήση δένδρων αποφάσεων κατηγοριοποίησης είναι πολλά και παρατίθενται παρακάτω:

Τα βασικά πλεονεκτήματα της ανάλυσης των δένδρων αποφάσεων είναι:

i. Αποτελεί τον καλύτερο τρόπο περιγραφής του προβλήματος γιατί παρουσιάζει κάθε ενέργεια (απόφαση), καθώς και τις αντίστοιχες δεδομένες εκβάσεις με σαφήνεια και απλότητα.
ii. Το μοντέλο του δένδρου αποφάσεων διακρίνεται για τη δυνατότητα προσαρμογής στις μεταβαλλόμενες συνθήκες του περιβάλλοντος. Ειδικότερα, διευκολύνει τη διενέργεια πειραματισμών ή την εκτέλεση τυχόν άλλων δραστηριοτήτων, καθώς και την προσθήκη άλλων πιθανών εκβάσεων (καταστάσεων της φύσης) κάτω από το φως νέων πληροφοριών.
iii. Διευκολύνει τον εντοπισμό των ευαίσθητων σημείων των διαφόρων ενεργειών (στρατηγικών) που χρειάζονται ιδιαίτερη προσοχή και αντιμετώπιση. Μ’ αυτόν τον τρόπο συμβάλλει στην άσκηση «διοίκησης με βάση τις εξαιρέσεις».
iv. Στην τεχνική αυτή μας δίνεται η δυνατότητα ανάλυσης τόσο ονοματικών, όσο και αριθμητικών δεδομένων. Σε συνδυασμό και με την ευκολία στο να κατανοηθεί από τον άνθρωπο, γίνεται εύκολα αντιληπτό ότι μπορεί να εφαρμοστεί σε πολλά και ποικίλα προβλήματα από οποιοδήποτε διοικητικό φορέα.
v. Τα δέντρα αποφάσεων αναγκάζουν τους αναλυτές να μελετήσουν τη σειρά των αποφάσεων. Πολύ εύκολα μπορεί κάποιος να εξακριβώσει ότι μια συνθήκη δεν μπορεί να υπάρξει παρά μόνο εάν υπάρχει ήδη κάποια άλλη συνθήκη και έχει διευθετηθεί με μια απόφαση. Έτσι καθορίζουμε ακόμη και τον χρόνο και την σειρά που θα λάβει χώρα κάθε συνθήκη και θα ληφθεί κάθε απόφαση.

Δε λείπουν ωστόσο και τα μειονεκτήματα από τη χρήση δένδρων απόφασης μερικά από τα οποία είναι:

i. Δε χειρίζονται εύκολα δεδομένα, τα γνωρίσματα των οποίων αποτελούνται από συνεχείς τιμές.
ii. Υπάρχει η πιθανότητα υπερ-προσπαρμογής ενός δένδρου στα σύνολα δεδομένων εκπαίδευσης.
iii. Δέντρα αποφάσεων στηριζόμενα σε αριθμητικά δεδομένα μπορεί να είναι ιδιαιτέρως πολύπλοκα.
iv. Οι σχετικοί αλγόριθμοι έχουν αποδειχθεί εξαιρετικά ασταθείς.

2.2.4 Νευρωνικά Δίκτυα

Πέρα από τις μεθόδους ταξινόμησης που βασίζονται στα δέντρα και τους κανόνες απόφασης, τα τεχνητά νευρωνικά δίκτυα (artificial neural networks) είναι επίσης μια διαδεδομένη μέθοδος ταξινόμησης (Michie et al, 1995)[11]. Εμπνευσμένα από το βιολογικό νευρικό σύστημα, και ειδικότερα από τον ανθρώπινο εγκέφαλο, διαθέτουν αξιοσημείωτα χαρακτηριστικά, όπως τη δυνατότητα τους να αναπαριστούν σύνθετες εξαρτήσεις ή
την ικανότητα τους να προβλέπουν την κλάση άγνωστων παρατηρήσεων.

Το νευρωνικό δίκτυο είναι ένα δίκτυο από απλούς υπολογιστικούς κόμβους (νευρώνες), διασυνδεδεμένους μεταξύ τους. Οι νευρώνες είναι τα δομικά στοιχεία του δικτύου. Κάθε τέτοιος κόμβος δέχεται ένα σύνολο αριθμητικών εισόδων από διαφορετικές πηγές (είτε από άλλους νευρώνες, είτε από το περιβάλλον), επιτελεί έναν υπολογισμό με βάση αυτές τις εισόδους και παράγει μία έξοδο. Η εν λόγω έξοδος είτε κατευθύνεται στο περιβάλλον, είτε τροφοδοτείται ως είσοδος σε άλλους νευρώνες του δικτύου. Στην ουσία αποτελούν διασυνδεδεμένα υπολογιστικά στοιχεία που έχουν την ικανότητα να ανταποκρίνονται σε ερεθίσματα που δέχονται σαν είσοδο και να μαθαίνουν να προσαρμόζονται στο περιβάλλον τους.

[11] **D. Michie, D.J. Spiegelhalter, C.C. Taylor**, Machine Learning, Neural and Statistical Classification, 1994

Οι νευρώνες ενός δικτύου χωρίζονται σε τρείς βασικές κατηγορίες:

1) **Τους νευρώνες εισόδου (input neurons):** οι οποίοι δέχονται τις πληροφορίες που θα υποστούν επεξεργασία. Δεν επιτελούν κανέναν υπολογισμό, μεσολαβούν απλώς ανάμεσα στις περιβαλλοντικές εισόδους του δικτύου και στους υπολογιστικούς νευρώνες

2) **Τους νευρώνες εξόδου (output neurons):** στους οποίους καταλήγουν τα αποτελέσματα της γενικότερης επεξεργασίας και εκμάθησης του αλγορίθμου και διοχετεύουν στο περιβάλλον τις τελικές αριθμητικές εξόδους του δικτύου.

3) **Τους ενδιάμεσους ή υπολογιστικούς νευρώνες:** οι οποίοι βρίσκονται μεταξύ των νευρώνων εισόδου και εξόδου. Οι τελευταίοι εναλλακτικά ονομάζονται και κρυφοί νευρώνες (hidden neurons). Οι υπολογιστικοί νευρώνες πολλαπλασιάζουν κάθε είσοδό τους με το αντίστοιχο συνοπτικό βάρος και υπολογίζουν το ολικό άθροισμα των γινομένων. Το άθροισμα αυτό τροφοδοτείται ως όρισμα στη συνάρτηση ενεργοποίησης, την οποία υλοποιεί εσωτερικά κάθε κόμβος. Η τιμή που λαμβάνει η συνάρτηση για το εν λόγω όρισμα είναι και η έξοδος του νευρώνα για τις τρέχουσες εισόδους και βάρη.

Ουσιαστικά, οι νευρώνες σε ένα δίκτυο είναι αφενός ένα σύνολο εισερχόμενων τιμών και των αντίστοιχων βαρών τους και αφετέρου μια συνάρτηση που αθροίζει τα παραπάνω βάρη, αντιστοιχώντας τα αποτελέσματα σε ένα νευρώνα εξόδου (Aggarwal & Yu, 1999).

Ένα νευρωνικό δίκτυο οφείλει την υπολογιστική του ισχύ κατά πρώτον στην παράλληλη, κατανεμημένη δομή του και κατά δεύτερον στην ικανότητά του να μαθαίνει και να γενικεύει τη γνώση που λαμβάνει. Τα νευρωνικά δίκτυα έχουν την ικανότητα να εξάγουν κάποιο συμπέρασμα από πολύπλοκα ή μη ακριβή δεδομένα και μπορούν να χρησιμοποιηθούν για να εξάγουν πρότυπα και να προσδιορίζουν τάσεις οι οποίες είναι πολύπλοκες για να προσδιοριστούν από ανθρώπους ή από άλλες υπολογιστικές τεχνικές. Τα νευρωνικά δίκτυα είναι μια τεχνική ισχυρά καθοδηγούμενη από τα δεδομένα. Αυτό σημαίνει ότι δεν επιβάλλουν αυθαίρετες υποθέσεις και ότι τα μοντέλα τους πηγάζουν από την επεξεργασία των δεδομένων. Εξαιτίας και της γενικότερης αρχής που τα διέπουν, χρησιμοποιούν ένα σύνολο από στοιχεία επεξεργασίας (κόμβους) ανάλογους με τους

νευρώνες στο ανθρώπινο μυαλό. Τα στοιχεία αυτά διασυνδέονται μεταξύ τους σε ένα δίκτυο το οποίο μπορεί να αναγνωρίζει πρότυπα μέσα σε ένα σύνολο δεδομένων μόλις αυτά παρουσιαστούν μέσα στα δεδομένα, δηλαδή το δίκτυο μπορεί να μαθαίνει από την εμπειρία όπως ακριβώς κάνουν και οι άνθρωποι.

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 (π.χ. η σταδιακή προσέγγιση μίας συνάρτησης). 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.

Πέραν όμως από την ικανότητα μάθησης ή εκπαίδευσής τους. τα νευρωνικά δίκτυα χαρακτηρίζονται και από την αρχιτεκτονική τους και από τη λειτουργία που επιτελούν. Η αρχιτεκτονική του δικτύου καθορίζει τη διάταξη των συνδέσεων των νευρώνων καθώς και τον αριθμό και τον τύπο τους. Οι νευρώνες οργανώνονται σε μορφή επιπέδων και ο τρόπος με τον οποίο είναι δομημένοι σχετίζεται με τον αλγόριθμο μάθησης που χρησιμοποιείται για την εκπαίδευση αυτού του δικτύου. Υπάρχουν δύο κατηγορίες αρχιτεκτονικών δομών των νευρωνικών δικτύων:

1. **Δίκτυα πρόσθιας τροφοδότησης (feedforward):** Δίκτυα στα οποία δεν υπάρχουν συνδέσεις επανατροφοδοσίας (feedback), δηλαδή δεν υπάρχουν συνδέσεις μεταξύ νευρώνων ενός επιπέδου και νευρώνων προηγούμενου ή του ίδιου επιπέδου. Το πρώτο στρώμα λαμβάνει τα εισερχόμενα σήματα και τα κατανέμει στο δεύτερο στρώμα. Το δεύτερο στρώμα ονομάζεται κρυφό στρώμα (hidden layer), επειδή δεν διαθέτει συνδέσεις προς τα έξω. Το τρίτο στρώμα εξόδου παραδίδει τα αποτελέσματα στον εξωτερικό κόσμο. Τα Feed-forward νευρωνικά δίκτυα επιτρέπουν τη διέλευση σημάτων μόνο προς μία κατεύθυνση, από την είσοδο στην έξοδο, δηλ. όλα τα σήματα προωθούνται μόνον

από το πρώτο στρώμα και δεν υπάρχουν αναδράσεις (από εκεί προέρχεται και η ονομασία τους). Δεν υπάρχουν προς τα πίσω loops και έτσι η έξοδος οποιουδήποτε στρώματος δεν επηρεάζει το ίδιο το στρώμα. Τα Feed-forward δίκτυα δεν έχουν μνήμη, έτσι η έξοδος τους καθορίζεται πάντα από την παρούσα είσοδο και από τις τιμές των βαρών.

2. **Αναδρομικά δίκτυα (recurrent neural network):** Δίκτυα στα οποία επιτρέπεται η ανατροφοδότηση, υπάρχουν δηλαδή συνδέσεις μεταξύ νευρώνων ενός επιπέδου και νευρώνων προηγούμενου επιπέδου και έχουμε αμφίδρομη διέλευση σημάτων με loops. Είναι πολύ δυναμικά και σε πολλές περιπτώσεις μπορεί να είναι πολύπλοκα. Η κατάστασή τους αλλάζει συνεχώς έως ότου να φτάσουν σε ένα σημείο ηρεμίας. Παραμένουν σ' αυτό το σημείο έως ότου αλλάξει η είσοδος. Τα δίκτυα αυτά ονομάζονται αλλιώς και Feedback νευρωνικά δίκτυα.

Σε ένα νευρωνικό δίκτυο πρόσθιας τροφοδότησης, τα κύρια βήματα για την κατασκευή ενός μοντέλου ταξινόμησης, είναι τα εξής (Βαζιργιάννης & Χαλκίδη, 2003)[12]:

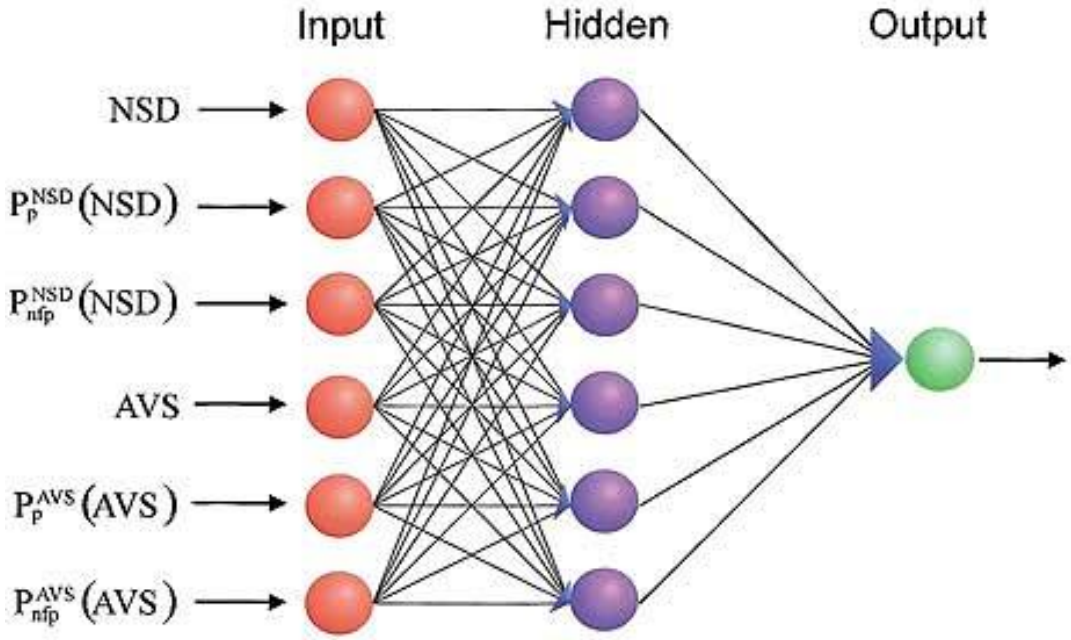
- ✓ Η αναγνώριση των χαρακτηριστικών εισόδου και εξόδου
- ✓ Η κατασκευή ενός δικτύου με την κατάλληλη τοπολογία
- ✓ Η επιλογή του σωστού συνόλου εκπαίδευσης το οποίο περιλαμβάνει δεδομένα που είναι ορισμένα ανά ζεύγη
- ✓ Η εκπαίδευση του δικτύου στην οποία τα δεδομένα εισέρχονται στο νευρωνικό δίκτυο ένα ένα. Το νευρωνικό δίκτυο μαθαίνει συγκρίνοντας τα αποτελέσματα ταξινόμησης ενός αντικειμένου με την γνωστή πραγματική ταξινόμηση αυτού. Τα λάθη από την αρχική ταξινόμηση του πρώτου αντικειμένου χρησιμοποιούνται για να διορθωθεί το δίκτυο μέσω της τροποποίησης των συναρτήσεων των νευρώνων. Η παραπάνω διαδικασία είναι επαναληπτική. Η επαναληπτική φύση ωστόσο της διαδικασίας εκπαίδευσης σημαίνει ότι ένα νευρωνικό δίκτυο είναι αρκετά αργό.

[12] **Μιχάλης Βαζιργιάννης, Μαρία Χαλκίδη**, Εξόρυξη Γνώσης από Βάσεις Δεδομένων και τον Παγκόσμιο Ιστό, 2η έκδοση, 2005

- ✓ Ο έλεγχος του δικτύου χρησιμοποιώντας ένα σύνολο ελέγχου, το οποίο είναι ανεξάρτητο από το σύνολο εκπαίδευσης.

Τα νευρωνικά δίκτυα δεν είναι προγραμματισμένα να λειτουργούν όπως τα συμβατικά προγράμματα υπολογιστών, αλλά μαθαίνουν από την εμπειρία. Ένα νευρωνικό δίκτυο μαθαίνει κατά τη διάρκεια μιας φάσης κατάρτισης στην οποία οι γνωστές είσοδοι και έξοδοι (αποτελέσματα) δείχνονται στο δίκτυο διαδοχικά και επανειλημμένα. Ένας αλγόριθμος κατάρτισης ρυθμίζει τα βάρη σε κάθε σύνδεση με στόχο να μειώσει το σφάλμα. Αρχικά τα αποτελέσματα που παράγονται από το δίκτυο είναι κάπως αυθαίρετα. Αλλά καθώς περνά χρόνος, δεδομένου ότι οι περιπτώσεις είναι επανειλημμένα επανεισαγμένες (εκατοντάδες ή χιλιάδες φορές), το δίκτυο αρχίζει να παίρνει σωστές μερικές των απαντήσεων. Ο αλγόριθμος κατάρτισης συνεχίζει να αλλάζει τα βάρη έως ότου οι περισσότερες από τις απαντήσεις είναι σωστές οπότε και η κατάρτιση του δικτύου σταματά. Η επόμενη φάση είναι να εξεταστεί ή να επικυρωθεί το νευρωνικό δίκτυο που αναπτύχθηκε. Ανάλογα με την απόδοση του δικτύου σε αυτό το σημείο καθορίζεται εάν αυτό έχει εκπαιδευτεί και διδαχθεί σωστά.

Στο παρακάτω σχήμα παραθέτουμε μία χαρακτηριστική απεικόνιση ενός νευρωνικού δικτύου, όπου διακρίνονται οι νευρώνες εισόδου, οι κρυμμένοι νευρώνες και οι νευρώνες εξόδου όπως περιγράφτηκαν παραπάνω.



Εικόνα 2.4: Τεχνητό νευρωνικό δίκτυο με διαδικασίες εισόδου και εξόδου δεδομένων. (πηγή http://www.logistics-management.gr/news/272)

2.3 Συσταδοποίηση

Η Συσταδοποίηση (Clustering) είναι μια από τις πιο χρήσιμες διεργασίες της εξόρυξης δεδομένων για την ομαδοποίηση των αντικειμένων σε συστάδες (clusters). Πιο συγκεκριμένα, το πρόβλημα της συσταδοποίησης σχετίζεται με τη διαμέριση (partitioning ή clustering) ενός συνόλου δεδομένων σε συστάδες, έτσι ώστε τα στοιχεία που ανήκουν σε μία συστάδα να είναι περισσότερο όμοια μεταξύ τους από ότι είναι με τα στοιχεία των άλλων συστάδων.

Έστω για παράδειγμα, τα προφίλ των χρηστών ενός ηλεκτρονικού καταστήματος. Θα μπορούσαμε να ομαδοποιήσουμε τους χρήστες με βάση τις αγοραστικές τους προτιμήσεις και να κατατάξουμε σε μια συστάδα εκείνους που παρουσιάζουν όμοιες αγοραστικές συνήθειες. Επομένως, σκοπός αυτής της διαδικασίας είναι η οργάνωση των στοιχείων σε «λογικές» ομάδες, έτσι ώστε να ανακαλύψουμε ομοιότητες και διαφορές μεταξύ των στοιχείων, αλλά και να εξάγουμε χρήσιμα συμπεράσματα για αυτά.

