ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

DATA MINING – ΑΠΟ ΤΗΝ ΑΝΑΚΑΛΥΨΗ ΣΤΗΝ ΕΦΑΡΜΟΓΗ

ΘΕΟΧΑΡΗ ΑΙΚΑΤΕΡΙΝΗ Α.Μ.1917

ΒΑΜΒΑΚΟΥΣΗΣ ΒΑΣΙΛΕΙΟΣ Α.Μ.1881

ΕΠΙΒΛΕΠΩΝ: ΑΣΑΡΙΔΗΣ ΗΛΙΑΣ

ΑΝΤΙΡΡΙΟ 2019

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή

Αντίρριο, 01-02-2019

ΕΠΙΤΡΟΠΗ ΑΞΙΟΛΟΓΗΣΗΣ

1.
2.
3.

ΠΕΡΙΛΗΨΗ

Η συλλογή και καταγραφή δεδομένων έχει ξεκινήσει από την αρχή της ανθρωπότητας, έχοντας πάντα στόχο την διευκόλυνση του ανθρώπινου είδους. Ωστόσο, εξαιτίας των γρήγορα αυξανόμενων τεχνολογικών δραστηριοτήτων που εκτυλίσσονται γύρω μας καθημερινά, η ταχύτητα με την οποία δημιουργούνται είναι φαινομενική και ο όγκος απερίγραπτος. Αυτά τα τεράστια σύνολα δεδομένων μας έχουν φέρει στην εποχή των *Big Data (Μεγάλα Δεδομένα)*, μια από τις πιο σημαντικές πτυχές στην ιστορία ανάπτυξης.

Παρόλα αυτά, η εύρεση και αποθήκευση δεδομένων δεν σημαίνει ότι μας παρέχει χρήσιμες και απαραίτητες πληροφορίες. Όσο ο κόσμος μεγαλώνει σε πολυπλοκότητα, τα δεδομένα είναι πλέον δύσκολο να επεξεργαστούν και να αναλυθούν με παραδοσιακούς τρόπους και χρειάζονται τα κατάλληλα εργαλεία για την ανακάλυψη αυτής της κρυμμένης πολύτιμης γνώσης. Αυτό το χάσμα που έχει δημιουργηθεί μεταξύ πληροφορίας και δεδομένων μπορεί να μειωθεί με την εφαρμογή του *Data Mining (Εξόρυξη Δεδομένων)*.

Αυτός ο κλάδος θεωρείται σήμερα ένας από τους ταχύτερα αναπτυσσόμενους στην βιομηχανία των υπολογιστών, καθώς και μια από τις πιο υποσχόμενες διεπιστημονικές εξελίξεις στην τεχνολογία των πληροφοριών. Οργανισμοί, επιχειρήσεις και ιδρύματα αφιερώνουν πολλούς πόρους για την ανάλυση των κολοσσιαίων *Βάσεων Δεδομένων* τους, ώστε να μπορέσουν να επιβιώσουν στον σημερινό ανταγωνιστικό κόσμο. Η εύρεση χρήσιμων δεδομένων, η ταυτοποίηση κρυμμένων προτύπων και η σωστή εκμετάλλευση της γνώσης που εξάγεται, είναι αυτό που φέρνει την Εξόρυξη Δεδομένων στην πρώτη γραμμή των νέων επιχειρησιακών τεχνολογιών.

ABSTRACT

The collection and recording of data has started from the beginning of humanity, always aiming at facilitating the human species. Nevertheless, due to the rapidly growing technological activities unfolding around us every day, the speed at which they are created is phenomenal and the volume is indescribable. These huge datasets have brought us into the era of Big Data, one of the most important aspects in story growth.

However, finding and storing data does not mean that it provides us with useful and necessary information. As the world grows in complexity, data is now hard to process and to analyze in traditional ways and they need the right tools to discover this hidden valuable knowledge. This gap created between information and data can be reduced by implementing Data Mining.

This discipline is currently considered one of the fastest growing in the computer industry, as well as one of the most promising interdisciplinary developments in information technology. Organizations, businesses and institutions devote a lot of resources to analyze their colossal databases so they can survive in today's competitive world. Finding useful data, identifying hidden patterns and exploiting the knowledge that is being extracted is what brings Data Mining to the forefront of new business technologies.

ΕΥΧΑΡΙΣΤΙΕΣ

Η παρούσα πτυχιακή εργασία εκπονήθηκε για το τμήμα Μηχανικών Πληροφορικής Τ.Ε. του Τ.Ε.Ι. Δυτικής Ελλάδος. Η ιδέα αυτής της εργασίας προήλθε από τον καθηγητή μας κ. Ηλία Ασαρίδη, ο οποίος με την βοήθεια και καθοδήγηση του μας ώθησε στο να βγάλουμε το καλύτερο δυνατό αποτέλεσμα της.

Ιδιαίτερες ευχαριστίες θα θέλαμε να απευθύνουμε στις οικογένειες και φίλους μας, οι οποίοι στάθηκαν δίπλα μας και μας στήριξαν όχι μόνο αυτόν τον χρόνο αλλά καθ'όλη την διάρκεια των σπουδών μας.

Πίνακας περιεχομένων

Περιεχόμενα εικόνων

ΕΙΣΑΓΩΓΗ

Αυτή τη στιγμή ζούμε στην εποχή των μεγάλων δεδομένων. Ο όρος "μεγάλα δεδομένα" φαίνεται ότι χρησιμοποιήθηκε για πρώτη φορά, σύμφωνα με την επίκαιρη έννοια του, στα τέλη της δεκαετίας του 1990. Το πρώτο ακαδημαϊκό έγγραφο παρουσιάστηκε το 2000 και δημοσιεύθηκε το 2003 από τον Francis X. Diebolt – "Big Data Dynamic Factor Models for Macroeconomic Measurement and Forecasting" – αλλά ο έπαινος δίνεται σε μεγάλο βαθμό στον John Mashey, τον επικεφαλή επιστήμονα της SGI, ως το πρώτο άτομο που χρησιμοποίησε τον όρο "μεγάλα δεδομένα". Στα τέλη της δεκαετίας του 1990, ο Mashey έδωσε μια σειρά συνομιλιών σε μικρές ομάδες σχετικά με αυτό το παλιρροιακό κύμα μεγάλων δεδομένων που έρχεται. Η εποχή των μεγάλων δεδομένων είναι μια εποχή που περιγράφεται από την ταχέως αναπτυσσόμενη ποσότητα δεδομένων, η οποία είναι πολύ μεγαλύτερη από αυτήν που οι περισσότεροι άνθρωποι θα φανταζόντουσαν ποτέ.

Αυτό που ορίζει το σήμερα ως την εποχή των big data είναι ότι οι εταιρείες, οι κυβερνήσεις και οι μη κερδοσκοπικές οργανώσεις έχουν βιώσει μια αλλαγή στη συμπεριφορά. Θέλουν πλέον να αρχίσουν να χρησιμοποιούν όλα τα δεδομένα που είναι δυνατόν να συλλέξουν, για έναν τρέχοντα ή μελλοντικό άγνωστο σκοπό, ώστε να βελτιώσουν την επιχείρησή τους. Πιστεύεται, μέσω έρευνας και μελετών, ότι οι οργανισμοί που χρησιμοποιούν τα δεδομένα για την λήψη αποφάσεων, με την πάροδο του χρόνου παίρνουν όντως καλύτερες αποφάσεις, γεγονός που οδηγεί σε μια ισχυρότερη, πιο βιώσιμη επιχείρηση. Δεδομένου ότι η ταχύτητα με την οποία δημιουργούνται τα δεδομένα αυξάνεται με τόσο ταχύ ρυθμό, οι εταιρείες ανταποκρίθηκαν διατηρώντας κάθε στοιχείο που θα μπορούσαν ενδεχομένως να καταγράψουν και εκτιμώντας τις μελλοντικές δυνατότητες αυτών των δεδομένων υψηλότερα από ό, τι είχαν στο παρελθόν.

Το 1995, η Ευρωπαϊκή Ένωση στη νομοθεσία περί απορρήτου, όρισε τα προσωπικά δεδομένα ως οποιαδήποτε πληροφορία που θα μπορούσε να προσδιορίσει ένα πρόσωπο, άμεσα ή έμμεσα. Η Διεθνής Εταιρεία Δεδομένων (IDC) εκτιμά ότι 2,8 zettabytes δεδομένων δημιουργήθηκαν το 2012 και ότι ο όγκος των δεδομένων που παράγονται κάθε χρόνο θα διπλασιαστεί μέχρι το 2015. Με έναν τόσο μεγάλο αριθμό είναι δύσκολο να κατανοηθεί πόσα από αυτά τα δεδομένα αφορούν έναν και μόνο άνθρωπο. Ουσιαστικά, περίπου 5 gigabytes δεδομένων ημερησίως αναλογούν στον μέσο Αμερικανό υπάλληλο. Αυτά τα δεδομένα αποτελούνται από τα μηνύματα ηλεκτρονικού ταχυδρομείου, τις ταινίες που μεταφορτώνονται, τα υπολογιστικά φύλλα του Excel κλπ. Σε αυτά τα δεδομένα περιλαμβάνονται επίσης και εκείνα που παράγονται καθώς μεταφέρονται πληροφορίες σε όλο

το Διαδίκτυο. Πολλά από αυτά τα δεδομένα που δημιουργούνται δεν εμφανίζονται απευθείας στον χρήστη, αλλά αποθηκεύονται. Μερικά παραδείγματα είναι το βίντεο της κάμερας κυκλοφορίας, οι συντεταγμένες GPS από τα κινητά ή οι συναλλαγές διοδίων μέσω E-ZPass.

Πριν ξεκινήσει η μεγάλη εποχή δεδομένων, οι επιχειρήσεις απέδιδαν σχετικά χαμηλή αξία στα δεδομένα που αποκτούσαν και δεν είχαν άμεση αξία. Αργότερα βέβαια, η επένδυση αυτή στη συλλογή και αποθήκευση δεδομένων για τη δυνητική μελλοντική της αξία άλλαξε και οι οργανώσεις κατέβαλαν συνειδητή προσπάθεια για να κρατήσουν κάθε πιθανό κομμάτι δεδομένων. Η επιτυχία στην εύρεση της αξίας οδήγησε σε συλλογή περισσότερων δεδομένων και ούτω καθεξής. Μερικά από τα δεδομένα που αποθηκεύτηκαν ήταν αδιέξοδα, αλλά πολλές φορές επιβεβαιώθηκε ότι όσο περισσότερα δεδομένα έχουμε, τόσο το καλύτερο. Άλλη σημαντική αλλαγή που έλαβε χώρα, ήταν η ταχεία ανάπτυξη, η δημιουργία και η ωρίμανση των τεχνολογιών αποθήκευσης, χειρισμού και ανάλυσης αυτών των δεδομένων με νέους και αποτελεσματικούς τρόπους.

Παρόλα αυτά, η πρόκληση δεν είναι η συλλογή των δεδομένων αλλά η εύρεση των σωστών δεδομένων και η χρήση υπολογιστών για την αύξηση της γνώσης στον τομέα μας, καθώς και ο προσδιορισμός προτύπων που δεν είδαμε ή δεν μπορούσαμε να βρούμε προηγουμένως.

Ορισμένες βασικές τεχνολογίες και διαταραχές της αγοράς μας οδήγησαν σε αυτό το σημείο όπου ο όγκος των δεδομένων που συλλέγονται, αποθηκεύονται και εξετάζονται σε αναλυτικές δραστηριότητες έχει αυξηθεί με τεράστιο ρυθμό. Αυτό οφείλεται σε πολλούς παράγοντες, όπως είναι το πρωτόκολλο Internet Protocol 6 (IPv6), ο βελτιωμένος τηλεπικοινωνιακός εξοπλισμός, οι τεχνολογίες όπως η RFID, το μειωμένο κόστος ανά μονάδα παραγωγής ηλεκτρονικών ειδών, τα κοινωνικά μέσα και το Διαδίκτυο. (Dean, 2014, pp. 1-5)

Κεφάλαιο 1ο

Big Data – Μεγάλα Δεδομένα

1.1. Τι είναι τα Big Data;

Τα *μεγάλα δεδομένα* παράγονται από μια αυξανόμενη πληθώρα πηγών, συμπεριλαμβανομένων των κλικ στο διαδίκτυο, των συναλλαγών μέσω κινητού τηλεφώνου, του περιεχομένου που δημιουργείται από τους χρήστες και των κοινωνικών μέσων ενημέρωσης, καθώς και του σκόπιμα δημιουργούμενου περιεχομένου μέσω δικτύων αισθητήρων ή επιχειρηματικών συναλλαγών, όπως συναλλαγές αγοράς. Η υγειονομική περίθαλψη, η μηχανολογία, η διαχείριση των επιχειρήσεων, το βιομηχανικό διαδίκτυο και η οικονομία προσθέτουν σημαντικά στην επιπλέον εξάπλωση τους. Αυτά τα δεδομένα απαιτούν τη χρήση ισχυρών υπολογιστικών τεχνικών, για να αποκαλύψουν τις τάσεις και τα πρότυπα που υπάρχουν μέσα και μεταξύ αυτών των εξαιρετικά μεγάλων κοινωνικοοικονομικών συνόλων δεδομένων. Οι νέες πληροφορίες που αντλήθηκαν από την εξαγωγή αυτών των δεδομένων, μπορούν να συμπληρώσουν ουσιαστικά τις στατιστικές, τις έρευνες και τις πηγές αρχειακών δεδομένων που παραμένουν σε μεγάλο βαθμό στατικές, προσθέτοντας βάθος και διορατικότητα από συλλογικές εμπειρίες, περιορίζοντας έτσι τόσο τις πληροφορίες όσο και τις χρονικές διαφορές.

Ίσως ο λάθος χαρακτηρισμός να βρίσκεται στο μέγεθος (bigness) των μεγάλων δεδομένων, που πάντα προσελκύει την προσοχή των ερευνητών στο μέγεθος του συνόλου δεδομένων. Μεταξύ των επαγγελματιών, υπάρχει μια αναδυόμενη συζήτηση ότι το "μεγάλο" δεν είναι πλέον η καθοριστική παράμετρος, αλλά μάλλον το πόσο "έξυπνο" είναι, δηλαδή τις πληροφορίες που ο όγκος των δεδομένων μπορεί λογικώς να παράσχει. Μπορεί να γίνει ανάλυση των κοινωνικών δικτύων και των κοινωνικών συμπεριφορών των ατόμων, χαρτογραφώντας μοτίβα κινητικότητας σε φυσικές διατάξεις των χώρων εργασίας με τη χρήση αισθητήρων, ή να γίνει ανάλυση της συχνότητας της χρήσης της αίθουσας συνεδριάσεων με απομακρυσμένους αισθητήρες που παρακολουθούν πρότυπα εισόδου και εξόδου, που θα μπορούσαν να παρέχουν πληροφορίες ανάλογα με τις ανάγκες στην επικοινωνία και στον συγχρονισμό βάσει της πολυπλοκότητας ενός project και της λήξης μιας προθεσμίας. Αυτά τα πολύ μικρά δεδομένα (micro data) παρέχουν έναν πλούτο ατομικών συμπεριφορών και δράσεων που δεν έχουν αξιοποιηθεί πλήρως στην έρευνα διαχείρισης. Είτε πρόκειται για μεγάλα είτε για έξυπνα δεδομένα, η χρήση δεδομένων μεγάλης κλίμακας για

την πρόβλεψη της ανθρώπινης συμπεριφοράς κερδίζει έδαφος στην πρακτική πολιτικής των επιχειρήσεων και της κυβέρνησης, καθώς και σε επιστημονικούς τομείς όπου συγκλίνουν οι φυσικές και κοινωνικές επιστήμες (social physics). (GEORGE, HAAS and PENTLAND, 2014, pp. 2-3)

Η ποσότητα των δεδομένων που παράγονται κάθε μέρα στον κόσμο εκρήγνυται. Ο αυξανόμενος όγκος των ψηφιακών και κοινωνικών μέσων και το διαδίκτυο των πραγμάτων (IoT), τροφοδοτεί ακόμη περισσότερο. Ο ρυθμός της αύξησης των δεδομένων είναι εκπληκτικός και τα δεδομένα αυτά έρχονται με ταχύτητα, με ποικιλία (όχι απαραίτητα δομημένα δεδομένα), και περιέχει πλούτο πληροφοριών που μπορεί να είναι ένα κλειδί για να κερδίσουμε ένα πλεονέκτημα σε ανταγωνιστικές επιχειρήσεις. Η ικανότητα ανάλυσης αυτού του τεράστιου όγκου δεδομένων φέρνει μια νέα εποχή της αύξησης της παραγωγικότητας, της καινοτομίας και του πλεονάσματος του καταναλωτή.

Συμπερασματικά, και σύμφωνα με όλα τα παραπάνω, ένας ορισμός που θα μπορούσε να δοθεί είναι ο εξής:

“Μεγάλα δεδομένα είναι ο όρος για μια συλλογή συνόλων δεδομένων τόσο μεγάλη και περίπλοκη που γίνεται δύσκολο να επεξεργαστεί χρησιμοποιώντας παραδοσιακά εργαλεία διαχείρισης βάσεων δεδομένων ή εφαρμογές επεξεργασίας δεδομένων. Οι προκλήσεις περιλαμβάνουν τους τομείς της σύλληψης, της επεξεργασίας, της αποθήκευσης, της αναζήτησης, της μεταφοράς, της ανάλυσης και της οπτικοποίησης αυτών των δεδομένων”. (Raste, 2014, p. 3)

1.2. Ιστορική αναδρομή

1991 : Το Internet ή ο Παγκόσμιος Ιστός γεννιέται. Το Hypertext Transfer Protocol (HTTP) καθίσταται ως ο τυπικός τρόπος για την ανταλλαγή πληροφοριών σε αυτό το νέο μέσο.

1995 : Η Sun κυκλοφορεί την πλατφόρμα Java. Η Java, η οποία εφευρέθηκε το 1991, έχει γίνει η δεύτερη πιο δημοφιλής γλώσσα, μετά την C. Κυριαρχεί στον χώρο των Web εφαρμογών και είναι το de facto πρότυπο για τις μεσαίου επιπέδου εφαρμογές. Αυτές οι εφαρμογές είναι η πηγή για την καταγραφή και την αποθήκευση του web traffic.

Το Global Positioning System (GPS) καθίσταται πλήρως λειτουργικό. Αναπτύχθηκε αρχικά από την DARPA (Defense Advanced Research Projects Agency) για στρατιωτικές

εφαρμογές στις αρχές της δεκαετίας του 1970. Αυτή η τεχνολογία υπάρχει πλέον παντού, όπως στις εφαρμογές για αυτοκίνητο και την αεροπορική πλοήγηση.

1998 : Ο Carlo Strozzi αναπτύσσει μια ανοιχτού κώδικα σχεσιακή βάση δεδομένων και την ονομάζει NoSQL. Δέκα χρόνια αργότερα, ένα κίνημα για την ανάπτυξη βάσεων δεδομένων NoSQL που θα μπορούν να δουλέψουν με μεγάλα και αδόμητα σύνολα δεδομένων αποκτά δυναμική.

Η Google ιδρύεται από τον Larry Page και τον Sergey Brin, οι οποίοι εργάστηκαν για περίπου ένα χρόνο σε ένα πρόγραμμα μηχανών αναζήτησης του Stanford που ονομάζεται BackRub.

1999 : Ο Kevin Ashton, συνιδρυτής του Auto-ID Center στο Ινστιτούτο Τεχνολογίας της Μασαχουσέτης (MIT), εφευρίσκει τον όρο "The Internet of Things".

2001 : Η παγκόσμια, ψηφιακή, διαδικτυακή, ελεύθερου περιεχομένου, εγκυκλοπαίδεια Wikipedia, λανσάρεται.

2002 : Η έκδοση 1.1 της Bluetooth προδιαγραφής κυκλοφορεί από το Institute of Electrical and Electronics Engineers (IEEE). Το Bluetooth είναι ένα πρότυπο ασύρματης τεχνολογίας για τη μεταφορά δεδομένων σε μικρές αποστάσεις. Η εξέλιξη αυτών των προδιαγραφών και η υιοθεσία της οδηγεί σε μια ολόκληρη σειρά φορητών συσκευών, που επικοινωνούν μεταξύ της συσκευής και άλλου υπολογιστή. Σήμερα, σχεδόν κάθε φορητή συσκευή διαθέτει δέκτη Bluetooth.

2003 : Σύμφωνα με μελέτες της IDC και της EMC, το ποσό των δεδομένων που δημιουργήθηκαν το 2003 ξεπερνά την ποσότητα των δεδομένων που είχαν δημιουργηθεί μέχρι τότε σε όλη την ανθρώπινη ιστορία. Εκτιμάται ότι 1,8 zettabytes (ZB) δημιουργήθηκαν μόνο το 2011.

Το LinkedIn, η δημοφιλής ιστοσελίδα κοινωνικής δικτύωσης για επαγγελματίες, ξεκινάει. Το 2013, ο ιστότοπος είχε περίπου 260 εκατομμύρια χρήστες.

2004 : Η Wikipedia φτάνει τα 500.000 άρθρα τον Φεβρουάριο. Επτά μήνες αργότερα φτάνει τα 1 εκατομμύριο.

Το Facebook, η υπηρεσία κοινωνικής δικτύωσης, ιδρύεται από τον Mark Zuckerberg και από άλλους στο Κέιμπριτζ της Μασαχουσέτης. Το 2013, ο ιστότοπος είχε περισσότερους από 1,15 δισεκατομμύρια χρήστες.

2005 : Το Apache Hadoop project δημιουργήθηκε από τον Doug Cutting και τον Mike Caferella. Ο πλέον διάσημος κίτρινος ελέφαντας γίνεται ένα θεμελιώδες μέρος σχεδόν όλων των στρατηγικών των μεγάλων δεδομένων.

Το Εθνικό Συμβούλιο Επιστημών προτείνει στην National Science Foundation (NSF) να δημιουργήσουν ένα νέο μονοπάτι για καριέρα, τους υψηλής ποιότητας data scientists, για να διαχειριστούν την αυξανόμενη συλλογή ψηφιακών πληροφοριών.

2007 : Η Apple κυκλοφορεί το iPhone και δημιουργεί μια ισχυρή αγορά καταναλωτών για smartphones.

2008 : Ο αριθμός των συσκευών που είναι συνδεδεμένες στο Διαδίκτυο υπερβαίνει το παγκόσμιο πληθυσμό.

2011 : Ο υπολογιστής Watson της IBM ανιχνεύει και αναλύει 4 terabytes (200 εκατομμύρια σελίδες) δεδομένων σε δευτερόλεπτα.

Η UnQL ξεκινάει, μια query language για τις βάσεις δεδομένων NoSQL.

2012 : Η κυβέρνηση Ομπάμα ανακοινώνει την Big Data Research και Development Initiative, αποτελούμενη από 84 προγράμματα σε έξι τμήματα.

Η IDC και η EMC εκτιμούν ότι θα δημιουργηθούν 2,8 ZB δεδομένων το 2012. Προβλέπεται ότι ο ψηφιακός κόσμος θα κατέχει μέχρι το 2020 40 ZB, 57 φορές ο αριθμός των σπόρων των άμμων σε όλες τις παραλίες του κόσμου.

2013 : Ξεκινά ο εκδημοκρατισμός των δεδομένων. Με τα smartphones, τα tablet, και τα Wi-Fi, ο καθένας παράγει δεδομένα σε εκπληκτικές τιμές. Περισσότερα άτομα έχουν πρόσβαση σε μεγάλους όγκους δημόσιων δεδομένων και χρησιμοποιούν τα δεδομένα πιο δημιουργικά. (Dean, 2014, pp. 5-8)

1.3. Τα Big Data σήμερα

Από τις παραστάσεις και τεχνολογικές δράσεις που λαμβάνουν χώρα γύρω μας καθημερινά, είναι προφανές ότι σήμερα βιώνουμε μια περίοδο που η έννοια και οι εφαρμογές των Μεγάλων Δεδομένων βρίσκονται σε πλήρη ισχύ και εφαρμογή.

Πιο συγκεκριμένα, μέσω της χρήσης κατάλληλων οργάνων (instrumentation), ο άνθρωπος βρίσκεται πλέον σε θέση να "επικοινωνήσει" με περισσότερα πράγματα και σε μεγαλύτερο εύρος, με αποτέλεσμα να υπάρχει και η δυνατότητα για αποθήκευση

μεγαλύτερου εύρους δεδομένων, όπως αυτή ανακύπτει από την δυνατότητα για επεξεργασία μεγαλύτερου όγκου πρώτης ύλης. Επιπρόσθετα, μέσω της αυξημένης δυνατότητας για επικοινωνία και της ανάπτυξης των αντίστοιχων τεχνολογιών, οι άνθρωποι και οι πληροφορίες σήμερα είναι δυνατό να βρίσκονται σε πλήρη και συνεχή διασύνδεση (interconnection) και διάδραση (interaction) . Υπό αυτό το πρίσμα, έχουμε πλέον φτάσει σε μια δυνατότητα για διασυνδεσιμότητα τύπου Machine to Machine, με τη διασύνδεση να παρουσιάζει χαρακτηριστικά διασύνδεσης όμοια με αυτά δύο τυποποιημένων μηχανών, δηλαδή συνέχεια αυξημένη διακίνηση πληροφορίας και αυξημένη δυνατότητα για ανάδραση. Αυτή η δυνατότητα άλλωστε είναι υπεύθυνη για τα γεωμετρικώς αυξανόμενα ποσοστά αποθήκευσης που παρατηρούνται κάθε χρόνο, όσον αφορά στα δεδομένα και κυρίως στα ψηφιακά δεδομένα.

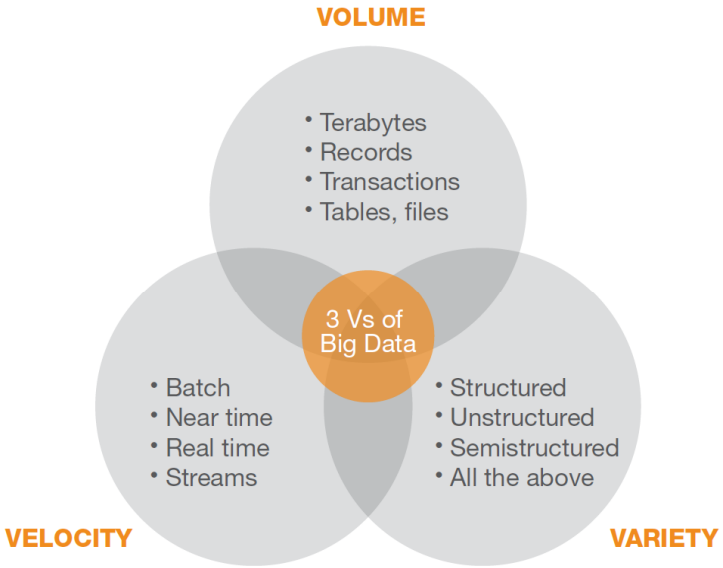
Προχωρώντας ένα βήμα πιο πέρα, με την αλματώδη ανάπτυξη της τεχνολογίας που έχει επιτελεστεί, πλέον υπάρχει δυνατότητα για αγορά σημαντικών διατάξεων που σχετίζονται άμεσα με εφαρμογές διακίνησης και ανταλλαγής πληροφοριών, με εξέχουσες τα ολοκληρωμένα κυκλώματα, σε πολύ χαμηλές τιμές. Αυτή η παράμετρος δίνει τη δυνατότητα να προστεθεί η έννοια της ευφυΐας σχεδόν σε κάθε εφαρμογή και δράση, με επακόλουθο να βελτιστοποιείται ακόμη περισσότερο η διαδικασία της ανταλλαγής και αποθήκευσης πληροφοριών και δεδομένων.

Με βάση την παραπάνω παρατήρηση, καταλήγουμε στη βασική διαπίστωση η οποία καθορίζει σήμερα την έννοια των Μεγάλων Δεδομένων στις παγκόσμιες εφαρμογές και δράσεις. Σύμφωνα με αυτή, στο σημερινό παγκόσμιο γίγνεσθαι, παρατηρούμε συνεχείς ενέργειες και αλληλεπιδράσεις στους διάφορους τομείς της ανθρώπινης δραστηριότητας, όπου η ταχύτητα ανταλλαγής πληροφοριών και δεδομένων, ο όγκος αυτών των ποσοτήτων, καθώς και η ποικιλία που τις χαρακτηρίζει, καθορίζουν σήμερα το βαθμό συνθετότητας και τις παραμέτρους της έννοιας των Μεγάλων Δεδομένων. (Τσικριτέας, 2015, pp. 12-13)

1.4. Τα 3Vs των Big Data

Οι περισσότεροι όροι των μεγάλων δεδομένων επικεντρώνονται στο μέγεθος των δεδομένων που είναι αποθηκευμένα. Το μέγεθος έχει σημασία, αλλά υπάρχουν και άλλες σημαντικές ιδιότητες των μεγάλων δεδομένων, όπως είναι η *ποικιλία (variety)* και η *ταχύτητα (velocity)*. Τα τρία Vs των μεγάλων δεδομένων (όγκος, ποικιλία και ταχύτητα) αποτελούν έναν περιεκτικό ορισμό, και καταρρίπτουν το μύθο ότι τα μεγάλα δεδομένα αφορούν μόνο

τον όγκο των δεδομένων. Επιπλέον, κάθε ένα από τα τρία Vs έχει τις δικές του διακλαδώσεις στον κλάδο της ανάλυσης. (Russom, 2011, p. 6)



Εικόνα 1.1 : Τα 3Vs των Big Data.

1.4.1. Όγκος – Volume

Ο όγκος είναι η πιο δύσκολη πτυχή των Big Data, καθώς επιβάλλει την ανάγκη για κλιμακούμενη αποθήκευση και μια κατανεμημένη προσέγγιση στο querying. Οι μεγάλες επιχειρήσεις έχουν ήδη συγκεντρώσει και αρχειοθετήσει μια μεγάλη ποσότητα δεδομένων τα τελευταία χρόνια. Θα μπορούσε να είναι με τη μορφή αρχείων καταγραφής συστήματος, διατήρησης αρχείων κλπ. Η ποσότητα αυτή των δεδομένων φτάνει εύκολα στο σημείο όπου συμβατικά συστήματα διαχείρισης βάσεων δεδομένων μπορεί να μην είναι σε θέση να την χειριστούν. Λύσεις που βασίζονται στις αποθήκες δεδομένων ενδέχεται να μην έχουν απαραίτητα τη δυνατότητα επεξεργασίας και ανάλυσης αυτών των δεδομένων λόγω έλλειψης παράλληλης αρχιτεκτονικής επεξεργασίας.

Πολλά μπορούν να προκύψουν από δεδομένα κειμένου (text data), από τοποθεσίες ή αρχεία καταγραφής (log files) όπως, για παράδειγμα, τα πρότυπα επικοινωνίας μέσω ηλεκτρονικού ταχυδρομείου, οι προτιμήσεις των καταναλωτών και οι τάσεις στα δεδομένα συναλλαγών, οι έρευνες ασφαλείας κ.α. Τα χωρικά και χρονικά (time-stamped) δεδομένα απορροφούν χώρο αποθήκευσης γρήγορα. Οι Big Data τεχνολογίες προσφέρουν μια λύση

στο να δημιουργήσουν αξία από αυτά τα μαζικά, και προηγουμένως αχρησιμοποίητα / δύσκολα να επεξεργαστούν, δεδομένα.

1.4.2. Ταχύτητα – Velocity

Τα δεδομένα ρέουν μέσα στους οργανισμούς με μεγάλη ταχύτητα. Τεχνολογίες διαδικτύου και κινητής τηλεφωνίας επέτρεψαν τη παραγωγή ροής δεδομένων στους παρόχους. Οι online αγορές έχουν φέρει επανάσταση στις αλληλεπιδράσεις μεταξύ των καταναλωτών και των παρόχων. Οι διαδικτυακοί λιανοπωλητές μπορούν τώρα να διατηρούν αρχεία καταγραφής και να έχουν πρόσβαση σε κάθε δράση των πελατών τους, καθώς επίσης μπορούν να διατηρούν το ιστορικό και να θέλουν να αξιοποιήσουν γρήγορα αυτές τις πληροφορίες, συνιστώντας προϊόντα και τοποθετώντας τον οργανισμό σε πρωτοποριακό επίπεδο. Οι online marketing οργανισμοί αποκομίζουν μεγάλο πλεονέκτημα με τη δυνατότητα να αποκτούν πληροφορίες στιγμιαία. Με την εφεύρεση της εποχής των smartphones, υπάρχουν ακόμα περισσότερα δεδομένα που βασίζονται στη τοποθεσία που παράγονται και γίνεται σημαντικό να μπορούμε να εκμεταλλευτούμε αυτό το τεράστιο ποσό δεδομένων.

1.4.3. Ποικιλία – Variety

Όλα αυτά τα δεδομένα που παράγονται από τα κοινωνικά και ψηφιακά μέσα είναι σπάνια δομημένα (structured data). Τα μη δομημένα έγγραφα κειμένου, τα βίντεο, τα ηχητικά δεδομένα, οι εικόνες, οι οικονομικές συναλλαγές, οι αλληλεπιδράσεις σε κοινωνικούς ιστότοπους, αποτελούν παραδείγματα μη δομημένων δεδομένων (unstructured data). Οι συμβατικές βάσεις δεδομένων υποστηρίζουν “μεγάλα αντικείμενα” (LOB's), αλλά έχουν τους περιορισμούς τους εάν δεν διανεμηθούν. Αυτά τα δεδομένα είναι δύσκολο να χωρέσουν στις συμβατικές, καθαρά σχετικές δομές διαχείρισης βάσεων δεδομένων και δεν είναι πολύ φιλικά δεδομένα ως προς την ενσωμάτωση, και για αυτό χρειάζονται πολύ χειρισμό προτού τα χρησιμοποιήσουν οι εφαρμογές. Και αυτό οδηγεί σε απώλεια πληροφοριών. Εάν τα δεδομένα χαθούν, τότε είναι μια απώλεια που δεν μπορεί να ανακτηθεί. Τα μεγάλα δεδομένα, από την άλλη πλευρά, τείνουν να διατηρούν όλα τα δεδομένα αφού τα περισσότερα από αυτά γράφονται μία φορά και διαβάζουν πολλές φορές τον τύπο των δεδομένων. Τα Big Data πιστεύουν ότι θα μπορούσαν να υπάρχουν πληροφορίες κρυμμένες σε κάθε κομμάτι δεδομένων. (Raste, 2014, pp. 4-5)

1.5. Αναλύοντας τα Big Data

Ως πηγή δεδομένων είναι εξίσου σημαντική η ανάλυση μεθοδολογιών και τα πρότυπα αποδεικτικών στοιχείων που θα ήταν αποδεκτά στους μελετητές του management για δημοσίευση. Όπως συμβαίνει με κάθε αναδυόμενη επιστήμη, υπάρχει πιθανότητα συμβιβασμού μεταξύ της θεωρητικής και της εμπειρικής συμβολής, καθώς και της αυστηρότητας με την οποία αναλύονται τα δεδομένα. Η τυπική στατιστική προσέγγιση του να βασίζεται σε p τιμές (probability values), για να καθιερωθεί η σημασία ενός ευρήματος, είναι απίθανο να είναι αποτελεσματική επειδή ο τεράστιος όγκος δεδομένων σημαίνει ότι σχεδόν όλα είναι σημαντικά. Χρησιμοποιώντας τα τυπικά στατιστικά εργαλεία μας για να αναλύσουμε τα Big Data, είναι πολύ εύκολο να λάβουμε ψευδείς συσχετισμούς. Ωστόσο, αυτό δεν σημαίνει απαραίτητα ότι πρέπει να κινηθούμε προς όλο και πιο περίπλοκες και εκλεπτυσμένες οικονομετρικές τεχνικές για την αντιμετώπιση αυτού του προβλήματος, γιατί δημιουργεί σοβαρό κίνδυνο υπέρ-συναρμολόγησης των δεδομένων. Αντιθέτως, βασικές Bayesian στατιστικές και βηματικές μέθοδοι παλινδρόμησης μπορεί να είναι κατάλληλες προσεγγίσεις. Πέρα από αυτές τις οικείες προσεγγίσεις, υπάρχει μια σειρά εξειδικευμένων τεχνικών για την ανάλυση μεγάλων δεδομένων που είναι σημαντικές για όσους εισέρχονται σε αυτό το πεδίο να κατανοήσουν. Αυτές οι τεχνικές αντλούν από διάφορες επιστήμες, όπως την στατιστική, την επιστήμη των υπολογιστών, τα εφαρμοσμένα μαθηματικά και τα οικονομικά. Περιλαμβάνουν την ανάλυση συστάδων (cluster analysis), την συγχώνευση και ενσωμάτωση δεδομένων, την *εξόρυξη δεδομένων*, τους γενετικούς αλγόριθμους, τη μηχανική μάθηση, τα νευρωνικά δίκτυα, την ανάλυση δικτύου, την επεξεργασία σήματος, την χωρική ανάλυση και την οπτικοποίηση.

Η πρόκληση όμως, είναι να απομακρυνθούμε από τις p τιμές και να επικεντρωθούμε στα μεγέθη των αποτελεσμάτων και τη διακύμανση που εξηγείται. Μια άλλη παγίδα του Big Data, ενισχυμένη και πάλι από τις κοινά χρησιμοποιούμενες στατιστικές τεχνικές, έγκειται στην υπερβολική της εστίαση στα αθροίσματα ή στους μέσους όρους, και πολύ λίγο στις αποκλίσεις (outliers). Σε πολλές περιπτώσεις, οι μέσοι όροι είναι πολύ σημαντικοί και συχνά αποκαλύπτουν τον τρόπο με τον οποίο οι άνθρωποι τείνουν να συμπεριφέρονται υπό συγκεκριμένες συνθήκες. Αλλά στο απέραντο σύμπαν των Big Data, οι αποκλίσεις μπορεί να είναι πιο ενδιαφέρουσες. Κρίσιμες καινοτομίες, τάσεις, αναταραχές ή επαναστάσεις μπορεί να συμβαίνουν εκτός των τάσεων του μέσου όρου, αλλά ακόμα απαιτούν αρκετούς ανθρώπους για να έχουν δραματικά αποτελέσματα με την πάροδο του χρόνου. Η λεπτή φύση των Big Data προσφέρει ευκαιρίες για τον εντοπισμό αυτών των πηγών αλλαγής – είτε

πρόκειται για επιχειρηματικές καινοτομίες, κοινωνικές τάσεις, οικονομικές κρίσεις ή για πολιτικές αναταραχές – καθώς συσσωρεύονται.

Μόλις εντοπιστούν υποσχόμενα στοιχεία, η επόμενη πρόκληση στην ανάλυση των μεγάλων δεδομένων είναι να προχωρήσουμε, πέρα από την ταυτοποίηση συσχετιστικών μοτίβων, στην εξερεύνηση αιτιολογίας. Δεδομένης της μη δομημένης φύσης των περισσότερων μεγάλων δεδομένων, η αιτιολογία δεν ενσωματώνεται στο σχεδιασμό τους και τα παρατηρούμενα πρότυπα είναι συχνά ανοιχτά σε ένα ευρύ φάσμα πιθανών αιτιολογικών εξηγήσεων. Υπάρχουν δύο βασικοί τρόποι προσέγγισης. Το πρώτο είναι να αναγνωρίσουμε την κεντρική σημασία της θεωρίας. Μια διαίσθηση σχετικά με τις αιτιώδεις διαδικασίες που παρήγαγαν τα δεδομένα, μπορεί να χρησιμοποιηθεί για να καθοδηγήσει την ανάπτυξη θεωρητικών επιχειρημάτων, που βασίζονται σε προηγούμενη έρευνα και πηγαίνουν πέρα από αυτή. Ο δεύτερος, συμπληρωματικός τρόπος, είναι να δοκιμάσουμε τότε αυτά τα θεωρητικά επιχειρήματα σε μεταγενέστερες έρευνες, ιδανικά μέσω πειραμάτων πεδίου. Φυσικά, τα εργαστηριακά πειράματα προσφέρουν το πλεονέκτημα του μεγαλύτερου ελέγχου, αλλά συνήθως επικεντρώνονται σε ένα πολύ περιορισμένο αριθμό μεταβλητών, και η φύση της έρευνας των Big Data είναι ότι μπορεί να υπάρχουν πολλοί παράγοντες που επηρεάζουν τα παρατηρούμενα συσχετιστικά πρότυπα.

Σε ένα πείραμα πεδίου μπορεί να συλλεχθεί ένα πλουσιότερο σύνολο δεδομένων σχετικά με τις συμπεριφορές και τις πεποιθήσεις, και για ένα παρατεταμένο χρονικό διάστημα. Για τους μελετητές, καθώς και για τους managers, που ενδιαφέρονται για την έρευνα δράσης (action research), υπάρχουν δελεαστικές ευκαιρίες για να ασχοληθούν με "management engineering", που ξεπερνάει την πιο τυπική διαχείριση της έρευνας φέρνοντας θεωρία και πρακτική μαζί, με πολύ πιο γρήγορους κυκλικούς χρόνους μεταξύ του προσδιορισμού μιας πολλά υποσχόμενης πληροφορίας και της δοκιμής αυτής της πληροφορίας με μια καλά σχεδιασμένη παρέμβαση, που μπορεί να βοηθήσει τόσο στην προώθηση των γνώσεων της διαχείρισης όσο και στην αντιμετώπιση πιεστικών πρακτικών ζητημάτων.

Τελικά, η υπόσχεση και ο στόχος μιας ισχυρής ερευνητικής διαχείρισης βασισμένης στα Big Data, δεν θα έπρεπε να είναι μόνο ο προσδιορισμός των συσχετίσεων και η καθιέρωση εύλογης αιτιολογίας, αλλά η σύγκλιση στοιχείων από πολλαπλές, ανεξάρτητες, μη σχετικές πηγές, που οδηγούν σε ισχυρά συμπεράσματα. Τα Big Data προσφέρουν συναρπαστικές νέες προοπτικές για την επίτευξη τέτοιας σύγκλισης λόγω του πρωτοφανή όγκου, του μικρού επιπέδου λεπτομέρειας, και του πολύπλευρου πλούτου. Η συντριπτική

πλειοψηφία των τωρινών ερευνών για τη διαχείριση βασίζεται στην επίπονη συλλογή μικρού αριθμού μέτρων που καλύπτουν μια σύντομη χρονική περίοδο. Σε αντίθεση, τα Big Data προσφέρουν τεράστιες ποσότητες δεδομένων σε πολλαπλές περιόδους (είτε είναι δευτερόλεπτα, λεπτά, ώρες, ημέρες, μήνες ή χρόνια).

Ενώ ορισμένα σύνολα μεγάλων δεδομένων είναι μονοδιάστατα ή ενός καναλιού, εστιάζοντας, για παράδειγμα, σε μια συγκεκριμένη συναλλαγή ή συμπεριφορά επικοινωνίας και βασιζόμενα σε αλληλεπιδράσεις ενός καναλιού (π.χ. μέσω τηλεφώνου ή ηλεκτρονικού ταχυδρομείου), υπάρχουν ολοένα και περισσότερες ευκαιρίες συλλογής και ανάλυσης πολυδιάστατων συνόλων δεδομένων που προσφέρουν πληροφορίες σε αστερισμούς συμπεριφορών, συχνά μέσω μιας ποικιλίας καναλιών (π.χ. αλληλεπιδράσεις πελάτη τηλεφωνικού κέντρου που αλλάζουν μεταξύ φωνής, διαδικτύου, συνομιλίας, κινητού τηλεφώνου, βίντεο κ.λπ.). Για τους management ερευνητές, το αποτέλεσμα αυτού του πλούτου είναι ότι υπάρχουν άνευ προηγουμένου ευκαιρίες για να παρατηρήσετε πιθανώς σημαντικές μεταβλητές που οι προηγούμενες μελέτες ίσως δεν κατάφεραν να εξετάσουν καθόλου, λόγω του κατ 'ανάγκη πιο εστιασμένου χαρακτήρα τους. Και όταν οι μεταβλητές αυτές αιχμαλωτίσουν την προσοχή ενός ερευνητή, οι σχέσεις μεταξύ τους μπορούν να διερευνηθούν και να εξεταστούν οι συμφραζόμενες συνθήκες υπό τις οποίες αυτές οι σχέσεις μπορούν ή όχι να μελετηθούν. (GEORGE, HAAS and PENTLAND, 2014, pp. 6-9)

Κεφάλαιο 2ο

Data Mining – Εξόρυξη Δεδομένων

2.1.Εισαγωγή στην Εξόρυξη Δεδομένων

Η σύγχρονη επιστήμη και η μηχανική βασίζονται στη χρήση μοντέλων πρώτης αρχής (first-principle models) για την περιγραφή φυσικών, βιολογικών και κοινωνικών συστημάτων. Μια τέτοια προσέγγιση ξεκινά με ένα βασικό επιστημονικό μοντέλο, όπως οι νόμοι κίνησης του Newton ή οι εξισώσεις του Maxwell στον ηλεκτρομαγνητισμό, και στη συνέχεια βασίζεται σε διάφορες εφαρμογές στη μηχανική ή στην ηλεκτρολογία. Σε αυτήν την προσέγγιση, τα πειραματικά δεδομένα χρησιμοποιούνται για την επαλήθευση των υποκείμενων μοντέλων πρώτης αρχής και για την εκτίμηση ορισμένων παραμέτρων που είναι δύσκολο, ή μερικές φορές αδύνατο, να μετρηθούν κατευθείαν. Ωστόσο, σε πολλούς τομείς οι υποκείμενες πρώτες αρχές είναι άγνωστες, ή τα υπό μελέτη συστήματα είναι πολύ περίπλοκα για να είναι μαθηματικά επισημοποιημένα. Με την αυξανόμενη χρήση υπολογιστών, υπάρχει ένα μεγάλο ποσό δεδομένων που παράγονται από τέτοια συστήματα. Με την απουσία των μοντέλων πρώτης αρχής, τέτοια διαθέσιμα δεδομένα μπορούν να χρησιμοποιηθούν για την εξαγωγή μοντέλων, εκτιμώντας χρήσιμες σχέσεις μεταξύ των μεταβλητών ενός συστήματος (δηλαδή, άγνωστες εξαρτήσεις εισόδου - εξόδου).

Όπως αναφέρθηκε και στο προηγούμενο κεφάλαιο, υπάρχουν εξαιρετικά μεγάλοι όγκοι δεδομένων που γεμίζουν τους υπολογιστές, τα δίκτυα και τις ζωές των ανθρώπων. Οι κρατικές υπηρεσίες, τα επιστημονικά ιδρύματα και οι επιχειρήσεις έχουν αφιερώσει τεράστιους πόρους για τη συλλογή και την αποθήκευση δεδομένων. Στην πραγματικότητα, σε πολλές περιπτώσεις, μόνο ένα μικρό ποσοστό αυτών των δεδομένων θα χρησιμοποιηθεί διότι οι όγκοι είναι απλά πολύ μεγάλοι για να μπορέσει ο άνθρωπος να τους διαχειριστεί, ή οι ίδιες οι δομές δεδομένων είναι πολύ περίπλοκες για να μπορούν να αναλυθούν αποτελεσματικά. Πώς θα μπορούσε να συμβεί αυτό; Ο πρωταρχικός λόγος είναι ότι η αρχική προσπάθεια δημιουργίας ενός συνόλου δεδομένων εστιάζεται συχνά σε θέματα όπως η αποδοτικότητα αποθήκευσης. Δεν περιλαμβάνει σχέδιο για τον τρόπο με τον οποίο τα δεδομένα θα χρησιμοποιηθούν και θα αναλυθούν τελικά.

Η ανάγκη να κατανοήσουμε μεγάλα, σύνθετα, πλούσια σε πληροφορία σύνολα δεδομένων είναι κοινή σε όλους τους τομείς των επιχειρήσεων, της επιστήμης και της μηχανικής. Στον επιχειρηματικό κόσμο, τα δεδομένα των εταιρειών και των πελατών

αναγνωρίζονται ως στρατηγικό πλεονέκτημα. Η ικανότητα εξαγωγής χρήσιμης γνώσης που κρύβεται σε αυτά τα δεδομένα και η δράση της στη γνώση αυτή, καθίσταται όλο και πιο σημαντική στο σημερινό ανταγωνιστικό κόσμο. Η όλη διαδικασία εφαρμογής μιας μεθοδολογίας που βασίζεται στον υπολογιστή, συμπεριλαμβανομένων νέων τεχνικών, για την ανακάλυψη γνώσης από δεδομένα, ονομάζεται *εξόρυξη δεδομένων*. (Kantardzic, 2011, pp. 1-2)

Οι άνθρωποι από τότε που άρχισε η ανθρώπινη ζωή αναζητούν πρότυπα στα δεδομένα. Οι κυνηγοί αναζητούν πρότυπα στη μεταναστευτική συμπεριφορά των ζώων, οι αγρότες αναζητούν πρότυπα στην ανάπτυξη των καλλιεργειών, οι πολιτικοί αναζητούν πρότυπα στις απόψεις των ψηφοφόρων. Η δουλειά ενός επιστήμονα είναι να κατανοήσει τα δεδομένα, να ανακαλύψει τα πρότυπα που διέπουν τον τρόπο με τον οποίο λειτουργεί ο φυσικός κόσμος και να τα ενσωματώσει σε θεωρίες που μπορούν να χρησιμοποιηθούν για να προβλέψει τι θα συμβεί σε νέες καταστάσεις. Η δουλειά του επιχειρηματία είναι να εντοπίσει ευκαιρίες, δηλαδή μοτίβα στην συμπεριφορά που μπορούν να μετατραπούν σε κερδοφόρα επιχείρηση, και να τα εκμεταλλευτεί.

Στην εξόρυξη δεδομένων, τα δεδομένα αποθηκεύονται ηλεκτρονικά και η αναζήτηση είναι αυτοματοποιημένη – ή τουλάχιστον ενισχυμένη – από έναν υπολογιστή. Ακόμα και αυτό δεν είναι ιδιαίτερα καινούργιο. Οι οικονομολόγοι, οι στατιστικολόγοι, οι μηχανικοί επικοινωνίας έχουν από καιρό εργαστεί με την ιδέα ότι τα πρότυπα στα δεδομένα μπορούν να αναζητηθούν αυτόματα, να αναγνωριστούν, να επικυρωθούν και να χρησιμοποιηθούν για πρόβλεψη. Αυτό που είναι καινούργιο είναι η εντυπωσιακή αύξηση των ευκαιριών για την εύρεση προτύπων στα δεδομένα. Η αχαλίνωτη ανάπτυξη των βάσεων δεδομένων τα τελευταία χρόνια, βάσεις δεδομένων για καθημερινές δραστηριότητες, όπως είναι οι επιλογές των πελατών, φέρνουν την εξόρυξη δεδομένων στην πρώτη γραμμή των νέων επιχειρησιακών τεχνολογιών. Καθώς η πλημμύρα δεδομένων διογκώνεται και οι μηχανές που μπορούν να αναλάβουν την αναζήτηση γίνονται συνηθισμένες, οι δυνατότητες της εξόρυξης δεδομένων αυξάνονται. Εκτιμάται πλέον ότι ο όγκος των δεδομένων που αποθηκεύονται στις παγκόσμιες βάσεις δεδομένων διπλασιάζεται κάθε 20 μήνες. Άρα η έξυπνη ανάλυση δεδομένων αποτελεί έναν πολύτιμο πόρο. Μπορεί να οδηγήσει σε νέες ιδέες και, όσο αναφορά το εμπορικό περιβάλλον, σε ανταγωνιστικά πλεονεκτήματα. Ενώ ο κόσμος μεγαλώνει σε πολυπλοκότητα και μας συντρίβει με τα δεδομένα που παράγει, η εξόρυξη δεδομένων γίνεται η μόνη ελπίδα για την αποσαφήνιση των μοτίβων που την στηρίζουν. (Witten and Frank, 2005, pp. 4-5)

2.2. Η ιστορία της Εξόρυξης Δεδομένων

Η διαδικασία της εξόρυξης δεδομένων εκτελείται εδώ και αιώνες. Μέθοδοι όπως το θεώρημα Bayes και η θεωρία ανάλυσης της παλινδρόμησης, ήταν από τα πρώτα ευρήματα που συνέβαλαν στον προσδιορισμό προτύπων και στην ανακάλυψη κρυφών σχέσεων. Καθώς οι συλλογές δεδομένων αυξήθηκαν σε όγκο και σε πολυπλοκότητα, η κουραστική, χρονοβόρα και χειρωνακτική τους ανάλυση δεν ήταν πλέον εφικτή. Ωστόσο, με την εξέλιξη της τεχνολογίας και τις ανακαλύψεις της επιστήμης των υπολογιστών, έφτασε η εποχή της αυτόματης επεξεργασίας δεδομένων. Καινούργιες μέθοδοι, όπως τα νευρωνικά δίκτυα, οι γενετικοί αλγόριθμοι (1950), τα δέντρα απόφασης (1960), η μηχανή υποστήριξης διανυσμάτων (1990), γεφυρώνουν την εφαρμοσμένη στατιστική και την τεχνητή νοημοσύνη με την διαχείριση των βάσεων δεδομένων, προσφέροντας της το μαθηματικό υπόβαθρο που χρειάζεται. (El.wikipedia.org, 2018)

Στις παρακάτω γραμμές παρουσιάζονται κάποια από τα πιο σημαντικά και "πρωταρχικά" ορόσημα στην ιστορία της εξόρυξης δεδομένων, καθώς και το πως εξελίσσεται και συνδυάζεται με την επιστήμη των δεδομένων και τα Big Data.

2.2.1. Ιστορική Αναδρομή

1763 : Το έγγραφο του Thomas Bayes δημοσιεύεται μεταθανάτια. Στη θεωρία πιθανοτήτων και στη στατιστική, το **θεώρημα Bayes** σχετίζει την τρέχουσα πιθανότητα με την αρχική πιθανότητα. Είναι θεμελιώδους σημασίας για την εξόρυξη δεδομένων καθώς επιτρέπει την κατανόηση σύνθετων πραγματικοτήτων που βασίζονται σε εκτιμώμενες πιθανότητες.

1805 : Ο Adrien-Marie Legendre και ο Carl Friedrich Gauss εφαρμόζουν την παλινδρόμηση για να καθορίσουν τις τροχιές των σωμάτων γύρω από τον Ήλιο (κομήτες και πλανήτες). Ο στόχος της ανάλυσης της παλινδρόμησης είναι να εκτιμηθούν οι σχέσεις μεταξύ των μεταβλητών, χρησιμοποιώντας την μέθοδο των ελάχιστων τετραγώνων. Η παλινδρόμηση είναι ένα από τα βασικά εργαλεία στην εξόρυξη δεδομένων.

1936 : Αυτή είναι η αρχή της εποχής των υπολογιστών, η οποία καθιστά δυνατή τη συλλογή και επεξεργασία μεγάλων ποσοτήτων δεδομένων. Σε ένα χαρτί του 1936, για τους αξιόπιστους αριθμούς, ο Alan Turing παρουσίασε την ιδέα μιας μηχανής (Universal Turing

Machine), ικανής να εκτελεί υπολογισμούς όπως οι σύγχρονοι υπολογιστές μας. Οι υπολογιστές της σύγχρονης εποχής βασίζονται στις ιδέες που πρωτοστάτησε ο Turing.

1943 : Ο Warren McCulloch και ο Walter Pitts ήταν οι πρώτοι που δημιούργησαν ένα εννοιολογικό μοντέλο ενός νευρικού δικτύου. Σε μια εργασία με τίτλο *"A logical calculus of the ideas immanent in nervous activity"*, περιγράφουν την ιδέα ενός νευρώνα σε ένα δίκτυο. Κάθε ένας από αυτούς τους νευρώνες μπορεί να κάνει 3 πράγματα: να λαμβάνει εισόδους, να εισάγει διαδικασίες και να παράγει αποτελέσματα.

1965 : Ο Lawrence J. Fogel δημιούργησε μια νέα εταιρεία αποκαλούμενη Decision Science, Inc. για εφαρμογές εξελικτικού προγραμματισμού. Ήταν η πρώτη εταιρεία που εφάρμοσε τον εξελικτικό υπολογισμό για την επίλυση πραγματικών προβλημάτων.

1970's : Με εξελιγμένα συστήματα διαχείρισης βάσεων δεδομένων, είναι δυνατή η αποθήκευση και η αναζήτηση terabytes και petabytes δεδομένων. Επιπλέον, οι αποθήκες δεδομένων επιτρέπουν στους χρήστες να μετακινούνται σε έναν πιο αναλυτικό τρόπο προβολής των δεδομένων. Ωστόσο, η εξόρυξη σύνθετων στοιχείων από αυτές τις αποθήκες δεδομένων πολυδιάστατων μοντέλων είναι πολύ περιορισμένη.

1975 : Ο John Henry Holland έγραψε το βιβλίο *"Adaptation in Natural and Artificial Systems"*, το πρωτοποριακό βιβλίο για τους γενετικούς αλγόριθμους. Είναι το βιβλίο που ξεκίνησε αυτόν τον τομέα σπουδών, παρουσιάζοντας τα θεωρητικά θεμέλια και διερευνώντας τις εφαρμογές.

1980's : Η εταιρία HNC δημιουργεί τη φράση "database mining". Αυτός ο όρος προοριζόταν να προστατεύει ένα προϊόν που ονομάζεται DataBase Mining Workstation. Ήταν ένα εργαλείο γενικού σκοπού για την οικοδόμηση μοντέλων νευρωνικών δικτύων το οποίο δεν είναι πλέον διαθέσιμο. Κατά τη διάρκεια αυτής της περιόδου οι εξελιγμένοι αλγόριθμοι μπορούν να "μαθαίνουν" τις σχέσεις από τα δεδομένα που επιτρέπουν στους εμπειρογνώμονες του αντικειμένου να διασαφηνίσουν τι σημαίνουν οι σχέσεις.

1989 : Ο όρος "Knowledge Discovery in Databases " δημιουργείται από τον Gregory Piatetsky-Shapiro. Επίσης, την ίδια περίοδο ιδρύει το πρώτο workshop που ονομάζεται επίσης KDD.

1990's : Ο όρος "εξόρυξη δεδομένων" δημιουργείται. Οι εταιρείες λιανικής και η χρηματοοικονομική κοινότητα χρησιμοποιούν την εξόρυξη δεδομένων για να αναλύσουν δεδομένα και να αναγνωρίσουν τάσεις για την αύξηση της πελατειακής τους βάσης, να προβλέψουν διακυμάνσεις των επιτοκίων, των τιμών των μετοχών, ζήτηση των πελατών κ.α.

1992 : Οι Bernhard E. Boser, Isabelle M. Guyon και Vladimir N. Vapnik πρότειναν μια βελτίωση στην αρχική Support Vector Machine που επιτρέπει τη δημιουργία μη γραμμικών ταξινομητών. Τα Support Vector Machines είναι μια προσέγγιση εποπτευόμενης μάθησης που αναλύει τα δεδομένα και αναγνωρίζει πρότυπα που χρησιμοποιούνται για ανάλυση ταξινόμησης και παλινδρόμησης.

1993 : Ο Gregory Piatetsky-Shapiro ξεκινάει το Knowledge Discovery Nuggets (KDnuggets) newsletter. Αρχικά σχεδιάστηκε για να ενώσει την ομάδα ερευνητών που παρευρισκόντουσαν στο KDD Workshop. Ωστόσο, το KDnuggets.com φαίνεται να έχει ένα ευρύτερο κοινό τώρα.

2001 : Αν και ο όρος “data science” υπήρχε από το 1960, ο William S. Cleveland την παρουσίασε ως ανεξάρτητο επιστημονικό κλάδο το 2001. Σύμφωνα με τις Build Data Science Teams, ο DJ Patil και ο Jeff Hammerbacher χρησιμοποίησαν τότε τον όρο για να περιγράψουν τους ρόλους τους στο LinkedIn και στο Facebook.