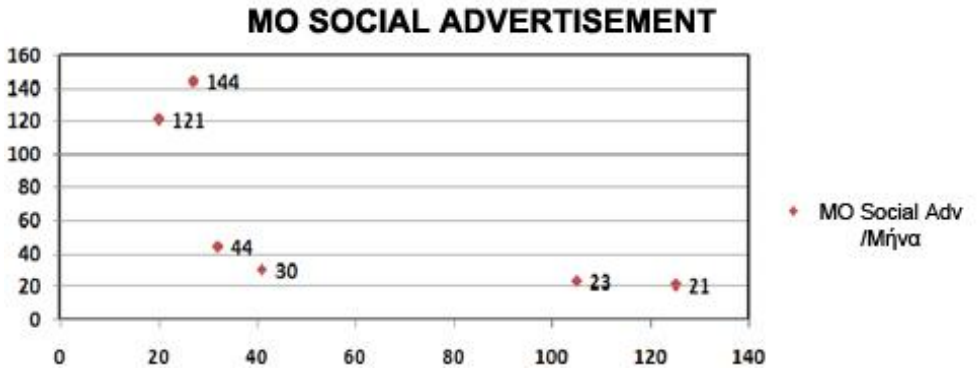
Ένα απλό παράδειγμα για την διαδικασία της συσταδοποίησης φαίνεται παρακάτω στον πίνακα 2.2: έχουμε έναν πίνακα με αναφορές από τους επιλεγμένους τρόπους διαφήμισης ορισμένων εταιριών, ώστε να ομαδοποιηθούν οι εταιρίες. σε κατηγορίες. Ως δεδομένα έχουμε τα ID των εταιριών, τον Μ.Ο. των μηνιαίων διαφημίσεων στα Social Media και τον Μ.Ο. των μηνιαίων διαφημίσεών τους στα υπόλοιπα μέσα διαφήμισης (πχ. τηλεόραση, ραδιόφωνο κλπ.).

ID ΕΤΑΙΡΙΑΣ	ΜΟ OFFINE ADV /ΜΗΝΑ	ΜΟ SOCIAL ADV/ΜΗΝΑ
1	27	144
2	32	44
3	41	30
4	125	21

5	105	23
6	**20**	121

Πίνακας 2.2: Δεδομένα παραδείγματος "Διαφήμιση Εταιρειών"

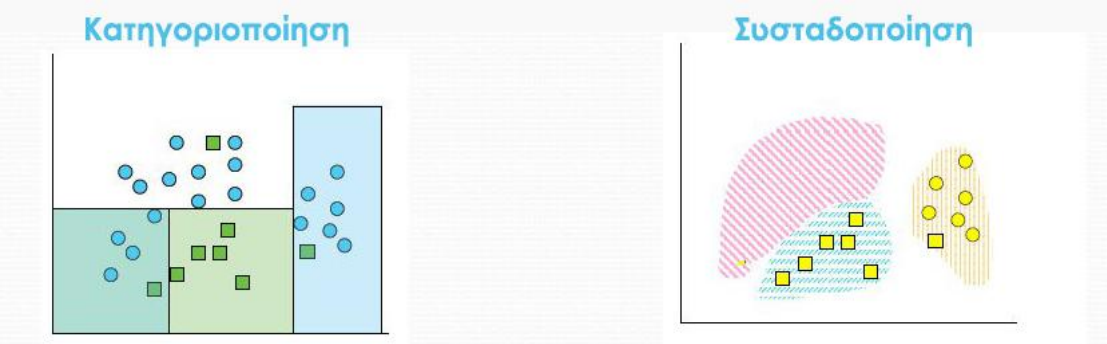
Όπως μπορούμε να δούμε στο παρακάτω σχήμα οι εταιρίες με ID 1 και ο 6 έχουν μικρή παρουσία σε offline τρόπους διαφήμισης και κατ' επέκταση μεγάλη προβολή στα Social Media. Με άλλα λόγια οι δύο αυτές εταιρίες θα λέγαμε ότι ανήκουν στην ίδια κατηγορία, όπου επιλέγουν τα Social Media ως τρόπο διαφήμισής τους. Για αυτό τον λόγο έχουν συγκεντρωθεί μαζί και έχουν φτιάξει μια ομάδα εταιριών η οποία μπορεί να λέγεται “Υψηλή Προβολή στα Social Media”. Επίσης, οι εταιρίες με ID 3 και 2 έχουν συγκεντρωθεί μαζί σχηματίζοντας την ομάδα "Εταιρίες με Τυπική Προβολή". Τέλος, οι 4 και 5 έχουν σχηματίσει την ομάδα "Υψηλή Προβολή σε Offline Μέσα". Όλα τα παραπάνω απεικονίζονται στην εικόνα 2.3 που ακολουθεί.



Εικόνα 2.5: Παράδειγμα Συσταδοποίησης

Στη διαδικασία της συσταδοποίησης δεν υπάρχουν προκαθορισμένες κατηγορίες, ούτε κάποια άλλη προηγούμενη γνώση σχετικά με τη σχέση μεταξύ των στοιχείων. Για το λόγο αυτό, η συσταδοποίηση είναι γνωστή και ως διαδικασία μη εποπτευόμενης μάθησης (unsupervised learning). Αντίθετως, η κατηγοριοποίηση είναι η διαδικασία με την οποία ένα σύνολο

αντικειμένων αντιστοιχίζεται σε ένα σύνολο προκαθορισμένων κατηγοριών εξετάζοντας τα χαρακτηριστικά κάθε αντικειμένου.



Εικόνα 2.6: Κατηγοριοποίηση vs Συσταδοποίηση

Στη συνέχεια, θα αναφερθούμε στα βήματα από τα οποία αποτελείται η διαδικασία της συσταδοποίησης.

Επιλογή Χαρακτηριστικών Γνωρισμάτων

Στόχος είναι η επιλογή κατάλληλων χαρακτηριστικών γνωρισμάτων στα οποία θα εφαρμόσουμε τη συσταδοποίηση. Η διαδικασία της προεπεξεργασίας είναι απαραίτητη σε αυτό το βήμα, έτσι ώστε τα δεδομένα να είναι σε κατάλληλη μορφή επεξεργασίας.

Αλγόριθμος Συσταδοποίησης

Σε αυτό το βήμα, γίνεται η επιλογή του κατάλληλου αλγόριθμου συσταδοποίησης. Η επιλογή του εξαρτάται από τα δεδομένα που πρόκειται να συσταδοποιηθούν και τις ανάγκες τις συγκεκριμένης εφαρμογής. Το μέτρο γειτνίασης και το κριτήριο συσταδοποίησης είναι αυτά που κυρίως χαρακτηρίζουν έναν αλγόριθμο συσταδοποίησης.

Α. Με το μέτρο γειτνίασης, υπολογίζεται η ομοιότητα μεταξύ των στοιχείων.

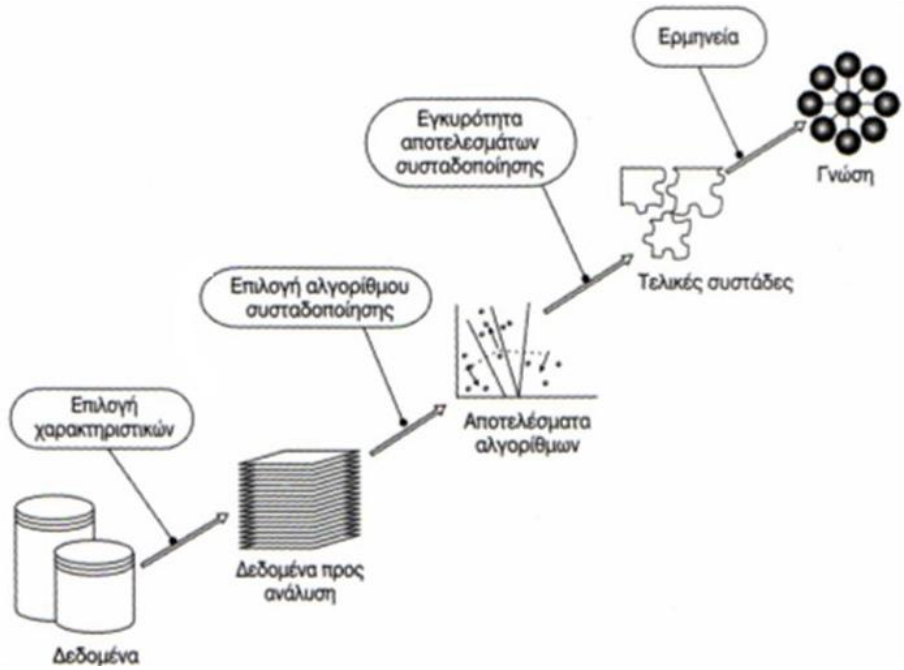
Β. Το Κριτήριο Συσταδοποίησης, εκφράζεται συνήθως μέσω μιας συνάρτησης κόστους ή κάποιου άλλου τύπου κανόνων. Το κριτήριο αυτό θα πρέπει να καθοριστεί έτσι ώστε να ταιριάζει με το σύνολο δεδομένων που θα συσταδοποιηθεί.

Επικύρωση Αποτελεσμάτων

Η επικύρωση των αποτελεσμάτων αποτελεί επίσης ένα κρίσιμο βήμα στη διαδικασία της συσταδοποίησης, καθώς στη συσταδοποίηση οι συστάδες που παράγονται από έναν αλγόριθμο συσταδοποίησης δεν είναι εκ των προτέρων γνωστές. Η ακρίβεια των αποτελεσμάτων μπορεί να επικυρωθεί χρησιμοποιώντας κατάλληλα κριτήρια και τεχνικές.

Ερμηνεία Αποτελεσμάτων

Τα αποτελέσματα της συσταδοποίησης θα πρέπει να συνδυαστούν με άλλα πειραματικά στοιχεία και αποτελέσματα προηγούμενων αναλύσεων των ίδιων στοιχείων προκειμένου να εξαχθούν σωστά συμπεράσματα.



Εικόνα 2.7: Βήματα Συσταδοποίησης (πηγή: Συσταδοποίηση Δεδομένων - http://slideplayer.gr/slide/2915809/)

2.4 Ανάλυση Συσχέτισης

Πολλές επιχειρήσεις συλλέγουν μεγάλες ποσότητες δεδομένων οι οποίες αναφέρονται στις αγορές των πελατών που μαζεύονται στα ταμεία των καταστημάτων. Ο κάθε πωλητής ενδιαφέρεται για την ανάλυση των δεδομένων για να μάθει για το τι αγοράζουν οι πελάτες, ποια προϊόντα αγοράζονται περισσότερο, κτλ. Με αυτά ασχολείται η ανάλυση καλαθιού αγοράς (market basket analysis), στο οποίο αναφορά γίνεται πιο κάτω. Αυτή η τεχνική χρησιμοποιείται κυρίως για την επεξεργασία πολλών δεδομένων. Επίσης, η εξόρυξη κανόνων συσχέτισης επιδιώκει να βρει κανόνες συσχέτισης οι οποίοι ικανοποιούν τις προκαθορισμένες απαιτήσεις υποστήριξης (support) και εμπιστοσύνης (confidence). Ως κανόνας συσχέτισης ορίζεται η μέθοδος ανακάλυψης σχέσεων μεταξύ των διαφόρων μεταβλητών σε μεγάλες βάσεις δεδομένων. Ένας άλλος ορισμός είναι μια έκφραση της μορφής X → Y όπου X και Y είναι στοιχεία συνόλου X ⊆ I, Y ⊆ I, X ∩ Y = ∅ [16] . Τέλος οι κανόνες συσχέτισης χρησιμοποιούνται αρκετά συχνά επειδή είναι αποτελεσματικοί και γρήγοροι στην επεξεργασία μεγάλου πλήθος δεδομένων.

Δύο ποσοτικά μεγέθη καθορίζουν πόσο ισχυρός είναι ο κανόνας X → Y. Τα μέτρα αυτά είναι η υποστήριξη (support) και η εμπιστοσύνη (confidence).

Η υποστήριξη του κανόνα X → Y είναι το ποσοστό των συναλλαγών (επί του συνόλου των συναλλαγών) που περιέχουν και το X και το Y. Μαθηματικά, αυτό ορίζεται με την παρακάτω σχέση:

$$supp\,(X \rightarrow Y) = P(X \cup Y)$$

Η εμπιστοσύνη του κανόνα X → Y είναι η δεσμευμένη πιθανότητα εμφάνισης του Y, όταν εμφανίζεται το X. Με απλούστερα λόγια, επιλέγονται μόνον οι συναλλαγές που περιέχουν το X και επί αυτών των συναλλαγών υπολογίζεται το ποσοστό εκείνων που περιέχουν το Y. Μαθηματικά, αυτό ορίζεται με την παρακάτω σχέση

$$conf(X \rightarrow Y) = P(Y|X)$$

Για την καλύτερη κατανόηση των εννοιών αυτών παρακάτω ακολουθεί ένα παράδειγμα κανόνων συσχέτισης.

2.4.1 Ανάλυση Καλαθιού Αγοράς

Ένα πρόβλημα που επιλύεται με του κανόνες συσχέτισης είναι η ανάλυση καλαθιού αγοράς (market basket analysis). Η ανάλυση καλαθιού αγοράς αναφέρεται στις πληροφορίες που αντλούνται για το τι αγοράζουν οι πελάτες, για να πάρουν μία πρώτη άποψη για το ποιοι είναι και γιατί κάνουν ορισμένες αγορές. Δείχνει ποια προϊόντα πρόκειται να αγοραστούν μαζί καθώς και ποια προωθούνται περισσότερο.

Αυτή η τεχνική έχει ευρέως εφαρμοστεί σε σουπερμάρκετ, όπως επίσης και τα τελευταία χρόνια στα online καταστήματα, τα οποία αναπτύσσονται ραγδαία στον Παγκόσμιο Ιστό. Αλγόριθμοι και τεχνικές πρόβλεψης καταναλωτικής συμπεριφοράς βασίζονται στην ανάλυση συσχετίσεων. Τα δεδομένα από την ανάλυση καλαθιού αγοράς στην πιο πρωτόγονη μορφή τους θα μπορούσαν να είναι μια λίστα συναλλαγής από τις αγορές των καταναλωτών αναφέροντας μόνο τα αντικείμενα που είναι μαζί και αναφέροντας τις τιμές τους.

Στόχος της ανάλυσης καλαθιού αγοράς είναι να βρει ποια αγαθά αγοράζονται μαζί, ώστε να τοποθετηθούν σε ένα συγκεκριμένο σημείο. Αυτό γίνεται, για να προωθηθούν περισσότερο τα προϊόντα σε σχέση με άλλα αγαθά. Για παράδειγμα ένας καταναλωτής θα αγοράσει καφέ, όμως μαζί με τον καφέ μπορεί να αγοράσει ζάχαρη, νερό ή και φίλτρα του καφέ για την καφετιέρα. Η ζάχαρη θα βρίσκεται δίπλα ακριβώς από τον καφέ ή στην περίπτωση των eshops (online καταστημάτων) θα βρίσκεται στα σχετικά ή προτεινόμενα προϊόντα, όπως όλοι θα έχουμε δει. Ο συγκεκριμένος διαχωρισμός και πρόταση προς τους χρήστες γίνεται για λόγους ευκολίας, αλλά και για λόγους τακτικής μάρκετινγκ. Τα δεδομένα για το κάθε προϊόν θα μπορούσαν να είναι μια λίστα για το τι έχουν αγοράσει οι πελάτες ή μια βάση δεδομένων με όλα τα χαρακτηριστικά της κάθε συναλλαγής. Δηλαδή, ημερομηνία, ώρα κλπ.

Η ανάλυση καλαθιού αγοράς στοχεύει στην παροχή εικόνας για συγγένειες προϊόντων. Φαίνονται σημαντικές πληροφορίες για τη διαφήμιση, για την αλλαγή στα ράφια ή τους καταλόγους, καθώς και για την ανάπτυξη ειδικών προσφορών προϊόντων ή υπηρεσιών. Επίσης στην παροχή προτάσεων του προϊόντος σύμφωνα με την συμπεριφορά του πελάτη. Με βάση λοιπόν τους κανόνες συσχέτισης μπορούμε να βρούμε τι περιέχει ένα καλάθι αγοράς και ανάλογα με το τι θέλει να επιτύχει η κάθε επιχείρηση να καταστρώσει μια νέα στρατηγική. Στην επόμενη ενότητα θα αναφερθούμε σε παραδείγματα κανόνων συσχέτισης.

2.4.2 Ορισμός Προβλήματος

Το κάθε πρόβλημα, το οποίο θέλουμε να επιλύσουμε με την τεχνική των κανόνων συσχέτισης αντιμετωπίζεται σαν δύο υπό-προβλήματα. Το ένα υπό-πρόβλημα είναι η εύρεση όλων των συχνών στοιχειοσυνόλων (Frequent Item sets) και το δεύτερο υπό-πρόβλημα είναι η δημιουργία κανόνων συσχέτισης. Ένας κανόνας συσχέτισης έχει δύο μέρη, την συνθήκη, η οποία καθορίζει το αποτέλεσμα, και το αποτέλεσμα. Ένα απλό παράδειγμα είναι:

{καφέ}→{ζάχαρη}

Αυτό ερμηνεύεται ως εξής: εάν ένας πελάτης αγοράσει πινέλα, κατ' επέκταση θα αγοράσει και χρώματα.

Ένα πιο ολοκληρωμένο παράδειγμα για τον ορισμό ενός προβλήματος κανόνων συσχέτισης ακολουθεί παρακάτω.

TID	ΕΜΠΟΡΕΥΜΑΤΑ
101	Ψωμί, Αυγά, Φυστικοβούτυρο
102	Ψωμί, Φυστικοβούτυρο
103	Ψωμί, Γάλα, Φυστικοβούτυρο
104	Μπύρα, Ψωμί
105	Μπύρα, Γάλα

Πίνακας 2.3: Συναλλαγές - Εμπορεύματα

Επομένως διαθέτουμε το στοιχειοσύνολο I ={Ψωμί, Γάλα, Αυγά, Φυστικοβούτυρο, Μπύρα}

Ακόμη, θεωρούμε τον κανόνα Ψωμί → Φυστικοβούτυρο, ο οποίος σημαίνει ότι όταν κάποιος αγοράζει το Ψωμί, τότε αγοράζει και το Φυστικοβούτυρο. Παρατηρούμε ότι σε τρεις από τις συνολικά πέντε συναλλαγές (101, 102, 103) πωλούνται ταυτόχρονα τα προϊόντα Ψωμί και Φυστικοβούτυρο. Η υποστήριξη του κανόνα είναι 3/5, δηλαδή 60%. Επίσης παρατηρούμε ότι το προϊόν Ψωμί εμφανίζεται σε τέσσερεις συναλλαγές (101, 102,103,104) και ότι σε τρεις από αυτές (101, 102, 103) εμφανίζεται και το προϊόν Φυστικοβούτυρο. Η εμπιστοσύνη του κανόνα είναι 3/4, δηλαδή 75%.

Για την ανακάλυψη Κανόνων Συσχέτισης, ο χρήστης προκαθορίζει ελάχιστες τιμές για την υποστήριξη και την εμπιστοσύνη. Στη συνέχεια, ο αλγόριθμος διατρέχει τη βάση δεδομένων, αναλύει τα δεδομένα και εντοπίζει όλους τους κανόνες που έχουν υποστήριξη και εμπιστοσύνη ίση ή μεγαλύτερη από τις προκαθορισμένες τιμές. Οι κανόνες αυτοί θεωρούνται ισχυροί.

Ορισμένοι πρόσθετοι όροι των Κανόνων Συσχέτισης είναι οι ακόλουθοι:

- **Στοιχειοσύνολο** (Itemset) ονομάζεται ένα σύνολο από στοιχεία (items). Στο παράδειγμα μας ένα σύνολο I, που περιέχει τα εμπορεύματα Ψωμί, Γάλα, Μπύρα (I={ Ψωμί, Γάλα, Μπύρα }), είναι ένα στοιχειοσύνολο.
- **k-Στοιχειοσύνολο** (k-Itemset) είναι ένα στοιχειοσύνολο που περιέχει k στοιχεία. Το στοιχειοσύνολο I={ Ψωμί, Γάλα, Μπύρα } είναι ένα 3-Στοιχειοσύνολο.
- **Συχνότητα Εμφάνισης** (frequency ή support count ή count) ενός στοιχειοσυνόλου είναι το πλήθος των συναλλαγών που περιέχουν το στοιχειοσύνολο. Η συχνότητα εμφάνισης του στοιχειοσυνόλου {Ψωμί,Φυστικοβούτυρο} είναι ίση με 3.