2003 : Το βιβλίο Moneyball, από τον Michael Lewis, δημοσιεύεται και αλλάζει τον τρόπο με τον οποίο πολλά major league front offices δουλεύουν. Η Oakland Athletics χρησιμοποίησε μια στατιστική προσέγγιση για να βρει ωφέλιμα χαρακτηριστικά σε παίκτες που ήταν υποτιμημένοι και φθηνότεροι. Με αυτόν τον τρόπο, συγκέντρωσαν με επιτυχία μια ομάδα που τους έφερε στα playoffs του 2002 και του 2003 με το 1/3 της μισθοδοσίας.

2015 : Τον Φεβρουάριο του 2015, ο DJ Patil έγινε ο πρώτος Chief Data Scientist του Λευκού Οίκου. Σήμερα, η εξόρυξη δεδομένων είναι ευρέως διαδεδομένη στις επιχειρήσεις, την επιστήμη, τη μηχανική, την ιατρική και σε πολλούς άλλους κλάδους. Η συλλογή δεδομένων γίνεται φθηνότερη και οι συσκευές συλλογής δεδομένων πολλαπλασιάζονται.

2016 : Μια από τις πιο δραστήριες τεχνικές που διερευνούνται σήμερα είναι η Deep Learning. Με την δυνατότητα να καταγράφει εξαρτήσεις και σύνθετα σχέδια αποτελεσματικότερα από άλλες τεχνικές, αναζωπυρώνει μερικές από τις μεγαλύτερες προκλήσεις στον κόσμο της εξόρυξης δεδομένων, της επιστήμης των δεδομένων και της τεχνητής νοημοσύνης. (Kdnuggets.com, 2018)

2.3. Τι είναι η Εξόρυξη Δεδομένων;

Με πολύ απλά λόγια, η εξόρυξη δεδομένων αναφέρεται στην εξαγωγή ή "εξόρυξη" γνώσης από μεγάλες ποσότητες δεδομένων. Ο όρος είναι στην πραγματικότητα μια

εσφαλμένη ονομασία. Όταν γίνεται εξόρυξη χρυσού από βράχους ή από άμμο αναφέρεται ως εξόρυξη χρυσού αντί για εξόρυξη πετρωμάτων ή άμμων. Έτσι, η εξόρυξη δεδομένων θα έπρεπε να έχει ονομαστεί πιο κατάλληλα "εξόρυξη γνώσης από δεδομένα", το οποίο δυστυχώς είναι μεγάλη φράση. Η "εξόρυξη γνώσης", ένας συντομότερος όρος, μπορεί να μην αντανακλά την έμφαση στην εξόρυξη μεγάλων ποσοτήτων δεδομένων. Ωστόσο, η εξόρυξη είναι ένας έντονος όρος, χαρακτηρίζοντας τη διαδικασία που βρίσκει ένα μικρό σύνολο πολύτιμων βόλων χρυσού από μεγάλης ποσότητας πρώτης ύλης. Επομένως, μια τέτοια εσφαλμένη ονομασία που φέρει τόσο "δεδομένα" όσο και "εξόρυξη" έγινε η δημοφιλής επιλογή. Πολλοί άλλοι όροι φέρουν ένα παρόμοιο ή ελαφρώς διαφορετικό νόημα της εξόρυξη δεδομένων, όπως εξόρυξη γνώσης από δεδομένα (knowledge mining from data), εξαγωγή γνώσης (knowledge extraction), ανάλυση δεδομένων / προτύπων (data/pattern analysis), αρχαιολογία δεδομένων (data archaeology) και βυθοκόρηση δεδομένων (data dredging).

Πολλοί άνθρωποι αντιμετωπίζουν την εξόρυξη δεδομένων ως συνώνυμο ενός άλλου ευρέως χρησιμοποιούμενου όρου, του KDD (Knowledge Discovery from Data), ή Ανακάλυψη Γνώσης από Δεδομένα. Εναλλακτικά, άλλοι θεωρούν την εξόρυξη δεδομένων ως απλώς ένα ουσιαστικό βήμα στη διαδικασία της ανακάλυψης της γνώσης. Η ανακάλυψη της γνώσης ως διαδικασία αποτελείται από μια επαναληπτική ακολουθία των ακόλουθων βημάτων :

1. **Καθαρισμός δεδομένων / Data cleaning** (για την εξάλειψη του θορύβου και των ασυνεπών δεδομένων).
2. **Ενσωμάτωση δεδομένων / Data integration** (όπου πολλαπλές πηγές δεδομένων μπορούν να συνδυαστούν).
3. **Επιλογή δεδομένων / Data selection** (όπου δεδομένα που σχετίζονται με την εργασία ανάλυσης ανακτώνται από τη βάση δεδομένων).
4. **Μετασχηματισμός δεδομένων / Data transformation** (όπου τα δεδομένα μετατρέπονται ή ενοποιούνται σε έντυπα κατάλληλα για εξόρυξη εκτελώντας διαδικασίες περίληψης ή συγκέντρωσης).
5. **Εξόρυξη δεδομένων / Data mining** (μια βασική διαδικασία όπου εφαρμόζονται έξυπνες μέθοδοι προκειμένου να εξαχθούν πρότυπα δεδομένων).
6. **Αξιολόγηση μοτίβων / Pattern evaluation** (για να προσδιοριστούν τα πραγματικά ενδιαφέροντα πρότυπα που αντιπροσωπεύουν τη γνώση με βάση ορισμένα μέτρα ενδιαφέροντος).

7. **Παρουσίαση γνώσης / Knowledge presentation** (όπου οι τεχνικές απεικόνισης και εκπροσώπησης της γνώσης χρησιμοποιούνται για να παρουσιάσουν την εξόρυξη γνώσης στον χρήστη).

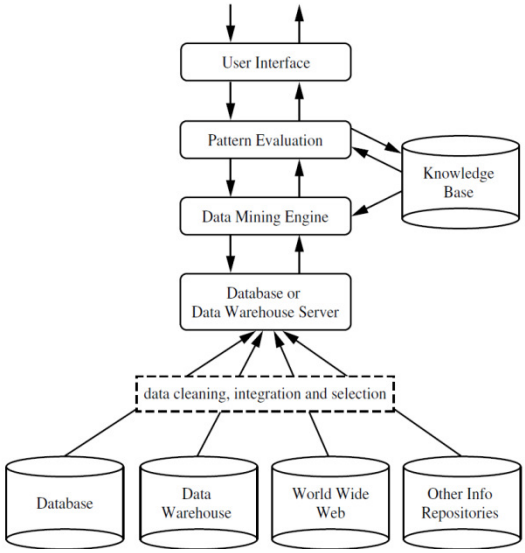
Τα βήματα 1 έως 4 είναι διαφορετικές μορφές προ-επεξεργασίας δεδομένων (data preprocessing), όπου τα δεδομένα προετοιμάζονται για εξόρυξη. Το βήμα εξόρυξης δεδομένων μπορεί να αλληλοεπιδρά με τον χρήστη ή με μια βάση γνώσεων. Τα ενδιαφέροντα πρότυπα παρουσιάζονται στον χρήστη και μπορούν να αποθηκευτούν ως νέες γνώσεις στη βάση γνώσεων. Σύμφωνα με αυτή την άποψη, η εξόρυξη δεδομένων είναι μόνο ένα βήμα στην ολόκληρη διαδικασία, αν και ουσιαστική επειδή αποκαλύπτει κρυμμένα πρότυπα για αξιολόγηση.

Παρόλα αυτά, στην βιομηχανία, στα μέσα μαζικής ενημέρωσης και στον περιβάλλον έρευνας της βάσης δεδομένων, ο όρος data mining γίνεται πιο δημοφιλής από αυτόν της KDD. Μια ευρεία άποψη της λειτουργικότητας της εξόρυξης δεδομένων είναι ότι αποτελεί μια διαδικασία ανακάλυψης ενδιαφέρουσας γνώσης από μεγάλες ποσότητες δεδομένων που αποθηκεύονται σε βάσεις δεδομένων, σε αποθήκες δεδομένων (data warehouses) ή σε άλλα αποθετήρια πληροφοριών. Με βάση αυτή την άποψη, η αρχιτεκτονική ενός τυπικού συστήματος εξόρυξης δεδομένων μπορεί να έχει τα παρακάτω κύρια συστατικά :

1. **Βάση δεδομένων, αποθήκη δεδομένων, WorldWideWeb ή άλλα αποθετήρια πληροφοριών** : Τεχνικές καθαρισμού δεδομένων και ενσωμάτωσης δεδομένων μπορούν να πραγματοποιηθούν στα δεδομένα.
2. **Διακομιστής βάσεων δεδομένων ή αποθήκης δεδομένων** : Είναι υπεύθυνος για την ανάκτηση των σχετικών δεδομένων, βάσει του αιτήματος εξόρυξης δεδομένων του χρήστη.
3. **Βάση γνώσεων / Knowledge Base** : Χρησιμοποιείται για την καθοδήγηση της αναζήτησης ή αξιολογεί το πόσο ενδιαφέροντα είναι τα πρότυπα που προκύπτουν.
4. **Μηχανή εξόρυξης δεδομένων** : Αυτό είναι απαραίτητο για το σύστημα εξόρυξης δεδομένων και ιδανικά αποτελείται από ένα σύνολο λειτουργικών ενοτήτων για εργασίες όπως είναι ο χαρακτηρισμός, η συσχέτιση και η ανάλυση συσχέτισης, η ταξινόμηση, η πρόβλεψη, η ανάλυση συστάδων κλπ.
5. **Μοντέλο αξιολόγησης προτύπου** : Αυτό το συστατικό συνήθως χρησιμοποιεί μέτρα ενδιαφέροντος και αλληλεπιδρά με μονάδες μέτρησης της εξόρυξης

δεδομένων έτσι ώστε η αναζήτηση να επικεντρωθεί προς ενδιαφέροντα πρότυπα.

6. **Διεπαφή χρήστη** : Επικοινωνεί μεταξύ των χρηστών και του συστήματος εξόρυξης δεδομένων, επιτρέποντας στον χρήστη να αλληλεπιδρά με το σύστημα καθορίζοντας μια ερώτηση ή ένα έργο εξόρυξης δεδομένων, παρέχοντας πληροφορίες για να βοηθήσει στην εστίαση της αναζήτησης και να διεξάγει διερευνητικά δεδομένα εξόρυξης με βάση τα ενδιάμεσα αποτελέσματα της εξόρυξης δεδομένων.



Εικόνα 2.1 : Αρχιτεκτονική ενός τυπικού συστήματος Εξόρυξης Δεδομένων.

Όσον αφορά την αποθήκη δεδομένων, η εξόρυξη δεδομένων μπορεί να θεωρηθεί ως ένα προχωρημένο στάδιο του Συστήματος Αναλυτικής Επεξεργασίας Συναλλαγών (on-line analytical processing - OLAP). Ωστόσο, η εξόρυξη δεδομένων υπερβαίνει το στενό πεδίο εφαρμογής του συνοπτικού στυλ αναλυτικής επεξεργασίας των συστημάτων αποθήκευσης δεδομένων ενσωματώνοντας πιο προηγμένες τεχνικές για την ανάλυση δεδομένων.

Παρόλο που υπάρχουν πολλά “συστήματα εξόρυξης δεδομένων” στην αγορά, δεν μπορούν όλα να εκτελέσουν πραγματική εξόρυξη δεδομένων. Ένα σύστημα ανάλυσης δεδομένων, που δεν χειρίζεται μεγάλα ποσά δεδομένων, θα πρέπει να είναι πιο κατάλληλα κατηγοριοποιημένο ως ένα σύστημα εκμάθησης μηχανών, ένα στατιστικό εργαλείο ανάλυσης δεδομένων ή ένα πειραματικό πρωτότυπο συστήματος. Ένα σύστημα που μπορεί να εκτελέσει μόνο ανάκτηση δεδομένων ή πληροφοριών, συμπεριλαμβανομένης της εύρεσης συγκεντρωτικών τιμών, ή που εκτελεί εξαναγκαστική απάντηση ερωτήματος σε μεγάλες

βάσεις δεδομένων θα πρέπει να κατηγοριοποιείται πιο κατάλληλα ως σύστημα βάσης δεδομένων, ένα σύστημα ανάκτησης πληροφοριών, ή ένα επαγωγικό σύστημα βάσεων δεδομένων.

Η εξόρυξη δεδομένων περιλαμβάνει μια ενσωμάτωση τεχνικών από πολλαπλούς κλάδους όπως είναι η τεχνολογία βάσεων δεδομένων και αποθήκης δεδομένων, η στατιστική, η μηχανική μάθηση, η υψηλή απόδοση της χρήσης του υπολογιστή, η αναγνώριση προτύπων, τα νευρωνικά δίκτυα, η οπτικοποίηση δεδομένων, η ανάκτηση πληροφοριών, η επεξεργασία εικόνας και σήματος και η ανάλυση χωρικών ή χρονικών δεδομένων. Με την εκτέλεση της εξόρυξης δεδομένων, ενδιαφέρουσες γνώσεις, κανονικότητες, ή υψηλού επιπέδου πληροφορίες μπορούν να εξαχθούν από βάσεις δεδομένων και να προβληθούν ή να περιηγηθούν από διαφορετικές γωνίες. Η ανακαλυφθείσα γνώση μπορεί να εφαρμοστεί στη λήψη αποφάσεων, στον έλεγχο της διαδικασίας, στην διαχείριση πληροφοριών και στην επεξεργασία ερωτημάτων. Ως εκ τούτου, η εξόρυξη δεδομένων θεωρείται μία από τα πιο σημαντικά σύνορα σε συστήματα βάσεων δεδομένων και πληροφοριών και μια από τις πιο υποσχόμενες διεπιστημονικές εξελίξεις στην τεχνολογία των πληροφοριών. (Han and Kamber, 2006, pp. 5-9)

Έχει καταστεί σαφές πλέον, ότι η εξόρυξη δεδομένων αποτελεί μια επαναληπτική διαδικασία στην οποία η πρόοδος ορίζεται από την ανακάλυψη, είτε με αυτόματες είτε με χειροκίνητες μεθόδους. Είναι πιο χρήσιμη σε ένα διερευνητικό σενάριο ανάλυσης στο οποίο δεν υπάρχουν προκαθορισμένες έννοιες για το τι θα αποτελέσει ένα "ενδιαφέρον" αποτέλεσμα. Είναι η αναζήτηση νέων, πολύτιμων και σημαντικών πληροφοριών σε μεγάλους όγκους δεδομένων. Είναι μια προσπάθεια συνεργασίας μεταξύ ανθρώπων και υπολογιστών. Τα καλύτερα αποτελέσματα επιτυγχάνονται εξισορροπώντας τη γνώση των ειδικών ανθρώπων στην περιγραφή των προβλημάτων και των στόχων με τις δυνατότητες αναζήτησης των υπολογιστών.

Στην πράξη, οι δύο βασικοί στόχοι της εξόρυξης δεδομένων τείνουν να είναι η *πρόβλεψη* και η *περιγραφή*. Η πρόβλεψη περιλαμβάνει τη χρήση ορισμένων μεταβλητών ή πεδίων στο σύνολο δεδομένων για την πρόβλεψη άγνωστων ή μελλοντικών τιμών άλλων μεταβλητών ενδιαφέροντος. Η περιγραφή, από την άλλη πλευρά, επικεντρώνεται στην εύρεση προτύπων που περιγράφουν τα δεδομένα που μπορούν να ερμηνευτούν από τον άνθρωπο. Επομένως, είναι δυνατό να τοποθετηθούν οι δραστηριότητες εξόρυξης δεδομένων σε μία από τις δύο κατηγορίες :

1. *Πρόβλεψη εξόρυξης δεδομένων*, η οποία *παράγει το μοντέλο* του συστήματος που περιγράφεται από το δεδομένο σύνολο δεδομένων, ή
2. *Περιγραφική εξόρυξη δεδομένων*, η οποία *παράγει νέες, σημαντικές πληροφορίες* βάσει του διαθέσιμου συνόλου δεδομένων.

Στο προγνωστικό τέλος του φάσματος, ο στόχος της εξόρυξης δεδομένων είναι να παραχθεί ένα μοντέλο, το οποίο εκφράζεται ως εκτελέσιμος κώδικας, που μπορεί να χρησιμοποιηθεί για την πραγματοποίηση ταξινόμησης, πρόβλεψης, εκτίμησης ή άλλων παρόμοιων εργασιών. Στο περιγραφικό τέλος του φάσματος, ο στόχος είναι να κατανοηθεί το σύστημα που αναλύεται, αποκαλύπτοντας πρότυπα και σχέσεις σε μεγάλα σύνολα δεδομένων. Η σχετική σημασία της πρόβλεψης και της περιγραφής για συγκεκριμένες εφαρμογές της εξόρυξης δεδομένων μπορεί να ποικίλει σημαντικά. Οι στόχοι της πρόβλεψης και της περιγραφής επιτυγχάνονται χρησιμοποιώντας τεχνικές εξόρυξης δεδομένων για τα ακόλουθα πρωταρχικά καθήκοντα αυτής :

1. **Ταξινόμηση / Classification** : Ταξινομεί ένα στοιχείο δεδομένων σε μία από τις διάφορες προκαθορισμένες κατηγορίες (predictive learning function).
2. **Παλινδρόμηση / Regression** : Χαρτογραφεί ένα στοιχείο δεδομένων σε μια πραγματικής αξίας μεταβλητή πρόβλεψης (predictive learning function).
3. **Ομαδοποίηση / Clustering** : Μια κοινή περιγραφική εργασία στην οποία κάποιος επιδιώκει να εντοπίσει ένα πεπερασμένο σύνολο κατηγοριών ή συμπλεγμάτων για να περιγράψει τα δεδομένα (descriptive task).
4. **Σύνοψη / Summarization** : Μια πρόσθετη περιγραφική εργασία που περιλαμβάνει μεθόδους για την εύρεση μιας συμπαγούς περιγραφής για ένα σύνολο (ή υποσύνολο) δεδομένων (descriptive task).
5. **Μοντέλο εξάρτησης / Dependency Modeling** : Εύρεση ενός τοπικού μοντέλου που περιγράφει σημαντικές εξαρτήσεις μεταξύ μεταβλητών ή μεταξύ των τιμών ενός χαρακτηριστικού σε ένα σύνολο δεδομένων ή σε ένα τμήμα ενός συνόλου δεδομένων.
6. **Ανίχνευση αλλαγής και απόκλισης / Change and Deviation Detection** : Ανακαλύπτει τις πιο σημαντικές αλλαγές στο σύνολο δεδομένων.

Η επιτυχία μιας εργασίας στην εξόρυξη δεδομένων εξαρτάται σε μεγάλο βαθμό από την ποσότητα της ενέργειας, της γνώσης και της δημιουργικότητας που προσφέρει ο σχεδιαστής. Ουσιαστικά, η εξόρυξη δεδομένων είναι σαν να προσπαθείς να λύσεις ένα πάζλ. Τα κομμάτια του πάζλ δεν αποτελούν σύνθετες κατασκευές από μόνα τους. Σαν ένα σύνολο

όμως, μπορούν να αποτελέσουν πολύ περίπλοκα συστήματα. Στην προσπάθεια να ξεδιπλωθούν αυτά τα συστήματα, η όλη διαδικασία θα φαίνεται αρκετά ενοχλητική, αλλά δουλεύοντας με τα κομμάτια γίνεται κατανοητό το πως πρέπει να χρησιμοποιηθούν. Στην αρχή, οι σχεδιαστές της data mining διαδικασίας πιθανότατα δεν γνώριζαν πολλά για τις πηγές δεδομένων, γεγονός που αν ίσχυε μάλλον δεν θα υπήρχε και μεγάλο ενδιαφέρον προς την εκτέλεση της. Μεμονωμένα, τα δεδομένα φαίνονται απλά, πλήρη και επεξηγηματικά. Αλλά συλλογικά, παίρνουν μια εντελώς νέα εμφάνιση που είναι εκφοβιστικό και δύσκολο να κατανοηθούν, όπως το πάζλ. Ως εκ τούτου, η ύπαρξη ενός αναλυτή και ενός σχεδιαστή σε μια διαδικασία εξόρυξης δεδομένων απαιτεί, πέρα από τις επαγγελματικές γνώσεις, τη δημιουργική σκέψη και την προθυμία να προβληθούν τα προβλήματα με διαφορετικό φως.

Η εξόρυξη δεδομένων είναι ένας από τους ταχύτερα αναπτυσσόμενους κλάδους της βιομηχανίας των υπολογιστών. Ένα από τα μεγαλύτερα πλεονεκτήματα της αντανακλάται στο ευρύ φάσμα μεθοδολογιών και τεχνικών που μπορούν να εφαρμοστούν σε ένα πλήθος από ομάδες προβλημάτων. Δεδομένου ότι η εξόρυξη δεδομένων είναι μια φυσική δραστηριότητα που πρέπει να διεξαχθεί σε μεγάλα σύνολα δεδομένων, ένας από τους μεγαλύτερους στόχους στην αγορά είναι ολόκληρη η κοινότητα της αποθήκευσης δεδομένων (data warehousing), της data mart και της υποστήριξης αποφάσεων, που περιλαμβάνει επαγγελματίες από τους κλάδους της λιανικής, της τηλεπικοινωνίας, της υγειονομικής περίθαλψης, την ασφάλισης και τη μεταφοράς. Στην επιχειρηματική κοινότητα, η εξόρυξη δεδομένων μπορεί να χρησιμοποιηθεί για να ανακαλύψει νέες τάσεις αγορών, να σχεδιάσει επενδυτικές στρατηγικές και να ανιχνεύσει μη εξουσιοδοτημένες δαπάνες στο λογιστικό σύστημα. Μπορεί να βελτιώσει τις εκστρατείες μάρκετινγκ και τα αποτελέσματα μπορούν να χρησιμοποιηθούν για να παρέχουν στους πελάτες πιο στοχευμένη υποστήριξη και προσοχή. Οι τεχνικές εξόρυξης δεδομένων μπορούν να εφαρμοστούν σε προβλήματα ανασχεδιασμού επιχειρησιακών διαδικασιών, στα οποία ο στόχος είναι να κατανοηθούν οι αλληλεπιδράσεις και οι σχέσεις μεταξύ επιχειρηματικών πρακτικών και οργανισμών.

Πολλές αρχές επιβολής του νόμου και ειδικές μονάδες έρευνας, των οποίων η αποστολή είναι να εντοπίζουν δόλιες δραστηριότητες και να ανακαλύπτουν εγκληματικές τάσεις, έχουν χρησιμοποιήσει επιτυχώς την εξόρυξη δεδομένων. Για παράδειγμα, αυτές οι μεθοδολογίες μπορούν να βοηθήσουν τους αναλυτές στην ταυτοποίηση κρίσιμων μοντέλων συμπεριφοράς, στις επικοινωνιακές αλληλεπιδράσεις σε οργανώσεις ναρκωτικών, στις νομισματικές συναλλαγές σε "ξέπλυμα" χρημάτων και στις συναλλαγές εμπιστευτικών πληροφοριών, στις κινήσεις των σειριακών δολοφόνων και στη στόχευση λαθρεμπόρων στα

σύνορα. Τέλος, οι τεχνικές εξόρυξης δεδομένων έχουν επίσης χρησιμοποιηθεί και σε δραστηριότητες που σχετίζονται με θέματα εθνικής ασφάλειας. (Kantardzic, 2011, pp. 2-4)

Συνοψίζοντας όλα τα παραπάνω και για το κλείσιμο της ενότητας, ο ορισμός που θα μπορούσε να δοθεί τελικά στο ερώτημα του τι είναι η εξόρυξη δεδομένων είναι ο εξής :

"Εξόρυξη Δεδομένων (Data Mining) είναι η ανάλυση (συνήθως τεράστιων) παρατηρούμενων (observational) συνόλων δεδομένων, έτσι ώστε να βρεθούν μη παρατηρηθείσες σχέσεις και να συνοψιστούν τα δεδομένα με καινοφανείς τρόπους, οι οποίοι να είναι κατανοητοί και χρήσιμοι στον κάτοχο των δεδομένων". (Hand, Mannila and Smyth, 2001)

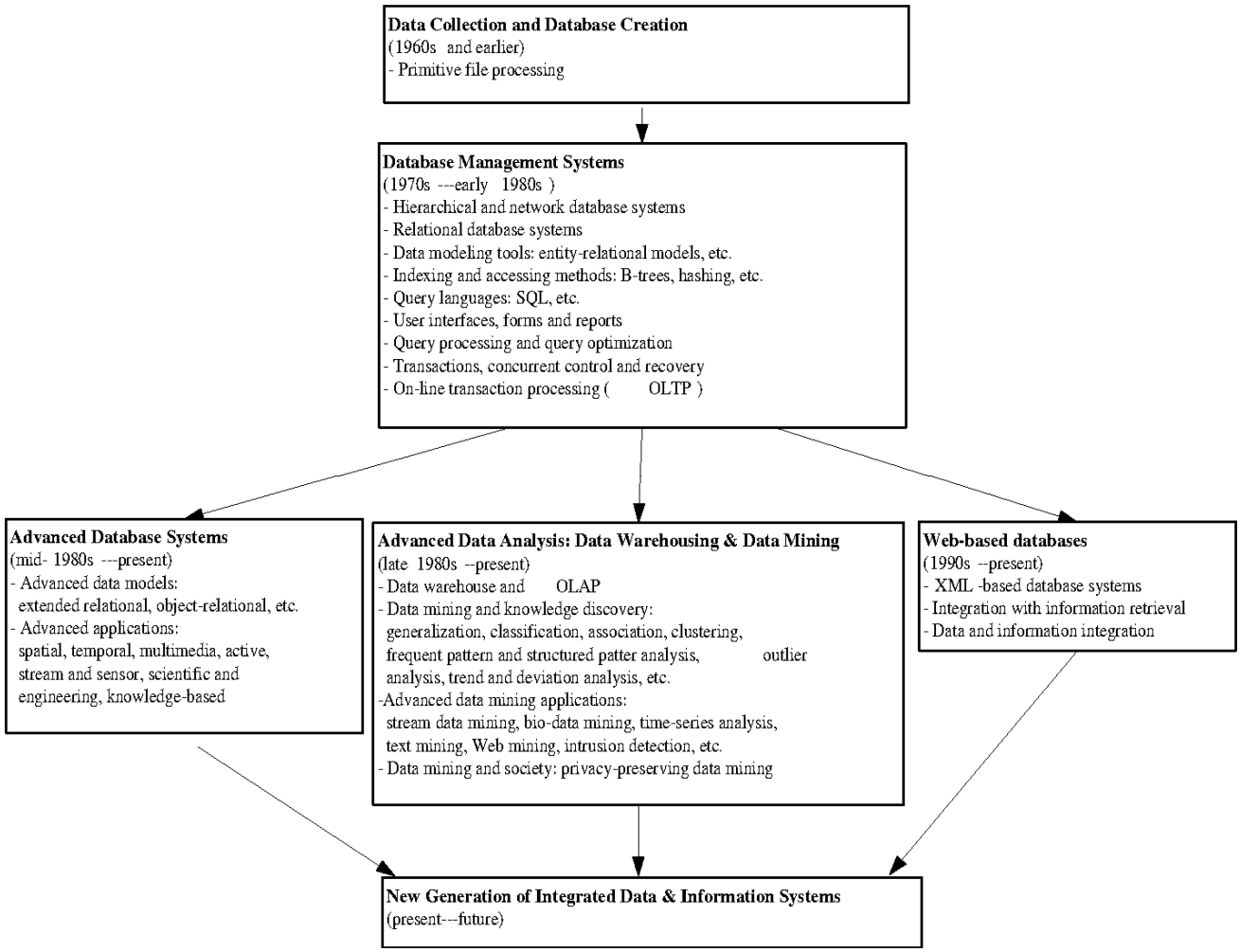
2.4. Γιατί είναι σημαντική

Η εξόρυξη δεδομένων μπορεί να θεωρηθεί ως αποτέλεσμα της φυσικής εξέλιξης της τεχνολογίας των πληροφοριών. Η βιομηχανία συστημάτων βάσεων δεδομένων έχει δει μια εξελικτική πορεία στην ανάπτυξη των ακόλουθων λειτουργιών :

- Συλλογή δεδομένων και δημιουργία βάσεων δεδομένων.
- Διαχείριση δεδομένων (συμπεριλαμβανομένης της αποθήκευσης και ανάκτησης δεδομένων και της επεξεργασίας συναλλαγών βάσεων δεδομένων).
- Προηγμένη ανάλυση δεδομένων.

Αυτή η πρώιμη ανάπτυξη των μηχανισμών συλλογής δεδομένων και δημιουργίας βάσεων δεδομένων αποτέλεσε προϋπόθεση για την μελλοντική ανάπτυξη αποτελεσματικών μηχανισμών αποθήκευσης και ανάκτησης δεδομένων. Με πλέον πολυάριθμα συστήματα βάσεων δεδομένων, η προηγμένη ανάλυση τους έχει γίνει τώρα ο επόμενος στόχος.

Από το 1960, τα συστήματα βάσεων δεδομένων και η τεχνολογία των πληροφοριών εξελίσσονται συστηματικά από πρωτόγονα συστήματα επεξεργασίας αρχείων σε προηγμένα και ισχυρά συστήματα βάσεων δεδομένων. Από το 1970, τα πρώτα ιεραρχικά και δικτυακά συστήματα βάσεων δεδομένων πέρασαν στην ανάπτυξη συστημάτων σχεσιακών βάσεων δεδομένων, σε εργαλεία μοντελοποίησης δεδομένων κ.α. Οι χρήστες έχουν αποκτήσει πλέον μια πιο εύκολη και ευέλικτη πρόσβαση σε δεδομένα (μέσω query languages, user interface



Εικόνα 2.2 : Η εξέλιξη της τεχνολογίας του συστήματος των Βάσεων Δεδομένων.

κ.τ.λ.) και μέθοδοι, όπως το Σύστημα Επεξεργασίας Συναλλαγών (online transaction processing - OLTP), συνέβαλαν στην εξέλιξη και στην ευρεία αποδοχή της σχεσιακής τεχνολογίας ως σημαντικό εργαλείο για την αποτελεσματική αποθήκευση, ανάκτηση και διαχείριση μεγάλων ποσών δεδομένων. Από το 1980, τα database systems βασίζονται σε πλέον προηγμένες εφαρμογές - όπως είναι οι εφαρμογές πολυμέσων, αισθητήρων και πολλές άλλες - και ζητήματα που σχετίζονται με την διανομή, την διαφοροποίηση και την ανταλλαγή δεδομένων μελετώνται εκτενώς. Τέλος, ετερογενή συστήματα βάσεων δεδομένων και παγκόσμια συστήματα πληροφοριών που βασίζονται στο Διαδίκτυο, όπως είναι το World Wide Web (WWW), αναδεικνύονται και παίζουν ζωτικό ρόλο στον κλάδο της πληροφορίας.

Κατά τις τρείς αυτές δεκαετίες παράλληλα είχε ανοδική πορεία και εξέλιξη ο κόσμος του hardware, ο οποίος όχι μόνο έκανε προσιτό τον υπολογιστή στην κοινωνία αλλά προσέφερε και τον εξοπλισμό συλλογής και τα μέσα αποθήκευσης που θα οδηγούσαν στην

ανάκτηση πληροφοριών και στην ανάλυση δεδομένων μέσα από τον τεράστιο αριθμό βάσεων δεδομένων που είχε δημιουργηθεί. Ένα από αυτά τα μέσα αποθήκευσης ή information repositories που προέκυψαν είναι και οι αποθήκες δεδομένων ή αλλιώς data warehouses. Μέσα από αυτές μπορεί να γίνει καθαρισμός και ενσωμάτωση των δεδομένων, καθώς και το Σύστημα Αναλυτικής Επεξεργασίας Συναλλαγών (online analytical processing - OLAP).

Αυτό που κάνει δύσκολη την αποτελεσματική ανάλυση δεδομένων και καθιστά σημαντική την χρήση του data mining είναι εφαρμογές όπως τα συστήματα παρακολούθησης, τηλεπικοινωνίας και τα δίκτυα αισθητήρων. Η ταχέως αναπτυσσόμενη ποσότητα δεδομένων έχει πλέον υπερβεί κατά πολύ την ανθρώπινη ικανότητα κατανόησης χωρίς ισχυρά και άκρως κατάλληλα εργαλεία με αποτέλεσμα τα δεδομένα που συλλέγονται να καθίστανται "ταφικά" ή "data tomps" - αρχεία δεδομένων δηλαδή που σπάνια επισκέπτονται. Κατά συνέπεια, σημαντικές αποφάσεις που πρέπει να ληφθούν, για παράδειγμα, σε μια επιστημονική ή ιατρική έρευνα, συνήθως βασίζονται στην διαίσθηση του εκάστοτε υπεύθυνου, μόνο και μόνο γιατί δεν διατίθενται τα εργαλεία για την εξόρυξη της πολύτιμης γνώσης που εμπεριέχεται στα τεράστια ποσά δεδομένων. Αυτή η διεύρυνση του χάσματος μεταξύ δεδομένων και πληροφοριών, μπορεί να μειωθεί με την συστηματική ανάπτυξη εργαλειών εξόρυξης δεδομένων, τα οποία θα μετατρέψουν τους "τάφους δεδομένων" σε πολύτιμη γνώση. (Han and Kamber, 2006, pp. 1-5)

2.5. Που εφαρμόζεται

2.5.1. Οικονομία

Ένας από τους κλάδους στους οποίους εφαρμόζεται η εξόρυξη δεδομένων είναι η οικονομία. Τα οικονομικά δεδομένα συλλέγονται κυρίως από τράπεζες και από άλλους οικονομικούς οργανισμούς. Συνήθως είναι αξιόπιστα, ολοκληρωμένα, έχουν υψηλή ποιότητα και απαιτούν συστηματική μέθοδο για την ανάλυση τους.

Ουσιαστικά, η εξόρυξη δεδομένων βοηθά τον εν λόγω τομέα στην συλλογή και κατανόηση των δεδομένων, στην βελτίωση τους (data refinement), καθώς και στην δημιουργία, στην εκτίμηση και στην ανάπτυξη ενός μοντέλου. Η σωστή ανάλυση των οικονομικών δεδομένων διευκολύνει στην λήψη καλύτερων αποφάσεων ενεργώντας, πάντα, σύμφωνα με την ανάλυση της αγοράς.

Ανάλογα με τα εργαλεία και τις τεχνικές της εξόρυξης δεδομένων που θα χρησιμοποιηθούν, τα οικονομικά δεδομένα μπορούν να αναλυθούν με τους εξής τρόπους :

- Τα δεδομένα που συλλέγονται από διάφορα οικονομικά ινστιτούτα, αρχικά συγκεντρώνονται σε μια αποθήκη δεδομένων (data warehouse). Έπειτα, χρησιμοποιούνται τεχνικές της πολυδιάστατης ανάλυσης στα δεδομένα που έχουν συλλεχθεί.
- Μέθοδοι της εξόρυξης όπως η επιλογή χαρακτηριστικών (feature selection), βοηθούν στην ταυτοποίηση ποικίλων χαρακτηριστικών, όπως το επίπεδο εισοδήματος του πελάτη, την εξόφληση ανάλογα με τα έσοδα, την πιστωτική του ιστορία κτλ. Επεξεργάζοντας αυτά τα χαρακτηριστικά, η τράπεζα μπορεί να αποφασίσει για τις πολιτικές δανειοδότησης βάσει των σχετικά χαμηλών κινδύνων. Τεχνικές, όπως η συσταδοποίηση (clustering) και η ταξινόμηση (classification), βοηθούν τα οικονομικά ινστιτούτα στην ομαδοποίηση των πελατών που έχουν κοινά χαρακτηριστικά. Η αποτελεσματική συσταδοποίηση και οι μέθοδοι φιλτραρίσματος βοηθούν τις τράπεζες να ταυτοποιούν μία ομάδα πελατών, να συσχετίζουν ένα νέο πελάτη με την παρούσα ομάδα και να τους παρέχουν κοινά οφέλη.
- Με την χρήση των εργαλείων της εξόρυξης δεδομένων γίνεται ανίχνευση απάτης και εγκλημάτων από παραποιημένα δεδομένα από τις διάφορες βάσεις δεδομένων, καθώς και από το ιστορικό συναλλαγών που έγιναν από τους πελάτες. Τεχνικές οπτικοποίησης βοηθούν στην παρουσίαση δεδομένων με διαφορετικές μορφές, όπως γράφοι που βασίζονται σε συγκεκριμένα γνωρίσματα. Προβάλλοντας τα δεδομένα από διάφορες οπτικές γωνίες, η τράπεζα μπορεί να διακρίνει τους πελάτες που έχουν επιχειρήσει παράνομες δραστηριότητες. Εν συνεχεία, κάνοντας μια λεπτομερής έρευνα αυτών των ύποπτων περιπτώσεων μπορεί να πραγματοποιηθεί η εξιχνίαση αυτών των απατών και εγκλημάτων. (El.wikipedia.org, 2018)

2.5.2. Marketing και Πωλήσεις

Σημαντικό και δραστήριο ρόλο παίζουν οι εφαρμογές της εξόρυξης δεδομένων στον τομέα του marketing και των πωλήσεων. Οι εταιρείες διαθέτουν τεράστιους όγκους καταγεγραμμένων δεδομένων τα οποία είναι εξαιρετικά πολύτιμα. Οι τράπεζες υιοθέτησαν νωρίς την τεχνολογία του data mining εξαιτίας των επιτυχιών τους στην χρήση της μηχανικής μάθησης (machine learning) για πιστωτική αξιολόγηση. Μέσω αυτής της τεχνολογίας μπορεί

τώρα να γίνει μείωση της φθοράς των πελατών εντοπίζοντας αλλαγές σε μεμονωμένα τραπεζικά πρότυπα, που μπορεί να δηλώσουν αλλαγή τράπεζας ή και αλλαγή ζωής - όπως μετακόμιση σε άλλη πόλη.

Το Market Basket Analysis (MBA) είναι η χρήση τεχνικών συσχέτισης για την εύρεση ομάδων από στοιχεία που τείνουν να εμφανίζονται μαζί στις συναλλαγές. Η ικανότητα αναγνώρισης και ταυτοποίησης μεμονωμένων πελατών αποτελεί μια αξία που επιτρέπει στους έμπορους λιανικής να παρακολουθούν τις αγορές που κάνουν κάθε φορά μέσω των εκπτωτικών καρτών τους. Τα προσωπικά δεδομένα που θα συγκεντρωθούν θα έχουν πολύ μεγαλύτερη αξία από την αξία των μετρητών της έκπτωσης λόγω του ότι αυτή η ταυτοποίηση επιτρέπει όχι μόνο την ιστορική ανάλυση των μοτίβων που έχουν δημιουργηθεί από τις αγορές αλλά και την αποστολή ειδικών προσφορών προς τους υποψήφιους πελάτες. (Witten and Frank, 2005, pp. 26-27)

2.5.3. Τηλεπικοινωνίες

Ο τομέας των τηλεπικοινωνιών είχε μια εξελικτική πορεία, από την προσφορά τηλεφωνικών υπηρεσιών σε τοπικές και μεγάλης απόστασης περιοχές στην παροχή ποικίλων υπηρεσιών επικοινωνίας, συμπεριλαμβανομένου του φαξ, του κινητού τηλεφώνου, των εικόνων, του e-mail κ.α. Η ενσωμάτωση της τηλεπικοινωνίας, του Internet και πολλών άλλων μέσων επικοινωνίας είναι επίσης σε εξέλιξη. Με την βοήθεια της εξόρυξης, ο υψηλά ανταγωνιστικός κλάδος της τηλεπικοινωνίας μπορεί να εντοπίσει τηλεπικοινωνιακά πρότυπα, να συλλάβει δραστηριότητες απάτης, να κάνει καλύτερη χρήση των πόρων και να βελτιώσει την ποιότητα των υπηρεσιών.

Η εξόρυξη δεδομένων μπορεί να βελτιώσει την βιομηχανία της τηλεπικοινωνίας με τους ακόλουθους τρόπους :

- **Ανάλυση μοτίβων απάτης και ταυτοποίηση ασυνήθιστων προτύπων :** Μια δραστηριότητα απάτης κοστίζει στην βιομηχανία εκατομμύρια δολάρια ετησίως. Είναι σημαντικό να αναγνωρίσει πιθανόν δόλιους χρήστες και την χρήση άτυπων μοτίβων τους, να ανιχνεύσει προσπάθειες δόλιων εισόδων σε λογαριασμούς πελατών και να ανακαλύψει ασυνήθιστα μοτίβα που μπορεί να χρειαστούν ιδιαίτερη προσοχή, όπως περιοδικές κλήσεις από εξοπλισμό, σαν το φαξ, που έχει προγραμματιστεί

εσφαλμένα. Πολλά από αυτά τα πρότυπα μπορούν να ανακαλυφθούν μέσω της πολυδιάστατης ανάλυσης και της ανάλυσης συσταδοποίησης.

- **Υπηρεσίες κινητής τηλεφωνίας :** Κινητές τηλεπικοινωνίες, το Διαδίκτυο και υπηρεσίες πληροφοριών αυξάνονται ανοδικά και γίνονται πιο κοινά στην δουλειά και στην ζωή του ανθρώπου. Ένα σημαντικό χαρακτηριστικό των δεδομένων κινητής τηλεπικοινωνίας είναι η συσχέτιση τους με τις χωροχρονικές πληροφορίες. Η εξόρυξη χωροχρονικών δεδομένων (spatiotemporal data) μπορεί να γίνει απαραίτητη στην εύρεση ορισμένων προτύπων. Για παράδειγμα, όταν υπάρχει ασυνήθιστη κίνηση σε ένα κινητό τηλέφωνο σε συγκεκριμένες περιοχές μπορεί να υποδεικνύει ότι συμβαίνει κάτι ασυνήθιστο σε αυτές τις περιοχές. Επιπλέον, η ευκολία χρήσης αποτελεί ζωτικής σημασίας στην προσέλκυση νέων πελατών για να υιοθετήσουν τις καινούργιες κινητές υπηρεσίες.
- **Χρήση εργαλείων οπτικοποίησης στην ανάλυση τηλεπικοινωνιακών δεδομένων :** Εργαλεία για OLAP visualization, outlier visualization, απεικόνιση σύνδεσης, απεικόνιση συσχέτισης και συσταδοποίηση έχουν αποδειχτεί πολύ χρήσιμα για την ανάλυση τηλεπικοινωνιακών δεδομένων. (Han and Kamber, 2006, pp. 652-653)

2.5.4. Εκπαίδευση

Τα τελευταία χρόνια, υπάρχει ένα αυξανόμενο ενδιαφέρον στη χρήση της εξόρυξης δεδομένων για τη διερεύνηση επιστημονικών ερωτημάτων μέσα στην εκπαιδευτική έρευνα. Αυτό το πεδίο ονομάζεται EDM (Educational Data Mining) και ορίζεται ως ο τομέας της επιστημονικής έρευνας γύρω από την ανάπτυξη μεθόδων ώστε να γίνουν ανακαλύψεις μέσα στα μοναδικά είδη δεδομένων που προέρχονται από εκπαιδευτικές τοποθεσίες. Χρησιμοποιώντας αυτές τις μεθόδους υπάρχει καλύτερη κατανόηση των μαθητών και των εγκαταστάσεων στις οποίες μαθαίνουν. Για παράδειγμα, στην εξόρυξη δεδομένων σχετικά με το πώς οι μαθητές επιλέγουν να χρησιμοποιούν εκπαιδευτικό λογισμικό, μπορεί να αξίζει να εξεταστούν ταυτόχρονα τα δεδομένα στο επίπεδο της πληκτρολόγησης, στο επίπεδο απάντησης, στο επίπεδο συνεδρίας, στο επίπεδο σπουδαστών, στο επίπεδο τάξης και στο επίπεδο σχολείου. (Baker)

Η αύξηση της χρήσης της τεχνολογίας στα εκπαιδευτικά συστήματα, έχει οδηγήσει στην αποθήκευση μεγάλων ποσοτήτων δεδομένων φοιτητών, γεγονός που καθιστά σημαντική

την χρήση του EDM για τη βελτίωση των διαδικασιών διδασκαλίας και της μάθησης. Είναι χρήσιμη σε πολλούς διαφορετικούς τομείς, συμπεριλαμβανομένου του εντοπισμού σπουδαστών που βρίσκονται σε κίνδυνο, της ταυτοποίησης των αναγκών μάθησης για διαφορετικές ομάδες σπουδαστών, της αύξησης του ποσοστού αποφοίτησης, της αποτελεσματικής αξιολόγησης της θεσμικής απόδοσης, της μεγιστοποίησης των πόρων της πανεπιστημιούπολης, και της βελτιστοποίησης της ανανέωσης του προγράμματος σπουδών. (Algarni, 2016)

Το Educational Data Mining αναφέρεται σε τεχνικές, εργαλεία και έρευνες που αποσκοπούν στην αυτόματη εξαγωγή γνώσης από μεγάλα αποθετήρια δεδομένων που παράγονται από ή σχετίζονται με τις μαθησιακές δραστηριότητες των ανθρώπων σε εκπαιδευτικά περιβάλλοντα. Πολύ συχνά, τα δεδομένα αυτά είναι εκτεταμένα, συγκεκραμένα και ακριβή. Για παράδειγμα, πολλά συστήματα διαχείρισης μάθησης (Learning Management Systems) παρακολουθούν πληροφορίες, όπως όταν κάποιος μαθητής είχε πρόσβαση σε κάποιο μαθησιακό αντικείμενο, πόσες φορές είχε πρόσβαση και πόσα λεπτά το μαθησιακό αντικείμενο εμφανίστηκε στην οθόνη του υπολογιστή του χρήστη.

Σε άλλες περιπτώσεις, τα δεδομένα είναι λιγότερο συγκεκριμένα. Για παράδειγμα, το πανεπιστημιακό αντίγραφο ενός φοιτητή μπορεί να περιέχει μια λίστα με τα μαθήματα που έχει πάρει ο σπουδαστής, τον βαθμό που πήρε σε κάθε μάθημα και πότε ο φοιτητής επέλεξε ή άλλαξε τον ακαδημαϊκό του κύκλο. Το EDM αξιοποιεί και τους δύο τύπους δεδομένων για να ανακαλύψει σημαντικές πληροφορίες σχετικά με τους διαφορετικούς τύπους μαθητών και τον τρόπο με τον οποίο μαθαίνουν, τη δομή του πεδίου γνώσης και την επίδραση των εκπαιδευτικών στρατηγικών που ενσωματώνονται σε διάφορα περιβάλλοντα εκμάθησης. Αυτές οι αναλύσεις παρέχουν νέες πληροφορίες που θα ήταν δύσκολο να διακριθούν εξετάζοντας τα ακατέργαστα δεδομένα. Για παράδειγμα, η ανάλυση δεδομένων από ένα LMS μπορεί να αποκαλύψει μια σχέση μεταξύ των αντικειμένων μάθησης που έχει πρόσβαση ένας φοιτητής κατά τη διάρκεια του μαθήματος και την τελική βαθμολογία του μαθήματος. Αυτές οι πληροφορίες παρέχουν γνώσεις για το σχεδιασμό των μαθησιακών περιβαλλόντων, που επιτρέπει στους φοιτητές, τους εκπαιδευτικούς, τους διευθυντές σχολείων και τους διαμορφωτές της εκπαιδευτικής πολιτικής να λαμβάνουν τεκμηριωμένες αποφάσεις σχετικά με τον τρόπο αλληλεπίδρασης, παροχής και διαχείρισης εκπαιδευτικών πόρων.

Οι Ryan S. Baker και Kalina Yacef προσδιόρισαν τους ακόλουθους τέσσερις στόχους της EDM:

1. **Πρόβλεψη της μελλοντικής μαθησιακής συμπεριφοράς των μαθητών** : Με τη χρήση της μοντελοποίησης μαθητών, ο στόχος αυτός μπορεί να επιτευχθεί με τη δημιουργία μαθησιακών μοντέλων που ενσωματώνουν τα χαρακτηριστικά του εκπαιδευόμενου, συμπεριλαμβανομένων λεπτομερών πληροφοριών όπως η γνώση, οι συμπεριφορές και τα κίνητρα για μάθηση. Η εμπειρία του χρήστη και η γενική ικανοποίησή του από τη μάθηση μετριούνται επίσης.
2. **Ανακάλυψη ή βελτίωση μοντέλων πεδίου** : Μέσω των διαφόρων μεθόδων και εφαρμογών του EDM, είναι δυνατή η ανακάλυψη νέων μοντέλων και η βελτίωση των ήδη υπαρχόντων. Παραδείγματα περιλαμβάνουν την απεικόνιση του εκπαιδευτικού περιεχομένου για την εμπλοκή των εκπαιδευομένων και τον προσδιορισμό των βέλτιστων εκπαιδευτικών ακολουθιών για την υποστήριξη του μαθησιακού στυλ του μαθητή.
3. **Μελετώντας τα αποτελέσματα της εκπαιδευτικής υποστήριξης** που μπορεί να επιτευχθεί μέσω των συστημάτων μάθησης.
4. **Προώθηση της επιστημονικής γνώσης σχετικά με τη μάθηση και τους μαθητευόμενους** με την οικοδόμηση και την ενσωμάτωση μαθησιακών μοντέλων, το πεδίο της έρευνας EDM και την τεχνολογία και το λογισμικό που χρησιμοποιείται.

Μια λίστα των πρωτογενών εφαρμογών του EDM παρέχεται από τους Cristobal Romero και Sebastian Ventura και είναι οι εξής :

- Ανάλυση και απεικόνιση δεδομένων.
- Παροχή σχολίων για τους εκπαιδευτές υποστήριξης.
- Συστάσεις για φοιτητές.
- Πρόβλεψη της απόδοσης των σπουδαστών.
- Μοντελοποίηση φοιτητών.
- Ανίχνευση ανεπιθύμητων μαθησιακών συμπεριφορών.
- Ομαδοποίηση μαθητών.
- Ανάλυση κοινωνικού δικτύου.
- Σχεδιασμός και προγραμματισμός.
- Κατασκευή μαθημάτων : Το EDM μπορεί να εφαρμοστεί στα συστήματα διαχείρισης μαθημάτων, όπως το Moodle ανοιχτού κώδικα. Το Moodle περιέχει δεδομένα χρήσης που περιλαμβάνουν διάφορες δραστηριότητες από χρήστες, όπως τα αποτελέσματα

των δοκιμών, το ποσό των αναγνώσεων που έχουν ολοκληρωθεί και τη συμμετοχή σε φόρουμ συζητήσεων. Τα εργαλεία εξόρυξης δεδομένων μπορούν να χρησιμοποιηθούν για να προσαρμόσουν τις μαθησιακές δραστηριότητες για κάθε χρήστη και να προσαρμόσουν το ρυθμό με τον οποίο ο φοιτητής ολοκληρώνει το μάθημα. Αυτό είναι ιδιαίτερα επωφελές για online μαθήματα με διαφορετικά επίπεδα ικανότητας.

Νέα έρευνα σε περιβάλλοντα κινητής μάθησης υποδεικνύει επίσης ότι η εξόρυξη δεδομένων μπορεί να είναι χρήσιμη. Η εξόρυξη δεδομένων μπορεί να χρησιμοποιηθεί για την παροχή εξατομικευμένου περιεχομένου στους χρήστες κινητών τηλεφώνων, παρά τις διαφορές στη διαχείριση περιεχομένου μεταξύ κινητών συσκευών και τυπικών υπολογιστών και προγραμμάτων περιήγησης ιστού.

Οι νέες εφαρμογές EDM θα επικεντρωθούν στο να επιτρέπουν στους μη τεχνικούς χρήστες να χρησιμοποιούν και να συμμετέχουν σε εργαλεία και δραστηριότητες εξόρυξης δεδομένων, καθιστώντας τη συλλογή και επεξεργασία δεδομένων πιο προσιτή σε όλους τους χρήστες EDM. Παραδείγματα περιλαμβάνουν εργαλεία στατιστικής και οπτικοποίησης που αναλύουν τα κοινωνικά δίκτυα και την επιρροή τους στα μαθησιακά αποτελέσματα και την παραγωγικότητα. (En.wikipedia.org, 2019)

2.5.5. Υγειονομική περίθαλψη

Η βιομηχανία υγειονομικής περίθαλψης παράγει τεράστιες ποσότητες δεδομένων, συμπεριλαμβανομένων του ηλεκτρονικού αρχείου υγείας (Electronic Health Record) ή των ηλεκτρονικών ιατρικών αρχείων (Electronic Medical Records), των δεδομένων σχετικά με την ανάπτυξη φαρμάκων και των δεδομένων των ασθενών. Η εξόρυξη δεδομένων στην υγειονομική περίθαλψη χρησιμοποιείται κυρίως για την πρόβλεψη διαφόρων ασθενειών και την παροχή συμβουλών στους γιατρούς κατά τη λήψη αποφάσεων. Χρησιμοποιώντας την εξόρυξη, ο κλάδος της υγείας μπορεί να επωφεληθεί σε πολλούς τομείς όπως είναι η ιατρική έρευνα, τα φαρμακευτικά προϊόντα, οι ιατρικές συσκευές, η διαχείριση των νοσοκομείων, η ασφάλιση υγείας, η ανίχνευση και η πρόληψη της απάτης.

Ο τομέας αυτός αντιμετωπίζει κάποια πίεση για τη μείωση του κόστους, ενώ ταυτόχρονα αυξάνει την ποιότητα των υπηρεσιών. Οι τεχνικές εξόρυξης δεδομένων

χρησιμοποιούνται ευρύτερα στους τομείς του διαβήτη, των καρδιακών παθήσεων και των καρδιακών προσβολών. Ορισμένες εφαρμογές εξόρυξης δεδομένων στο υγειονομική περίθαλψη περιλαμβάνουν τα ακόλουθα :

- Αποτελεσματική διαχείριση των νοσοκομειακών πόρων.
- Καλύτερη σχέση με τον πελάτη.
- Μείωση της απάτης στην ασφάλιση.
- Πιο έξυπνες τεχνικές θεραπείας.
- Βελτιωμένη φροντίδα των ασθενών.

Τα οφέλη της εξόρυξης δεδομένων στον τομέα της υγειονομικής περίθαλψης παρουσιάζονται ως εξής :

- Οι ασθενείς λαμβάνουν πιο προσιτές και καλύτερες υπηρεσίες υγειονομικής περίθαλψης.
- Οι πάροχοι υγειονομικής περίθαλψης χρησιμοποιούν εξόρυξη δεδομένων και ανάλυση δεδομένων για την εύρεση βέλτιστων πρακτικών.
- Ο ασφαλιστικός οργανισμός μπορεί πλέον να ανιχνεύσει την κατάχρηση της ιατρικής περίθαλψης και την απάτη.
- Ο πάροχος υγειονομικής περίθαλψης μπορεί να λάβει καλύτερα αποφάσεις σχετικά με τον ασθενή.

Η ιδιωτικότητα και η ασφάλεια των δεδομένων των ασθενών αποτελεί μεγάλη πρόκληση λόγω της ευαισθησίας των δεδομένων της υγειονομικής περίθαλψης. Δεδομένου ότι τα δεδομένα για την υγεία περιέχουν προσωπικές, ευαίσθητες πληροφορίες, υπάρχει ο κίνδυνος εισβολής προσωπικών δεδομένων. (N. O. Sadiku, G. Eze and M. Musa, 2018)

2.6. Διαδικασία Εξόρυξης Δεδομένων

Η λέξη “διαδικασία” είναι πολύ σημαντική εδώ. Ακόμη και σε μερικά επαγγελματικά περιβάλλοντα υπάρχει η πεποίθηση ότι η εξόρυξη δεδομένων αποτελείται απλώς από τη συλλογή και την εφαρμογή ενός εργαλείου που βασίζεται σε υπολογιστή για να ταιριάζει με το πρόβλημα που παρουσιάζεται και να αποκτά αυτόματα μια λύση. Αυτή είναι μια εσφαλμένη αντίληψη βασισμένη σε μια τεχνητή εξιδανίκευση του κόσμου και υπάρχουν πολλοί λόγοι για τους οποίους αυτό είναι λανθασμένο. Ένας λόγος είναι ότι η εξόρυξη δεδομένων δεν είναι απλά μια συλλογή από μεμονωμένα εργαλεία, το καθένα εντελώς

διαφορετικό από το άλλο και περιμένοντας να ταιριάξει στο πρόβλημα. Ένας δεύτερος λόγος έγκειται στην έννοια της αντιστοίχισης ενός προβλήματος σε μια τεχνική. Πολύ σπάνια μόνο μια και απλή εφαρμογή μιας μεθόδου θα είναι επαρκής. Στην πραγματικότητα, όπως αναφέρθηκε και πιο πάνω, γίνεται μια επαναληπτική διαδικασία. Κάποιος μελετά τα δεδομένα, εξετάζει χρησιμοποιώντας κάποια αναλυτική τεχνική, αποφασίζει να το εξετάσει με άλλο τρόπο, ίσως τροποποιώντας το και μετά επιστρέφει στην αρχή και εφαρμόζει ένα άλλο εργαλείο ανάλυσης δεδομένων, φτάνοντας σε είτε καλύτερα είτε διαφορετικά αποτελέσματα. Αυτό μπορεί να συμβαίνει πολλές φορές. Κάθε τεχνική χρησιμοποιείται για να διερευνηθούν ελαφρώς διαφορετικές πτυχές των δεδομένων. Αυτό που ουσιαστικά περιγράφεται εδώ είναι ένα ταξίδι της ανακάλυψης που κάνει την σύγχρονη εξόρυξη δεδομένων συναρπαστική. Ωστόσο, η εξόρυξη δεδομένων δεν αποτελεί μια τυχαία εφαρμογή στατιστικής και μηχανικής μάθησης μεθόδων και εργαλείων. Δεν είναι μια τυχαία βόλτα μέσα από το χώρο των αναλυτικών τεχνικών, αλλά μια προσεκτικά σχεδιασμένη και εξεταζόμενη διαδικασία που αποφασίζει τι θα είναι πιο χρήσιμο, πολλά υποσχόμενο και αποκαλυπτικό. Είναι σημαντικό να συνειδητοποιήσουμε ότι το πρόβλημα της ανακάλυψης ή της εκτίμησης των εξαρτήσεων από τα δεδομένα ή της ανακάλυψης εντελώς νέων δεδομένων είναι μόνο ένα μέρος της γενικής πειραματικής διαδικασίας που χρησιμοποιείται από επιστήμονες, μηχανικούς και άλλους που εφαρμόζουν τα συνήθη βήματα για να βγάλουν συμπεράσματα από τα δεδομένα. Η γενική πειραματική διαδικασία προσαρμοσμένη στα προβλήματα της εξόρυξης δεδομένων περιλαμβάνει τα ακόλουθα βήματα :

1. **Δήλωση προβλήματος και διατύπωση υπόθεσης** : Οι περισσότερες μελέτες μοντελοποίησης που βασίζονται σε δεδομένα εκτελούνται σε έναν συγκεκριμένο τομέα εφαρμογής. Ως εκ τούτου, είναι απαραίτητες ειδικές γνώσεις και εμπειρίες στον τομέα, προκειμένου να καταλήξουμε σε μια σημαντική δήλωση προβλημάτων. Δυστυχώς, πολλές μελέτες εφαρμογής τείνουν να επικεντρώνονται στην τεχνική της εξόρυξης δεδομένων σε βάρος μιας σαφούς δήλωσης του προβλήματος. Σε αυτό το βήμα, ένας σχεδιαστής συνήθως καθορίζει ένα σύνολο μεταβλητών για την άγνωστη εξάρτηση και, αν είναι δυνατόν, μια γενική μορφή αυτής της εξάρτησης ως αρχική υπόθεση. Μπορεί να υπάρχουν αρκετές υποθέσεις που διατυπώνονται για ένα μόνο πρόβλημα σε αυτό το στάδιο. Το πρώτο βήμα απαιτεί τη συνδυασμένη εξειδίκευση ενός τομέα εφαρμογής και ενός μοντέλου εξόρυξης δεδομένων. Στην πράξη, συνήθως σημαίνει στενή αλληλεπίδραση μεταξύ του ειδικού της

εξόρυξης δεδομένων και του ειδικού της εφαρμογής. Σε επιτυχημένες εφαρμογές εξόρυξης δεδομένων, η συνεργασία αυτή δεν σταματά στην αρχική φάση, αλλά συνεχίζεται κατά τη διάρκεια ολόκληρης της διαδικασίας της εξόρυξης δεδομένων.