- **Υποστήριξη (support)** ενός στοιχειοσυνόλου είναι το ποσοστό των συναλλαγών που περιέχουν το στοιχειοσύνολο. Η υποστήριξη του στοιχειοσυνόλου {Ψωμί, Γάλα} είναι ίση με 1/5=20%

2.4.3 Εντοπισμός συχνών στοιχειοσυνόλων - Αλγόριθμος Apriori

Ο Apriori αλγόριθμος εκτελείται σε δύο στάδια:

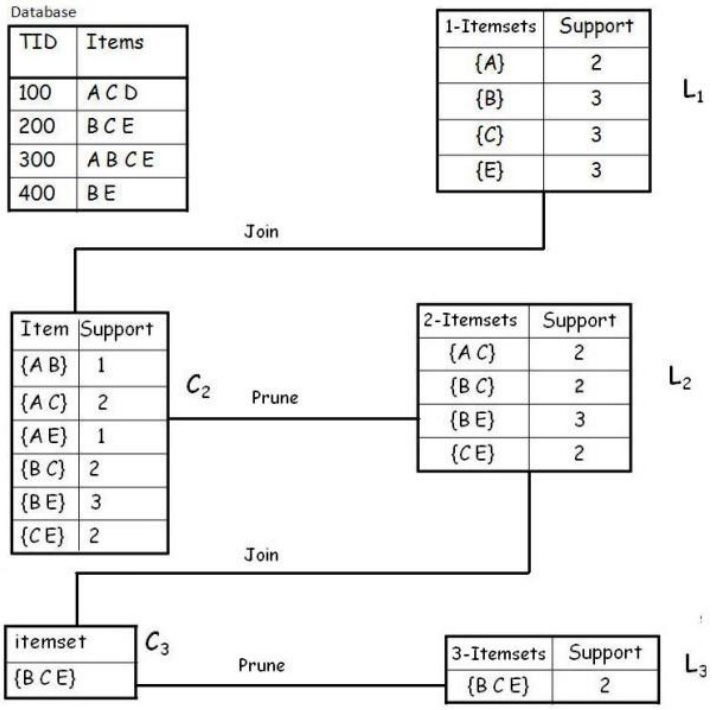
1. **Στο πρώτο στάδιο**, γίνονται οι υπολογισμοί για να επιλεγούν τα itemsets που εμφανίζονται με τη μεγαλύτερη συχνότητα. Ο αλγόριθμος, λοιπόν, υπολογίζει το support όλων των items, του καθενός ξεχωριστά. Τα items που έχουν support μεγαλύτερο ή ίσο από το ελάχιστο όριο υποστήριξης γίνονται δεκτά και συγκροτούν το σύνολο L1. Στη συνέχεια, παράγονται όλα τα δυνατά ζευγάρια των στοιχείων του συνόλου L1, δηλαδή συγκροτούνται δυάδες από items. Για κάθε ζευγάρι υπολογίζεται ξανά το support και όσα από τα ζευγάρια γίνονται δεκτά συγκροτούν το σύνολο L2. Κατόπιν, παράγονται όλα τα δυνατά ζευγάρια των στοιχείων του συνόλου L2, δηλαδή τριάδες από items. Τελικά, ο αλγόριθμος συνεχίζει να παράγει n-άδες από items, έως ότου η τιμή του n να γίνει ίση με την τιμή μιας παραμέτρου για το μέγιστο μέγεθος ενός itemset που έχουμε ορίσει.

2. **Στο δεύτερο στάδιο**, ο Apriori δημιουργεί τους κανόνες συσχέτισης. Από το τελευταίο σύνολο L που προκύπτει, ελέγχεται το confidence όλων των δυνατών κανόνων συσχέτισης που μπορεί να προκύψουν. Οι κανόνες που έχουν εμπιστοσύνη μεγαλύτερη από την ελάχιστη εμπιστοσύνη που έχει προσδιοριστεί γίνονται τελικά αποδεκτοί. Τέλος, οι κανόνες αυτοί ελέγχονται και ως προς τη σημαντικότητα τους.

Τα παραπάνω στάδια φαίνονται χαρακτηριστικά στην παρακάτω εικόνα, όπου ο συγκεκριμένος αλγόριθμος διαβάζει την αρχική μας βάση-πίνακα D διαδοχικές φορές. Η παραπάνω πρόταση για τον

αλγόριθμο Apriori έχει οριστεί από τους Μ. Χαλκίδης, Μ. Βαζιργιάννης (2005)[13].

Συνολικά η βάση D θα διαβαστεί -το πολύ- τόσες φορές όσες είναι το πλήθος των διαφορετικών items στον πίνακα. Στο πρώτο διάβασμα (πέρασμα) του πίνακα μετριέται η υποστήριξη των 1-itemsets και βρίσκεται ποια από αυτά ικανοποιούν την απαίτηση για ελάχιστη υποστήριξη. Σε κάθε επόμενο βήμα χρησιμοποιούνται τα itemsets του προηγούμενου περάσματος για να δημιουργηθούν καινούργια itemsets. Τα itemsets αυτά ονομάζονται υποψήφια (candidate itemsets) καθώς δεν γνωρίζουμε την υποστήριξή τους και κατ' επέκταση αν είναι συχνά (frequent). Για το λόγο αυτόν μετριέται η υποστήριξή τους μέσω ενός περάσματος από τον αρχικό πίνακα. Το κλειδί σε όλη αυτή τη διαδικασία είναι ότι σε κάθε βήμα γίνεται ακριβώς ένα μόνο πέρασμα από τον αρχικό πίνακα. Στο τέλος του κάθε βήματος αποφασίζεται ποια itemsets είναι συχνά ώστε να χρησιμοποιηθούν για το επόμενο βήμα. Αυτός είναι περιγραφικά ο τρόπος με τον οποίο ο αλγόριθμος Apriori παράγει τα frequent itemsets.

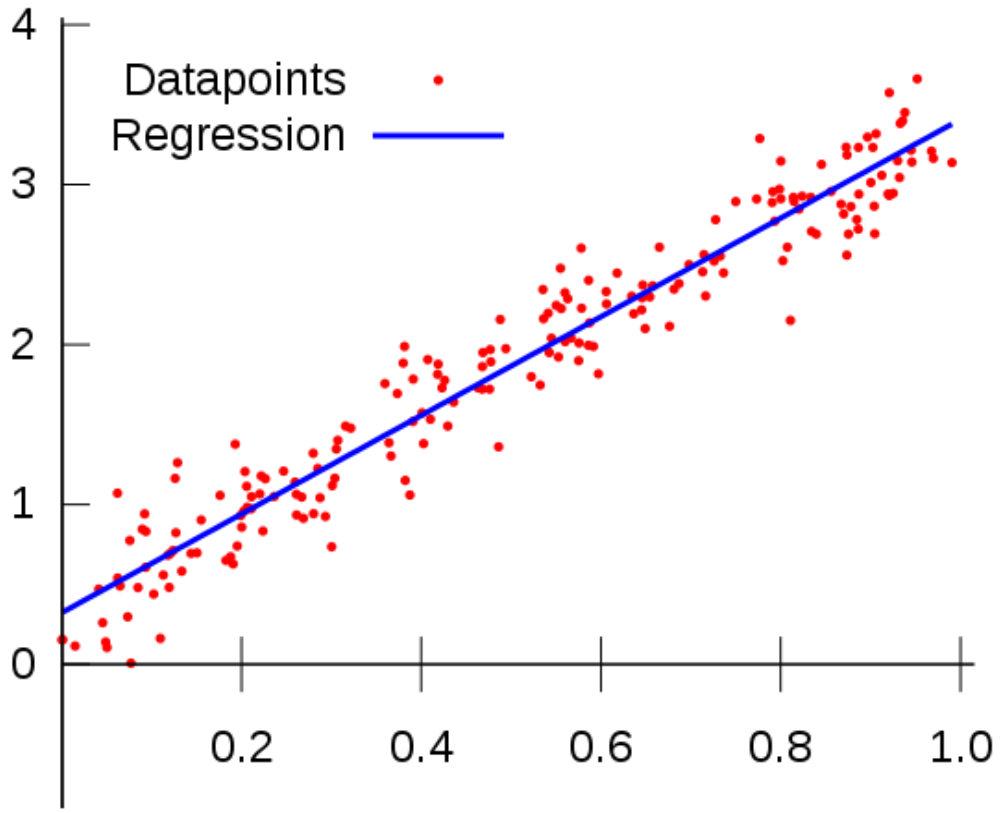


Εικόνα 2.8: Αλγόριθμος Apriori

[13] **Μιχάλης Βαζιργιάννης, Μαρία Χαλκίδη**, Εξόρυξη Γνώσης από Βάσεις Δεδομένων και τον Παγκόσμιο Ιστό, 2η έκδοση, 2005

2.5 Παλινδρόμηση

Η παλινδρόμηση είναι μια σχετική διαδικασία με την κατηγοριοποίηση. Στόχος της είναι η μάθηση ή αλλιώς η εκπαίδευση (training) μιας συνάρτησης, η οποία απεικονίζει ένα αντικείμενο σε μία πραγματική μεταβλητή. Πρόκειται για μια, επίσης, προγνωστική μέθοδο. Στόχος είναι με βάση κάποιες ανεξάρτητες μεταβλητές (independent variables) να προβλεφθούν οι τιμές μιας εξαρτημένης μεταβλητής (dependent variable).



Εικόνα 2.9: Παράδειγμα Γραμμικής Παλινδρόμησης

Στην παραπάνω εικόνα παρουσιάζουμε ένα απλό παράδειγμα γραμμικής παλινδρόμησης. Οι μεταβλητές είναι ο άξονας των X που αναπαριστά τα τετραγωνικά στρέμματα ενός σπιτιού και ο άξονας των Y που αναπαριστά την τιμή πώλησης σε εκατοντάδες χιλιάδες Ευρώ (τιμή * 100000). Η γραμμική παλινδρόμηση

προσαρμόζει μια ευθεία στα δείγματα του συνόλου δεδομένων, τα οποία σηματοδοτούνται με κόκκινο pointmark. Η προσαρμογή γίνεται με βάση μια συνάρτηση απόστασης ή συνάρτηση κόστους, την τιμή της οποία θέλουμε να ελαχιστοποιήσουμε. Έχοντας τη βέλτιστη ευθεία, δηλαδή την ευθεία που ελαχιστοποιεί την τιμή της συνάρτησης κόστους, μπορούμε να δώσουμε μια προσεγγιστικά καλή απάντηση σε ερωτήματα της μορφής: «Σε τι τιμές πωλούνται σπίτια των 0.15 τετραγωνικών στρεμμάτων;».

Κεφάλαιο 3

3.1 Ψάχνοντας πληροφορίες στον Παγκόσμιο Ιστό

Σε έναν κόσμο όπου όλα κινούνται γρήγορα, που ο χρόνος σε συνδυασμό με τη γνώση αποτελεί σημαντικό παράγοντα της απόδοσης, η γρήγορη και στοχευόμενη πληροφορία αποτελεί σημείο αναφοράς για την σωστή και ομαλή λειτουργία της καθημερινότητας. Η ανάγκη λοιπόν για προβολή συσσωρευμένης πληροφορίας με τρόπο ώστε να αναδεικνύονται τα σημαντικά τμήματα αυτής και να είναι ταχύτερη η επεξεργασία από τους ενδιαφερόμενους, είναι και το κλειδί για την κάλυψη του κόστους του χρόνου.

Η μηχανή αναζήτησης της Google αποτελεί στη σύγχρονη εποχή το μεγαλύτερο και πιο εμφανές παράδειγμα συλλογής, επεξεργασίας και παρουσίασης δεδομένων στον Παγκόσμιο Ιστό και όχι μόνο. Ενώ ο συνολικός όγκος των δεδομένων που πρέπει να επεξεργαστεί η συγκεκριμένη μηχανή σε κάθε αναζήτηση που πραγματοποιεί είναι τεράστιος, εντούτοις τα αποτελέσματα που παράγονται και η παρουσίαση αυτών δεν ξεπερνούν χρονικά τα δύο δευτερόλεπτα. Μέσα από τη μηχανή αναζήτησης της Google γίνεται εύκολα αντιληπτή η επιτυχία της εφαρμογής της διαδικασίας εξόρυξης δεδομένων στο διαδίκτυο, καθώς ο τελικός χρήστης λαμβάνει εύκολα και γρήγορα μόνο την προς αναζήτηση πληροφορία.

Η τεράστια και συνεχώς αυξανόμενη ποσότητα και διαθεσιμότητα πληροφοριών, που παρέχεται μέσω του Παγκόσμιου Ιστού (World Wide Web), έχει διαφοροποιήσει αρκετά τη ζωή των ανθρώπων και συνάμα τους έχει βοηθήσει όσον αφορά την ακρίβεια στην αναζήτηση πληροφοριών. Τις τελευταίες δεκαετίες, μια πληθώρα εμπορικών μηχανών αναζήτησης έχουν αναδυθεί και παρέχουν πλέον όλα τα απαραίτητα εργαλεία και εφόδια ώστε να είναι σχετικά απλή και αρκετά γρήγορη για τους χρήστες του Παγκόσμιου Ιστού η αναζήτηση πληροφοριών. Παρόλα ταύτα, η πληθώρα των πληροφοριών αυτών, έχει οδηγήσει στην ανάγκη επέκτασης των μηχανών αναζήτησης με σκοπό τη δημιουργία εξατομικευμένων και διαφορετικών για κάθε χρήστη πλαισίων

καθώς και την ανάπτυξη νέων τεχνικών και μεθοδολογιών με σκοπό την αποτελεσματική επεξεργασία των πληροφοριών.

Οι μηχανές αναζήτησης είναι ένα ανεκτίμητο εργαλείο για την ανάκτηση πληροφοριών από το διαδίκτυο. Απαντώντας στα ερωτήματα του χρήστη, επιστρέφουν μια λίστα με αποτελέσματα, ταξινομημένα κατά σειρά, με βάση τη συνάφεια του περιεχομένου τους προς το ερώτημα.

Το World Wide Web ή “Web” για συντομία, είναι μια τεράστια συλλογή από ψηφιακές σελίδες. Ένα μεγάλο υποσύνολο του λογισμικού του Διαδικτύου είναι αφιερωμένο στο περιεχόμενο εκπομπών, με τη μορφή σελίδων της HTML. Η ιστοσελίδα προβάλλεται με τη χρήση δωρεάν λογισμικού που ονομάζεται web browser. Γεννημένος το 1989, ο Παγκόσμιος Ιστός ή World Wide Web ή “Web” έχει χτιστεί χρησιμοποιώντας σαν βάση το πρωτόκολλο μεταφοράς υπερκειμένου (hypertext transfer protocol), μια γλώσσα που επιτρέπει σε όλους μας την μετάβαση μέσω hyperlinks (υπερ-σύνδεση) σε οποιαδήποτε δημόσια σελίδα του web. Σήμερα υπάρχουν περίπου γύρω στις 70 δισεκατομμύρια δημόσιες ιστοσελίδες στο Web.

3.2 Μηχανές αναζήτηση<u>ς</u>

3.2.1 Ιστορική αναδρομή

Ύστερα από την πρώτη σύλληψη της ιδέας του Διαδικτύου, από τον Αμερικανικό στρατό ως τρόπος ασφαλής και σταθερής επικοινωνίας και την υλοποίηση του από τον Timothy John Berners-Lee για το ερευνητικό ίδρυμα Cern, έως σήμερα, ο γνωστός σε όλους «Παγκόσμιος Ιστός» ή Internet, έχει εξελιχθεί σε ένα σύστημα όχι μόνο διηπειρωτικής επικοινωνίας, αλλά και παγκόσμιας «αποθήκης» δεδομένων.

Από το 1989 με τη χρήση του Διαδικτύου από το Cern για την παρουσίαση αποτελεσμάτων σε πολλούς χρήστες, καθώς η αποστολή μηνυμάτων ηλεκτρονικού ταχυδρομείου, που αποτελούσε τότε τρόπο επικοινωνίας και ενημέρωσης αρχικά των επιστημόνων και των ενδιαφερόμενων του ιδρύματος, αποδείχτηκε

αποτελεσματικότερος και απλούστερος. Όπως ήταν λογικό όμως, ο όγκος των δεδομένων άρχισε να αυξάνει με ταχύς ρυθμούς.

Η αύξηση του όγκου των δεδομένων, σήμαινε ταυτόχρονα και αύξηση του χρόνου αναζήτησης των πληροφοριών που αναζητούσε ο χρήστης. Η ιδέα, όπως φάνηκε από νωρίς, ήταν απλή. Θα έπρεπε ο όγκος των δεδομένων αυτών να είναι εύκολα προσβάσιμος και στοχευμένος για τους ενδιαφερόμενους. Άλλωστε, η «πολύ» πληροφορία χωρίς να είναι «σωστή» πληροφορία, παύει να είναι πληροφορία.

Για κάθε πρόβλημα όμως, συνήθως υπάρχει και η αντίστοιχη λύση. Η έννοια της αναζήτησης τότε δεν ήταν ξένη, αλλά όχι και απόλυτα ξεκάθαρη. Όταν ψάχνουμε να βρούμε έναν τηλεφωνικό αριθμό, συνήθως χρησιμοποιούμε τον τηλεφωνικό μας κατάλογο, καθώς είναι αδύνατο να θυμόμαστε όλους τους αριθμούς από μνήμη. Ο τηλεφωνικός κατάλογος είναι αντίστοιχα η λειτουργία που υλοποιεί μια μηχανή αναζήτησης στο διαδίκτυο. Ο τηλεφωνικός κατάλογος έχει δική του δομή όπου καταγράφονται τα τηλέφωνα και μπορούμε ταχύτερα να βρούμε το νούμερο που επιζητούμε. Έτσι και μια μηχανή αναζήτησης, έχει «καταγράψει» τα δεδομένα για τις σελίδες του διαδικτύου, οπότε μπορεί γρήγορα να βρει την σελίδα ή την πληροφορία που ψάχνουμε. Επίσης στους τηλεφωνικούς καταλόγους, συνήθως, έχει ανά σελίδα και τα γράμματα της αλφαβήτου, οπότε οι διαδικασίες εύρεσης αποτελέσματος μειώνουν τον χρόνο αναζήτησης. Κάτι αντίστοιχο με αυτή τη λειτουργία, υλοποιεί και μια μηχανή αναζήτησης βελτιώνοντας τους χρόνους απόκρισης των αποτελεσμάτων.

Η έννοια και λειτουργία της μηχανής αναζήτησης θα αποτελούσε σίγουρα αναπόσπαστο κομμάτι της λειτουργίας του διαδικτύου γι αυτό και αρκετά νωρίς έγινε και η παρουσίαση της πρώτης μηχανής αναζήτησης. Ενδεικτικά παρακάτω αναφέρονται κάποιες μηχανές αναζήτησης που αποτέλεσαν τον πρόδρομο για τις σημερινές σύγχρονες μηχανές.

- ✓ Archie

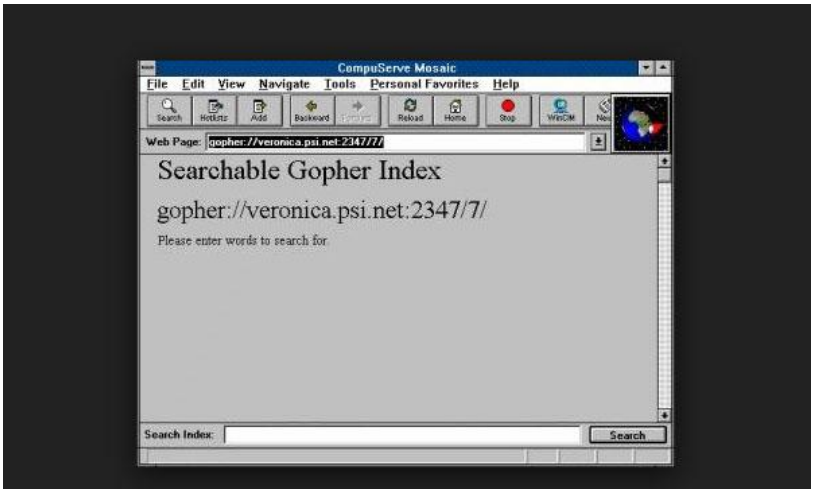
Στα τέλη του 1990 έγινε η παρουσίαση της πρώτης μηχανής αναζήτησης από τους Peter Deutsch, Alan Emtage και Bill Heelan.