2. **Συλλογή δεδομένων** : Αυτό το βήμα αφορά τον τρόπο με τον οποίο παράγονται και συλλέγονται τα δεδομένα. Σε γενικές γραμμές, υπάρχουν δύο διακριτές πιθανότητες. Η πρώτη είναι όταν η διαδικασία παραγωγής δεδομένων βρίσκεται υπό τον έλεγχο ενός ειδικού (modeler - σχεδιαστής). Αυτή η προσέγγιση είναι γνωστή ως σχεδιασμένο πείραμα (designed experiment). Η δεύτερη πιθανότητα είναι όταν ο ειδικός δεν μπορεί να επηρεάσει τη διαδικασία παραγωγής δεδομένων. Αυτή είναι γνωστή ως παρατηρητική προσέγγιση (observational approach). Ένα παρατηρητικό περιβάλλον, δηλαδή η παραγωγή τυχαίων δεδομένων, τοποθετείται στις περισσότερες εφαρμογές εξόρυξης δεδομένων. Συνήθως, η κατανομή δειγματοληψίας είναι εντελώς άγνωστη αφού συγκεντρωθούν τα δεδομένα, ή δίνεται εν μέρει στη διαδικασία συλλογής δεδομένων. Είναι πολύ σημαντικό, ωστόσο, να κατανοήσουμε πώς η συλλογή δεδομένων επηρεάζει τη θεωρητική κατανομή της, αφού μια τέτοια εκ των προτέρων γνώση μπορεί να είναι πολύ χρήσιμη για τη μοντελοποίηση και, στη συνέχεια, για την τελική ερμηνεία των αποτελεσμάτων. Επίσης, είναι σημαντικό να διασφαλιστεί ότι τα δεδομένα που χρησιμοποιούνται για την εκτίμηση ενός μοντέλου και τα δεδομένα που χρησιμοποιούνται αργότερα για τη δοκιμή και την εφαρμογή ενός μοντέλου προέρχονται από την ίδια άγνωστη κατανομή δειγματοληψίας. Αν αυτό δεν ισχύει, το εκτιμώμενο μοντέλο δεν μπορεί να χρησιμοποιηθεί με επιτυχία σε μια τελική εφαρμογή των αποτελεσμάτων.
3. **Προ-επεξεργασία δεδομένων** : Στο παρατηρησιακό περιβάλλον, τα δεδομένα συνήθως "συλλέγονται" από τις υπάρχουσες βάσεις δεδομένων, τις αποθήκες δεδομένων και τα data marts. Η προ-επεξεργασία δεδομένων συνήθως περιλαμβάνει τουλάχιστον δύο κοινές εργασίες:
    i. Ανίχνευση (και απομάκρυνση) / Outlier detection (and removal) : Οι αποκλίσεις (outliers) είναι ασυνήθιστες τιμές δεδομένων που δεν συμφωνούν με τις περισσότερες παρατηρήσεις. Συνήθως, οι αποκλίσεις προκύπτουν από σφάλματα μέτρησης, σφάλματα κωδικοποίησης και καταγραφής, και μερικές φορές είναι φυσικές, ασυνήθιστες τιμές. Αυτά

τα μη αντιπροσωπευτικά δείγματα μπορούν να επηρεάσουν σοβαρά το μοντέλο που παράγεται αργότερα. Υπάρχουν δύο στρατηγικές για την αντιμετώπιση των αποκλίσεων:

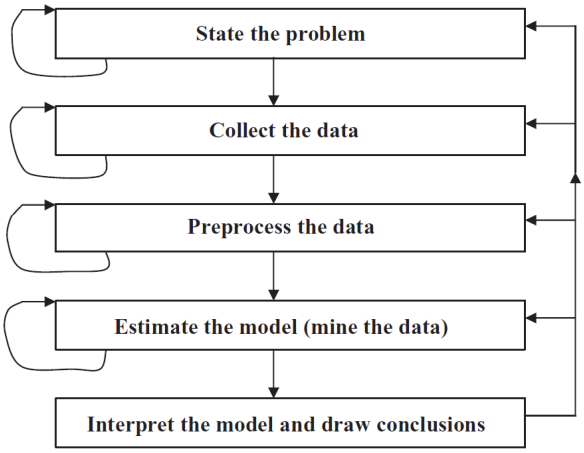
a. Ανίχνευση και αφαίρεση των αποκλίσεων ως μέρος της φάσης της προ-επεξεργασίας, ή
b. Ανάπτυξη ισχυρών μεθόδων μοντελοποίησης που δεν είναι ευαίσθητες στις αποκλίσεις.

ii. Κλιμάκωση, κωδικοποίηση και επιλογή χαρακτηριστικών.

Η προ-επεξεργασία δεδομένων περιλαμβάνει διάφορα βήματα, όπως μεταβλητή κλιμάκωση και διαφορετικούς τύπους κωδικοποίησης. Για παράδειγμα, ένα χαρακτηριστικό με το εύρος [0,1] και το άλλο με το εύρος [-100, 1000] δεν θα έχει το ίδιο βάρος στην εφαρμοζόμενη τεχνική. Επίσης, το καθένα θα επηρεάσει τα τελικά αποτελέσματα της εξόρυξης δεδομένων διαφορετικά. Ως εκ τούτου, συνιστάται να τα κλιμακωθούν και να έρθουν και τα δύο χαρακτηριστικά στο ίδιο βάρος για περαιτέρω ανάλυση. Επιπλέον, οι μέθοδοι κωδικοποίησης για συγκεκριμένες εφαρμογές συνήθως επιτυγχάνουν μείωση των διαστάσεων παρέχοντας έναν μικρότερο αριθμό πληροφοριακών στοιχείων για την επακόλουθη μοντελοποίηση δεδομένων. Αυτές οι δύο κατηγορίες εργασιών προ-επεξεργασίας είναι μόνο επεξηγηματικά παραδείγματα ενός μεγάλου φάσματος δραστηριοτήτων προ-επεξεργασίας σε μια διαδικασία εξόρυξης δεδομένων. Τα βήματα προ-επεξεργασίας δεδομένων δεν πρέπει να θεωρούνται εντελώς ανεξάρτητα από άλλες φάσεις εξόρυξης δεδομένων. Σε κάθε επανάληψη της διαδικασίας εξόρυξης δεδομένων, όλες οι δραστηριότητες, μαζί, θα μπορούσαν να καθορίσουν νέα και βελτιωμένα σύνολα δεδομένων για επακόλουθες επαναλήψεις. Γενικά, μια καλή μέθοδος προ-επεξεργασίας παρέχει μια βέλτιστη αναπαράσταση για μια τεχνική εξόρυξης δεδομένων, ενσωματώνοντας μια εκ των προτέρων γνώση στη μορφή συγκεκριμένης εφαρμογής κλιμάκωσης και κωδικοποίησης.

4. **Εκτίμηση μοντέλου** : Η επιλογή και η εφαρμογή της κατάλληλης τεχνικής εξόρυξης δεδομένων είναι η κύρια εργασία σε αυτή τη φάση. Αυτή η διαδικασία δεν είναι απλή. Συνήθως, στην πράξη, η εφαρμογή βασίζεται σε πολλά μοντέλα, και η επιλογή του καλύτερου είναι μια πρόσθετη εργασία.

5. **Ερμήνευση του μοντέλου και εξαγωγή συμπερασμάτων** : Στις περισσότερες περιπτώσεις, τα μοντέλα εξόρυξης δεδομένων πρέπει να βοηθούν στη λήψη αποφάσεων. Ως εκ τούτου, τα μοντέλα αυτά πρέπει να ερμηνεύονται προκειμένου να είναι χρήσιμα, επειδή οι άνθρωποι δεν είναι πιθανό να βασίζουν τις αποφάσεις τους σε πολύπλοκα μοντέλα "μαύρου κουτιού". Ωστόσο, οι στόχοι της ακρίβειας του μοντέλου και της ακρίβειας της ερμηνείας του είναι κάπως αντιφατικοί. Συνήθως, τα απλά μοντέλα είναι πιο ερμηνεύσιμα, αλλά είναι επίσης λιγότερο ακριβή. Οι σύγχρονες μέθοδοι εξόρυξης δεδομένων αναμένεται να αποδώσουν εξαιρετικά ακριβή αποτελέσματα χρησιμοποιώντας μοντέλα υψηλών διαστάσεων. Το πρόβλημα της ερμηνείας αυτών των μοντέλων (επίσης πολύ σημαντικό) θεωρείται ξεχωριστό έργο, με συγκεκριμένες τεχνικές για την επικύρωση των αποτελεσμάτων. Ένας χρήστης δεν θέλει εκατοντάδες σελίδες αριθμητικών αποτελεσμάτων. Δεν τις καταλαβαίνει, δεν μπορεί να τις συνοψίσει, να τις ερμηνεύσει και να τις χρησιμοποιήσει για επιτυχή λήψη αποφάσεων.



Εικόνα 2.3 : Η διαδικασία της Εξόρυξης Δεδομένων.

Όλες οι φάσεις, ξεχωριστά, και ολόκληρη η διαδικασία εξόρυξης δεδομένων, στο σύνολό τους, είναι ιδιαίτερα επαναληπτικές, όπως φαίνεται στο παραπάνω σχήμα. Η καλή κατανόηση της όλης διαδικασίας είναι σημαντική για κάθε επιτυχή εφαρμογή. Ανεξάρτητα από το πόσο ισχυρή είναι η μέθοδος εξόρυξης δεδομένων που χρησιμοποιήθηκε στο βήμα 4, το προκύπτον μοντέλο δεν θα είναι έγκυρο αν τα δεδομένα δεν συλλεχθούν και προ-επεξεργαστούν σωστά ή εάν η διατύπωση προβλημάτων δεν έχει νόημα. (Kantardzic, 2011, pp. 6-9).

2.7. Εργασίες της Εξόρυξης δεδομένων

2.7.1. Εκτίμηση - Estimation

Η κατηγοριοποίηση ασχολείται με διακριτά αποτελέσματα: ναι ή όχι. ιλαρά, ερυθρά, ή ανεμοβλογιά. Η εκτίμηση ασχολείται με συνεχώς εκτιμημένα αποτελέσματα. Δίνοντας ορισμένα δεδομένα εισόδου, η εκτίμηση βρίσκει μια τιμή για κάποια άγνωστη συνεχή μεταβλητή όπως το εισόδημα, το ύψος ή το υπόλοιπο μιας πιστωτικής σας κάρτας.

Στην πράξη, η εκτίμηση χρησιμοποιείται συχνά για την εκτέλεση μιας εργασίας της ταξινόμησης. Μια εταιρεία πιστωτικών καρτών που επιθυμεί να πουλήσει διαφημιστικό χώρο στους χρεωστικούς φακέλους της σε έναν κατασκευαστή για μπότες του σκι, θα μπορούσε να δημιουργήσει ένα μοντέλο ταξινόμησης που θέτει όλους τους κατόχους καρτών του σε μία από τις δύο κατηγορίες, σκιέρ ή μη. Μια άλλη προσέγγιση είναι να κατασκευαστεί ένα μοντέλο που αναθέτει σε κάθε κάτοχο κάρτας μια "τάση προς το σκορ". Αυτό μπορεί να είναι μια τιμή από το 0 έως το 1 που υποδεικνύει την εκτιμώμενη πιθανότητα ότι ο κάτοχος είναι σκιέρ. Η εργασία ταξινόμησης τώρα καταλήγει στην δημιουργία ενός ορίου στο σκορ. Όσοι έχουν βαθμολογία μεγαλύτερη ή ίση με το όριο ταξινομούνται ως σκιέρ, και οποιοσδήποτε με χαμηλότερη βαθμολογία θεωρείται ότι δεν είναι σκιέρ.

Η προσέγγιση της εκτίμησης έχει το μεγάλο πλεονέκτημα ότι τα μεμονωμένα αρχεία μπορεί να έχουν σειρά κατάταξης σύμφωνα με την εκτίμηση. Για να δούμε τη σημασία αυτού, φανταστείτε ότι η εν λόγω εταιρεία, που φτιάχνει μπότες του σκι, έχει στον προϋπολογισμό της μια ταχυδρομική αποστολή των 500.000 κομματιών. Εάν χρησιμοποιείται η προσέγγιση ταξινόμησης και ταυτοποιούνται 1,5 εκατομμύρια σκιέρ, τότε θα μπορούσε απλά να τοποθετήσει μια διαφήμιση στους λογαριασμούς 500.000 ατόμων που επιλέχθηκαν τυχαία από εκείνη την ομάδα. Εάν, από την άλλη πλευρά, κάθε κάτοχος κάρτας έχει τάση προς το σκορ, μπορεί να στείλει τη διαφήμιση στους 500.000 πιο πιθανούς υποψηφίους.

Παραδείγματα εργασιών εκτίμησης περιλαμβάνουν:

- Εκτίμηση του αριθμού των παιδιών σε μια οικογένεια.
- Εκτίμηση του συνολικού οικογενειακού εισοδήματος μιας οικογένειας.
- Εκτίμηση της διάρκειας ζωής ενός πελάτη.

Μοντέλα παλινδρόμησης και νευρωνικά δίκτυα είναι κατάλληλα για τις εργασίες εκτίμησης. Η ανάλυση επιβίωσης (survival analysis) είναι επίσης κατάλληλη, όπου ο στόχος είναι να εκτιμηθεί ο χρόνος σε ένα συμβάν, όπως είναι η διακοπή ενός πελάτη.

2.7.2. Πρόβλεψη - Prediction

Η πρόβλεψη είναι ίδια με την ταξινόμηση ή την εκτίμηση, με εξαίρεση ότι τα αρχεία ταξινομούνται σύμφωνα με κάποια προβλεπόμενη μελλοντική συμπεριφορά ή εκτιμώμενη μελλοντική αξία. Σε μια εργασία πρόβλεψης, ο μόνος τρόπος για να ελεγχθεί η ακρίβεια της ταξινόμησης είναι να περιμένουμε μέχρι να δούμε. Ο πρωταρχικός λόγος για την αντιμετώπιση της πρόβλεψης ως ξεχωριστής εργασίας από την ταξινόμηση και την εκτίμηση, είναι ότι στο μοντέλο πρόβλεψης υπάρχουν επιπρόσθετα ζητήματα σχετικά με τη χρονική σχέση των μεταβλητών εισόδου ή των προγνωστικών για τη στοχευμένη μεταβλητή.

Οποιαδήποτε από τις τεχνικές που χρησιμοποιούνται για ταξινόμηση και εκτίμηση μπορούν να προσαρμοστούν για χρήση στην πρόβλεψη χρησιμοποιώντας εκπαιδευτικά παραδείγματα όπου η τιμή της προβλεπόμενης μεταβλητής είναι ήδη γνωστή, μαζί με τα ιστορικά δεδομένα για αυτά παραδείγματα. Τα ιστορικά δεδομένα χρησιμοποιούνται για την κατασκευή ενός μοντέλου που εξηγεί την τρέχουσα παρατηρούμενη συμπεριφορά. Όταν το μοντέλο αυτό εφαρμόζεται στις τρέχουσες εισόδους, το αποτέλεσμα είναι μια πρόβλεψη της μελλοντικής συμπεριφοράς.

Παραδείγματα εργασιών πρόβλεψης περιλαμβάνουν:

- Πρόβλεψη των πελατών που θα αποχωρήσουν εντός των επόμενων 6 μηνών.
- Προβλέψεις για το ποιοι συνδρομητές τηλεφώνου θα παραγγείλουν υπηρεσία προστιθέμενης αξίας όπως τριμερής κλήση ή ηχογραφημένο μήνυμα.

Οι περισσότερες τεχνικές εξόρυξης δεδομένων είναι κατάλληλες για χρήση στην πρόβλεψη, όσο τα δεδομένα εκπαίδευσης είναι διαθέσιμα στην κατάλληλη μορφή. Η επιλογή της τεχνικής εξαρτάται από τη φύση των δεδομένων εισόδου, τον τύπο της αξίας που πρέπει να προβλεφθεί και τη σημασία που αποδίδεται στην ερμηνεία της πρόβλεψης. (Berry and Linoff, 2004, pp. 9-11)

2.7.3. Παλινδρόμηση - Regression

Ένας από τους κύριους σκοπούς της προσαρμογής καμπυλών είναι η εκτίμηση της εξαρτημένης μεταβλητής από την ανεξάρτητη μεταβλητή. Η μέθοδος ή η διαδικασία εκτίμησης ονομάζεται παλινδρόμηση. Ουσιαστικά, χρησιμοποιείται για να απεικονιστεί ένα στοιχειώδες δεδομένο σε μια πραγματική μεταβλητή υπό πρόβλεψη. Η παλινδρόμηση προϋποθέτει ότι τα σχετικά δεδομένα ταιριάζουν με μερικά γνωστά είδη συναρτήσεων (π.χ. γραμμική, λογαριθμική κλπ) και καθορίζει τη συνάρτηση που μοντελοποιεί καλύτερα τα δεδομένα που έχουν δοθεί. Η κύρια διαφορά της παλινδρόμησης με την κατηγοριοποίηση είναι ότι το υπό πρόβλεψη χαρακτηριστικό παίρνει συνεχείς τιμές.

2.7.4. Ανάλυση Χρονολογικών Σειρών - Time Series Analysis

Μελετά την τιμή ενός γνωρίσματος καθώς μεταβάλλεται στο χρόνο με κάποια περιοδικότητα (π.χ. ημερήσια, εβδομαδιαία, μηνιαία κλπ). Ως χρονολογική σειρά ορίζεται η ακολουθία των τιμών μιας μεταβλητής οι οποίες λαμβάνονται σε προκαθορισμένα χρονικά σημεία που συνήθως ισαπέχουν ή αναφέρονται σε διαδοχικές περιόδους ίδιας διάρκειας. Η γραφική απεικόνιση των χρονολογικών σειρών, οι οποίες εκφράζονται σε απόλυτα ή σε σχετικά μεγέθη, γίνεται βάσει ειδικών διαγραμμάτων, τα λεγόμενα χρονογράμματα ή χρονοδιαγράμματα. Υπάρχουν τρείς βασικές λειτουργίες που χρησιμοποιούνται στην ανάλυση χρονοσειρών. Η πρώτη περιλαμβάνει τη χρησιμοποίηση μονάδων μέτρησης απόστασης ώστε να καθοριστούν οι ομοιότητες ανάμεσα σε διαφορετικές χρονοσειρές. Η δεύτερη λειτουργία εξετάζει τη δομή της χρονοσειράς για να κατηγοριοποιήσει τη συμπεριφορά της. Τέλος, η τρίτη λειτουργία χρησιμοποιεί διαγράμματα χρονοσειρών για την πρόβλεψη μελλοντικών τιμών.

2.7.5. Σύνοψη - Summarization

Απεικονίζει τα δεδομένα σε υποσύνολα με συνοδευτικές απλές περιγραφές και χαρακτηρίζει τα περιεχόμενα της βάσης δεδομένων. Εξάγει αντιπροσωπευτικές πληροφορίες για την βάση δεδομένων και παράγει τους ονομαζόμενους *χαρακτηριστικούς κανόνες (characteristic rules)*. Καθώς ένας κύβος δεδομένων περιέχει συγκεντρωμένα δεδομένα, οι απλές λειτουργίες OLAP ταιριάζουν στον σκοπό της σύνοψης.

2.7.6. Ανακάλυψη Ακολουθιών - Sequence Discovery

Χρησιμοποιείται για τον καθορισμό προτύπων σε σειριακά δεδομένα. Σειρές διακριτών τιμών ή καταστάσεων γνωρισμάτων δεδομένων συνθέτουν μια ακολουθία. Τα δεδομένα, τόσο στην ανακάλυψη ακολουθιών όσο και στην ανάλυση χρονολογικών σειρών, περιέχουν γειτονικές παρατηρήσεις που αλληλεξαρτώνται, με τη μόνη διαφορά ότι στην πρώτη περίπτωση τα δεδομένα είναι διακριτά ενώ στη δεύτερη είναι συνεχή. Επίσης, η διαφορά της ανακάλυψης ακολουθιών με τους κανόνες συσχετίσεων έγκειται στο γεγονός ότι τα μοντέλα ακολουθίας θεωρούν ότι τα προϊόντα αγοράζονται με κάποια σειρά ενώ τα μοντέλα συσχετίσεων θεωρούν ότι κάθε προϊόν έχει την ίδια πιθανότητα να αγοραστεί και δεν εξαρτάται από τις άλλες αγορές. (Καρασιώτου, 2010, pp. 31-32)

2.8. Απαιτήσεις της Εξόρυξης Δεδομένων

Για να υπάρξει ένα ολοκληρωμένο αποτέλεσμα από μια διαδικασία εξόρυξης δεδομένων, πρέπει αρχικά να ελεγχθούν τα χαρακτηριστικά που αναμένουμε να έχει το σύστημα εξόρυξης δεδομένων, καθώς και τις απαιτήσεις για την εφαρμογή των τεχνικών.

Με βάση τους Chen et al. (1996), τα κυριότερα ζητήματα που πρέπει να λαμβάνονται υπόψη κάθε φορά είναι :

1. **Χειρισμός διαφορετικών τύπων δεδομένων** : Είναι ξεκάθαρο ότι ένα σύστημα εξόρυξης δεδομένων θα πρέπει να μπορεί να εφαρμόζεται σε διαφορετικούς τύπους δεδομένων, καθώς συχνά χρησιμοποιούνται διαφορετικοί τύποι και βάσεις δεδομένων σε διαφορετικές εφαρμογές. Επίσης, παρατηρείται συχνά η ύπαρξη συγγενών (relational) βάσεων δεδομένων. Επομένως, πρέπει ένα σύστημα εξόρυξης δεδομένων να είναι σε θέση να υποστηρίζει τεχνικές για αποδοτική και αποτελεσματική ανάλυση συγγενικών δεδομένων. Τέλος, θα έπρεπε να λειτουργεί ανεξάρτητα από τύπους δεδομένων, καθώς πολλά σύγχρονα συστήματα βάσεων δεδομένων περιέχουν σύνθετους τύπους δεδομένων (δομές δεδομένων και σύνθετα αντικείμενα, υπερκείμενο και στοιχεία πολυμέσων, χωροχρονικά στοιχεία κ.λπ.). Η ποικιλία των τύπων δεδομένων και οι διαφορετικοί στόχοι της εξόρυξης δεδομένων κάνουν πιο απίθανη την ύπαρξη ενός τέτοιου συστήματος που να μπορεί να χειριστεί όλα αυτά τα είδη δεδομένων. Καλό θα ήταν να

διαμορφωθούν εξειδικευμένα συστήματα για εξόρυξη γνώσης πάνω σε συγκεκριμένους τύπους δεδομένων όπως βάσεις δεδομένων πολυμέσων, συστήματα που ασχολούνται αποκλειστικά με την εξόρυξη γνώσης από σχεσιακές βάσεις δεδομένων, χωροχρονικές βάσεις δεδομένων, κ.λπ.

2. **Απόδοση και εξελιξιμότητα των αλγορίθμων της Εξόρυξης Δεδομένων** : Για να έχουμε αποτελεσματική εξόρυξη γνώσης από μεγάλα σύνολα δεδομένων, πρέπει να έχουμε αλγορίθμους κατάλληλα προσαρμοσμένους σε αυτά. Επομένως, ο χρόνος εκτέλεσης των αλγορίθμων πρέπει να είναι αποδεκτός και αναμενόμενος για μεγάλες βάσεις δεδομένων. Να σημειωθεί εδώ ότι οι αλγόριθμοι με εκθετική ή πολυωνυμική πολυπλοκότητα δεν θεωρούνται πρακτικοί στη χρήση.
3. **Χρησιμότητα, βεβαιότητα, εκφραστικότητα των αποτελεσμάτων της Εξόρυξης Δεδομένων** : Η εξορυγμένη γνώση πρέπει να παρουσιάζει με ακριβή τρόπο τα περιεχόμενα των βάσεων δεδομένων και να είναι χρήσιμη για συγκεκριμένες εφαρμογές. Η ακρίβεια των αποτελεσμάτων θα μπορούσε να εκφραστεί μέσω κάποιων μέτρων βεβαιότητας, προσεγγιστικά ή ποσοτικά. Εξαιρέσεις όπως ο θόρυβος και οι αποκλίσεις (outliers) πρέπει να αντιμετωπιστούν από τα συστήματα εξόρυξης δεδομένων. Το γεγονός αυτό δίνει το κίνητρο για μια συστηματική μελέτη της ποιότητας της εξορυγμένης γνώσης, κατασκευάζοντας στατιστικά ή αναλυτικά μοντέλα, μοντέλα προσομοίωσης, καθώς και τα εργαλεία αυτών.
4. **Εκφράσεις διαφορετικού τύπου για τα αποτελέσματα** : Λογικά, από μεγάλα σύνολα δεδομένων μπορούν να προκύψουν διαφορετικοί τύποι γνώσεων. Έτσι, θα ήταν πολύ χρήσιμο να μπορεί να ελεγχθεί η γνώμη από διαφορετικές απόψεις και να την εκφραστεί σε διάφορες μορφές. Θεωρείται ότι θα ήταν πολύ καλό να μπορούν να εκφραστούν τα ερωτήματα της εξόρυξης δεδομένων και η εξορυγμένη γνώση σε γλώσσες υψηλού επιπέδου ή μέσω γραφικών διεπαφών των χρηστών. Έτσι, η εξόρυξη δεδομένων θα μπορούσε να είναι εφαρμόσιμη και από μη ειδικούς και η εξορυγμένη γνώση θα χρησιμοποιούταν άμεσα από όλους. Τέλος, απαιτείται το σύστημα να υιοθετήσει εκφραστικές τεχνικές αναπαράστασης της γνώσης, έτσι ώστε να επιτευχθεί η αποτελεσματική παρουσίαση της γνώσης.
5. **Διαλογική ανακάλυψη γνώσης στα πολλαπλά εννοιολογικά επίπεδα** : Είναι δύσκολο να προβλεφθεί αυτό που θα μπορούσε να ανακαλυφθεί ακριβώς από

μια βάση δεδομένων. Για αυτό, θα μπορούσε να καθοριστεί μια σειρά ερωτήσεων της εξόρυξης δεδομένων, προκειμένου να διαμορφωθεί η εστίαση στα δεδομένα, να δημιουργηθεί ένα λεπτομερέστερο επίπεδο εξόρυξης δεδομένων και να παρατηρηθούν τα αποτελέσματα της σε πολλαπλά επίπεδα και από διαφορετικές πτυχές. Όλα αυτά μπορούν να επιτευχθούν μέσω της διαλογικής ανακάλυψης της γνώσης.

6. **Εξόρυξη πληροφορίας από διαφορετικές πηγές δεδομένων** : Σε σχέση με τη σύνδεση των διάφορων πηγών δεδομένων, υπάρχει προβάδισμα της ευρέως διαθέσιμης σύνδεσης υπολογιστών σε τοπικό και ευρύτερο δίκτυο, συμπεριλαμβανομένου του διαδικτύου. Αυτό οδηγεί στη δημιουργία μεγάλων κατανεμημένων και ετερογενών βάσεων δεδομένων. Επιπλέον, το τεράστιο μέγεθος των βάσεων δεδομένων, η υψηλή κατανομή των δεδομένων και η υπολογιστική πολυπλοκότητα ορισμένων μεθόδων εξόρυξης δεδομένων, οδηγούν στην ανάπτυξη παράλληλων και κατανεμημένων αλγορίθμων.
7. **Προστασία ιδιωτικότητας και ασφάλεια δεδομένων** : Η προστασία και αποκλειστικότητα των δεδομένων απειλείται στην περίπτωση που αυτά μπορούν να παρατηρηθούν από πολλές διαφορετικές σκοπιές. Είναι σημαντικό να μελετηθεί το πότε μπορεί να οδηγηθούμε σε μια εισβολή στην ιδιωτικότητα μέσω της KDD και τι μέτρα ασφαλείας μπορούν να αναπτυχθούν για να εμποδιστεί η αποκάλυψη των ευαίσθητων πληροφοριών.