Ήταν ουσιαστικά ένα σύστημα καταγραφής διακομιστών FTP και των αρχείων που περιλάμβαναν.

Εικόνα 3.1: Archie, μηχανή αναζήτησης

- ✓ Veronica

Το 1992 οι Steven Foster και Fred Barrie από το πανεπιστήμιο της Νεβάδα, παρουσιάζουν την μηχανή αναζήτησης Veronica (Very Easy Rodent-Oriented Net-wide Index to Computer Archives) βασισμένη στο πρωτόκολλο Gopher σχεδιασμένο για αναζήτηση, αποστολή και μετάδοση αρχείων μέσω του διαδικτύου.



Εικόνα 3.2: Veronica, μηχανή αναζήτησης

- ✓ Jughead

Το 1993 παρουσιάζεται το Jughead από τον Rhett Jones, πανεπιστήμιο της Γιούτα. Χρησιμοποιεί επίσης το πρωτόκολλο

Gopher με την διαφορά ότι κάνει αναζήτηση δεδομένων στον διακομιστή σε πραγματικό χρόνο.

Εικόνα 3.3: Jughead, Μηχανή Αναζήτησης

✓ Excite

Τέλη του 1995 και φαίνεται ότι η μηχανή αναζήτησης στο διαδίκτυο δεν αποτελεί απλά μια μηχανή με συγκεκριμένη λειτουργία. Το Excite, δημιούργημα φοιτητών από το πανεπιστήμιο του Στάνφορντ, αποτελεί σημείο ορισμού για τις μοντέρνες μηχανές. Αποτελεί μία διαδικτυακή πύλη με μηχανή αναζήτησης, προσφέροντας στους χρήστες λειτουργίες ενημέρωσης ειδήσεων, καιρού, ηλεκτρονικό ταχυδρομείο, IM, προσωπική σελίδα χρήστη κλπ. Ίσως ένα προάγγελος της Google.

Εικόνα 3.4: Excite, μηχανή αναζήτησης

✓ Lycos

Το 1994 παρουσιάζεται το Lycos. Η μηχανή αναζήτησης προσέφερε λειτουργίες ηλεκτρονικού ταχυδρομείου, κοινωνικής δικτύωσης και σελίδες διακσέδασης.

Εικόνα 3.5: Lycos, μηχανή αναζήτησης

- ✓ Altavista

Το 1996 παρουσιάζεται η μηχανή Altavista όπου αποτέλεσε και βασικό πάροχο αναζήτησης της Yahoo.

Εικόνα 3.6: Altavista, μηχανή αναζήτησης

3.2.2 Γνωστές σύγχρονες μηχανές αναζήτησης

Από τους προδρόμους των μηχανών αναζήτησης έως σήμερα, το διαδίκτυο αποτελεί αναπόσπαστο κομμάτι της καθημερινής ζωής και ο βασικός του συνοδοιπόρος οι μηχανές αναζήτησης. Το 1998

δύο φοιτητές οι Larry Page και Sergey Brin του πανεπιστημίου Στάνφορντ, παρουσίασαν την καινούργια μηχανή αναζήτησης που σήμερα κατέχει το μεγαλύτερο ποσοστό αναζήτησης στο διαδίκτυο, την Google. Οι δύο φοιτητές εφάρμοσαν στην αναζήτηση τους ένα νέο σύστημα αξιολόγησης σελίδων, όπου σύντομα θα αποτελούσε εμφανή καινοτομία στην ανάπτυξη αντίστοιχων εφαρμογών καθώς και εξέλιξης του ίδιου του διαδικτύου. Μάλιστα, στη δική μας εποχή, το διαδίκτυο και οι ανάπτυξη των εφαρμογών του, είναι πλέον αυτό που παραμετροποιείται ώστε να λειτουργήσει αρμονικά με τις διαθέσιμες μηχανές αναζήτησης.

Πώς άλλωστε θα μπορούσε να γίνει αλλιώς καθώς το διαδίκτυο αποτελεί πλέον χώρο πληροφόρησης, εργασίας, διασκέδασης και πολλά άλλα! Παρακάτω παρουσιάζονται σύγχρονες μηχανές αναζήτησης που χρησιμοποιούνται στον κόσμο από εκατομμύρια χρήστες.

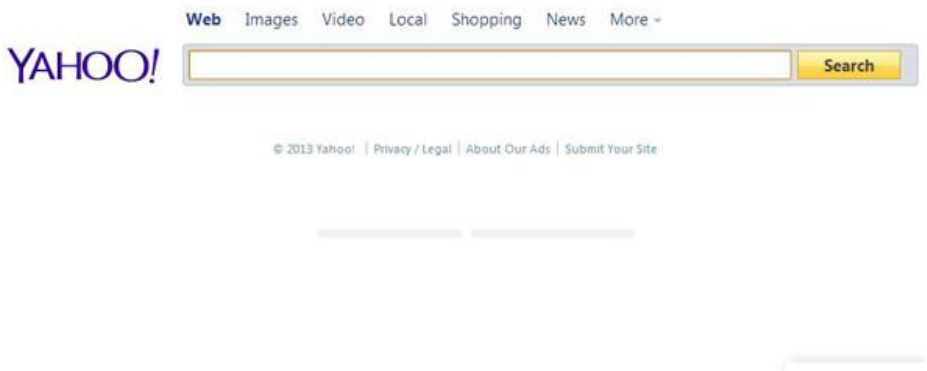
- ✓ Google

Το 1998 οι δύο φοιτητές από το Στάνφορντ παρουσιάζουν το Google. Η καινούργια αυτή μηχανή, βασισμένη σε καινοτόμο σύστημα αξιολόγησης των σελίδων σύμφωνα με το περιεχόμενο τους, θα αποτελέσει την διασημότερη μηχανή αναζήτησης μέχρι και σήμερα. Αρχικά βασισμένη σε αλγορίθμους consumer-producer κι έπειτα mapreduce, εξασφαλίζει όχι μόνο την απόδοση και την σωστή αξιολόγηση των διαδικτυακών σελίδων, αλλά και ταχύτητα ανανέωσης των περιεχομένων της καθώς και της απόκρισης αποτελεσμάτων. Επιπλέον, εισάγει στο διαδίκτυο την έννοια του SEO(Search engine optimization), που αναφέρεται σε βέλτιστες τεχνικές ανάπτυξης της σελίδας για την υποβοήθηση της επεξεργασίας της, από την μηχανή αναζήτησης.

Εικόνα 3.7: Η No1 σύγχρονη μηχανή αναζήτησης της Google

- ✓ Yahoo

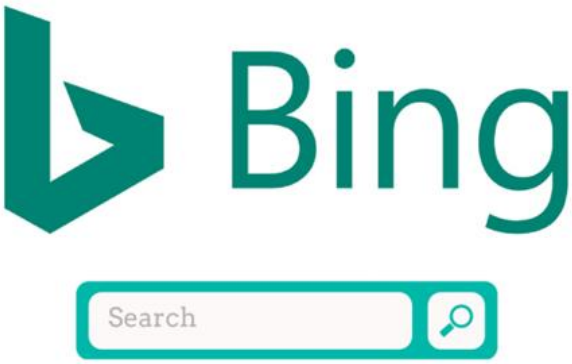
Το 1995 παρουσιάζεται η μηχανή Yahoo(Yet Another Hierarchical Officious Oracle). Αρχικά υλοποιημένη το 1994 από τους φοιτητές του Στάνφορντ, Γιανγκ και Φίλο με ονομασία «Jerry's guide to the World Wide Web», αποτελούσε μηχανή για την ιεραρχική προβολή των διαδικτυακών σελίδων. Το 2000 η Yahoo ξεκινάει να χρησιμοποιεί την Google για την αναζήτηση στον Ιστό, ενώ τέσσερα χρόνια αργότερα εισάγει την δική της μηχανή αναζήτησης.



Εικόνα 3.8: Μηχανή Αναζήτησης Yahoo

- ✓ Bing

Η μηχανή αναζήτησης Bing παρουσιάστηκε από την Microsoft το 2009 καθώς ανακοινώνουν με την Yahoo την κοινή τους συνεργασία. Η μηχανή Bing αντικαθιστά την μηχανή της Yahoo αφού πρώτα έχει γίνει γνώστη στον διαδικτυακό χώρο ως Live Search(2007), Windows Live Search(2006), και MSN Search(1998). Βασικός πλέον ανταγωνιστής της Google, η μηχανή Bing ανακοινώνει την καινούργια υποδομή αναζήτησης και τεχνολογία αρχειοθέτησης το 2011.



Εικόνα 3.9: Μηχανή Αναζήτησης Bing

- ✓ Yandex

Η μηχανή αναζήτησης Yandex παρουσιάζεται το 2010 από την ομώνυμη Ρώσικη εταιρεία και κατέχει το μεγαλύτερο ποσοστό αναζήτησης χρηστών στην Ρωσία.

Εικόνα 3.10: Μηχανή Αναζήτησης Yandex

✓ Baidu

Το 2000 η ιδρύεται η εταιρία Baidu με έδρα την Κίνα από τους Ρόμπιν Λι και Έρικ Χιου. Η ομώνυμη μηχανή αναζήτησης το 2010 καταλαμβάνει τις μισές αναζητήσεις στο διαδίκτυο στην Κίνα ενώ από το 2011 κρατάει σταθερά την πρώτη θέση στην αντίστοιχη χώρα με το μεγαλύτερο μέρισμα του ποσοστού.

Εικόνα 3.11: Μηχανή Αναζήτησης Baidu

✓ Duck Duck Go

Τον Ιούλιο του 2010 παρουσιάζεται η ιδέα της μηχανής DuckDuckGo αρχικά σαν κοινότητα όπου οι χρήστες παρουσίαζαν τα προβλήματα τους σχετικά με μηχανές αναζήτησης, επιπλέον λειτουργίες κ.α. Η μηχανή αναζήτησης χρησιμοποιεί πήγες από άλλες μηχανές αναζήτησης όπως Yahoo, Bing, Yandex και

WolframAlpha καθώς και δικά της δεδομένα τα οποία παρουσιάζονται στους τελικούς χρήστες.

Εικόνα 3.12: Μηχανή Αναζήτησης Duck Duck Go

3.3 Τρόπος λειτουργίας των μηχανών αναζήτησης

Παρόλο το μεγάλο πλήθος των μηχανών αναζήτησης, μιας και αυξάνονται πλέον όχι μόνο για γενικούς αλλά και ειδικούς σκοπούς, η βασική αρχή της λειτουργίας τους παραμένει ίδια. Μια μηχανή αναζήτησης βασίζει την ταχύτητά απόκρισης της αλλά και την ορθότητα των αποτελεσμάτων της στα ήδη προ αποθηκευμένα δεδομένα του διαδικτύου. Άλλωστε, η αναζήτηση σε πραγματικό χρόνο των δεδομένων θα καθιστούσε ανυπόφορη την λειτουργία τους.

Πέραν λοιπόν των διαφορετικών αλγορίθμων για την εύρεση, εξαγωγή και αποθήκευση των διαδικτυακών σελίδων, μπορούμε να επικεντρωθούμε σε τρία βασικά στάδια λειτουργίας για την

α. περισυλλογή τους (Crawling),

β. επεξεργασία και αποθήκευση(Indexing) και

γ. προβολή των αποτελεσμάτων.

3.3.1 Web Crawling

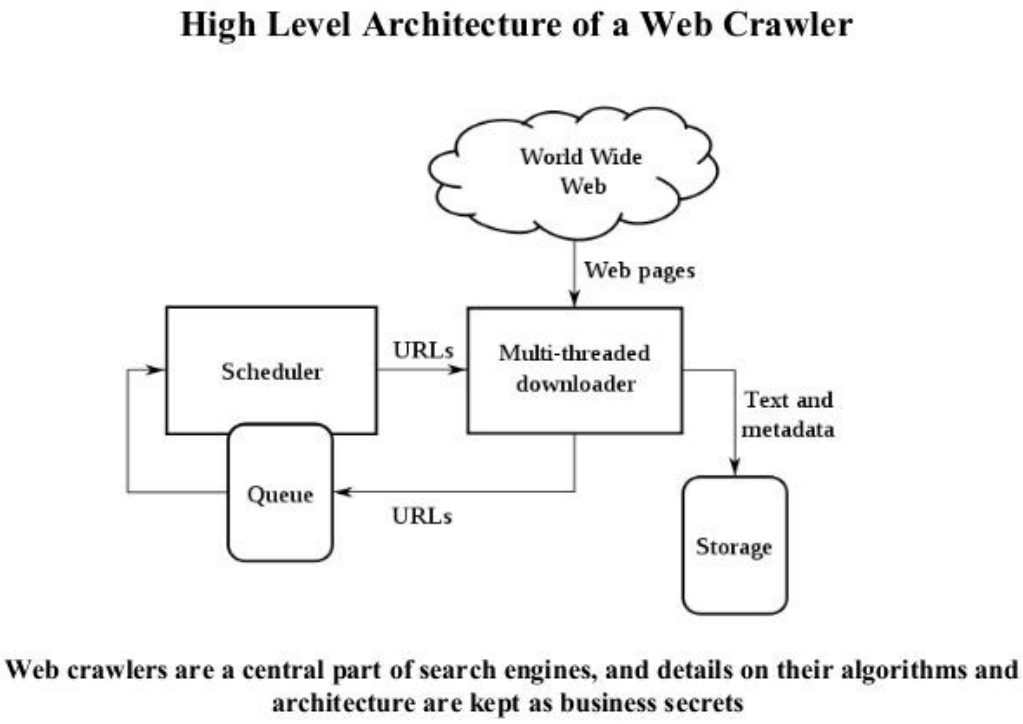
Ο Web Crawler ή Web Spider αναφέρεται στην διαδικασία περισυλλογής των δεδομένων. Αποτελεί στην πραγματικότητα μια εφαρμογή, ή αλλιώς όπως είναι γνωστό «Internet bot», το οποίο ακολουθώντας εσωτερικές διαδικασίες επισκέπτεται διαδικτυακές διευθύνσεις ανά τακτικό ή μη, χρονικό διάστημα, για την εξαγωγή δεδομένων.

Η αρχή λειτουργίας ενός Web Crawler, βασίζεται στην αρχική τροφοδοσία του με διαδικτυακές διευθύνσεις. Με μεθόδους κλήσης HTTP GET, μπορεί να ζητήσει από τον εκάστοτε εξυπηρετητή τον HTML κώδικα, από τον οποίο μπορεί να εξάγει περεταίρω διευθύνσεις που είτε αφορούν την ίδια διαδικτυακή εφαρμογή είτε είναι διευθύνσεις που παραπέμπουν σε άλλες εφαρμογές. Το σημαντικό πρόβλημα σε αυτή την λειτουργία, είναι ο κίνδυνος ο Crawler να πέσει σε ατέρμονα βρόγχο και να εγκλωβιστεί μέσω των ανακατευθύνσεων στην ίδια εφαρμογή, ζητώντας από τον εξυπηρετητή τις ίδιες σελίδες. Για την επίλυση αυτού του προβλήματος, χρησιμοποιούνται τεχνικές κανονικοποίησης των κλήσεων στις διευθύνσεις. Τέτοιες τεχνικές περιλαμβάνουν την απαλοιφή σημείων στίξης, την μετατροπή των γραμμάτων σε κεφαλαία ή μόνο πεζά, δομή της διεύθυνσης με συμπλήρωση των πεδίων όταν αυτά λείπουν, π.χ. το http://example.gr/ θα μετατραπεί σε http://www.example.gr. Έτσι λοιπόν εξασφαλίζουμε ότι ο Crawler δεν θα εγκλωβιστεί στη ίδια εφαρμογή και δεν θα ζητήσει την ίδια σελίδα από τον εξυπηρετητή.

- ✓ Η Αρχιτεκτονική

Η σωστή λειτουργία ενός Web Cawler βασίζεται τόσο στην απόδοση για την σωστή ζήτηση των σελίδων από τον εξυπηρετητή όσο και στην ταχύτητα επεξεργασίας των δεδομένων αυτών. Είναι άλλωστε γνωστή στρατηγική, η κάθε εταιρία να διατηρεί μέρος ή και ολόκληρο τον αλγόριθμο λειτουργίας κρυφό ώστε να αποφθεχθούν περιπτώσεις αντιγραφής. Για την επιτυχία στην ταχύτητα ζήτησης και επεξεργασίας δεδομένων, συχνά χρησιμοποιείται παραλληλισμός ως προς την λειτουργία του

ρομπότ. Όπως παρουσιάζεται στην παρακάτω εικόνα η λειτουργία έχει 3 βήματα, πρώτον το ρομπότ ζητάει μια σελίδα από τον εξυπηρετητή, δεύτερον περνάει από επεξεργασία και τυχόν καινούργιες διευθύνσεις μπαίνουν στην ουρά και τέλος επιλέγει προσεκτικά την επόμενη κλήση στον εξυπηρετητή.



Εικόνα 3.13: Αρχιτεκτονική Υψηλού Επιπέδου ενός Web Crawler (Πηγή: http://www.slideshare.net/eibeed/the-development-of-web-archiving-3)

Για την αποδοτικότερη λειτουργία του, ο Web Crawler θα πρέπει να λειτουργεί με συγκεκριμένη στρατηγική ώστε να πετύχει σωστά αποτελέσματα. Η επαναλαμβανόμενη ζήτηση ίδιων σελίδων αλλά και μία απλή αρχιτεκτονική για προσπέλαση μερικών δεκάδων σελίδων το λεπτό δεν αποτελεί βέλτιστη λειτουργία. Θα πρέπει λοιπόν ο Crawler μέσα από μία λίστα από διαθέσιμες επισκέψεις να διαλέξει την κατάλληλη η οποία και δεν έχει ξαναζητηθεί αλλά και θα αποφέρει τα αναμενόμενα αποτελέσματα.

Μοντέρνες τεχνικές πλέον προσπαθούν κα κατανοήσουν το περιεχόμενο της σελίδας από την ίδια την διεύθυνση. Αυτό αποδίδει ευελιξία και ταχύτητα καθώς περιορίζει την αναζήτηση

των σελίδων μόνο σε αυτές που θα εν δυνάμει «μεταφέρουν» την σωστή πληροφορία.

Ο Crawler αποτελεί τον κινητήριο μοχλό για την μηχανή αναζήτησης καθώς είναι αυτός όπου θα περισυλλέξει τα δεδομένα από το διαδίκτυο. Η διαδικασία όμως ενημέρωσης των δεδομένων αποτελεί και το σημαντικότερο πρόβλημα της λειτουργίας μιας μηχανής αναζήτησης. Πολλές φορές απαιτούνται εβδομάδες ή ακόμα και μήνες για την ολοκλήρωση προσπέλασης ενός μόνο τμήματος του διαδικτύου. Όλο αυτό το διάστημα είναι πολύ πιθανόν οι πληροφορίες που έχουν συλλεχθεί να μην είναι έγκυρες, να έχουν επεξεργαστεί ή ακόμα και διαγραφεί.

Από την πλευρά της μηχανής αναζήτησης υπάρχει κάποιο κόστος που συνδέει την ανίχνευση και εύρεση κάποιου γεγονότος που προκύπτει από την μη εγκυρότητα του καταγεγραμμένου πόρου. Για την αξιοπιστία και την ορθότητα των αποτελεσμάτων, χρησιμοποιούνται δύο βασικές έννοιες, το Freshness που σαν αποτέλεσμα δυαδικής τιμής, εκφράζει αν το αντίγραφο είναι έγκυρο συναρτήσει του χρόνου και η τιμή Age, που παρουσιάζει σε χρόνο την «γήρανση» του αποθηκευμένου αντίγραφου. Για την διατήρηση υψηλού επιπέδου Freshness, θα πρέπει τα δεδομένα που αποθηκεύονται να μην αλλάζουν συχνά, πρακτικά αυτό σημαίνει στατικές HTML σελίδες όπου το περιεχόμενο τους είναι κατά βάση μη δυναμικό και για την διατήρηση του όρου Age χαμηλού, η συχνή επίσκεψη, με βάση παραμέτρους χρόνου, σε σελίδες όπου το περιεχόμενό τους μεταβάλλεται πιο συχνά.

- ✓ Για την υποβοήθηση του Web Crawler στη λήψη απόφασης του χρόνου επανεπισκεψιμότητας στη σελίδα, χρησιμοποιούνται meta tags σε επίπεδο HTML κώδικα, που εκφράζουν την συχνότητα ανανέωσης του περιεχομένου της σελίδας.

<u>Ηθικά ζητήματα Crawling</u>

Όπως είναι εύκολα κατανοητό, η λειτουργία εξερεύνησης του διαδικτύου από έναν Web Crawler είναι διαδικασία ταχύτερη από το απλό «σερφάρισμα» του χρήστη από σελίδα σε σελίδα για να εξάγει επιθυμητά αποτελέσματα. Η ταχύτητα των λειτουργιών αυτών πολλαπλασιάζεται όταν η μηχανή εκμεταλλεύεται τεχνικές παραλληλοποίησης, άρα και ο όγκος ζήτησης δεδομένων από τον εξυπηρετητή πολλαπλασιάζεται αντίστοιχα. Ας σκεφτούμε επίσης ότι δεν υπάρχει μόνο μία μηχανή αναζήτησης για το διαδίκτυο αλλά ο καθένας μπορεί να λειτουργεί μια για γενικούς ή ειδικούς σκοπούς.

Οι παραπάνω λειτουργίες απαιτούν πόρους, όχι τόσο από τον ίδιο τον Crawler αλλά από το ίδιο το διαδίκτυο και τους εξυπηρετητές των εκάστοτε εφαρμογών. Πρακτικά, αυτό σημαίνει φόρτος για την εφαρμογή και καθυστερημένη εξυπηρέτηση αιτήσεων σε πραγματικούς χρήστες. Τέλος, ο Web Crawler σαν εφαρμογή μπορεί να επηρεάσει την απόδοση της σελίδας, καθώς μπορεί να ζητήσει παραπάνω δεδομένα απ’ ότι μπορεί να επεξεργαστεί, δημιουργώντας έτσι αδικαιολόγητο φόρτο στον εξυπηρετητή. Αν λάβουμε υπ όψιν μας, ότι ο καθένας μπορεί να λειτουργήσει έναν Web Crawler για ιδιωτικό σκοπό, τότε ο φόρτος αυτόματα πολλαπλασιάζεται, ειδικά όταν η υλοποίηση της μηχανής δεν είναι σωστά δομημένη και κάνει άσκοπη χρήση πόρων.

Επιπλέον, σε μια εφαρμογή, υπάρχουν αρχεία που οι χρήστες δεν χρειάζεται να γνωρίζουν ότι υπάρχουν ή δεν έχουν καμία αξία δεδομένων για την αποθήκευση και διαμοιρασμό πληροφορίας. Είναι λογικό, αυτά τα αρχεία να μην προσπελάζονται από τον Crawler, καθώς θα δημιουργήσουν αδικαιολόγητο φόρτο. Για την σωστή λειτουργία των Web Crawler, υπάρχει σε κάθε εξυπηρετητή αρχείο με όνομα «robots.txt». Στο αρχείο αυτό αναγράφονται οδηγίες για το πώς πρέπει να συμπεριφέρεται ο Crawler κατά την επίσκεψη του στον διαδικτυακό τόπο. Μπορεί δηλαδή κάποιος να αποκλείσει συγκεκριμένες μηχανές ή να δηλώσει μόνο τις μηχανές που επιθυμεί να δεχτεί για αναζήτηση. Μπορεί να ορίσει περιοχές που δεν είναι αναγκαίο ή δεν πρέπει η μηχανή να εξερευνήσει ή ακόμα, όπως πλέον υποστηρίζουν οι σύγχρονες μηχανές, να

ορίσει τα χρονικά διαστήματα ζήτησης αρχείων από τον εξυπηρετητή.

Παράδειγμα αρχείου robots.txt από τον διαδικτυακό χώρο της Google.gr

User-agent: *

Disallow: /search

Disallow: /sdch

Disallow: /groups

Disallow: /images

Disallow: /catalogs

Allow: /catalogs/about

Allow: /catalogs/p?

…

3.3.2 Αποθήκευση και Indexing

Στην προηγούμενη ενότητα έγινε αναφορά στις μηχανές Web Crawling που αποτελούν την ραχοκοκαλιά μιας μηχανής αναζήτησης, καθώς είναι οι εφαρμογές όπου θα ζητήσουν τα δεδομένα από τους εξυπηρετητές. Η ιδέα όμως της λειτουργίας μιας μηχανής αναζήτησης δεν βασίζεται στην αναζήτηση σε πραγματικό χρόνο αλλά σε αναζήτηση ήδη προ-αποθηκευμένων δεδομένων που σχετίζονται με τα πραγματικά δεδομένα του διαδικτύου. Επίσης για την πετυχημένη αναζήτηση, τόσο σε ταχύτητα όσο και σε ορθότητα δεδομένων, θα πρέπει τα δεδομένα να αποθηκευτούν σε ειδικές δομές, κατάλληλα επεξεργασμένα, με αφαιρετική έννοια, ώστε ο όγκος τους να είναι όσο το δυνατόν μικρότερος και να επιτευχτεί ταχύτερη επεξεργασία πριν την τελική τους επισκοπή.

Η αποθήκευση όπως εκφράζει και η ίδια η λέξη, αναφέρεται στην διαδικασία διατήρησης της πληροφορίας, πιθανότατα σε κάποιο τρίτο λογισμικό όπως π.χ. μια βάση δεδομένων. Κατά την αποθήκευση τα δεδομένα συνήθως διατηρούνται αυτούσια χωρίς να υποστούν κάποιο είδος επεξεργασίας. Είναι δεδομένα τα οποία η μηχανή αναζήτησης μπορεί να ανακαλέσει για προβολή ή και μετέπειτα επεξεργασία, για την εξαγωγή συμπερασμάτων σχετικά με τις σελίδες αλλά και ακόμα για αξιολόγηση και βελτιστοποίηση της λειτουργίας αλγορίθμων. Τα δεδομένα αυτά είναι συνήθως το περιεχόμενο των HTML Tags που σχετίζονται με τον τίτλο, την περιγραφή, λέξεις κλειδιά κωδικοποίηση κ.α., καθώς και μέρος ή και ολόκληρο το «καθαρό» περιεχόμενο της σελίδας ή και ολόκληρος ο HTML κώδικας.

Στις σύγχρονες μηχανές αναζήτησης, που το αποτέλεσμα αποτελεί σημαντικότερο κομμάτι από την ταχύτητα απόκρισης, συν του ότι οι αλγόριθμοι αξιολόγησης των σελίδων σύμφωνα με το περιεχόμενο τους, καθιστούν αναγκαία την διατήρηση και επεξεργασία της πληροφορίας, πέρα από το περιεχόμενο της κάθε σελίδας, διατηρούνται δεδομένα που έχουν να κάνουν με την προέλευσή της, τους χρόνους ζήτησης και επεξεργασίας από τους εξυπηρετητές, συσχετισμός της σελίδας με άλλους διαδικτυακούς τόπους, το μέγεθος της, δεδομένα που η ίδια εφαρμογή μπορεί να δημιουργήσει ανάλογα με την σελίδα κλπ.

Πέρα από την αποθήκευση των δεδομένων όπως αναφέρθηκε παραπάνω, που συντελεί στην πλουσιότερη απόδοση πληροφορίας, είναι εξίσου αναγκαία και η ταχύτητα απόδοσης της πληροφορίας αυτής. Η αναζήτηση σε ακατέργαστα δεδομένα των βάσεων δεδομένων και λειτουργία αλγορίθμων για την εξαγωγή συμπερασμάτων σε πραγματικό χρόνο, όπως εύκολα είναι αντιληπτό, θα αποτελούσε μειονέκτημα για την γρήγορη λειτουργία της μηχανής. Για τον λόγω αυτό χρησιμοποιούνται τεχνικές «Indexing» σε μη επεξεργασμένα δεδομένα. Η διαδικασία αυτή αποτελεί κομμάτι αναγκαίας λειτουργίας για την μηχανή, καθώς κατά την ζήτηση και αποθήκευση των δεδομένων, εκτελούνται συγκεκριμένοι αλγόριθμοι και διαδικασίες για την απόδοση συμπερασμάτων, ενώ αφηρημένη αλλά και συγκεκριμένη

πληροφορία επεξεργάζεται με την διαδικασία του Indexing, ώστε κατά την αναζήτηση, ο όγκος που θα πρέπει να ελεγχθεί είναι υποπολλαπλάσια μικρότερος και η πληροφορία στοχευόμενη και σύντομη.

Οι τεχνικές Indexing όπως θα αναφέρουμε και παρακάτω, έχουν σαν βασική λειτουργία την «κωδικοποίηση» λέξεων ή φράσεων, όπου με κατάλληλες διαδικασίες, βαθμολογούν την προς αναζήτηση λέξη η φράση και επιστρέφουν ιεραρχικές λίστες αποτελεσμάτων. Ένα πολύ γνωστό παράδειγμα τέτοιας εφαρμογής ανοιχτού κώδικα, είναι το Lucene.

3.3.3 Αναζήτηση και Εμφάνιση αποτελεσμάτων

Το τελικό βήμα έπειτα από την εξερεύνηση, αποθήκευση και επεξεργασία των αποτελεσμάτων, είναι και η προβολή αυτών στους χρήστες του διαδικτύου. Το διαδίκτυο αποτελεί έναν ζωντανό οργανισμό ο οποίος εξελίσσεται, καινοτόμες λειτουργίες εισάγονται καθημερινά, άλλες διαδικασίες ξεχνιούνται και δεν θα ήταν λογικό οι χρήστες του να μην συμβαδίζουν με αυτό! Οι χρήστες είναι απαιτητικοί, επιζητούν ταχύτητα και ορθότητα των δεδομένων και οι Κολοσσοί όπως η Google, Microsoft και Yahoo φρόντισαν να δημιουργήσουν αυτά τα στάνταρ, που οποιαδήποτε αντίστοιχη εφαρμογή πρέπει να ακολουθεί ώστε να έχει μερίδιο επιτυχίας στο διαδίκτυο.

Ενώ η αρχική μορφή της προς αναζήτησης πρότασης αποτελούταν από σύντομες ή και μονολεκτικές φράσεις και πολλά πεδία ελέγχου σε επίπεδο HTML κώδικα (radio buttons, check boxes), πλέον οι μηχανές αναζήτησης έχουν αποκτήσει επίσης ευφυΐα και στο κομμάτι της έκφρασης των χρηστών. Πρώτη η Google καινοτόμησε με το να «διορθώνει» τους χρήστες στα λεκτικά τους και εισήγαγε την έννοια «Μήπως εννοείται…?» ή και να συμπληρώνει αυτόματα τις λέξεις/προτάσεις. Για την υλοποίηση της συγκεκριμένης διαδικασίας, αλγόριθμοι τρέχουν σε πραγματικό χρόνο για την «εξερεύνηση» των προτάσεων αναζήτησης και δόμησης της με κατάλληλο τρόπο ώστε να επιτευχθεί αποτελεσματικότερη αναζήτηση.

Η προβολή των αποτελεσμάτων δεν θα μπορούσε να έχει άλλη πορεία από αυτή που χάραξαν οι πρώτες μηχανές αναζήτησης. Το διαδίκτυο, αποτελεί πλέον μέρος της καθημερινότητας. Η συνήθειες των ανθρώπων και η προσαρμογή τους σε καινούργιες καταστάσεις αποτελούν δύο αντικρουόμενες διαδικασίες. Για την εξασφάλιση λοιπόν της ομαλής μετάβασης σε μια καινούργια εφαρμογή, αλλά να διατηρείται η προηγούμενη γνώση των χρηστών σε αντίστοιχες λειτουργίες, εισήχθησαν στο διαδίκτυο και στην ανάπτυξη των εφαρμογών του τεχνικές UX (User Experience). Η μελέτη στην προβολή των αποτελεσμάτων αναζήτησης έδειξε ότι η απλή παρουσίαση τους σε συγκεκριμένα χρώματα εξυπηρετεί τους χρήστες να κατανοήσουν, να επεξεργαστούν και να χρησιμοποιήσουν την διδόμενη απάντηση. Παρακάτω παρουσιάζεται η μορφή των αποτελεσμάτων σε διαφορετικές μηχανές αναζήτησης. Σημαντικό είναι να παρατηρήσουμε τον τρόπο απόδοσης αυτών στον χρήστη.

3.4 Η Γλώσσα HTML

Ο όρος HTML προκύπτει από τα αρχικά των λέξεων HyperText Markup Language και σημαίνει γλώσσα περιγραφής σελίδων του διαδικτύου.

Βασικό στοιχείο της γλώσσας αυτής είναι έννοια του “tag” (ετικέτα). Κάθε στοιχείο των σελίδων HTML εμφανίζεται ανάμεσα σε ετικέτες (tags) και οι ετικέτες αυτές καθορίζουν την τοποθεσία μέσα στη σελίδα των στοιχείων αυτών και τη μορφή με την οποία θα εμφανίζονται και θα φαίνονται.

Όλες οι σελίδες HTML ξεκινούν με την ετικέτα <html> και τελειώνουν με την αντίστοιχη ετικέτα τέλους </html>. Επίσης κάθε σελίδα HTML αποτελείται από δύο τμήματα. Το πρώτο τμήμα καθορίζεται από τις ετικέτες <head> και </head> και το δεύτερο από τις <body> και </body>.

Το πρώτο τμήμα αποτελεί και την κεφαλή του κειμένου και καθορίζει διάφορες παραμέτρους της συγκεκριμένης σελίδας. Για

παράδειγμα καθορίζει τον τίτλο της, τη σχέση της με άλλες σελίδες, τη μορφή που μπορεί να έχει ή ακόμα και τη scripting γλώσσα που

μπορεί να χρησιμοποιήσει. Από τα παραπάνω στοιχεία αυτό που χαρακτηρίζει κατά κάποιο τρόπο της συγκεκριμένη σελίδα είναι ο τίτλος της ο οποίος εμφανίζεται ανάμεσα στις ετικέτες <title> και </title>.

Το δεύτερο τμήμα αποτελεί το σώμα της σελίδας και είναι αυτό που περιέχει όλη την πληροφορία που επιθυμεί ο δημιουργός της σελίδας να παρουσιάσει, είτε με τη μορφή κειμένου, είτε εικόνων, είτε ακόμα και με διασυνδέσεις προς άλλες σελίδες δικές του ή άλλων ατόμων. Από τα στοιχεία που μπορούν να τοποθετηθούν στο σώμα μίας σελίδας HTML τα περισσότερα μπορούν να ενταχθούν σε δύο κατηγορίες.

Η πρώτη αποτελείται από τα στοιχεία ορισμού περιοχής τα οποία είναι οι επικεφαλίδες, οι παράγραφοι και οι οριζόντιες γραμμές. Σημαντικό ενδιαφέρον παρουσιάζουν οι επικεφαλίδες, οι οποίες καθώς τονίζονται από το δημιουργό της σελίδας υποδηλώνουν ότι περιέχουν κάποια πληροφορία η οποία πρέπει να προσεχτεί. Υπάρχουν έξι επίπεδα από επικεφαλίδες, το H1 είναι το πιο σημαντικό και το H6 το λιγότερο σημαντικό. Το κείμενο που εμφανίζεται σαν επικεφαλίδα περιέχεται ανάμεσα σε ετικέτες της μορφής <h1> και </h1>, δηλαδή ανάλογα με την επικεφαλίδα καθορίζεται και η ετικέτα.

Η δεύτερη βασική κατηγορία αποτελείται από τα στοιχεία ορισμού κειμένου τα οποία ορίζουν τύπους χαρακτήρων στο κείμενο. Βασική υποκατηγορία αυτών των στοιχείων αποτελούν τα στοιχεία τύπου γραμματοσειράς. Ανάλογα με τα στοιχεία που επιλέγει ο δημιουργός της σελίδας μπορεί να παρουσιάσει κάποιο κείμενο, είτε με πιο έντονα γράμματα ,είτε με πλάγιους χαρακτήρες, είτε να είναι υπογραμμισμένο. Ένα κείμενο για να εμφανίζεται με έντονα γράμματα πρέπει να βρίσκεται είτε ανάμεσα στις ετικέτες και είτε ανάμεσα στις και . Για να εμφανίζεται με πλάγιους χαρακτήρες πρέπει να βρίσκεται ανάμεσα στις ετικέτες και . Τέλος για να είναι υπογραμμισμένο πρέπει να βρίσκεται ανάμεσα στις ετικέτες <u> και </u>.

Όπως αναφέρθηκε και παραπάνω ο δημιουργός της σελίδας μπορεί να εμφανίζει στη σελίδα του διάφορες εικόνες. Για να το επιτύχει αυτό χρειάζεται ένα άλλο στοιχείο της γλώσσας HTML, το img, το οποίο ανήκει στα στοιχεία ορισμού κειμένου. Το στοιχείο αυτό καθορίζεται από την ετικέτα <img> και δεν έχει ετικέτα τέλους. Για να εμφανιστεί μία εικόνα σε μία σελίδα HTML πρέπει να υπάρχει στον κώδικα της σελίδας το στοιχείο img με την παρακάτω μορφή: <img src=”imagepath/imagename.jpeg” alt=”comment”>. Στο πεδίο «src» του στοιχείου img δίνεται πρώτα το μονοπάτι που δείχνει τον κατάλογο στον οποίο είναι αποθηκευμένη η εικόνα που θα εμφανιστεί και μετά δίνεται το όνομα της εικόνας με την κατάληξη του τύπου της. Στο πεδίο «alt» δίνεται το κείμενο που επιθυμεί ο δημιουργός να φαίνεται μέχρι να εμφανιστεί η εικόνα, είναι εμφανές ότι το κείμενο αυτό περιγράφει κατά κάποιο τρόπο την εικόνα και το θέμα της.

Τέλος ένα ακόμα πολύ χρήσιμο και θεμελιώδες στοιχείο της γλώσσας HTML είναι οι υπερδεσμοί. Οι υπερδεσμοί περιέχονται ανάμεσα στις ετικέτες <a> και και έχουν την ακόλουθη μορφή: hyperlink - text. Στο πεδίο «href» του

στοιχείου a δίνεται η διεύθυνση της σελίδας προς την οποία δείχνει ο συγκεκριμένος υπερδεσμός. Ανάμεσα στις δύο ετικέτες δίνεται το κείμενο που παρουσιάζει ο υπερδεσμός και η σελίδα προς την οποία δείχνει, επομένως κατά κάποιο τρόπο παρουσιάζει και το περιεχόμενο της σελίδας αυτής.

```
1  <!DOCTYPE html>
2  <html>
3      <head>
4          <title>Example<title>
5          <link rel="stylesheet" href="styl
6      </head>
7      <body>
8          <h1>
9              <a href="/">Header</a>
10         </h1>
11         <nav>
12             <a href="one/">One</a>
13             <a href="two/">Two</a>
14             <a href="three/">Three</a>
15         </nav>
```

Εικόνα 3.14: Παράδειγμα γλώσσας HTML

Κεφάλαιο 4

Η διαδικασία της εξόρυξης δεδομένων διαθέτει ποικιλία εφαρμογών, όπως προαναφέρθηκε. Η σύγχρονη ψηφιακή αγορά, τα Κοινωνικά Δίκτυα και μεγάλες εταιρίες και διαδικτυακές πλατφόρμες βασίζονται στις τεχνικές και στους αλγόριθμους του τομέα της εξόρυξης δεδομένων για την λειτουργία τους και την ολοένα αυξανόμενη εξέλιξή και καλύτερη εξυπηρέτηση των χρηστών τους.