Μερικές από τις απαιτήσεις που αναφέρθηκαν παραπάνω μπορεί να φέρουν αντικρουόμενους στόχους. Για παράδειγμα, ο στόχος της προστασίας της ασφάλειας δεδομένων μπορεί να αντικρούει στην απαίτηση για διαλογική εξόρυξη πολυεπίπεδης γνώσης από διαφορετικές σκοπιές. Η παρουσίαση των απαιτήσεων αυτών γίνεται στα πλαίσια του ενδιαφέροντός μας για την ανάπτυξη αποτελεσματικών και εξελίξιμων αλγορίθμων. Για το λόγο αυτό, έγιναν συγκεκριμένες ομαδοποιήσεις των απαιτήσεων ώστε να γίνει μια γενική απεικόνιση. (Σταυλιώτης, 2008, pp. 6-8)

2.9. Αποθήκες Δεδομένων - Data Warehouses

Οι Αποθήκες Δεδομένων γενικεύουν και ενοποιούν τα δεδομένα σε πολυδιάστατο χώρο. Ουσιαστικά, αποτελούν ένα σύνολο τεχνολογιών που παρέχει τη δυνατότητα στους αναλυτές ενός οργανισμού - επιχείρησης να σχεδιάσουν την πολιτική του έχοντας αποδοτική

πρόσβαση στα δεδομένα του οργανισμού - επιχείρησης. Η υλοποίηση μια αποθήκης δεδομένων περιλαμβάνει το σχεδιασμό μιας κεντρικής βάσης δεδομένων με σκοπό τη συγκέντρωση ετερογενών πηγών πληροφοριών σε μια τοποθεσία και παράλληλα την αποφυγή σύγκρουσης μεταξύ συστημάτων επεξεργασίας συναλλαγών (OLTP) και συστημάτων αναλυτικής επεξεργασίας δεδομένων (OLAP). Η σχεδίαση της αποθήκης δεδομένων έχει σαν στόχο την αποδοτική απάντηση πολύπλοκων ερωτήσεων που δημιουργούνται κατά την αναλυτική επεξεργασία των δεδομένων και συντελεί στην αύξηση της αποδοτικότητας των εφαρμογών για τη λήψη αποφάσεων και τη χάραξη στρατηγικού σχεδιασμού. Η δημιουργία και η συντήρηση μιας αποθήκης δεδομένων είναι μια πολύπλοκη διαδικασία και εξαρτάται από τους στόχους που θέτει κάθε οργανισμός κατά την αναλυτική επεξεργασία των δεδομένων του. Πολλοί οργανισμοί επιδιώκουν τη δημιουργία αποθήκης δεδομένων στην οποία να συγκεντρώνεται η αναλυτική πληροφορία από τις δραστηριότητες του οργανισμού, γεγονός που αυξάνει σημαντικά το κόστος υλοποίησης της αποθήκης. Ενίοτε, η αποθήκη δεδομένων ενός οργανισμού συμπληρώνεται από εξειδικευμένα θεματικά υποσύνολα - επιμέρους συλλογές δεδομένων (data marts) για περαιτέρω απόδοση των OLAP εφαρμογών, καθώς πρόκειται για πιο ευέλικτα συστήματα στη δημιουργία τους, που όμως δεν παρέχουν ενιαία λύση, ενώ η μακροχρόνια χρήση τους δημιουργεί προβλήματα. (Καρασιώτου, 2010, p. 12)

DATA WAREHOUSE	*DATA MART*
Total company sales information Πληροφορίες για τις συνολικές πωλήσεις	**Sales for a simple location** Πωλήσεις για μία τοποθεσία ,εστία της αγοράς
Store sales / Hourly sales Συνολικό ποσοστό πωλήσεων	**Credit card sales only** Πωλήσεις / αγορές που έγιναν με πιστωτική κάρτα
Customer profiles Εικόνα / Προφίλ πελατών	**Credit card customer profiles** Εικόνα πελατών σχετικά με την χρησιμοποίηση της πιστωτικής κάρτας
	Credit card purchasing history Πορεία πιστωτικής κάρτας

Εικόνα 2.4 : Data Warehouse vs Data mart. (Παγουρόπουλος, 2006, p. 11)

Μια αποθήκη δεδομένων σημαίνει διαφορετικά πράγματα σε διάφορους ανθρώπους. Κάποιοι ορισμοί περιορίζονται στα δεδομένα, ενώ άλλοι αναφέρονται σε ανθρώπους, διαδικασίες, λογισμικό, εργαλεία και δεδομένα. Ένας από τους παγκόσμιους ορισμούς παρόλα αυτά είναι ο ακόλουθος:

"Η αποθήκη δεδομένων είναι μια συλλογή ολοκληρωμένων, αντικειμενοστραφών (subject - oriented) βάσεων δεδομένων σχεδιασμένων για να υποστηρίξουν τις λειτουργίες απόφασης - υποστήριξης (Decision-Support Functions), όπου κάθε μονάδα δεδομένων σχετίζεται με κάποια χρονική στιγμή."

Με βάση αυτόν τον ορισμό, μια αποθήκη δεδομένων μπορεί να θεωρηθεί ως μια αποθήκη δεδομένων ενός οργανισμού, που δημιουργήθηκε για να υποστηρίξει τη λήψη στρατηγικών αποφάσεων. Η λειτουργία της αποθήκης δεδομένων είναι να αποθηκεύσει τα ιστορικά δεδομένα ενός οργανισμού με έναν ολοκληρωμένο τρόπο που αντανακλά τις διάφορες πτυχές της οργάνωσης και της επιχείρησης. Τα δεδομένα σε μια αποθήκη δεν ενημερώνονται ποτέ αλλά χρησιμοποιούνται μόνο για να απαντήσουν σε ερωτήματα από end users που είναι γενικά οι υπεύθυνοι λήψης αποφάσεων. Τυπικά, οι αποθήκες δεδομένων είναι τεράστιες, αποθηκεύοντας δισεκατομμύρια αρχεία.

Σε αυτό το πρώιμο χρονικό διάστημα στην εξέλιξη των αποθηκών δεδομένων, δεν αποτελεί έκπληξη το γεγονός ότι πολλά έργα "αγωνίζονται" χωρίς επιτυχία λόγω της βασικής παρεξήγησης για το τι είναι μια αποθήκη δεδομένων. Αυτό που προκαλεί έκπληξη είναι το μέγεθος και η κλίμακα αυτών των έργων. Πολλές εταιρείες σφάλλουν επειδή δεν ορίζουν ακριβώς τι είναι μια αποθήκη δεδομένων, τα επιχειρησιακά προβλήματα που θα λύσει και τις χρήσεις υπό τις οποίες θα τεθεί. Δύο πτυχές της αποθήκης δεδομένων είναι πιο σημαντικές για την καλύτερη κατανόηση της διαδικασίας σχεδιασμού της. Η πρώτη είναι οι συγκεκριμένοι τύποι ταξινόμησης των δεδομένων που αποθηκεύονται σε μια αποθήκη δεδομένων και η δεύτερη είναι το σύνολο μετασχηματισμών που χρησιμοποιείται για την προετοιμασία των δεδομένων στην τελική μορφή έτσι ώστε να είναι χρήσιμα για τη λήψη αποφάσεων. Μια αποθήκη δεδομένων περιλαμβάνει τις ακόλουθες κατηγορίες δεδομένων, όπου η ταξινόμηση προσαρμόζεται στις χρονικά εξαρτώμενες πηγές δεδομένων :

1. *Παλιά λεπτομερή δεδομένα / Old detail data.*
2. *Τωρινά (καινούργια) λεπτομερή δεδομένα / Current (new) detail data.*
3. *Ελαφρά συνοπτικά δεδομένα / Lightly summarized data.*
4. *Εξαιρετικά συνοπτικά δεδομένα / Highly summarized data.*

5. *Μεταδεδομένα / Meta-data (the data directory or guide).*

Για την προετοιμασία αυτών των πέντε στοιχειωδών ή συμπληρωματικών τύπων δεδομένων σε μια αποθήκη δεδομένων, οι βασικοί τύποι μετασχηματισμού δεδομένων είναι συγκεκριμένοι. Υπάρχουν τέσσερις κύριοι τύποι μετασχηματισμών και ο καθένας έχει τα δικά του χαρακτηριστικά :

1. *Απλοί μετασχηματισμοί / Simple Transformations* : Αυτοί οι μετασχηματισμοί είναι τα δομικά στοιχεία όλων των άλλων πιο περίπλοκων μετασχηματισμών. Αυτή η κατηγορία περιλαμβάνει τη χειραγώγηση δεδομένων που επικεντρώνονται σε ένα πεδίο κάθε φορά, χωρίς να λαμβάνονται υπόψη οι τιμές τους σε σχετικά πεδία. Παραδείγματα περιλαμβάνουν την αλλαγή του τύπου δεδομένων ενός πεδίου ή την αντικατάσταση μιας κωδικοποιημένης τιμής πεδίου με μια αποκωδικοποιημένη τιμή.
2. *Καθαρισμός / Cleansing and Scrubbing* : Αυτοί οι μετασχηματισμοί εξασφαλίζουν σταθερή μορφοποίηση και χρήση ενός πεδίου ή σχετικών ομάδων πεδίων. Αυτό μπορεί να περιλαμβάνει μια σωστή μορφοποίηση των πληροφοριών διεύθυνσης, για παράδειγμα. Αυτή η κατηγορία μετασχηματισμών περιλαμβάνει επίσης ελέγχους για έγκυρες τιμές σε ένα συγκεκριμένο πεδίο, συνήθως ελέγχοντας το εύρος ή επιλέγοντας από μια απαριθμημένη λίστα.
3. *Ενσωμάτωση / Integration* : Πρόκειται για μια διαδικασία λήψης επιχειρησιακών δεδομένων από μία ή περισσότερες πηγές και χαρτογραφώντας τα, από πεδίο σε πεδίο, σε μια νέα δομή δεδομένων στην αποθήκη δεδομένων. Το πρόβλημα του κοινού identifier είναι ένα από τα πιο δύσκολα ζητήματα ενσωμάτωσης στην κατασκευή μιας αποθήκης δεδομένων. Ουσιαστικά, η κατάσταση αυτή συμβαίνει όταν υπάρχουν πολλές πηγές συστήματος για τις ίδιες οντότητες, και δεν υπάρχει σαφής τρόπος να προσδιοριστούν οι ίδιες οι οντότητες. Πρόκειται για ένα δύσκολο πρόβλημα και σε πολλές περιπτώσεις δεν μπορεί να λυθεί αυτοματοποιημένα. Συχνά απαιτεί εξελιγμένους αλγόριθμους για τη σύζευξη πιθανών ζευγαριών. Ένα άλλο πολύπλοκο σενάριο ενσωμάτωσης δεδομένων προκύπτει όταν υπάρχουν πολλές πηγές για το ίδιο στοιχείο δεδομένων. Στην πραγματικότητα, είναι συνηθισμένο ότι ορισμένες από αυτές τις τιμές είναι αντιφατικές και η επίλυση μιας αντίθεσης δεν είναι μια απλή διαδικασία. Επίσης δύσκολο είναι να μην

υπάρχει τιμή για ένα στοιχείο δεδομένων σε μια αποθήκη. Όλα αυτά τα προβλήματα και οι αντίστοιχες αυτόματες ή ημιαυτόματες λύσεις εξαρτώνται πάντα από το συγκεκριμένο τομέα.

4. *Συσσωμάτωση και Σύνοψη / Aggregation and Summarization* : Οι όροι συσσωμάτωση και σύνοψη έχουν ελαφρώς διαφορετικές έννοιες στο πλαίσιο της αποθήκευσης δεδομένων. Η σύνοψη είναι μια απλή προσθήκη τιμών κατά μήκος μίας ή περισσοτέρων διαστάσεων δεδομένων όπως, για παράδειγμα, προσθέτοντας ημερήσιες πωλήσεις για την παραγωγή μηνιαίων πωλήσεων. Η συσσωμάτωση αναφέρεται στην προσθήκη διαφορετικών επιχειρηματικών στοιχείων σε ένα κοινό σύνολο. Εξαρτάται σε μεγάλο βαθμό από τον τομέα (domain). Για παράδειγμα, η συσσωμάτωση προσθέτει καθημερινές πωλήσεις προϊόντων και μηνιαίες συμβουλευτικές πωλήσεις για να ληφθεί το συνδυαστικό μηνιαίο σύνολο.

Αυτοί οι μετασχηματισμοί είναι ο κύριος λόγος για τον οποίο προτιμούμε μια αποθήκη ως πηγή δεδομένων για μια διαδικασία εξόρυξης δεδομένων. Εάν η αποθήκη δεδομένων είναι διαθέσιμη, η φάση προ-επεξεργασίας στην εξόρυξη δεδομένων είναι σημαντικά μειωμένη, μερικές φορές ακόμη και εξαλειμμένη. Αυτή η προετοιμασία δεδομένων είναι η πιο χρονοβόρα φάση. Μια τριών σταδίων διαδικασία ανάπτυξης αποθήκης δεδομένων συνοψίζεται στα ακόλουθα βασικά βήματα:

1. *Μοντελοποίηση /Modeling* : Με απλά λόγια, η αφιέρωση χρόνου για την κατανόηση επιχειρηματικών διαδικασιών, οι απαιτήσεις πληροφοριών αυτών των διαδικασιών και οι αποφάσεις που γίνονται επί του παρόντος μέσα στις διαδικασίες.
2. *Κατασκευή / Building* : Δημιουργία απαιτήσεων για εργαλεία που να ταιριάζουν με τους τύπους υποστήριξης λήψης αποφάσεων που απαιτούνται για την στοχευμένη επιχειρηματική διαδικασία. Δημιουργία ενός μοντέλου δεδομένων που θα συμβάλει περαιτέρω στον προσδιορισμό απαιτήσεων των πληροφοριών. Αποσύνθεση προβλημάτων σε προδιαγραφές δεδομένων και στο πραγματικό απόθεμα δεδομένων (data store), το οποίο, στην τελική του μορφή, θα αντιπροσωπεύει είτε ένα data mart είτε μια πληρέστερη αποθήκη δεδομένων.
3. *Ανάπτυξη / Deploying* : Να υλοποιήσει. σχετικά σύντομα τη συνολική διαδικασία, τη φύση των δεδομένων που πρόκειται να αποθηκευτούν και τα

διάφορα εργαλεία επιχειρηματικής ευφυΐας που πρέπει να χρησιμοποιηθούν και να ξεκινήσει με την εκπαίδευση των χρηστών. Το στάδιο ανάπτυξης περιέχει ρητά ένα χρονικό διάστημα κατά το οποίο οι χρήστες διερευνούν τόσο το αποθετήριο όσο και τις πρώτες εκδόσεις της πραγματικής αποθήκης δεδομένων. Αυτό μπορεί να οδηγήσει σε μια εξέλιξη της αποθήκης δεδομένων, η οποία περιλαμβάνει την προσθήκη περισσότερων δεδομένων, την επέκταση ιστορικών περιόδων ή την επιστροφή στο στάδιο κατασκευής για να επεκτείνει το πεδίο της αποθήκης δεδομένων μέσω ενός μοντέλου δεδομένων.

Η εξόρυξη δεδομένων αντιπροσωπεύει μία από τις σημαντικότερες εφαρμογές αποθήκευσης δεδομένων, δεδομένου ότι η μοναδική λειτουργία μιας αποθήκης δεδομένων είναι η παροχή πληροφοριών στους end users για υποστήριξη αποφάσεων. Σε αντίθεση με άλλα εργαλεία ερωτήσεων και συστήματα εφαρμογών, η διαδικασία εξόρυξης δεδομένων παρέχει στον end user τη δυνατότητα να εξαγάγει κρυφές, σημαντικές πληροφορίες. Οι πληροφορίες αυτές, παρόλο που είναι πιο δύσκολο να εξαχθούν, μπορούν να προσφέρουν μεγαλύτερα επιχειρηματικά και επιστημονικά πλεονεκτήματα και να αποφέρουν υψηλότερες αποδόσεις στις επενδύσεις "αποθήκευσης δεδομένων και εξόρυξης δεδομένων". (Kantardzic, 2011, pp. 14-16)

2.9.1. OLTP και OLAP

Επειδή οι περισσότεροι άνθρωποι είναι εξοικειωμένοι με τα εμπορικά συστήματα σχεσιακών βάσεων δεδομένων, είναι αρκετά εύκολο να κατανοήσουμε τι είναι μια αποθήκη δεδομένων συγκρίνοντας συστήματα OLTP/OLAP.

Ένα Σύστημα Επεξεργασίας Συναλλαγών (OLTP) αποτελεί ένα πλήρες σύστημα που περιέχει εργαλεία για τον προγραμματισμό των εφαρμογών, την εκτέλεση και την διαχείριση των συναλλαγών. Είναι μια εφαρμογή που δουλεύει συνεχώς, εξελίσσεται συνεχώς, είναι συνήθως κατανεμημένη και περιλαμβάνει μια βάση δεδομένων, κάποιο δίκτυο και τα αντίστοιχα προγράμματα για την εφαρμογή. Από την άλλη πλευρά ένα Σύστημα Αναλυτικής Επεξεργασίας Συναλλαγών (OLAP) παρέχει ευέλικτη, υψηλής απόδοσης πρόσβαση και ανάλυση μεγάλου όγκου σύνθετων δεδομένων από διαφορετικές εφαρμογές, συμμετοχή αθροιστικών και ιστορικών δεδομένων σε πολύπλοκες ερωτήσεις, μεταβολή της “οπτικής γωνίας” παρουσίασης των δεδομένων (π.χ. από πωλήσεις ανά περιοχή σε πωλήσεις ανά

τμήμα κλπ), συμμετοχή πολύπλοκων υπολογισμών (π.χ. στατιστικές αναρτήσεις) και γρήγορες απαντήσεις σε οποιαδήποτε χρονική στιγμή τεθεί ένα ερώτημα (On-Line).

Τα βασικά χαρακτηριστικά που διαφοροποιούν τα OLTP συστήματα από τα OLAP είναι τα εξής :

- *Χρήστες και προσανατολισμός του συστήματος* : Ένα OLTP σύστημα προσανατολίζεται στις απαιτήσεις του πελάτη και χρησιμοποιείται από διοικητικούς υπαλλήλους και διαχειριστές της βάσης δεδομένων του οργανισμού. Ένα OLAP σύστημα προσανατολίζεται στις απαιτήσεις της αγοράς και χρησιμοποιείται από διευθυντικά στελέχη και αναλυτές.
- *Περιεχόμενα Δεδομένων* : Ένα OLTP σύστημα διαχειρίζεται τρέχοντα - καθημερινά δεδομένα μεγάλης λεπτομέρειας τα οποία εύκολα μπορούν να αναζητηθούν και απαντούν σε απλές ερωτήσεις. Ένα OLAP σύστημα διαχειρίζεται μεγάλες ποσότητες ιστορικής πληροφορίας και παρέχει αποδοτική πρόσβαση στα δεδομένα για λήψη αποφάσεων.
- *Σχεδιασμός βάσης δεδομένων* : Ένα OLTP σύστημα σχεδιάζεται για να διατηρεί την ακεραιότητα των δεδομένων και εξασφαλίζει ταχύτητα στην αποθήκευση των καθημερινών συναλλαγών του οργανισμού, επομένως η βάση δεδομένων του συστήματος είναι κανονικοποιημένη βάσει κάποιου μοντέλου Οντοτήτων - Συσχετίσεων (Entity – Relationship model). Ένα OLAP σύστημα σχεδιάζεται για να παρέχει ταχύτητα στην ανάλυση και η βάση δεδομένων του συστήματος είναι από-κανονικοποιημένη βάσει κάποιου μοντέλου αστέρα ή χιονονιφάδας (star/snowflake schema) καθώς οι εφαρμογές OLAP επιταχύνονται αν τα δεδομένα οργανωθούν με μη παραδοσιακούς τρόπους.
- *Πρότυπα πρόσβασης (access patterns)* : Τα δεδομένα ενός OLTP συστήματος υπόκεινται σε λειτουργίες τροποποίησης (π.χ. επεξεργασία συναλλαγών, ανάνηψη, έλεγχος συνδρομικότητας). Από την άλλη πλευρά, τα OLAP συστήματα περιέχουν ιστορική πληροφορία που δεν μεταβάλλεται και επομένως η πρόσβαση σε αυτά επιτρέπει λειτουργίες μόνο για ανάγνωση (read-only). (Καρασιώτου, 2010, pp. 13-15)

	OLTP	OLAP
Δομή	Files/DBMS's	RDBMS
Πρόσβαση	SQL/COBOL/…	SQL & επεκτάσεις
Ανάγκες που καλύπτουν	Αυτοματισμός καθημερινών εργασιών	Άντληση και επεξεργασία πληροφορίας για χάραξη στρατηγικής
Τύπος Δεδομένων	Λεπτομερή, Λειτουργικά	Συνοπτικά, αθροιστικά
Όγκος Δεδομένων	από 100MB έως GB	από 100GB έως TB
Φύση Δεδομένων	Δυναμικά, τρέχοντα	Στατικά, ιστορικά
I/O Τύποι	Περιορισμένο I/O συχνές αναζητήσεις στο δίσκο	Εκτεταμένο I/O συχνές σαρώσεις του δίσκου
Τροποιήσεις	Συνεχείς	Περιοδικές ενημερώσεις
Μέτρηση Απόδοσης	Μέσος Ρυθμός Αποθήκευσης Εγγραφών - Troughput	Χρόνος Απόκρισης
Φόρτος	Συναλλαγές με πρόσβαση λίγων εγγραφών	Ερωτήσεις που σαρώνουν εκατομμύρια εγγραφών
Σχεδίαση Βάσης Δεδομένων	Κατευθυνόμενη από εφαρμογή	Κατευθυνόμενη από περιεχόμενο
Τυπικοί Χρήστες	Χαμηλόβαθμοι Υπάλληλοι, π.χ. διοικητικοί υπάλληλοι, διαχειριστές βάσης δεδομένων	Υψηλόβαθμοι Υπάλληλοι, π.χ. διευθυντικά στελέχη, αναλυτές
Χρήση	Μέσω προκατασκευασμένων φορμών	Ad-hoc
Αριθμός Χρηστών	Χιλιάδες	Δεκάδες
Εστίαση	Εισαγωγή Δεδομένων	Εξαγωγή Πληροφοριών

Εικόνα 2.5 : OLTP vs OLAP.

Κεφάλαιο 3ο

Τεχνικές Data mining

3.1. Ταξινόμηση – Classification

Μια υπάλληλος τραπεζικών δανείων χρειάζεται να κάνει ανάλυση των δεδομένων της προκειμένου να μάθει ποιοι αιτούντες είναι "ασφαλείς" και ποιοι είναι "επικίνδυνοι" για την τράπεζα. Σε μια εταιρεία, ο διευθυντής στο τμήμα μάρκετινγκ χρειάζεται να κάνει μια ανάλυση δεδομένων για να μάθει αν ένας πελάτης, με ένα συγκεκριμένο προφίλ, θα μπει στην διαδικασία να αγοράσει ένα νέο προϊόν. Ένας ιατρικός ερευνητής θέλει να κάνει ανάλυση των στοιχείων που διαθέτει για τον καρκίνο του μαστού, ώστε να μπορέσει να προβλέψει ποια από τις τρεις συγκεκριμένες θεραπείες θα πρέπει να λάβει ένας ασθενής. Σε κάθε μια από αυτές τις περιπτώσεις, η τεχνική ανάλυσης δεδομένων που θα χρησιμοποιηθεί είναι η ταξινόμηση, όπου ένα μοντέλο ή ταξινομητής (classifier) κατασκευάζεται για να προβλέψει κατηγορηματικές ετικέτες, όπως για παράδειγμα "ασφαλής" ή "επικίνδυνη" για τα δεδομένα της αίτησης δανείου, ναι ή όχι για τα δεδομένα μάρκετινγκ ή θεραπεία Α, θεραπεία Β ή θεραπεία Γ για τα ιατρικά δεδομένα. (Han and Kamber, 2006, pp. 285-286)

Ο άνθρωπος συνεχώς κατατάσσει, κατηγοριοποιεί και ταξινομεί, ώστε να μπορέσει να κατανοήσει και να επικοινωνήσει για τον κόσμο. Η **ταξινόμηση** ή **κατηγοριοποίηση (classification)** αποτελεί μία από τις πιο κοινές τεχνικές εξόρυξης δεδομένων διότι φαίνεται να είναι απολύτως αναγκαία. Αποτελείται από την εξέταση των χαρακτηριστικών ενός πρόσφατα παρουσιαζόμενου αντικειμένου και την εκχώρηση του σε ένα από τα προ-καθορισμένα σύνολα κλάσεων. Τα αντικείμενα που πρέπει να ταξινομηθούν αντιπροσωπεύονται γενικά από εγγραφές σε έναν πίνακα βάσεων δεδομένων ή αρχείο και η δουλειά της ταξινόμησης είναι να προσθέσει μια νέα στήλη με κάποιο κωδικό κλάσης. Ουσιαστικά, η λειτουργία αυτής της τεχνικής είναι να κατασκευάσει ένα μοντέλο που να μπορεί να εφαρμοστεί στα μη ταξινομημένα δεδομένα ώστε να τα ταξινομήσει.

Κάποια παραδείγματα προβλημάτων που έχουν αντιμετωπιστεί χρησιμοποιώντας την ταξινόμηση ως τεχνική είναι:

- Η επιλογή περιεχομένου που θα εμφανίζεται σε μια ιστοσελίδα.
- Ο προσδιορισμός των αριθμών τηλεφώνου που αντιστοιχούν σε μηχανές φαξ.
- Η ταξινόμηση αιτούντων πίστωσης ως χαμηλού, μεσαίου ή υψηλού κινδύνου.

Σε όλα αυτά τα παραδείγματα, υπάρχει ένας περιορισμένος αριθμός κλάσεων. Τα δέντρα αποφάσεων είναι τεχνική κατάλληλη για ταξινόμηση, καθώς και τα Νευρωνικά δίκτυα σε ορισμένες περιπτώσεις.

Η ταξινόμηση ασχολείται με διακριτά αποτελέσματα: ναι ή όχι, ιλαρά ή ανεμοβλογιά. Η **εκτίμηση (estimation)** ασχολείται με συνεχώς εκτιμημένα αποτελέσματα. Δίνοντας μερικά δεδομένα εισόδου, η εκτίμηση παρέχει μια τιμή για κάποια άγνωστη συνεχή μεταβλητή όπως το εισόδημα, το ύψος ή το υπόλοιπο της πιστωτικής κάρτας. Τα μοντέλα παλινδρόμησης και τα νευρωνικά δίκτυα είναι κατάλληλα για τα καθήκοντα εκτίμησης. (Berry and Linoff, 2004, pp. 8-9)

3.2. Αλγόριθμοι Ταξινόμησης

Οι αλγόριθμοι ταξινόμησης εφαρμόζονται σε διακριτά δεδομένα τα οποία έχουν προ-ταξινομηθεί σε συγκεκριμένες κατηγορίες ή κλάσεις με στόχο την εξαγωγή κανόνων, οι οποίοι πιθανόν να χρησιμοποιηθούν αργότερα για την κατηγοριοποίηση καινούργιων δεδομένων στις ίδιες κλάσεις. Ένα σύνολο εξαγόμενων κανόνων ονομάζεται ταξινομητής (classifier).

Ο όρος ταξινομητής αναφέρεται στη μαθηματική συνάρτηση, που εφαρμόζεται από έναν αλγόριθμο ταξινόμησης, ο οποίος χαρτογραφεί δεδομένα εισόδου σε μια κατηγορία. (El.wikipedia.org, 2018) Μετά από την εκπαίδευση του αλγορίθμου, ο ταξινομητής που έχει ήδη προκύψει μπορεί να χρησιμοποιηθεί και σε άλλες εγγραφές, οι οποίες δεν έχουν κατηγοριοποιηθεί. Επιπλέον, με βάση ένα δεύτερο προ-ταξινομημένο σύνολο εγγραφών, το σύνολο ελέγχου (test set), μπορεί να ελεγχθεί η ακρίβεια ταξινόμησης ενός ταξινομητή. Έπειτα, το εν λόγω σύνολο ελέγχου ταξινομείται εκ νέου, με την χρήση φυσικά του ταξινομητή και τέλος πραγματοποιείται μέτρηση του ποσοστού των λανθασμένων ταξινομήσεων (error rate). (Μουζάκη, 2006, p. 10)

Όσον αναφορά τους αλγορίθμους, υπάρχουν τρείς βασικές λειτουργίες που ακολουθούν:

1. Ένα σύνολο από δεδομένα εισάγεται στον αλγόριθμο.
2. Ο αλγόριθμος “μαθαίνει” και κατανοεί τον τρόπο και τους κανόνες βάσει των οποίων κατηγοριοποιήθηκαν τα εισαγόμενα στοιχεία.

3. Ακολουθώντας τους κανόνες έχει πλέον την ικανότητα να ταξινομήσει καινούργια δεδομένα.

Επιπρόσθετα, ανάλογα με το είδος του ταξινομητή που παράγει, ο κάθε αλγόριθμος διακρίνεται σε δύο βασικούς τύπους:

1. **Αλγόριθμοι που παράγουν λίστες αποφάσεων :** Αποτελούν μια σχετικά καινούργια μορφή αλγορίθμων (π.χ. Clark). Έχουν την μορφή λογικών κανόνων που βγάζουν ανάλογα συμπεράσματα.
2. **Αλγόριθμοι που παράγουν δένδρα αποφάσεων :** Αποτελούν ίσως και την πιο παλιά μορφή της τεχνικής του Data Mining (π.χ. Quinlan). Έχουν στην ρίζα τους και στους ενδιάμεσους κόμβους τους τις τιμές των διάφορων πεδίων και στα φύλλα τους έχουν τις τιμές του πεδίου κλάσης. Ο κάθε κόμβος του δένδρου διακλαδώνεται προς τα κάτω έχοντας για κάθε διακριτή τιμή ένα κλαδί του πεδίου ενώ σε περίπτωση συνεχούς αριθμητικού πεδίου, το εύρος του πεδίου χωρίζεται σε διαστήματα και ο κόμβος διακλαδώνεται με βάση αυτά. Συνήθως, αυτοί οι αλγόριθμοι ακολουθούν αναλυτική προσέγγιση (top-down), όπου ουσιαστικά δημιουργούν τα δένδρα από την ρίζα και προχωρούν προς τα κάτω. (Παγουρόπουλος, 2006, p. 13)

3.2.1. ID3 Αλγόριθμος

Ο ID3 (Iterative Dichotomiser 3), αποτελεί έναν από τους βασικότερους αλγορίθμους ταξινόμησης που παράγει δένδρα αποφάσεων. Ανήκει στην οικογένεια των συστημάτων μάθησης TDIDT (Top-Down Induction of Decision Trees), ακολουθώντας την αναλυτική προσέγγιση. Παρουσιάστηκε ολοκληρωμένα από τον J.R. Quinlan το 1986, έχοντας διεξάγει μια πρωτοποριακή και πλήρη μελέτη.