Η δημιουργία αγοραστικών προφίλ χρηστών (User Profiles), η καταγραφή των προτιμήσεών τους, η χρησιμοποίηση συστημάτων προτάσεων (Recommendations Systems), η αποδοτικότερη και οικονομικότερη λειτουργία των διαδικτυακών πλατφόρμων και η προσέλκυση μεγαλύτερου πλήθους χρηστών αποτελούν μόνο μερικά επιτεύγματα της χρήσης τεχνικών εξόρυξης δεδομένων στον Παγκόσμιο Ιστό.

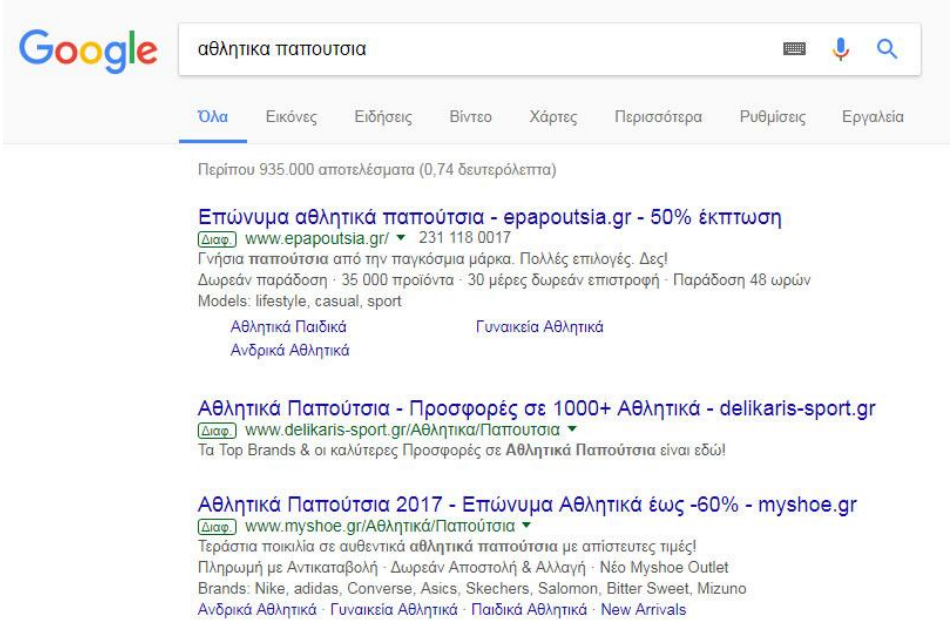
Όπως θα δούμε και στη συνέχεια εταιρίες κολοσσοί της σύγχρονης ψηφιακής αγοράς, όπως η Google, το TripAdvisor κ.ά. βασίστηκαν στην εφαρμογή τεχνικών της εξόρυξης δεδομένων για την ύπαρξή τους και την εξέλιξή τους.

4.1 Digital Marketing

Ένας από τους κυριότερους λόγους χρήσης τεχνικών εξόρυξης δεδομένων είναι και η στοχευόμενη διαφήμιση. Ως στοχευόμενη διαφήμιση καλείται η οπτικοακουστική παρουσίαση προσφορών στους χρήστες με συγκεκριμένες προτιμήσεις.

Η σύγχρονη αγορά του διαδικτύου χαρακτηρίζεται από τον τεράστιο αριθμό διαφημίσεων ποικίλης ποιότητας. Αυτό έχει δημιουργήσει έναν τεράστιο ανταγωνισμό και δεν είναι τυχαίο ότι αρκετές εταιρίες και ιστοσελίδες έχουν προχωρήσεις στη δημιουργία νέων πλατφόρμων και αλγορίθμων για την, όσο το δυνατόν, πιο στοχευόμενη διαφήμιση. Χαρακτηριστικά παραδείγματα αποτελούν τα σύγχρονα εργαλεία διαφήμισης της Google, όπως το Google AdSense και το Google AdWords. Μία

διαφήμιση στοχευόμενη στον σωστό αποδέκτη μπορεί να κάνει την διαφορά ανάμεσα στο σωστό μάρκετινγκ και την συνεχή και άσχετη ανεπιθύμητη αλληλογραφία.



Εικόνα 4.1: Εφαρμογή του συστήματος διαφημίσεων Google AdWords στην μηχανή αναζήτησης της Google

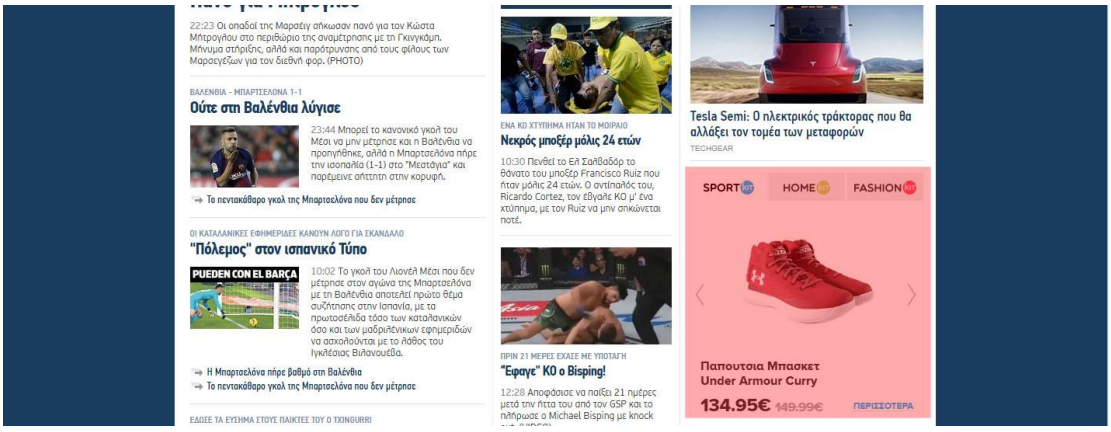
Σε αυτό το σημείο εισέρχεται και η έννοια και οι τεχνικές της εξόρυξης δεδομένων και γνώσης, με τις οποίες μπορεί να γίνει αντιληπτό το μέσο πλήθος επισκεπτών κάθε σελίδας, καθώς και τα ενδιαφέροντά τους. Με σύγχρονες τεχνικές ανάλυσης και καταγραφής προφίλ και τη χρήση συστημάτων προτάσεων (Recommendations Systems) η στοχευόμενη και αποτελεσματικότερη διαφήμιση γίνεται πιο εύκολη. Έτσι, προβάλλονται διαφημίσεις και προσφορές που έχουν μεγαλύτερη πιθανότητα να προσελκύσουν συγκεκριμένους χρήστες μιας ιστοσελίδας. Για παράδειγμα, banners με διαφημίσεις καλλυντικών θα έχουν μικρή απήχηση σε ένα αθλητικό ειδησεογραφικό portal. Καλύτερη επιλογή διαφήμισης για το συγκεκριμένο portal θα ήταν αθλητικά είδη ή ενεργειακά ποτά.

Υπάρχουν πολλοί τρόποι για να προσδιοριστεί το κοινό μιας σελίδας. Συνήθως η διαδικασία αυτοματοποιείται χρησιμοποιώντας web crawlers, οι οποίοι ανιχνεύουν μια ιστοσελίδα για «ετικέτες»,

κομμάτια κώδικα δηλαδή που αναφέρουν λέξεις κλειδιά σχετικά με τα περιεχόμενα.

Έπειτα τα δεδομένα αυτά ομαδοποιούνται σε μεγάλες κατηγορίες όπως άνδρες, αθλητικά, μοντελισμός, παπούτσια και συνδυάζονται με έτοιμες διαφημίσεις με παρόμοιες «ετικέτες» για τη αποδοτικότερη προβολή τους.

Αν και ο μηχανισμός λειτουργίας είναι απλός, εντούτοις υπάρχουν και ορισμένα προβλήματα όπως διαφημιστές να χρησιμοποιούν «ετικέτες» άσχετες με το θέμα που διαφημίζουν, εκμεταλλευόμενοι την αυτοματοποιημένη διαδικασία. Τέτοιες διαφημίσεις όμως γρήγορα αποσύρονται καθώς καταγγέλλονται από τους ιδιοκτήτες των σελίδων. Επιπλέον, συχνό φαινόμενο είναι banners τα οποία είναι ιδιαίτερα ενοχλητικά παίζοντας δυνατή μουσική ή βίντεο, εικόνες πορνογραφικού περιεχομένου σε άσχετες σελίδες είτε τέλος με την τοποθέτησή τους σε ζωτικά σημεία της ιστοσελίδας όπως το πλήκτρο λήψης, αναγκάζοντας τον χρήστη να τις επισκεφθεί.



Εικόνα 4.2: Προβολή στοχευόμενης διαφήμισης εντός αθλητικού portal

Η λογική μιας διαφήμισης είναι φυσικά να προσελκύσει τον αγοραστή. Οι σύγχρονες τεχνικές εξόρυξης δεδομένων κάνουν την διαφήμιση να εμφανίζεται ταυτόχρονα σε χιλιάδες διαφορετικές σελίδες, οι οποίες τις περισσότερες φορές πληρώνονται για την προβολή τους. Υπάρχουν δύο διαφορετικοί τρόποι πληρωμής. Ο ένας αφορά την πληρωμή pay-per-click, όπου ο χρήστης πρέπει

να κάνει κλικ στην διαφήμιση για να χρεωθεί ένα ποσό στο διαφημιζόμενο και ο δεύτερος αφορά την πληρωμή για την απλή προβολή της.

Μια άλλη ενδιαφέρουσα πτυχή της στοχευόμενης διαφήμισης είναι η χρήση έξυπνων αλγόριθμων οι οποίοι δεν περιορίζονται στην απλή ταυτοποίηση «ετικετών», αλλά συλλέγουν δεδομένα σε μεγάλες ποσότητες, δημιουργώντας λογικές σχέσεις ανάμεσα σε έννοιες, κάνοντας δυνατή την αποστολή ειδικευμένων διαφημίσεων σε χρήστες χωρίς οι ίδιοι να έχουν επισκεφτεί μια σχετική σελίδα. Για παράδειγμα, αν χιλιάδες χρήστες αρχίζουν να αναζητούν συγκεκριμένες μάρκες κινητών τηλεφώνων και στην συνέχεια φορτιστές, νέοι πελάτες που αναζητούν τις ίδιες συσκευές θα βλέπουν διαφημίσεις για φορτιστές χωρίς να τους αναζητούν, με βάση προηγούμενες συμπεριφορές άλλων χρηστών. Η μηχανική αυτή μάθηση είναι η κορυφαία τεχνολογία στον χώρο του Digital Marketing.

Η σύγχρονη διαφήμιση στον Παγκόσμιο Ιστό αποτελεί έναν από τις βασικές εφαρμογές της εξόρυξης δεδομένων σε αυτόν και μία από τις πιο προφανείς χρήσης της. Η δυνατότητα μεγιστοποίησης του κέρδους με την εφαρμογή μερικών απλών βημάτων κάνει το μήνυμα της διαφήμισης να φτάνει σε περισσότερους καταναλωτές, σε περισσότερες ιστοσελίδες, με λιγότερη προσπάθεια. Η εξ' ολοκλήρου αυτοματοποίηση της διαδικασίας έχει θετικές και αρνητικές πλευρές, αλλά σαν σύνολο λειτουργεί και θα συνεχίσει έτσι για πολύ καιρό ακόμα.

4.2 Google Analytics & Βελτιστοποίηση Ιστοσελίδων για τις Μηχανές Αναζήτησης (SEO)

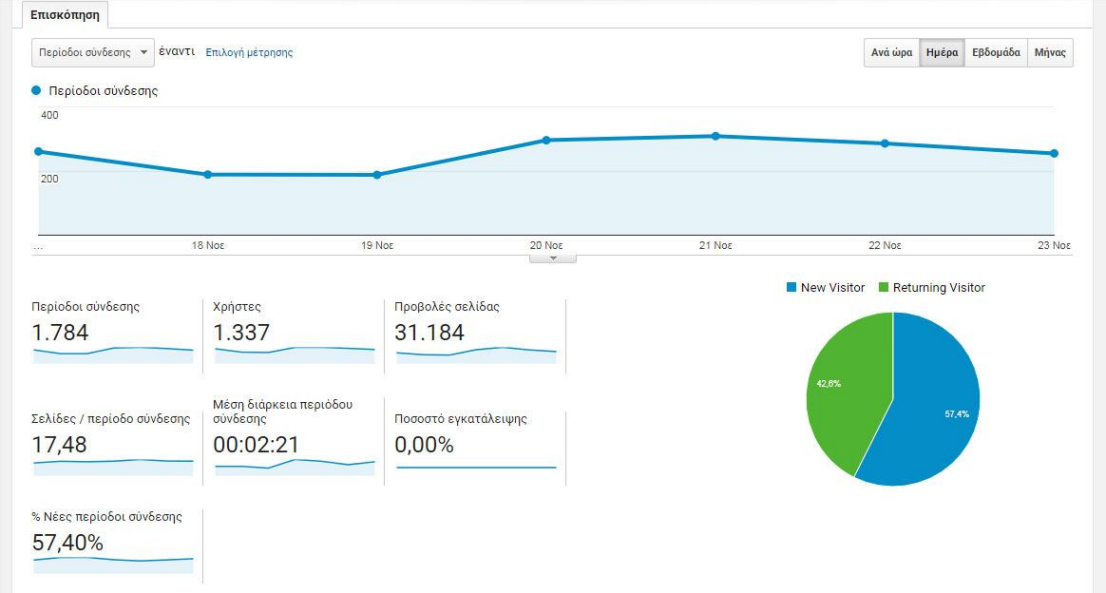
Ίσως ο κυριότερος λόγος ύπαρξης της επιστήμης της εξόρυξης γνώσης είναι η στατιστική. Η ανάγκη για γνώση και πρόβλεψη σε συνδυασμό με την ευκολία παρουσίασης και ανάλυσης των στατιστικών δημιουργούν έναν εξαιρετικό συνδυασμό.

Ουσιαστικά, ο σκοπός της στατιστικής είναι η μελέτη των δεδομένων. Οι χρήσεις της ποικίλουν από την αναγνώριση

μοτίβων και την απεικόνιση γραφημάτων μέχρι την εξαγωγή συμπερασμάτων για την πρόβλεψη μελλοντικών εισόδων, κάτι με το οποίο ασχολείται και η επιστήμη της εξόρυξης δεδομένων, της οποίας ο σκοπός είναι η επεξεργασία των εισόδων με κατάλληλο τρόπο ώστε να προκύψουν νέα και πρωτότυπα συμπεράσματα. Η ποιοτική διαφορά της εξόρυξης γνώσης από την στατιστική όμως, είναι η ικανότητα της πρώτης να φιλτράρει τον θόρυβο, δηλαδή άσχετα δεδομένα, αλλά και να συνθέσει αποδεκτά αποτελέσματα για να αντικαταστήσει εγγραφές που λείπουν (σύνηθες φαινόμενο σε βάσεις δεδομένων).

Επομένως, η στατιστική και η εξόρυξη δεδομένων είναι αλληλένδετες, με κοινό στόχο και διαφορετικά εργαλεία. Για παράδειγμα, μια στατιστική μπορεί να απεικονίσει τις τάσεις της επισκεπτών μιας ιστοσελίδας ή ενός ηλεκτρονικού καταστήματος σε σχέση με ένα προϊόν ή υπηρεσία τους και να καταγράψει το πόσο επιτυχημένο είναι. Προσθέτοντας τώρα τις τεχνικές εξόρυξης δεδομένων μπορούμε να μάθουμε τις ηλικίες που απευθύνεται, τις χώρες που αυτό είναι δημοφιλέστερο, τον μέσο όρο ζωής του προϊόντος ή της υπηρεσίας, και τι μπορεί να αλλάξει ώστε να γίνει πιο επιτυχημένο και να προσελκύσει περισσότερους επισκέπτες-πελάτες. Ενώνοντας λοιπόν τις δύο επιστήμες έχουμε την μεγιστοποίηση της κατανόησης των αποτελεσμάτων και την ικανότητα να προβλέψουμε τα μελλοντικά.

Αυτό το ρόλο διαδραματίζει ένα από τα πιο διαδεδομένα παγκοσμίως εργαλεία της Google, το Google Analytics. Το συγκεκριμένο εργαλείο αποτελεί το πιο ενδεικτικό παράδειγμα σύνδεσης της στατιστικής με την Εξόρυξη Δεδομένων. Εκτός του ότι μας επιτρέπει να μετράμε πωλήσεις και μετατροπές, μας δίνει και μία νέα, διαφορετική ματιά σχετικά με τον τρόπο που χρησιμοποιούν οι επισκέπτες έναν ιστότοπο, πώς έφτασαν στον σε αυτόν και πώς ο διαχειριστής του μπορεί να τους προτρέψει να επιστρέψουν στον ιστότοπό του.



Εικόνα 4.3: Στατιστικά επισκεψιμότητας ιστοσελίδας από το εργαλείο Google Analytics

Με την εξέταση αυτών των δεδομένων ο ιδιοκτήτης μιας συγκεκριμένης ιστοσελίδας γνωρίζει πώς επισκέπτονται τον ιστότοπο οι διάφοροι χρήστες και από πού τον επισκέπτονται, όπως χώρα, τύπος browser που χρησιμοποιήθηκε κ.ά. Ωστόσο ο ιδιοκτήτης του συγκεκριμένου ιστότοπου δεν λαμβάνει προτάσεις ή ιδέες για τη βελτίωση του. Για αυτόν τον λόγο ο ιδιοκτήτης του ιστότοπου θα πρέπει να προβεί στην εξόρυξη και ανάλυση των συγκεκριμένων δεδομένων. Η στατιστική ανάλυση ιστοσελίδων είναι κάτι αρκετά σημαντικό και όσον αφορά το SEM (search engine marketing) είναι κάτι απολύτως απαραίτητο. Επίσης, η βελτιστοποίηση μιας ιστοσελίδας για να είναι φιλική προς τις μηχανές αναζήτησης (Search Engine Optimization - SEO) αποτελεί σημαντικό στοιχείο του σύγχρονου διαδικτύου.

Οι τυποποιημένες αναλυτικές λύσεις εξετάζουν συνήθως και αξιολογούν τα τρέχοντα δεδομένα και συχνά σε ένα ευρύ φάσμα. Ενώ αυτό οδηγεί σε κρίσιμες πληροφορίες για τις εταιρείες, εξακολουθεί να βασίζεται στην ανθρώπινη κατανόηση και γενικά στα σύνολα δεδομένων. Εφαρμόζοντας όμως τεχνικές εξόρυξης δεδομένων αυτών των πληροφοριών, μπορεί να αναλυθεί και να επαληθευτεί ένα συγκεκριμένο σύνολο κανόνων.

Η εξόρυξη αυτών των δεδομένων σε συνδυασμό με τεχνικές Μηχανικής Μάθησης (Machine Learning) θα οδηγήσει τους διαχειριστές των ιστοσελίδων σε βελτιστοποίησή. Υπάρχουν διάφορες τεχνικές που χρησιμοποιούνται για την εξόρυξη δεδομένων από το Google Analytics, αλλά στον πυρήνα τους είναι οι στατιστικές, η τεχνητή νοημοσύνη και η μηχανική μάθηση.