Ο αλγόριθμος αυτός βασίζεται στην επιστημονική αρχή του Occam’s Razor (Ξυράφι του Όκαμ), καθώς προτιμά τα μικρότερα δένδρα απόφασης - κατά συνέπεια και την απλούστερη θεωρία - από τα μεγαλύτερα. Παρόλα αυτά, αυτό δεν σημαίνει ότι παράγει μόνο τα μικρότερα δένδρα και για αυτό είναι ευρετικός. (El.wikipedia.org, 2018)

Ο ID3 δέχεται σαν είσοδο ένα σύνολο εκπαίδευσης, όπου οι εγγραφές του έχουν προ-ταξινομηθεί σε κατηγορίες. Στην αρχική του μορφή θεωρεί δύο διακριτές τιμές κλάσης - η P (positive) και η N (negative) - αν και μπορεί να επεκταθεί και σε περισσότερες από δυο.

Ωστόσο, αντί να εξάγει ολόκληρο ένα σύνολο εκπαίδευσης από το δένδρο, ο αλγόριθμος χρησιμοποιεί ένα υποσύνολο εγγραφών. Έτσι, με το δένδρο που προκύπτει ταξινομείται ολόκληρο το σύνολο εκπαίδευσης και ελέγχεται και η ακρίβεια της ταξινόμησης. Αν η ταξινόμηση έχει διεξαχθεί σωστά, τότε το δένδρο γίνεται αποδεκτό και ο αλγόριθμος τερματίζει, αλλιώς προστίθενται κι άλλες εγγραφές στο υποσύνολο και η διαδικασία επαναλαμβάνεται μέχρι όλες οι εγγραφές να ταξινομηθούν σωστά.

Σημαντική παράμετρο αποτελεί και το ποσοστό των εγγραφών που θα περιέχει το υποσύνολο και με τι ρυθμό θα μεγαλώνει εφόσον δεν είναι επαρκές. Εξίσου σημαντική παράμετρο αποτελεί και το κριτήριο επιλογής κάθε κόμβου, με το οποίο θα γίνει η διακλάδωση. Ως κριτήριο επιλογής χρησιμοποιείται η Εντροπία. Το μέγεθος Εντροπία εκτιμά το πόσο λανθασμένα χωρίζεται κάθε φορά το σύνολο εκπαίδευσης, με βάση το συγκεκριμένο πεδίο. Ουσιαστικά, το πεδίο που παρουσιάζει την μικρότερη Εντροπία, χωρίζει καλύτερα το σύνολο εκπαίδευσης.

Συνοπτικά, ο αλγόριθμος ID3 μπορεί να αναλυθεί ως εξής:

1. Γίνεται επιλογή πεδίου για ρίζα του δένδρου και σχηματισμός διακλάδωσης με ένα φύλλο για κάθε διαφορετική τιμή ή διάστημα αυτού του πεδίου.
2. Το δένδρο απόφασης, που έχει κατασκευασθεί, χρησιμοποιείται για την ταξινόμηση του συνόλου εκπαίδευσης. Αν όλες οι εγγραφές που ταξινομούνται σε ένα συγκεκριμένο φύλλο ανήκουν στην ίδια κλάση, τότε ονομάζουμε το φύλλο με αυτήν την κλάση. Όταν έχουν ονομαστεί όλα τα φύλλα με κάποια κλάση, τότε ο αλγόριθμος τελειώνει.
3. Αν κάποιο φύλλο δεν έχει ονομαστεί με κάποια κλάση, ξαναγίνεται επιλογή πεδίου το οποίο δεν έχει επιλεχθεί στο μονοπάτι από το φύλλο ως την ρίζα, ονομάζουμε το φύλλο (κόμβος) με αυτό το πεδίο και γίνεται σχηματισμός διακλάδωσης με ένα φύλλο για κάθε διαφορετική τιμή ή διάστημα αυτού του πεδίου. Τέλος, ξαναγυρνάμε στο δεύτερο βήμα.

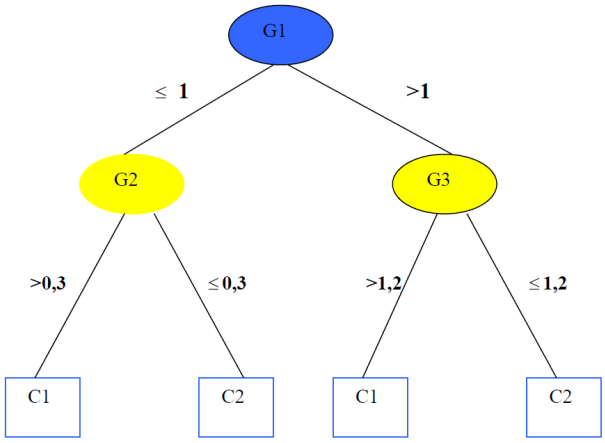
Μια άλλη εξήγηση της υλοποίησης του ID3 παρουσιάζεται στον αλγόριθμο ταξινόμησης C4.5, ο οποίος αποτελεί εξέλιξη του ID3. (Παγουρόπουλος, 2006, pp. 13-14)

3.2.2. C4.5 Αλγόριθμος

Ο αλγόριθμος αυτός αποτελεί μία από τις πλέον γνωστές τεχνικές στο χώρο της μηχανικής μάθησης. Αναπτύχθηκε από τον Quinlan (1993) και αποτελεί εξέλιξη του αλγορίθμου ID3. Τα βασικά πλεονεκτήματα που προκύπτουν είναι :

1. Δυνατότητα επεξεργασίας και διαχείρισης ποσοτικών κριτηρίων.
2. Δυνατότητα διαχείρισης δεδομένων με ελλιπή στοιχεία.
3. Αποφυγή της μεγάλης προσαρμογής στα δεδομένα του δείγματος εκμάθησης (overfitting).

Ο αλγόριθμος ξεκινώντας από ένα σύνολο παραδειγμάτων, τα οποία διαμορφώνουν το δείγμα εκμάθησης, οδηγεί στην ανάπτυξη ενός συνόλου κανόνων αποφάσεων για την ταξινόμηση εναλλακτικών δραστηριοτήτων. Έπειτα οι κανόνες αυτοί οργανώνονται και δημιουργούν ένα δέντρο αποφάσεων. Ο κάθε κόμβος του δέντρου περιλαμβάνει το κριτήριο αξιολόγησης το οποίο ελέγχεται βάσει των συνθηκών που καθορίζουν τα κλαδιά του δέντρου. Τα φύλλα με την σειρά τους υποδεικνύουν την κατηγορία στην οποία πρέπει να ενταχθεί μια εναλλακτική δραστηριότητα η οποία επαληθεύει την συνθήκη του κλάδου που καταλήγει στο φύλλο.



Εικόνα 3.1 : Δέντρο απόφασης ενός C4.5 αλγορίθμου.

Αυτό το δέντρο απόφασης, αναπτύσσεται μέσω μίας επαναληπτικής διαδικασίας όπου κάθε στάδιο αυτής περιλαμβάνει τα εξής τρία βήματα:

1. Αξιολόγηση της διακριτικής ικανότητας των κριτηρίων αξιολόγησης στην ταξινόμηση των εναλλακτικών δραστηριοτήτων.
2. Επιλογή του κριτηρίου αξιολόγησης με την υψηλότερη διακριτική ικανότητα.
3. Διαχωρισμός των εναλλακτικών δραστηριοτήτων σε υποσύνολα αντίστοιχα με το πλήθος των τιμών του επιλεγμένου κριτηρίου αξιολόγησης (στην περίπτωση ποιοτικών κριτηρίων) ή των σημείων διαχωρισμού (cut-points, εάν τα κριτήρια είναι ποσοτικά).

Η παραπάνω διαδικασία επαναλαμβάνεται για κάθε υποσύνολο εναλλακτικών δραστηριοτήτων που σχηματίζεται στο τρίτο βήμα, μέχρι τελικά να επιτευχθεί η σωστή ταξινόμηση όλων των εναλλακτικών δραστηριοτήτων του δείγματος εκμάθησης. (Παγουρόπουλος, 2006, pp. 14-16)

Ορισμένοι χώροι καθοδηγούν αυτόν τον αλγόριθμο με τον εξής τρόπο :

- Αν όλες οι περιπτώσεις είναι της ίδιας κατηγορίας, τότε το δέντρο είναι ένα φύλλο και έτσι το φύλλο επιστρέφεται επισημασμένο με αυτήν την κλάση.
- Για κάθε χαρακτηριστικό, γίνεται υπολογισμός των πιθανών πληροφοριών που παρέχονται από τη δοκιμή του χαρακτηριστικού (με βάση των πιθανοτήτων της κάθε περίπτωσης να έχει μια συγκεκριμένη τιμή για το χαρακτηριστικό). Επίσης, υπολογίζουμε το κέρδος σε πληροφορίες που θα προέκυπταν από τη δοκιμή στο χαρακτηριστικό (με βάση τις πιθανότητες κάθε περίπτωσης με μια συγκεκριμένη τιμή για το χαρακτηριστικό που είναι μιας συγκεκριμένης κλάσης).
- Ανάλογα με το τρέχον κριτήριο επιλογής, βρίσκουμε το καλύτερο χαρακτηριστικό. (Korting, 2014, p. 2)

3.2.3. Support Vector Machine Αλγόριθμος

Η SVM (Support Vector Machine) αποτελεί μια πολλά υποσχόμενη νέα μέθοδο, η οποία αφορά τη ταξινόμηση γραμμικών και μη γραμμικών δεδομένων. Ουσιαστικά, είναι ένας αλγόριθμος που λειτουργεί χρησιμοποιώντας μια μη γραμμική χαρτογράφηση για να μετασχηματίσει τα αρχικά δεδομένα εκπαίδευσης σε μια υψηλότερη διάσταση. Μέσα σε αυτή τη νέα διάσταση, ψάχνει για ένα "όριο απόφασης" που διαχωρίζει τις πλειάδες μιας τάξης από την άλλη (linear optimal separating hyperplane). Με μια κατάλληλη μη γραμμική χαρτογράφηση σε μια αρκετά μεγάλη διάσταση, δεδομένα από δύο κατηγορίες μπορούν

πάντα να διαχωριστούν από ένα hyperplane. Η SVM βρίσκει αυτό το hyperplane χρησιμοποιώντας διανύσματα υποστήριξης, support vectors, ("βασική" εκπαίδευση πλειάδων) και τα περιθώρια, margins, (που ορίζονται από τους φορείς υποστήριξης).

Τον τελευταίο καιρό, οι support vector μηχανές έχουν προσελκύσει πολύ μεγάλη προσοχή. Το πρώτο δημοσίευμα παρουσιάστηκε το 1992 από τον Vladimir Vapnik, τον Bernhard Boser και την Isabelle Guyon, παρόλο που η ιδέα υπάρχει από τη δεκαετία του 1960. Αν και ο χρόνος εκπαίδευσης ακόμη και των ταχύτερων SVM μπορεί να είναι εξαιρετικά αργή, είναι τρομερά ακριβείς, λόγω της ικανότητάς τους να μοντελοποιούν πολύπλοκα μη γραμμικά όρια απόφασης. Επιπλέον, είναι λιγότερο επιρρεπείς στην υπερφόρτωση σε σχέση με άλλες μεθόδους. Τέλος, πέρα από την ταξινόμηση μπορεί να χρησιμοποιηθεί και για πρόβλεψη. (Han and Kamber, 2006, p. 337)

Ο SVM αλγόριθμος μπορεί να χρησιμοποιηθεί για την επίλυση διάφορων πραγματικών προβλημάτων στον κόσμο. Δηλαδή :

- Βοηθά στην κατηγοριοποίηση των κειμένων και των υπερκειμένων.
- Μπορεί να χρησιμοποιηθεί για την ταξινόμηση εικόνων. Τα αποτελέσματα που εμφανίζονται από την διεξαγωγή πειραμάτων δείχνουν ότι οι SVM επιτυγχάνουν σημαντικά υψηλότερη ακρίβεια αναζήτησης. Αυτό ισχύει και για τα συστήματα κατακερματισμού εικόνας, συμπεριλαμβανομένων εκείνων που χρησιμοποιούν μια τροποποιημένη έκδοση SVM που χρησιμοποιεί την προνομιακή προσέγγιση όπως προτείνεται από τον Vapnik.
- Οι χειρόγραφοι χαρακτήρες μπορούν να αναγνωριστούν χρησιμοποιώντας SVM.
- Ο αλγόριθμος SVM έχει εφαρμοστεί ευρέως στις βιολογικές και σε άλλες επιστήμες. Έχει χρησιμοποιηθεί για την ταξινόμηση πρωτεϊνών με έως και 90% των ενώσεων να έχουν ταξινομηθεί σωστά. (En.wikipedia.org, 2018)

3.3. Συμπερασματική ανάλυση των αλγορίθμων της Ταξινόμησης

Όπως αναφέρθηκε προηγουμένως, ο ID3 αποτελεί έναν από τους βασικότερους αλγόριθμους της ταξινόμησης λόγω του ότι είναι η πρώτη – από τις τρεις – ολοκληρωμένη υλοποίηση του Quinlan, όσον αφορά τα Decision Trees. Χρησιμοποιείται για την παραγωγή ενός δένδρου απόφασης, πιο συχνά στον τομέα της μηχανικής μάθησης και της επεξεργασίας της φυσικής γλώσσας (Natural Language Processing). Ωστόσο, ο C4.5 αλγόριθμος είναι

αυτός που χρησιμοποιείται και προτιμάται πιο συχνά, ειδικά στην ταξινόμηση. Ο λόγος που συμβαίνει αυτό είναι διότι προσφέρει κάποια σημαντικά πλεονεκτήματα – όντας εξέλιξη του ID3 – τα οποία τον έφεραν στην πρώτη θέση στους Top 10 Αλγόριθμους στην Εξόρυξη Δεδομένων το 2008. Από την άλλη, ο SVM αποτελεί εξίσου έναν ευρέως χρησιμοποιούμενο αλγόριθμο, ο οποίος όμως ενώ χρησιμοποιείται για ταξινόμηση – αλλά και ανάλυση παλινδρόμησης – το κάνει αναλύοντας δεδομένα, γραμμικά και μη.

Συγκεκριμένα, η χρήση των αλγορίθμων με βάση τα Δένδρα Αποφάσεων προσφέρει μια πιο κατανοητή μορφή ανάγνωσης προς τον χρήστη από την διαθέσιμη γνώση, ενώ η χρήση του SVM προσφέρει υψηλή ακρίβεια. Παρόλα αυτά, η επιλογή ανάμεσα σε αυτούς τους αλγορίθμους κρίνεται από τα δεδομένα που έχουμε στην διάθεση μας.

3.4. Συσταδοποίηση – Clustering

Η **ομαδοποίηση** ή **συσταδοποίηση (clustering)** είναι η τεχνική που διαχωρίζει έναν ετερογενή πληθυσμό σε έναν αριθμό των πιο ομοιογενών υποομάδων ή ομάδων (clusters). Αυτό που διαφοροποιεί την ομαδοποίηση από την ταξινόμηση είναι ότι η ομαδοποίηση δεν βασίζεται σε προκαθορισμένες κατηγορίες. Στην ταξινόμηση, κάθε εγγραφή έχει εκχωρηθεί σε μια προκαθορισμένη κλάση με βάση ένα μοντέλο που αναπτύχθηκε μέσω της εκπαίδευσης σε προ-ταξινομημένα παραδείγματα.

Στην ομαδοποίηση, δεν υπάρχουν προκαθορισμένες κατηγορίες και δεν υπάρχουν παραδείγματα. Τα αρχεία συγκεντρώνονται με βάση την ομοιότητα. Είναι στο χέρι του χρήστη να προσδιορίσει τι νόημα, αν υπάρχει, θα προσκολληθεί στις ομάδες (clusters) που προέκυψαν. Συστάδες συμπτωμάτων μπορεί να υποδεικνύουν διάφορες ασθένειες Συστάδες χαρακτηριστικών πελατών μπορεί να υποδεικνύουν διαφορετικά τμήματα της αγοράς.

Η ομαδοποίηση συχνά γίνεται ως προοίμιο σε κάποια άλλη μορφή εξόρυξης δεδομένων ή μοντέλου. Για παράδειγμα, η ομαδοποίηση μπορεί να είναι το πρώτο βήμα σε μια προσπάθεια διαχωρισμού της αγοράς: Αντί να βρεθεί ένα μέγεθος για όλους στο “σε ποιο είδος προώθησης ανταποκρίνονται οι πελάτες με τον καλύτερο τρόπο ", κατατάσσουμε πρώτα τον πελάτη σε ομάδες ή άτομα με παρόμοιες αγοραστικές συνήθειες και, στη συνέχεια πραγματοποιείται η ερώτηση για κάθε σύμπλεγμα. (Berry and Linoff, 2004, p. 11)

Οι κανόνες Ομαδοποίησης είναι αρκετά διαδεδομένοι. Σήμερα είναι ιδιαίτερα σημαντικό για της επιχειρήσεις να μπορούν να ομαδοποιούν τους πελάτες τους σε

συγκεκριμένες κατηγορίες. Σύμφωνα με αυτές τις κατηγορίες μπορούν να αξιολογούν έναν νέο πελάτη με βάση την ομάδα στην οποία κατατάσσεται ή ακόμα να προσδιορίσουν τα χαρακτηριστικά των πελατών που αποφέρουν μεγάλα κέρδη στην εταιρεία. Από αυτόν τον διαχωρισμό των πελατών μπορούν να προσανατολίσουν την στρατηγική της εταιρείας στην εξειδικευμένη εξυπηρέτηση ορισμένων πελατειακών ομάδων.

Για να μπορέσει να γίνει η επιλογή του κατάλληλου αλγορίθμου απαραίτητη προϋπόθεση είναι η μελέτη των δεδομένων που θα χρησιμοποιηθούν για τον προσδιορισμό κυρίως του κριτηρίου ομοιότητας των εγγραφών μίας ομάδας. Γενικά, η τεχνική της ομαδοποίησης μπορεί να είναι:

- **Στατιστική ή Αριθμητική (statistical/numerical clustering)** : Σε αυτήν την περίπτωση χρησιμοποιούνται διάφορα αριθμητικά κριτήρια ομοιότητας. Έτσι οι ομάδες που προκύπτουν περιγράφονται από αριθμητικές τιμές.
- **Εννοιολογική (conceptual clustering)** : Σε αυτήν την περίπτωση ο προσδιορισμός των ομάδων βασίζεται στο νόημα και στις έννοιες που τα διάφορα αριθμητικά στοιχεία αντιπροσωπεύουνε. Έτσι οι τιμές που έχουμε είναι κατηγορικές και όχι αριθμητικές. Πολλοί από τους αλγόριθμους ομαδοποίησης απαιτούν το σύνολο εκπαίδευσης που επεξεργάζονται να είναι αριθμητικό (πχ k-means) είτε κατηγορικό (πχ k-modes).Υπάρχουν και αλγόριθμοι βέβαια που επιτρέπουν μικτό σύνολο εκπαίδευσης (πχ ο k-prototypes). (Παγουρόπουλος, 2006, pp. 21-22)

3.5. Μέθοδοι Συσταδοποίησης

Το να γίνει μια καθαρή κατηγοριοποίηση των μεθόδων της ομαδοποίησης αποτελεί ένα δύσκολο έργο, εξαιτίας του ότι αυτές οι κατηγορίες μπορεί να επικαλύπτουν η μια την άλλη, με αποτέλεσμα μια μέθοδος να έχει χαρακτηριστικά από πολλές από αυτές. Παρόλα αυτά, είναι χρήσιμο να παρουσιαστεί μια οργανωμένη εικόνα των διαφόρων μεθόδων της εν λόγω τεχνικής. Η ταξινόμηση του μπορεί να γίνει ως εξής:

3.5.1. Μέθοδοι Διαχωρισμού – Partitioning Methods

Δίνοντας μια βάση δεδομένων από n αντικείμενα ή πλειάδες δεδομένων, μια μέθοδος διαχωρισμού κατασκευάζει k χωρίσματα των δεδομένων, όπου κάθε χώρισμα

αντιπροσωπεύει ένα σύμπλεγμα και το k είναι μικρότερο ίσο από το n. Ουσιαστικά, ταξινομεί τα δεδομένα σε ομάδες k, οι οποίες ικανοποιούν τις ακόλουθες απαιτήσεις :

1. Κάθε ομάδα πρέπει να περιέχει τουλάχιστον ένα αντικείμενο, και
2. Κάθε αντικείμενο πρέπει να ανήκει σε μία ακριβώς ομάδα.

Το γενικό κριτήριο ενός καλού διαχωρισμού είναι ότι τα αντικείμενα που ανήκουν στο ίδιο σύμπλεγμα είναι "κοντά" ή σχετίζονται μεταξύ τους, ενώ αντικείμενα διαφορετικών συστάδων είναι πολύ "απομακρυσμένα" ή πολύ διαφορετικά.

Για να επιτευχθεί η παγκόσμια βελτιστοποίηση στην ομαδοποίηση που βασίζεται στην μέθοδο του διαχωρισμού απαιτείται η απαρίθμηση όλων των πιθανών διαχωρισμών, κάτι το οποίο είναι εξαντλητικό. Αντί αυτού, οι περισσότερες εφαρμογές χρησιμοποιούν μεθόδους όπως τον αλγόριθμο k-means, όπου κάθε σύμπλεγμα αντιπροσωπεύεται από την μέση τιμή των αντικειμένων στο σύμπλεγμα, και τον αλγόριθμο k-medoids, όπου κάθε σύμπλεγμα αντιπροσωπεύεται από ένα από τα αντικείμενα που βρίσκεται κοντά στο κέντρο του συμπλέγματος.

3.5.2. Ιεραρχικές Μέθοδοι – Hierarchical Methods

Αυτή η μέθοδος δημιουργεί μια ιεραρχική αποσύνθεση του δεδομένου συνόλου αντικειμένων δεδομένων (data objects). Μια ιεραρχική μέθοδος μπορεί να χαρακτηριστεί ως είτε αθροιστική ή διαιρετική, με βάση τον τρόπο με τον οποίο διαμορφώνεται η ιεραρχική αποσύνθεση. Η αθροιστική (agglomerative or bottom-up) προσέγγιση ξεκινά με κάθε αντικείμενο να σχηματίζει μια ξεχωριστή ομάδα. Συγχωνεύει διαδοχικά τα αντικείμενα ή τις ομάδες που βρίσκονται κοντά μεταξύ τους, μέχρις ότου όλες οι ομάδες συγχωνευθούν σε ένα (το ανώτατο επίπεδο της ιεραρχίας), ή μέχρις ότου τεθεί σε ισχύ ένας όρος τερματισμού. Η διαιρετική προσέγγιση (divisive or top-down), ξεκινά με όλα τα αντικείμενα στο ίδιο σύμπλεγμα. Σε κάθε διαδοχική επανάληψη, ένα σύμπλεγμα χωρίζεται σε μικρότερες ομάδες, μέχρι τελικά κάθε αντικείμενο να είναι σε ένα σύμπλεγμα ή μέχρις ότου διατηρηθεί μια κατάσταση τερματισμού.

Οι ιεραρχικές μέθοδοι υποφέρουν από το γεγονός ότι μόλις πραγματοποιηθεί ένα βήμα (συγχώνευση ή διάσπαση) δεν μπορεί ποτέ να ανατραπεί. Αυτή η ακαμψία είναι χρήσιμη διότι οδηγεί σε μικρότερους υπολογισμούς του κόστους, χωρίς να ανησυχεί κανείς για μια συνδυαστική σειρά διαφορετικών επιλογών. Ωστόσο, τέτοιες τεχνικές δεν μπορούν να

διορθώσουν εσφαλμένες αποφάσεις. Υπάρχουν δύο προσεγγίσεις για τη βελτίωση της ποιότητας της ιεραρχικής ομαδοποίησης :

1. Να γίνει προσεκτική ανάλυση των αντικειμένων "linkages" σε κάθε ιεραρχική ομαδοποίηση, όπως στο Chameleon, ή
2. Να ενσωματωθεί ιεραρχική συσσώρευση και άλλες προσεγγίσεις χρησιμοποιώντας πρώτα έναν ιεραρχικό αθροιστικό αλγόριθμο για την ομαδοποίηση αντικειμένων σε μικρο-ομάδες (microclusters) και στη συνέχεια να εκτελεσθεί μικρο-ομαδοποίηση επί των μικρο-ομάδων χρησιμοποιώντας μια άλλη μέθοδο ομαδοποίησης όπως επαναληπτική μετατόπιση (iterative relocation), όπως στο BIRCH.

3.5.3. Μέθοδοι με βάση την πυκνότητα – Density-based Methods

Οι περισσότερες μέθοδοι διαχωρισμού ομαδοποιούν αντικείμενα με βάση την απόσταση μεταξύ των αντικειμένων. Τέτοιες μέθοδοι μπορούν να βρουν μόνο ομάδες σφαιρικού σχήματος και αντιμετωπίζουν δυσκολίες στην ανακάλυψη ομάδων αυθαίρετων σχημάτων. Άλλες μέθοδοι ομαδοποίησης έχουν αναπτυχθεί με βάση της έννοιας της πυκνότητας. Η γενική ιδέα τους είναι να συνεχίσουν την ανάπτυξη της δεδομένης ομάδας όσο η πυκνότητα (αριθμός αντικειμένων ή δεδομένων σημεία) στην "γειτονιά" υπερβαίνει κάποιο όριο, δηλαδή για κάθε σημείο δεδομένων μέσα σε ένα δεδομένο σύμπλεγμα, η γειτονιά μιας δεδομένης ακτίνας πρέπει να περιέχει τουλάχιστον έναν ελάχιστο αριθμό σημείων. Μια τέτοια μέθοδος μπορεί να χρησιμοποιηθεί για να φιλτράρει τον θόρυβο (απόκλιση) και για την ανίχνευση ομάδων αυθαίρετου σχήματος.

Το DBSCAN και η επέκτασή του, OPTICS, είναι τυπικές μέθοδοι που βασίζονται στην πυκνότητα που αναπτύσσουν συστάδες-ομάδες σύμφωνα με μια ανάλυση συνδεσιμότητας βάσει πυκνότητας.

3.5.4. Μέθοδοι με βάση το Πλέγμα – Grid-based Methods

Οι μέθοδοι με βάση το πλέγμα κβαντοποιούν τον χώρο του αντικειμένου σε έναν πεπερασμένο αριθμό των κυττάρων που σχηματίζουν μια δομή πλέγματος. Όλες οι λειτουργίες ομαδοποίησης εκτελούνται στη δομή πλέγματος (δηλ. στον κβαντισμένο χώρο). Το κύριο πλεονέκτημα αυτής της προσέγγισης είναι ο γρήγορος χρόνος επεξεργασίας, ο

οποίος είναι συνήθως ανεξάρτητος από τον αριθμό των αντικειμένων δεδομένων (data objects) και εξαρτάται μόνο από τον αριθμό των κελιών σε κάθε διάσταση στο κβαντισμένο χώρο.

Το STING είναι ένα τυπικό παράδειγμα μιας μεθόδου με βάση το πλέγμα. Το WaveCluster εφαρμόζει έναν κυματισμό μετασχηματισμού για ανάλυση ομαδοποίησης και βασίζεται τόσο στο πλέγμα όσο και στην πυκνότητα.

3.5.5. Μέθοδοι που βασίζονται σε Μοντέλα – Model-based Methods

Οι μέθοδοι που βασίζονται σε μοντέλα υποθέτουν ένα μοντέλο για κάθε ομάδα και βρίσκουν την καλύτερη προσαρμογή των δεδομένων στο συγκεκριμένο μοντέλο. Ένας αλγόριθμος με βάση το μοντέλο μπορεί να εντοπίσει ομάδες κατασκευάζοντας μια συνάρτηση πυκνότητας που αντικατοπτρίζει τη χωρική κατανομή των σημείων των δεδομένων. Επίσης οδηγεί σε έναν τρόπο αυτόματου προσδιορισμού του αριθμού των συστάδων που βασίζονται σε τυποποιημένα στατιστικά στοιχεία, λαμβάνοντας υπόψη το θόρυβο ή τις αποκλίσεις και συνεπώς παρέχοντας ισχυρές μεθόδους ομαδοποίησης.

Ο EM είναι ένας αλγόριθμος που εκτελεί ανάλυση αναμενόμενης μεγιστοποίησης με βάση την στατιστική μοντελοποίηση. Ο COBWEB είναι ένας εννοιολογικός αλγόριθμος εκμάθησης που εκτελεί ανάλυση πιθανοτήτων και παίρνει τις έννοιες (concepts) ως μοντέλο για τις συστάδες. Ο SOM είναι ένας αλγόριθμος βασισμένος σε νευρωνικό δίκτυο που ομαδοποιεί χαρτογραφώντας υψηλών διαστάσεων δεδομένα σε 2-D ή 3-D, το οποίο είναι επίσης χρήσιμο για την οπτικοποίηση δεδομένων.

Η επιλογή του αλγόριθμου ομαδοποίησης εξαρτάται τόσο από τον τύπο των διαθέσιμων δεδομένων όσο και από τον συγκεκριμένο σκοπό της εφαρμογής. Αν η ανάλυση ομάδων χρησιμοποιείται ως περιγραφικό ή διερευνητικό εργαλείο, είναι δυνατόν να γίνει δοκιμή αρκετών αλγορίθμων στα ίδια δεδομένα για να παρατηρηθεί στο τι ενδέχεται να αποκαλύψουν.

Μερικοί αλγόριθμοι ομαδοποίησης ενσωματώνουν τις ιδέες πολλών μεθόδων ομαδοποίησης, οπότε είναι μερικές φορές δύσκολο να ταξινομηθεί ένας αλγόριθμος που ανήκει αποκλειστικά και μόνο σε μια κατηγορία μεθόδου ομαδοποίησης. Επιπλέον, ορισμένες εφαρμογές ενδέχεται να έχουν κριτήρια ομαδοποίησης που απαιτούν την ενσωμάτωση πολλών τεχνικών ομαδοποίησης. (Han and Kamber, 2006, pp. 398-400)

3.6. Αλγόριθμοι Συσταδοποίησης

3.6.1. K-means Αλγόριθμος

Ο k-means διαχωρίζει τα δεδομένα του συνόλου εκπαίδευσης (σύνολο εγγραφών) σε k ομάδες, όπου το k καθορίζεται από τον χρήστη. Η λειτουργία του βασίζεται σε διαδοχικές επαναλήψεις κατά τις οποίες τα δεδομένα κατατάσσονται σε κάποια ομάδα με βάση την ομοιότητα που παρουσιάζουν με το μέσο αυτής της ομάδας.