Η Μηχανική Μάθηση αποτελεί και αυτή ένα από τα πολυπλοκότερα ζητήματα. Αναφέρεται στο κατά πόσο μπορεί ένα πρόγραμμα να είναι ικανό για εκμάθηση, είτε μέσω εμπειρίας, είτε μέσω παραδειγμάτων αλλά και περιγραφών σε φυσική γλώσσα. Αυτός ο τομέας εξελίσσεται δυναμικά και σε αυτόν βασίζονται και οι τεχνικές της Εξόρυξη Δεδομένων (Data Mining). Η Μηχανική Μάθηση αποσκοπεί στη διερεύνηση των μηχανισμών και των υπολογιστικών διαδικασιών, μέσω των οποίων είναι δυνατή η εξαγωγή και οργάνωση της γνώσης από την υπάρχουσα εμπειρία. Μεταξύ των διαφόρων μορφών συστημάτων μάθησης, η επαγωγική μάθηση μέσω παραδειγμάτων (inductive learning) έχει γνωρίσει τη μεγαλύτερη διάδοση. Στο συγκεκριμένο πεδίο σημαντικό ρόλο παίζει και η ποιότητα των κανόνων που εξάγονται από την υπάρχουσα γνώση, καθώς και η βελτίωσή της.

Η εφαρμογή τεχνικών μηχανικής μάθησης μπορεί να προβλέψει τις τάσεις των επισκεπτών και μπορεί να βοηθήσει στην αυτοματοποίηση ενεργειών πωλήσεων και μάρκετινγκ, με έξυπνη προσαρμογή στα συνεχώς μεταβαλλόμενα δεδομένα.

Αρκετοί αλγόριθμοι, προγράμματα και πλατφόρμες σε γλώσσα προγραμματισμού R έχουν δημιουργηθεί, ώστε η εξόρυξη των δεδομένων από το εργαλείο Google Analytics να βοηθήσει επιχειρήσεις και ιστοσελίδες να βελτιώσουν την παρουσία τους και την επισκεψιμότητά τους στο Διαδίκτυο.

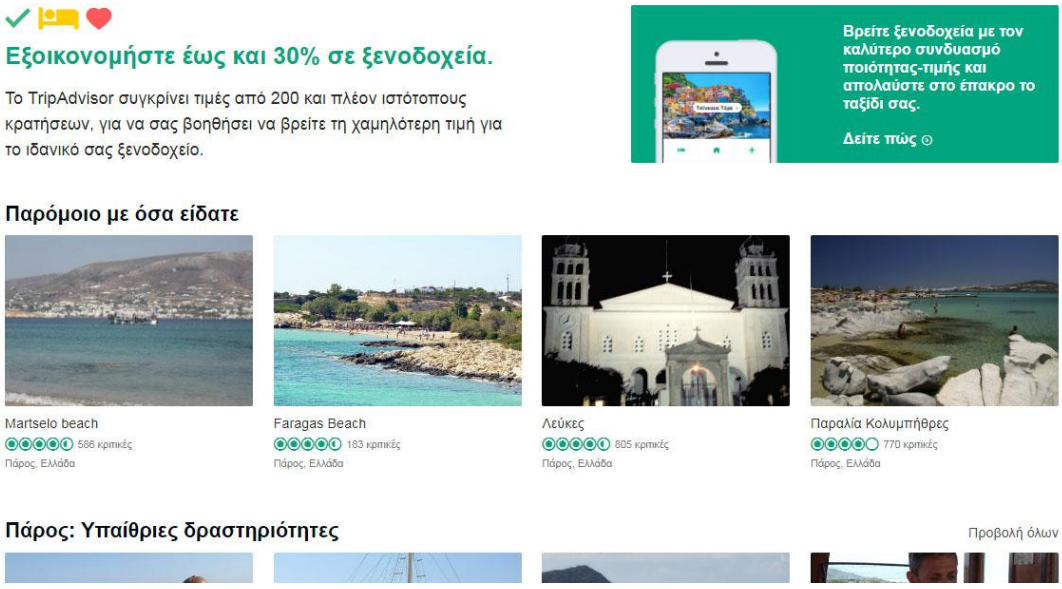
4.3 Εφαρμογή σε επιχειρήσεις του κλάδου φιλοξενίας (TripAdvisor)

Η φιλοξενία μεταβάλλεται ταχύτατα, με τρόπους που ο κλάδος δεν αντιλαμβάνεται ακόμα πλήρως. Η σύγχρονη αγορά του

συγκεκριμένου κλάδου επιβάλει οι ιδιοκτήτες επιχειρήσεων να γνωρίζουν ποιες πρακτικές είναι αποτελεσματικές, ποιες θα είναι σύντομα και ποιες ήταν κάποτε αλλά δεν είναι πια. Αυτό μπορεί να επιτευχθεί μόνο μέσω των χρήσιμων γνώσεων που μπορούν οι σύγχρονοι επαγγελματίες να αποκομίσουν από την εφαρμογή της εξόρυξης και της μελέτης των δεδομένων.

Βελτιστοποίηση τιμολόγησης, αποτελεσματικές ειδικές προσφορές, κανάλια διαφήμισης είναι μόνο μερικά παραδείγματα και εν δυνάμει εργαλεία που έχουν στη διάθεσή τους οι επαγγελματίες του κλάδου και η μελέτη των δεδομένων των οποίων μπορεί να τους βοηθήσει να πάρουν καλύτερες αποφάσεις για όλα αυτά τα ζητήματα. Πλέον και οι μικρές επιχειρήσεις έχουν πρόσβαση σε ένα μεγάλο όγκο δεδομένων.

Χαρακτηριστικό παράδειγμα σε αυτόν τον τομέα αποτελεί η γνωστή πλατφόρμα του TripAdvisor. Με περισσότερες από 435 εκατομμύρια κριτικές και γνώμες που καλύπτουν 6,8 εκατομμύρια επιχειρήσεις του κλάδου φιλοξενίας, το TripAdvisor επεξεργάζεται έναν τεράστιο όγκο δεδομένων. Επίσης, η συγκεκριμένη πλατφόρμα αποτελεί ακόμη ένα παράδειγμα χρήσης Recommendations Systems. Το TripAdvisor επομένως στράφηκε στο Big Data για να προσφέρει τις καλύτερες συστάσεις και περιεχόμενο στους χρήστες του. Είναι επομένως σαφές ότι το TripAdvisor δημιουργεί τεράστια ποσά δεδομένων, τα οποία χρησιμοποιούνται για να βελτιώσουν τις υπηρεσίες του και να παραμείνει ο μεγαλύτερος ταξιδιωτικός ιστότοπος στον κόσμο.



Εικόνα 4.4: Χρήση συστημάτων προτάσεων από την πλατφόρμα του TripAdvisor

Εφαρμογή Data Mining στην Ανανέωση των συνδρομών

Το TripAdvisor έχει χτίσει προηγμένα μοντέλα Big Data για να προβλέψει το ποσό των κλικ που χρειάζονται ώστε ξενοδοχείο να ανανεώσει τη συνδρομή του. Υπολογίζουν το ποσό της δραστηριότητας μάρκετινγκ που απαιτείται για την επίτευξη ενός ορισμένου επιπέδου κλικ, ο οποίος θα οδηγήσει σε ανανέωση συνδρομής. Χρησιμοποιώντας προηγμένες αναλύσεις, γνωρίζουν αν είναι οικονομικά συνετό να αυξήσουν τις δραστηριότητες μάρκετινγκ. Ο στόχος του TripAdvisor είναι να βοηθήσει τους χρήστες να βρουν καλύτερο περιεχόμενο πιο γρήγορα. Αυτό επιτυγχάνεται αναλύοντας την ατομική συμπεριφορά τους στον ιστότοπο καθώς και τη δραστηριότητα του πλήθους των επισκεπτών στο σύνολό του. Προσφέροντας μια καλύτερη εμπειρία, είναι πιο εύκολο για τους χρήστες στο να βρουν αυτό που ψάχνουν, με αποτέλεσμα, φυσικά, περισσότερα έσοδα από διαφημίσεις. Χρησιμοποιούν αρκετές τεχνικές εξόρυξης δεδομένων για να το επιτύχουν, που κυμαίνονται από αναλύσεις μεγάλης κλίμακας σε πραγματικό χρόνο, προγνωστικές αναλύσεις, εξόρυξη δεδομένων και στατιστική μοντελοποίηση. Στο επίκεντρο των τεχνικών της συγκεκριμένης πλατφόρμας βρίσκεται η παροχή καλύτερων εξατομικευμένων συστάσεων για τον επισκέπτη.

Καταπολέμηση ψευδών κριτικών

Μια άλλη εφαρμογή του τομέα της εξόρυξης δεδομένων που χρησιμοποιεί το TripAdvisor είναι στο να μπορέσει να περιορίσει και να "κερδίσει" τις ψευδείς κριτικές. Οι ψευδείς κριτικές είναι επιβλαβείς για τους επιχειρηματίες, είναι άχρηστες για τους επισκέπτες και μακροπρόθεσμα θα επηρεάσουν αρνητικά το TripAdvisor. Μια λύση για την καταπολέμηση των ψεύτικων σχολίων είναι οι εξακριβωμένες κριτικές, όπου επιβεβαιώνεται, μέσω 3rd party εργαλείων, ότι ένας χρήστης έμεινε στην πραγματικότητα στο κατάλυμα που αξιολόγησε ή σχολίασε. Το 2014, το TripAdvisor και η American Express ήρθαν σε συμφωνία που επιτρέπει στους κατόχους καρτών της Amex να συνδέσουν την κάρτα τους με το προφίλ τους στο TripAdvisor. Κάθε φορά που αφήνουν ένα σχόλιο ή μια αξιολόγηση στον ιστότοπο, είναι η αγορά ως "Επανεξέταση μέλους του Amex Card" που επιβεβαιώνει ότι η κάρτα χρησιμοποιήθηκε για να κάνει μια αγορά σε αυτές τις τοποθεσίες.

Μια πραγματικά προσαρμοσμένη πλατφόρμα μεγάλων δεδομένων

Προκειμένου να επεξεργαστούν και να αναλυθούν όλα τα δεδομένα που η πλατφόρμα του TripAdvisor έχει στη διάθεσή της, να αναπτυχθούν εξατομικευμένα συστήματα προτάσεων και να βελτιωθούν οι αλγόριθμοι καταπολέμησης των ψευδών κριτικών, το TripAdvisor έχει αναπτύξει μια προσαρμοσμένη πλατφόρμα Big Data. Χρησιμοποιεί μια πληθώρα τεχνολογιών και τεχνικών εξόρυξης δεδομένων, συμπεριλαμβανομένων των Hadoop (για την αποθήκευση και επεξεργασία δεδομένων στο διαδίκτυο), SQL Servers (για την αναφορά ενάντια σε συγκεντρωτικά δεδομένα), Hive (για την αναζήτηση των δεδομένων και την τοποθέτησή τους σε πίνακες), τεχνικές βελτιστοποίησης της ιστοσελίδας καθώς και άλλες τεχνολογίες όπως το Redshift και οι γλώσσες προγραμματισμού R και Python.

Το TripAdvisor διαθέτει ένα τεράστιο όγκο δεδομένων από επισκέπτες και τις κριτικές. Ο σωστός συνδυασμός τους μπορεί να έχει ως αποτέλεσμα μια ολοκληρωμένη εμπειρία για τους επισκέπτες και επιπλέον έσοδα για την πλατφόρμα του

TripAdvisor. Για το TripAdvisor, καθώς και για άλλους ταξιδιωτικούς κολοσσούς, η εξόρυξη δεδομένων είναι ο μόνος τρόπος να προσφέρουν συνεχώς στους επισκέπτες τους μια ολοκληρωμένη εμπειρία στον ταξιδιωτικό κλάδο.

4.4 Εξόρυξη Δεδομένων στα Κοινωνικά Δίκτυα

Η Εξόρυξη στα Κοινωνικά Δίκτυα αποτελεί έναν κλάδο της Εξόρυξης του Διαδικτύου, ενώ πολλές φορές χρησιμοποιούνται οι ίδιες μέθοδοι ανακάλυψης πληροφορίας. Οι αναλύσεις σε αυτό το επίπεδο αποσκοπούν στη μελέτη των κοινωνικών οντοτήτων και των αλληλεπιδράσεων και σχέσεων μεταξύ αυτών. Οι σχέσεις αυτές αναπαρίστανται με τη χρήση ενός δικτύου ή γράφου, όπου κάθε κόμβος αντιστοιχεί σε μία οντότητα και κάθε ακμή σε μία σχέση. Μέσα σε αυτούς τους γράφους μπορούμε να διακρίνουμε υπό-γράφους οι οποίοι αναπαριστούν κοινότητες μέσα στα Κοινωνικά Δίκτυα.

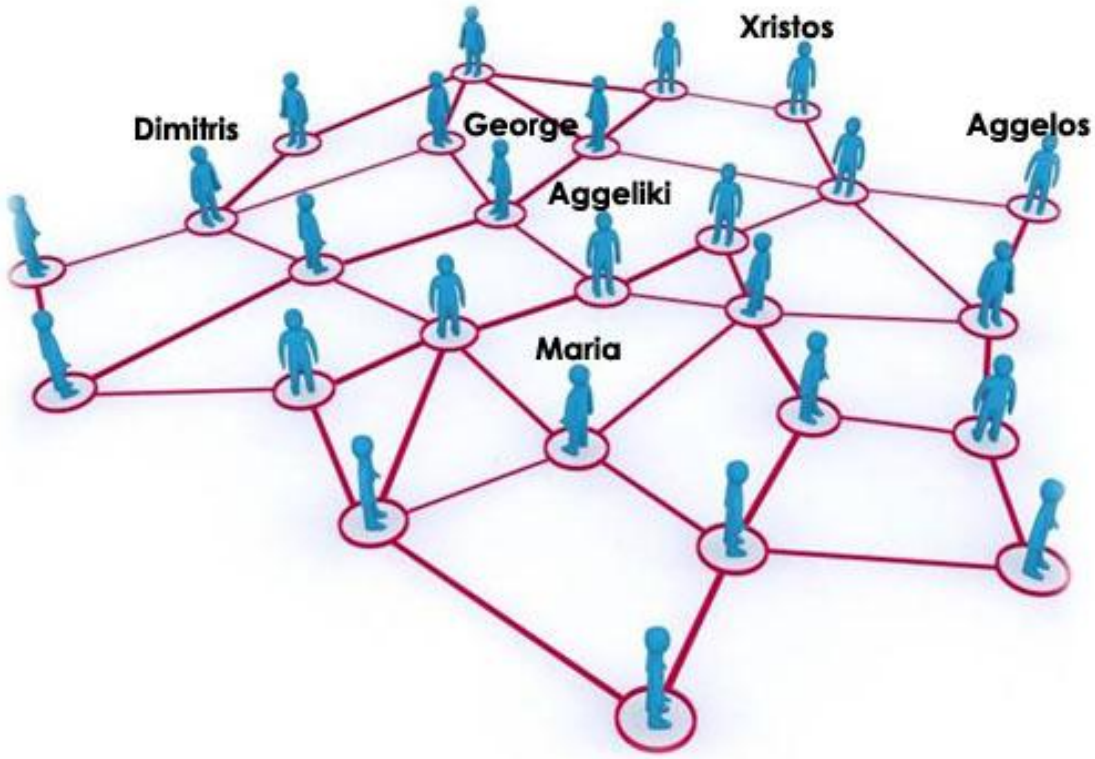
Η εξόρυξη στα Κοινωνικά Δίκτυα είναι ιδιαίτερα σημαντική για την κατανόηση της Εξόρυξης στο Διαδίκτυο, το οποίο πρακτικά αποτελεί μία εικονική κοινωνία ή εικονικό κοινωνικό δίκτυο, όπου κάθε σελίδα αντιστοιχεί σε μία οντότητα και κάθε υπερσύνδεσμος σε μία σχέση. Όπως γίνεται αντιληπτό από αυτή τη δομή, είναι δυνατό να επιτύχουμε την ανάλυση των Κοινωνικών Δικτύων με τη χρήση μεθόδων Εξόρυξης Διαδικτύου, όπως η εξόρυξη δομής και περιεχομένου.

Η συνεχώς αυξανόμενη χρήση των Κοινωνικών Δικτύων, έχει οδηγήσει και τις Μηχανές Αναζήτησης να στρέψουν την προσοχή τους προς την ανάλυση αυτών, προκειμένου να αυξήσουν την αποτελεσματικότητά τους. Δεν είναι τυχαίο ότι σε αρκετές αναζητήσεις αποτελέσματα από τα Κοινωνικά Δίκτυα του Facebook και του Twitter βγαίνουν στις πρώτες θέσεις.

Έχουν αναπτυχθεί νέες μέθοδοι ανάλυσης στα Κοινωνικά Δίκτυα, οι οποίες βασίζονται στις έννοιες της Κεντρικότητας (Centrality) και του Κύρους (Prestige).

Κεντρικότητα (Centrality)

Οι Κοινωνικές Οντότητες οι οποίες είναι εκτενέστερα διασυνδεδεμένες με άλλες Οντότητες, λογίζονται ως σημαντικές ή διακεκριμένες. Στα πλαίσια ενός οργανισμού, άτομα με περισσότερες διασυνδέσεις, επαφές ή επικοινωνίες θεωρούνται πιο σημαντικά από άτομα με λιγότερες. Αυτά τα άτομα ονομάζονται Κεντρικές Οντότητες και ξεχωρίζουν στο γράφο, όπως φαίνεται στο παρακάτω σχήμα.



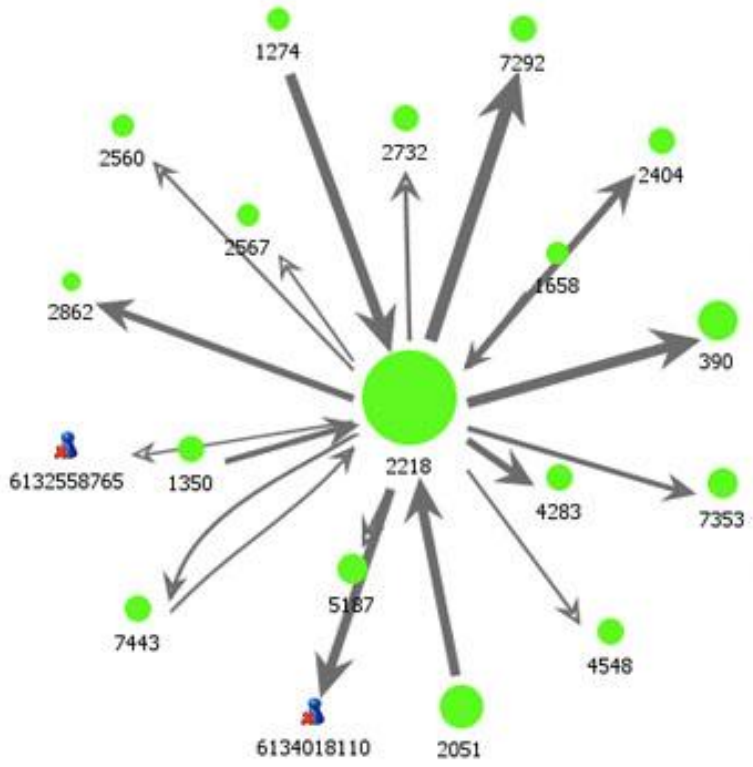
Εικόνα 4.5: Κεντρικότητα. Η Οντότητα Aggeliki ονομάζεται Κεντρική

Κύρος (Prestige)

Το Κύρος αποτελεί μία πιο εξειδικευμένη μέθοδο μέτρησης της σημαντικότητας μίας Κοινωνικής Οντότητας. Εδώ οι διασυνδέσεις δεν προσμετρούνται ενιαία, αλλά διακρίνονται σε εισερχόμενες (in-links) και εξερχόμενες (out-links). Μία Οντότητα λέμε ότι έχει κύρος, όταν υπάρχουν πολλές εισερχόμενες διασυνδέσεις. Για τη

μέτρηση του Κύρους δε λαμβάνονται υπόψη οι εξερχόμενες, οι οποίες είναι σημαντικές μόνο για την Κεντρικότητα.

Γίνεται άμεσα αντιληπτό, ότι προκειμένου να υπάρχουν μετρήσεις Κύρους, είναι απαραίτητο ο γράφος να είναι κατευθυνόμενος, όπως φαίνεται και στο παρακάτω σχήμα.



Εικόνα 4.6: Η Οντότητα 2218 έχει μεγάλο Κύρος

Οι έννοιες της Κεντρικότητας και του Κύρους είναι πολύ σημαντικές και κατά τη διαδικασία της Συναισθηματικής Ανάλυσης στα Κοινωνικά Δίκτυα. Τα μηνύματα που προέρχονται από κεντρικές οντότητες με μεγάλο κύρος είναι ικανά να επηρεάσουν τη γνώμη των υπολοίπων ατόμων του δικτύου τους.

4.5 Γεωγραφικά Συστήματα Πληροφοριών (G.I.S.) και Εξόρυξη Δεδομένων

Αρχικά, όλες οι πληροφορίες που επεξεργάζονται τα συστήματα αυτά, είναι συνδεδεμένες με γεωαναφορές (geo-references). Τυπικές βάσεις δεδομένων μπορεί να περιέχουν δεδομένα τοποθεσίας, όπως για παράδειγμα μία διεύθυνση, ή έναν ταχυδρομικό κώδικα.

Στα Γεωγραφικά Συστήματα Πληροφοριών (G.I.S.), το κύριο μέσο αποθήκευσης και ανάκλησης δεδομένων αποτελούν οι γεωαναφορές, οι οποίες συνήθως εκφράζονται με τη μορφή των γεωγραφικών συντεταγμένων.

Επίσης, τα G.I.S. αποτελούν συστήματα τα οποία ενσωματώνουν τεχνολογίες. Σε αντίθεση με άλλα συστήματα τα οποία είτε αναλύουν φωτογραφίες, είτε επεξεργάζονται δορυφορικά σήματα και δημιουργούν μοντέλα ή χάρτες, τα Γεωγραφικά Συστήματα Πληροφοριών ενσωματώνουν όλες αυτές τις δυνατότητες σε ένα ενιαίο σύστημα.

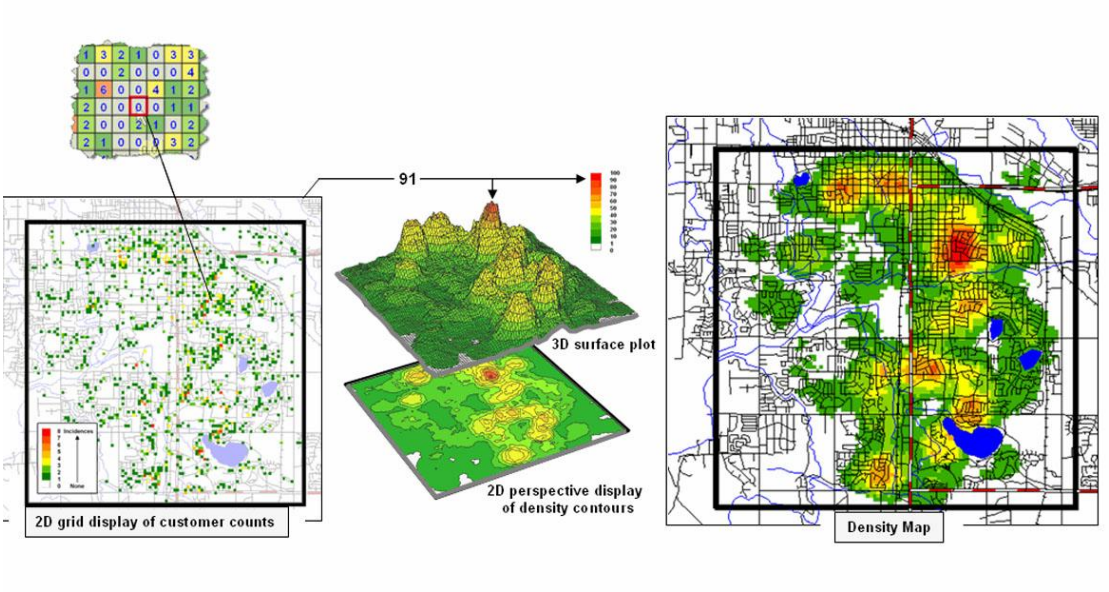
Ο τρόπος με τον οποίο τα δεδομένα συλλέγονται, αποθηκεύονται και επεξεργάζονται μέσα σε ένα G.I.S. σύστημα, υποδεικνύει την κατεύθυνση στην οποία πρέπει να κινηθεί ο ερευνητής ή η επιχείρηση προκειμένου να πάρει μία απόφαση.

Η αυτόματη ανάλυση των χωρικών δεδομένων (spatial data) πρέπει να διαθέτει τις τεχνικές για την ερμηνεία και την αναγνώριση των δομών. Τέτοιες διαδικασίες απαιτούνται ιδιαίτερα σε συστήματα G.I.S. και στην ψηφιακή χαρτογραφία, προκειμένου να αυτοματοποιηθεί η χρονοβόρα αναπροσαρμογή στοιχείων και να παραχθούν απεικονίσεις των στοιχείων σε πολλά επίπεδα. Για να προκύψουν οι σημαντικές πληροφορίες από ένα σύνολο στοιχείων, οι ομοιογενείς δομές σε ένα σύνολο δεδομένων πρέπει να σκιαγραφηθούν. Για την αντιμετώπιση αυτού του προβλήματος έχουν εφαρμοστεί διαφορετικές προσεγγίσεις, όπως είναι π.χ. η ερμηνεία βασισμένη στο μοντέλο, η ερμηνεία βασισμένη στον κανόνα ή η μέθοδος της συσταδοποίησης, οι οποίες αποτελούν μέρος της επιστήμης της Εξόρυξης Δεδομένων (Data Mining).

Στο G.I.S. και στην ψηφιακή χαρτογραφία παρατηρείται αυξανόμενη ζήτηση για τέτοιου είδους τεχνικές. Προκειμένου να επιταχυνθούν οι κύκλοι ανανέωσης και να παραδίδονται οι

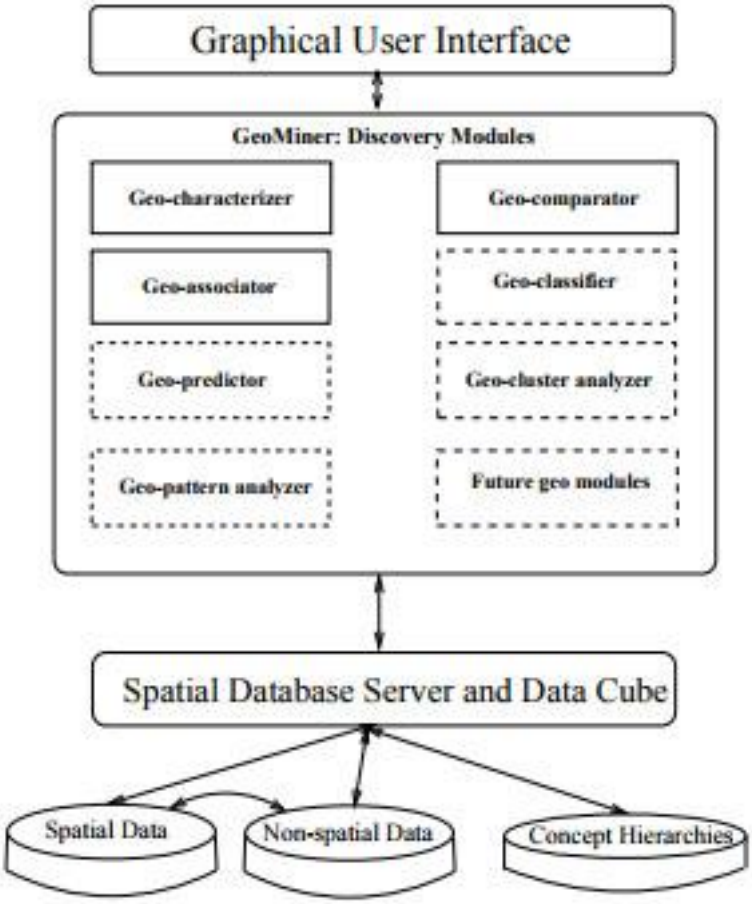
σημαντικές πληροφορίες πιο άμεσα, απαιτούνται εργαλεία και τεχνικές για την αυτοματοποίηση της αρχικής συλλογής και ανανέωσης δεδομένων.

Στο πλαίσιο της συνάθροισης στοιχείων εφαρμόζονται πολλές προσεγγίσεις. Η προαναφερθείσα συσταδοποίηση είναι μια γνωστή τεχνική για την κατανόηση των δεδομένων. Απαιτεί συνήθως πρωταρχικές πληροφορίες που πρέπει να δοθούν από τον χρήστη (input), π.χ. για τη στατιστική κατανομή των στοιχείων ή τον αριθμό συστάδων που θα ανιχνευθούν. Οι υπάρχοντες αλγόριθμοι συσταδοποίησης, όπως οι K-means, PAM, CLARAN, DBSCAN, ROCK κ.α. είναι κατασκευασμένοι έτσι ώστε να συσταδοποιούν δεδομένα σε στατικά μοντέλα. Όλοι αυτοί οι αλγόριθμοι μπορεί να διακοπούν, εάν η επιλογή των αρχικών παραμέτρων είναι λάθος ή είναι ανακριβής.



Εικόνα 4.7: Απεικόνιση δεδομένων από ένα G.I.S. σύστημα

Ένα πρωτότυπο πρόγραμμα για εξόρυξη δεδομένων από ένα χωρικό σύστημα είναι το GeoMiner. Το πρόγραμμα περιλαμβάνει ένα γραφικό τμήμα για την απεικόνιση στοιχείων το οποίο απευθύνεται στον χρήστη (GUI), ενότητες για την εκτέλεση της διερευνητικής ανάλυσης στοιχείων (EDA) και εξόρυξης δεδομένων σε χωρικά δεδομένα, καθώς και έναν κεντρικό υπολογιστή (server) χωρικών βάσεων δεδομένων.



Εικόνα 4.8: Αρχιτεκτονική συστήματος GeoMiner (πηγή: Data Mining in Spatial Data Sets, Hemant Kumar Jerath)

Τα τελευταία χρόνια έχουν χρησιμοποιηθεί σε αρκετές περιπτώσεις εργαλεία εξόρυξης γνώσης από γεωγραφικά συστήματα πληροφοριών.

Χαρακτηριστικό παράδειγμα αποτελεί η χρήση τεχνικών Data Mining σε ένα γεωγραφικό σύστημα που πραγματοποιήθηκε σε Πανεπιστήμιο της Γερμανίας. Το γεωγραφικό σύστημα υλοποιήθηκε για τη βιβλιοθήκη του Πανεπιστημίου και περιελάμβανε χαρτογραφική αναπαράσταση αυτής. Η χρήση εργαλείων Data Mining κατέληξε σε αρκετά χρήσιμα αποτελέσματα.

Σε αυτήν τη βιβλιοθήκη κάθε εγγεγραμμένος χρήστης εξοπλιζόταν με ένα μηχάνημα (ένα είδος beeper) με οθόνη κειμένου. Καθώς προχωρούσε στη βιβλιοθήκη, ανάλογα με τον όροφο, τους διαδρόμους και τα ράφια, λάμβανε ειδοποιήσεις σχετικά με βιβλία

που είχε δανειστεί στο παρελθόν και με βιβλία που είναι πολύ πιθανόν να τον ενδιαφέρουν σε συνάρτηση με την εθνικότητά του, την ηλικία του, το γνωστικό του αντικείμενο κτλ.

Γενικό Συμπέρασμα

Η εργασία είχε σκοπό την βαθύτερη κατανόηση κατ' αρχάς των τεχνικών και αλγορίθμων της εξόρυξης δεδομένων. Μετά από την ανάλυσή τους το συμπέρασμα είναι ότι υπάρχει ένας αλγόριθμος για κάθε πρόβλημα αναζήτησης, προβολής και κατανόησης δεδομένων, και διαφορετικοί τρόποι αντιμετώπισης ταιριάζουν σε διαφορετικές ανάγκες των χρηστών.

Επίσης, καταγράφηκαν οι γνωστότερες τεχνικές εξόρυξης δεδομένων που οι μεγάλες εταιρείες του διαδικτύου αλλά και του σύγχρονου επιχειρηματικού κόσμου γενικότερα χρησιμοποιούν για να κατανοήσουν τις αγοραστικές τάσεις των πελατών τους, να μειώσουν το κόστος λειτουργίας τους και να αυξήσουν τους χρήστες-πελάτες τους.

Σημαντικό ρόλο στον σύγχρονο Παγκόσμιο Ιστό παίζουν και οι μηχανές αναζήτησης. Μια σύγχρονη επιχείρηση θα πρέπει να συμβαδίζει με τις τάσεις και τις τεχνικές των μηχανών αναζήτησης, να είναι φιλική προς αυτές (Search Engine Friendly) και να εφαρμόζει τεχνικές βελτιστοποίησης. Ο σύγχρονος τομέας της εξόρυξης δεδομένων στοχεύει και βοηθάει τις σύγχρονες ψηφιακές πλατφόρμες και ιστοσελίδες προς αυτή την κατεύθυνση. Η εξόρυξη, αξιοποίηση και κατανόηση των διαθέσιμων δεδομένων είναι το σημαντικότερο και το πιο απαραίτητο στοιχείο για την επιτυχία μιας επιχείρησης στον Παγκόσμιο Ιστό. Η διαφοροποίηση μιας επιχείρησης και η αύξηση του ανταγωνιστικού της προφίλ μπορεί να επιτευχθεί μόνο μέσω της αξιοποίησης των δεδομένων.

Νέες πλατφόρμες, νέοι τρόποι διαφήμισης και νέοι τρόποι επικοινωνίας δημιουργήθηκαν χάρη στις εφαρμογές της εξόρυξης δεδομένων. Η χρήση των ιστοσελίδων και γενικά των διαδικτυακών πλατφόρμων έγινε πιο εύκολη τόσο για τους απλούς χρήστες του Παγκόσμιου Ιστού, όσο και για τις ίδιες τις επιχειρήσεις.

Εν κατακλείδι, ο τομέας της Εξόρυξης Δεδομένων στον Παγκόσμιο Ιστό (Web Data Mining) αποτελεί το βασικότερο κομμάτι ύπαρξης

και λειτουργίας του. Ο διαδικτυακός κόσμος οφείλει την ύπαρξή του στην Εξόρυξη Δεδομένων και δεν θα μπορούσε να λειτουργεί αποδοτικά, χωρίς την ύπαρξή του συγκεκριμένου τομέα. Μεγάλες διαδικτυακές πλατφόρμες και σύγχρονα κοινωνικά δίκτυα επικοινωνίας, όπως το Facebook και το Twitter, δεν θα λειτουργούσαν ή η λειτουργία τους θα παρέμενε σε αρκετά πρώιμο επίπεδο χωρίς να υπάρχουν λειτουργίες που τα ξεχωρίζουν και τα διατηρούν στην κορυφή των δημοφιλέστερων ιντερνετικών πλατφόρμων. Το κόστος διαφήμισης των εταιριών θα ήταν πολύ μεγαλύτερο. Η απόκτηση πελατών και η στόχευση του κατάλληλου κοινού μέσω του Παγκόσμιου Ιστού θα ήταν πολύ πιο δύσκολη. Η περιήγηση των χρηστών και η γρήγορη εύρεση του προϊόντος ή της υπηρεσίες που επιθυμούν θα ήταν πιο δύσκολη και χρονοβόρα και γενικά ο κόσμος του Διαδικτύου θα ήταν πολύ διαφορετικός από ότι τον γνωρίζουμε στη σύγχρονη εποχή.

Βιβλιογραφία

Ahlemeyer-Stubbe Andrea, Coleman Shirley, A Practical Guide to Data Mining for Business and Industry, 2014.

Cios, K.J., Pedrycz, W., Swiniarski, R.W., Kurgan, L., Data Mining - A Knowledge Discovery Approach, 2007.

Cooley, R., Web usage mining: discovery and application of interesting patterns from web data, Ph.D. thesis, 2000.

Dunham M.H., Data Mining: Introductory and Advanced Topics, 2004.

D. Michie, D.J. Spiegelhalter, C.C. Taylor, Machine Learning, Neural and Statistical Classification, 1994

Fayyad, U., Piatsesky-Shapiro, G., Smyth, P., & Uthurusamy, R., Advances in Knowledge Discovery and Data Mining, 1996.

Han Jiawei, Kamber Micheline, Jian Pei, Data Mining - Concepts and Techniques 3rd Edition, 2011.

Han Jiawei, Koperski Krzysztof, Stefanovic Nebojsa, GeoMiner: A system Prototype for Spatial Data Mining, 1997.

Haykin, S., Neural Networks: A Comprehensive Foundation 2nd Edition, 1999.

Inmon, H.W., The data warehouse and data mining, 1996.

J.R. Quinlan, Induction of Decision Trees, 2007

Kantardzic Mehmed, Data Mining: Concepts, Models, Methods, and Algorithms, 2003.

Larose T. Daniel, Discovering knowledge in data, 2005.

Markov Zdravko, Data mining the Web, 2006.

Michael J. A. Berry, Gordon S. Linoff, Data Mining Techniques, 1997.

Michael J.A. Berry, Gordon S. Linoff, Data Mining Techniques For Marketing, Sales, and Customer Relationship Management Second Edition, 2004

Mohammed J. Zaki, Wagner Meira, Jr., Data Mining and Analysis, 2014.

Nanopoulos Alexandros, Katsaros Dimitrios, and Manolopoulos Yannis, Exploiting Web Log Mining for Web Cache Enhancement, 2002.

Shuai Yuan, Ahmad Zainal Abidin, Marc Sloan, Jun Wang, Internet Advertising: An Interplay among Advertisers, Online Publishers, Ad Exchanges and Web Users, 2012.

Srivastava, G., Srivastava, R., & Vaishya, R., Distributed GIS Based Decision Support System for Efficiency Evaluation of Education System: A Case Study of Primary School Education System of Bundelkhand Zone, Uttar Pradesh, India, 2015.

Tan, P. N., Steinbach, M., and Kumar, V., Introduction to Data Mining, 2006.

Xiangliang Zhang, King Abdullah University of Science and Technology, AMCS/CS 340: Data Mining Classification I: Decision Tree, 2015.

Wikipedia, Data Mining, https://en.wikipedia.org/wiki/Data_mining

Wikipedia, Geographic Information System, https://en.wikipedia.org/wiki/Geographic_information_system

https://www.tripadvisor.com.gr

https://www.clickatell.com/articles/technology/analytics-data-mining-ai/

A. Silberschatz, H. F. Korth, S. Sudarshan, «Συστήματα Βάσεων Δεδομένων, η Πλήρης Θεωρία των Βάσεων Δεδομένων», 2011.

Βαζιργιάννης Μιχάλης, Χαλκίδη Μαρία, Εξόρυξη Γνώσης από Βάσεις Δεδομένων, 2005.

Βλάχος, Π., Δρόσος, Δ., Νέες Τεχνολογίες και Διαφήμιση**,** 2004.

Κουρής, Γ., Πανεπιστήμιο Πατρών, Εφαρμογή Τεχνικών Data Mining σε Συστήματα Ηλεκτρονικού Εμπορίου, Διδακτορική Διατριβή, 2006.

Μανωλόπουλος Ιωάννης, Παπαδόπουλος Απόστολος, «Συστήματα Βάσεων Δεδομένων – Θεωρία και Πρακτική Εφαρμογή», 2006.

Μαρκέλλου Πηνελόπης, Πολυτεχνική Σχολή Πατρών, Τεχνικές και συστήματα διαχείρισης γνώσης στο διαδίκτυο, 2005.

Πασχαλάκης Θρήσκος, Λάρισα, Εισαγωγή στην Εξόρυξη Δεδομένων, 2016.

Σταλίδης Γέωργιος, Σταλίδης Παναγιώτης, Διαμαντάρας Κωνσταντίνος, Καραπιστόλης Δημήτριος, Αλεξανδρούπολη, Εξόρυξη γνώσης από σχόλια σε τουριστικές ιστοσελίδες και παραγοντική ανάλυση του αισθήματος ικανοποίησης των πελατών για το ξενοδοχείο τους, 2015.