Ουσιαστικά, ο αλγόριθμος στηρίζεται σε κάποια αντιπροσωπευτικά δείγματα (means) κάθε ομάδας. Κάθε μία από τις k ομάδες που θα δημιουργηθούν θα περιέχει ένα αντιπροσωπευτικό δείγμα το οποίο θα αντιπροσωπεύει την ομάδα, καθώς θα αποτελεί μια μέση περιγραφή αυτής. Αυτό το αντιπροσωπευτικό δείγμα θεωρείται ότι είναι το κέντρο βάρους της ομάδας. Έπειτα, ο αλγόριθμος προσπαθεί να κατατάξει τις εγγραφές στις διάφορες ομάδες έτσι ώστε μετά τον τερματισμό, κάθε εγγραφή να ανήκει σε εκείνη την ομάδα από της οποίας το αντιπροσωπευτικό δείγμα απέχει λιγότερο σε σχέση με αυτά των άλλων ομάδων. (Παγουρόπουλος, 2006, p. 22)

Τα βήματα του αλγορίθμου είναι τα εξής:

1. Προσδιορίζουμε το k.
2. Παίρνουμε τα αρχικά k αντιπροσωπευτικά δείγματα (π.χ. παίρνουμε τις πρώτες k αντιπροσωπευτικές εγγραφές).
3. Επαναλαμβάνουμε.
4. Βρίσκουμε την απόσταση της κάθε εγγραφής από τα αντιπροσωπευτικά δείγματα και θεωρούμε ότι ανήκει στην ομάδα του πιο κοντινού αντιπροσωπευτικού δείγματος.
5. Υπολογίζουμε τα νέα αντιπροσωπευτικά δείγματα (κέντρα βάρους) των ομάδων.
6. Εκτελούμε μέχρι να μην γίνονται αλλαγές.

3.6.2. K-modes Αλγόριθμος

Μία βελτιωμένη έκδοση του αλγορίθμου k-means αποτελεί ο αλγόριθμος k-modes. Είναι ένας αλγόριθμος κατηγοριοποίησης δεδομένων, ο οποίος πραγματεύεται και εφαρμόζεται σε κατηγορικά δεδομένα. Ο συγκεκριμένος αλγόριθμος απαιτεί από τον χρήστη να καθορίσει από την αρχή τον αριθμό των ομάδων που επιθυμεί να εξαχθούν και να παραχθούν και ο αλγόριθμος με την σειρά του προχωράει σε αυτό.

Κάθε ομάδα (cluster) έχει ένα κέντρο ή αλλιώς μέσο (mode) που σχετίζεται με αυτήν. Υποθέτουμε ότι τα αντικείμενα του συνόλου που έχουμε στην διάθεσή μας περιγράφονται από m κατηγορικά πεδία. Το κέντρο της κάθε ομάδας είναι ένα διάνυσμα Q = $(q_1,q_2,\ldots\ldots,q_m)$ όπου το στοιχείο q_i είναι εκείνο με την μεγαλύτερη συχνότητα όσον αφορά την τιμή του για το ιοστό πεδίο στην ομάδα των αντικειμένων.

Δίνοντας ένα σύνολο δεδομένων και ορίζοντας τον αριθμό των ομάδων, ο αλγόριθμος λειτουργεί ως εξής:

1. Αρχικά επιλέγουμε k κέντρα (modes) για k ομάδες (clusters).
2. Για κάθε στοιχείο x :
    - Υπολογίζουμε την ομοιότητα μεταξύ του αντικειμένου / στοιχείου x και των κέντρων όλων των ομάδων.
    - Εισάγουμε το στοιχείο x στην ομάδα c της οποίας το κέντρο / μέσο είναι πιο κοντινό, όμοιο με το x.
    - Κάνουμε update στο κέντρο της ομάδας c.
3. Επανεξετάζουμε την ομοιότητα των στοιχείων σε σχέση με τα υπάρχοντα κέντρα (modes) των ομάδων. Στην περίπτωση που ένα στοιχείο είναι πιο κοντά σε ένα κέντρο / μέσο που ανήκει σε άλλη ομάδα παρά στην δική του, το επανατοποθετούμε σε εκείνη την ομάδα και επαναπροσδιορίζουμε τα κέντρα / μέσα των ομάδων.
4. Επαναλαμβάνουμε το βήμα 3 έως ότου κανένα ή ελάχιστα αντικείμενα να αλλάζουν ομάδες . (Παγουρόπουλος, 2006, p. 22-23)

3.6.3. EM Αλγόριθμος

Ο αλγόριθμος EM λειτουργεί εκτιμώντας τα δεδομένα που λείπουν (E-step) και έπειτα εκτιμώντας τις παραμέτρους του μοντέλου με την μεγαλύτερη ομοιότητα (M-step). Η προσέγγιση αυτή απαιτεί η συλλογή αντικειμένων και οι ομάδες τους (clusters) να αναπαρίστανται από ένα στατιστικό μοντέλο. Τα δεδομένα θεωρούνται σαν ένα τυχαίο δείγμα από ένα μίγμα πιθανοτικών κατανομών (distributions). Αυτές καθορίζουν και τα clusters. (Παγουρόπουλος, 2006, p. 32)

3.6.4. DENCLUE Αλγόριθμος

Ο συγκεκριμένος αλγόριθμος χρησιμοποιεί συναρτήσεις οι οποίες είναι γνωστές και σαν influence function για να μοντελοποιήσει τον αντίκτυπο ενός αντικειμένου μέσα στον χώρο / γειτονιά όπου βρίσκεται. Η πυκνότητα τότε του διαστήματος εκείνου υπολογίζεται σαν το άθροισμα από τις influence functions από όλα τα αντικείμενα. Έπειτα οι ομάδες / clusters (που αποκαλούνται density attractors) καθορίζονται από το τοπικό μέγιστο της ολικής density function. (Παγουρόπουλος, 2006, p. 38)

3.7. Συμπερασματική ανάλυση των αλγορίθμων της Συσταδοποίησης

Ο k-means αλγόριθμος είναι αυτός που χρησιμοποιείται πιο συχνά στην συσταδοποίηση. Παρόλα αυτά, διαθέτει έναν πολύ σημαντικό περιορισμό διότι εφαρμόζεται μόνο πάνω σε αριθμητικά δεδομένα (numeric data). Ο k-modes, ως μια πιο βελτιωμένη έκδοση του k-means, τον καταρρίπτει αυτόν τον περιορισμό και μπορεί να εφαρμοστεί πάνω σε μη-αριθμητικά (non-numeric) ή κατηγορικά (categorical) δεδομένα.

Από την άλλη, ο EM αλγόριθμος ενώ χρησιμοποιείται σε πολλές και διαφορετικού τύπου εφαρμογές, καθώς και αρκετά συχνά στην συσταδοποίηση δεδομένων, προτιμάται από τον χρήστη διότι ασχολείται με τα "χαμένα" δεδομένα από μια μήτρα δεδομένων. Τέλος, ο DENCLUE αποτελεί μια καινούργια προσέγγιση και ιδιαίτερη περίπτωση, που ουσιαστικά αφορά μια πιο γρήγορη εφαρμογή της συσταδοποίησης (fast clustering).

3.8. Κανόνες Συσχέτισης – Association Rules

Αυτή η τεχνική χρησιμοποιείται για την ανακάλυψη προτύπων που περιγράφουν σημαντικές αλληλεξαρτήσεις μεταξύ των διαφόρων πεδίων - χαρακτηριστικών ενός συνόλου δεδομένων. Εφαρμόζεται σε καταστήματα λιανικής πώλησης και βοηθά στο μάρκετινγκ, στη διαφήμιση, στον έλεγχο του καταλόγου απογραφής κλπ. Η πιο συνηθισμένη εφαρμογή της είναι "η ανάλυση του καλαθιού της νοικοκυράς" ή Market Basket Analysis (MBA), καθώς σκοπός της είναι να αναγνωρισθούν τα αγαθά που αγοράζονται μαζί.

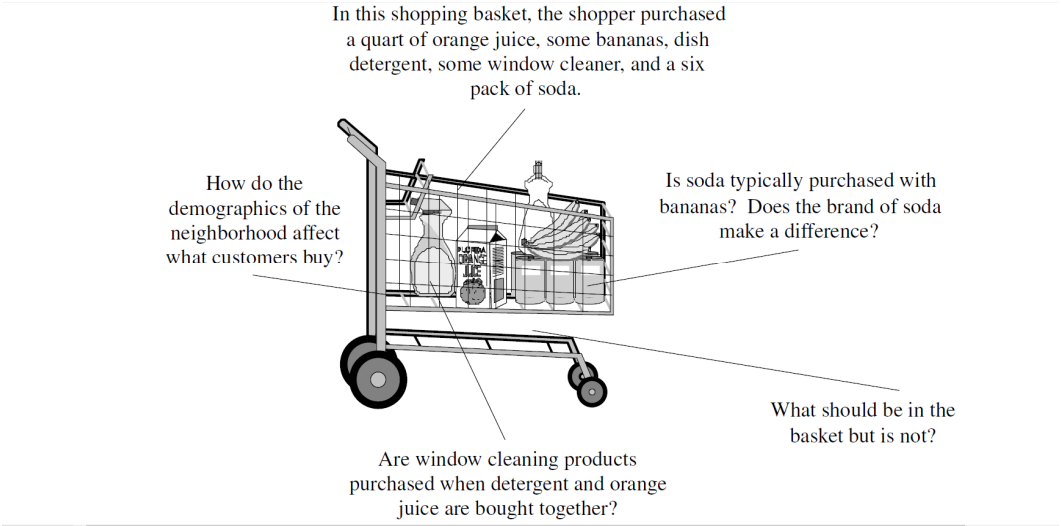
Οι **κανόνες συσχέτισης** μελετούν το πρόβλημα της εύρεσης συχνών συνόλων αντικειμένων ή στοιχειοσυνόλων (frequent itemsets) σε βάσεις δεδομένων. Βασίζονται σε ένα κατώφλι που ονομάζεται υποστήριξη (support), το οποίο αναγνωρίζει τα στοιχειοσύνολα.

Ένα άλλο κατώφλι είναι η εμπιστοσύνη (confidence), η οποία εκφράζει την υπό συνθήκη πιθανότητα ότι ένα αντικείμενο εμφανίζεται σε μια δοσοληψία όταν επίσης εμφανίζεται ένα άλλο αντικείμενο και χρησιμοποιείται για τον εντοπισμό των κανόνων συσχέτισης.

Αυτό το πρόβλημα εύρεσης όλων των κανόνων συσχέτισης που πληρούν τις επιθυμητές τιμές υποστήριξης και εμπιστοσύνης μπορεί να διαιρεθεί σε δύο υποπροβλήματα :

- Στην εύρεση όλων των συνδυασμών των προϊόντων που έχουν υποστήριξη πάνω από την ελάχιστη υποστήριξη. Αυτοί οι συνδυασμοί ονομάζονται μεγάλες λίστες από προϊόντα (large itemsets).
- Στην χρήση όλων των μεγάλων λιστών από προϊόντα για εξόρυξη των κανόνων συσχέτισης που ικανοποιούν την ελάχιστη εμπιστοσύνη. (Καρασιώτου, 2010, pp. 34-35)

Οι κανόνες συσχέτισης αντιπροσωπεύουν πρότυπα στα δεδομένα χωρίς συγκεκριμένο στόχο. Ως εκ τούτου, αποτελούν παράδειγμα μη κατευθυνόμενης εξόρυξης δεδομένων. Το αν τα πρότυπα βγάζουν νόημα αφήνεται στο πως θα το ερμηνεύσει ο άνθρωπος.



Εικόνα 3.2 : Market Basket Analysis.

Αρχικά, η τεχνική αυτή προέκυψε από point-of-sale δεδομένα, που περιγράφουν τα προϊόντα που αγοράζονται μαζί. Αν και η ρίζες της είναι στην ανάλυση των point-of-sale συναλλαγών, οι κανόνες συσχέτισης μπορούν να εφαρμοστούν και εκτός της λιανικής βιομηχανίας για να βρεθούν σχέσεις μεταξύ άλλων τύπων "καλαθιών". Ορισμένα παραδείγματα πιθανών εφαρμογών είναι :

- Στοιχεία που αγοράζονται με πιστωτική κάρτα, όπως είναι τα ενοικιαζόμενα αυτοκίνητα και τα δωμάτια ξενοδοχείων, παρέχουν πληροφορίες για το επόμενο προϊόν που είναι πιθανό να αγοράσουν οι πελάτες.
- Προαιρετικές υπηρεσίες που αγοράζονται από τηλεπικοινωνιακούς πελάτες (αναμονή κλήσης, προώθηση κλήσης, DSL κλπ.), συμβάλλουν στον προσδιορισμό του τρόπου δέσμευσης των υπηρεσιών αυτών προκειμένου να μεγιστοποιηθούν τα έσοδα.
- Τραπεζικές υπηρεσίες που χρησιμοποιούνται από πελάτες λιανικής (λογαριασμοί χρηματαγοράς, CD, επενδυτικές υπηρεσίες, δάνεια αυτοκινήτων κλπ.) προσδιορίζουν τους πελάτες που πιθανόν να επιθυμούν και άλλες υπηρεσίες.
- Ασυνήθιστοι συνδυασμοί ασφαλιστικών απαιτήσεων μπορεί να αποτελούν ένδειξη απάτης και μπορεί να προκαλέσουν μια περαιτέρω έρευνα.
- Ιατρικές ιστορίες ασθενών μπορούν να δώσουν ενδείξεις πιθανών επιπλοκών που βασίζονται σε ορισμένους συνδυασμούς θεραπειών.

Οι κανόνες σύνδεσης συχνά δεν ανταποκρίνονται στις προσδοκίες. Για παράδειγμα, δεν είναι καλή επιλογή για την κατασκευή μοντέλων πολλαπλών πωλήσεων σε κλάδους όπως η λιανική τραπεζική, επειδή οι κανόνες καταλήγουν να περιγράφουν προηγούμενες προσφορές μάρκετινγκ. Επίσης, στη λιανική τραπεζική, οι πελάτες αρχίζουν συνήθως με έναν τρεχούμενο λογαριασμό και στη συνέχεια έναν λογαριασμό ταμιευτηρίου. Η διαφοροποίηση μεταξύ των προϊόντων δεν εμφανίζεται έως ότου οι πελάτες να έχουν περισσότερα προϊόντα.

Ποια είναι τα πιο δημοφιλή αντικείμενα; Γνωρίζοντας τις πωλήσεις ενός μεμονωμένου αντικειμένου είναι μόνο η αρχή. Υπάρχουν σχετικές ερωτήσεις όπως :

- Ποιο είναι το πιο συνηθισμένο αντικείμενο που βρέθηκε σε παραγγελία ενός αντικειμένου;
- Ποιο είναι το πιο συνηθισμένο αντικείμενο που βρέθηκε σε μια παραγγελία πολλών αντικειμένων;
- Ποιο είναι το πιο συνηθισμένο αντικείμενο που βρέθηκε μεταξύ των πελατών που είναι συχνοί αγοραστές;
- Πώς άλλαξε η δημοτικότητα συγκεκριμένων αντικειμένων με την πάροδο του χρόνου;
- Πώς διαφέρει η δημοτικότητα ενός αντικειμένου σε περιφερειακό επίπεδο;

Οι τρεις πρώτες ερωτήσεις είναι ιδιαίτερα ενδιαφέρουσες επειδή μπορεί να προτείνουν ιδέες για την ανάπτυξη σχέσεων με τους πελάτες. Οι κανόνες συσχέτισης μπορούν να παρέχουν απαντήσεις σε αυτές τις ερωτήσεις, ιδίως όταν χρησιμοποιούνται με εικονικά αντικείμενα για να αντιπροσωπεύουν το μέγεθος της παραγγελίας ή τον αριθμό των παραγγελιών που έχει κάνει ένας πελάτης.

Οι δύο τελευταίες ερωτήσεις αναδεικνύουν τις διαστάσεις του χρόνου και της γεωγραφίας, οι οποίες είναι πολύ σημαντικές για τις εφαρμογές της ανάλυσης του καλαθιού αγοράς (Market Basket Analysis). Διαφορετικά προϊόντα έχουν διαφορετικές συγγένειες σε διάφορες περιοχές - κάτι που οι λιανοπωλητές είναι πολύ εξοικειωμένοι. Είναι επίσης δυνατό να χρησιμοποιηθούν κανόνες συσχέτισης για να αρχίσουν να κατανοούν αυτούς τους τομείς, εισάγοντας εικονικά στοιχεία για την περιοχή και την εποχικότητα.

Μια έφεση των κανόνων συσχέτισης είναι η σαφήνεια και η χρησιμότητα των αποτελεσμάτων, τα οποία έχουν τη μορφή κανόνων για τις ομάδες προϊόντων. Υπάρχει μια διαισθητική έκκληση σε έναν κανόνα συσχέτισης επειδή εκφράζει το πώς απτά προϊόντα και υπηρεσίες ομαδοποιούνται. Ένας κανόνας όπως "εάν ο πελάτης αγοράσει την υπηρεσία τριπλής κλήσης, τότε αυτός ο πελάτης θα αγοράσει επίσης αναμονή κλήσης", είναι σαφής. Ακόμα καλύτερα, θα μπορούσε να υποδείξει μια συγκεκριμένη πορεία δράσης, όπως για παράδειγμα η ομαδοποίηση τριπλών κλήσεων με αναμονή κλήσεων σε ένα ενιαίο πακέτο υπηρεσιών.

Ενώ οι κανόνες συσχέτισης είναι εύκολο να κατανοηθούν, δεν είναι πάντοτε χρήσιμοι. Οι ακόλουθοι τρεις κανόνες είναι παραδείγματα πραγματικών κανόνων που παράγονται από πραγματικά δεδομένα :

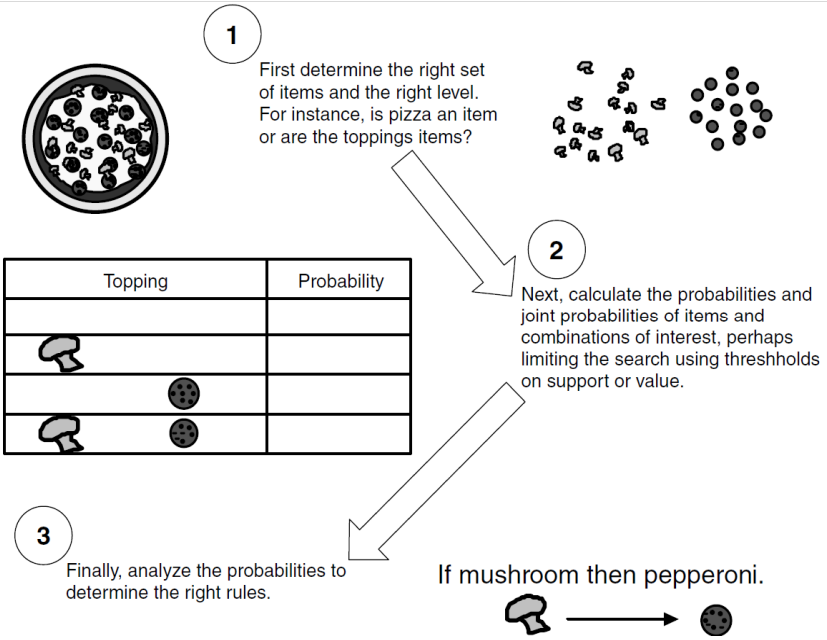
- Πελάτες της Wal-Mart που αγοράζουν κούκλες της Barbie έχουν 60% πιθανότητα να αγοράσουν έναν από τους τρεις τύπους καραμελών.
- Πελάτες που αγοράζουν συμφωνίες συντήρησης είναι πολύ πιθανό να αγοράσουν μεγάλες συσκευές.
- Όταν ανοίγει ένα νέο κατάστημα εργαλείων, ένα από τα πιο συχνά πωλούμενα αντικείμενα είναι τα καθαριστικά λεκάνης τουαλέτας.

Τα δύο τελευταία παραδείγματα αποτελούν παραδείγματα που έχουμε δει πραγματικά στα δεδομένα. Το πρώτο είναι ένα παράδειγμα που αναφέρθηκε στο Forbes στις 8 Σεπτεμβρίου 1997. Αυτά τα τρία παραδείγματα απεικονίζουν τους τρεις κοινούς τύπους

κανόνων που παράγονται από τους κανόνες συσχέτισης: το πραγματοποιήσιμο, το ασήμαντο και το ανεξήγητο.

Η βασική διαδικασία για την εύρεση κανόνων συσχέτισης απεικονίζεται στο παρακάτω σχήμα. Υπάρχουν τρεις σημαντικές ανησυχίες για τη δημιουργία κανόνων συσχέτισης :

- Η επιλογή του σωστού συνόλου αντικειμένων.
- Η δημιουργία κανόνων αποκρυπτογραφώντας τις μετρήσεις στην συνυπάρχουσα μήτρα (matrix).
- Η υπέρβαση των πρακτικών ορίων που επιβάλλονται από χιλιάδες ή δέκα χιλιάδες αντικείμενα. (Berry and Linoff, 2004, pp. 287-303)



Εικόνα 3.3 : Τα βασικά βήματα εύρεσης των κανόνων συσχέτισης.

3.8.1. Apriori Αλγόριθμος

Ο Apriori παρουσιάστηκε αρχικά το 1994 και είναι ο βασικότερος αλγόριθμος των κανόνων συσχέτισης. Δέχεται ως είσοδο ένα σύνολο αγορών (transactions) που αποτελεί και το σύνολο εκπαίδευσης. Κάθε αγορά είναι ουσιαστικά μία λίστα (itemset) από αγαθά (items) που αγοράστηκαν μαζί. Για κάθε αγορά υπάρχει ένας και μοναδικός κωδικός αναγνώρισης, ο οποίος ονομάζεται TID (transaction identifier).

Όπως αναφέρθηκε στην προηγούμενη ενότητα, οι κανόνες συσχέτισης μελετούν το πρόβλημα της εύρεσης συχνών συνόλων αντικειμένων. Η εύρεση μεγάλων λιστών από προϊόντα, για να αποφεύγει ένα εξαντλητικό ψάξιμο όλων των συνδυασμών βασίζεται στο γεγονός ότι μία λίστα είναι μεγάλη όταν κάθε υποσύνολό της είναι μεγάλη λίστα από προϊόντα. Ο αλγόριθμος Apriori εντοπίζει τις μεγάλες λίστες από προϊόντα έχοντας πρόσβαση πολλές φορές στο σύνολο εκπαίδευσης. Στην πρώτη προσπέλαση υπολογίζεται η επιβεβαίωση - υποστήριξη κάθε διαφορετικού προϊόντος ξεχωριστά και στην συνέχεια καθορίζεται ποια από αυτές είναι μεγάλες λίστες από προϊόντα. Σε κάθε επόμενη προσπέλαση, αρχίζουμε έχοντας τις μεγάλες λίστες από προϊόντα που βρέθηκαν στην προηγούμενη προσπέλαση. Από αυτές δημιουργούμε νέες πιθανές μεγάλες λίστες από προϊόντα, που καλούνται υποψήφιες (candidate) μεγάλες λίστες από προϊόντα. Έπειτα, μετράμε την ακριβή επιβεβαίωση - υποστήριξη αυτών και καθορίζουμε ποιες είναι οι πραγματικά μεγάλες λίστες από προϊόντα. Οι τελευταίες αποτελούν βάση για το επόμενο βήμα. Η αποτελεσματικότητα στην εύρεση μεγάλων λιστών από προϊόντα αποτελεί κριτήριο για την αποτελεσματικότητα συνολικά ενός αλγορίθμου εύρεσης κανόνων συσχέτισης.

Τα βήματα του αλγορίθμου Apriori είναι τα εξής:

1. Βρίσκουμε τα αγαθά που εμφανίζονται περισσότερο από την ελάχιστη επιβεβαίωση (minimum support), δηλαδή το σύνολο L_1 = μεγάλες λίστες από ένα αγαθό (large 1-item sets).
2. Από k = 2 και όσο το L_{k-1} δεν είναι κενό :
    A. Βρίσκουμε το σύνολο C_k των υποψήφιων μεγάλων λιστών από k αγαθά (canditate large k-item sets) με βάση το L_{k-1}.
    B. Βρίσκουμε ποια από αυτά εμφανίζονται περισσότερο από την ελάχιστη επιβεβαίωση και φτιάχνουμε το σύνολο L_k = μεγάλες λίστες από k αγαθά.
3. Για κάθε στοιχείο των L_1, … , L_n βρίσκουμε ποια ικανοποιούν την ελάχιστη αξιοπιστία (minimum confidence).

Στο πρώτο βήμα, ο αλγόριθμος μετρά τις εμφανίσεις του κάθε προϊόντος ξεχωριστά για να καθοριστούν οι μεγάλες λίστες μεγέθους ενός προϊόντος. Το δεύτερο βήμα είναι μία επαναλαμβανόμενη σειρά από υποβήματα. Κάθε επανάληψη, έστω η k, περιλαμβάνει δύο υποβήματα. Αρχικά, μεγάλες λίστες από k-1 προϊόντα L_{k-1}, που βρέθηκαν στην προηγούμενη k-1 επανάληψη χρησιμοποιούνται για να δημιουργηθούν οι υποψήφιες μεγάλες λίστες από k προϊόντα (C_k), με βάση την αρχή ότι μια λίστα από προϊόντα είναι μεγάλη, αν κάθε

υποσύνολό της είναι μεγάλη λίστα από προϊόντα. Στην συνέχεια, σαρώνουμε το σύνολο εκπαίδευσης για να βρούμε την επιβεβαίωση των υποψήφιων μεγάλων λιστών από k προϊόντα. Οι επαναλήψεις σταματούν όταν δεν υπάρχουν υποψήφιες μεγάλες λίστες από προϊόντα. Τότε, στο τελευταίο βήμα, από κάθε μία μεγάλη λίστα από προϊόντα προκύπτουν κανόνες από τους οποίους γίνονται τελικά αποδεκτοί όσοι έχουν μεγαλύτερη από την ελάχιστη αξιοπιστία. (Παγουρόπουλος, 2006, pp. 41-42)

3.9. Συμπερασματική ανάλυση των αλγορίθμων των Κανόνων Συσχέτισης

Υπάρχουν πολλοί αλγόριθμοι για την παραγωγή των Κανόνων Συσχέτισης, όπως είναι για παράδειγμα ο Eclat, ο FP-Growth και άλλοι. Τον βασικότερο αποτελεί ο Apriori λόγω και της ιστορικής του σημασίας. Παρόλα αυτά, ο συγκεκριμένος αλγόριθμος σκανάρει την βάση δεδομένων πάρα πολλές φορές, κάτι το οποίο μειώνει την όλη εκτέλεση. Επίσης, η πολυπλοκότητα του χώρου και του χρόνου είναι εξαιρετικά υψηλή. Αυτό όμως που τον καθιστά τόσο βασικό είναι η ευκολία που παρέχει στην κατανόηση του αλλά και στην εφαρμογή του, καθώς και το ότι μπορεί να χρησιμοποιηθεί σε μεγάλα στοιχεία (large itemsets).

Συμπέρασμα

Είναι πλέον σαφές ότι η Εξόρυξη Δεδομένων είναι ένας από τους πιο σημαντικούς τομείς της εποχής μας. Τα δεδομένα είναι ασταμάτητα και η άχρηστη πληροφορία κατέχει το μεγαλύτερο ποσοστό μέσα σε αυτά. Η έξυπνη ανάλυση τους αποτελεί πολύτιμο πόρο αλλά είναι ένα δύσκολο έργο. Στις περισσότερες περιπτώσεις, η χρήση μιας τεχνικής, μιας μεθόδου ή ενός αλγορίθμου δεν είναι αρκετή. Οι δομές δεδομένων είναι περίπλοκες και οι βάσεις δεδομένων ξεχειλίζουν από ακατέργαστη πληροφορία, ώστε είναι αρκετά προκλητική και σκληρή διαδικασία για να αναλυθούν αποτελεσματικά. Η ανάγκη όμως να κατανοήσουμε αυτά τα πλούσια σε γνώση σύνολα δεδομένων, είναι κοινός στόχος σε όλους τους τομείς, συμπεριλαμβανομένων και των επιχειρήσεων, της επιστήμης και της μηχανικής.

Με την χρήση τεχνικών, αλγορίθμων και άλλων εργαλείων, η εξαγωγή γνώσης καθίσταται δυνατή. Δεν υπάρχει καλύτερος αλγόριθμος ή κάποια καλύτερη τεχνική που χρησιμοποιείται ώστε να λάβουμε το επιθυμητό και καλύτερο δυνατό αποτέλεσμα. Συγκριμένα, η επιλογή των αλγορίθμων γίνεται ανάλογα με την γνώση που θέλουμε να εξάγουμε, τα δεδομένα που διαθέτουμε, την προτίμηση ανάμεσα στον ρυθμό και την ακρίβεια, καθώς και πάνω σε ποιες εφαρμογές θέλουμε να τους χρησιμοποιήσουμε. Ουσιαστικά, υπάρχουν πολλές παράμετροι και πολλοί παράγοντες που πρέπει να ληφθούν υπόψη στην επιλογή του κατάλληλου – για εμάς – αλγορίθμου.

Βιβλιογραφικές Παραπομπές

1. **Algarni, A.** (2016). Data Mining in Education. *International Journal of Advanced Computer Science and Applications*, 7(6).
2. **Baker, R. S. J. d.** *Data Mining for Education.*
3. **Berry, M. and Linoff, G.** (2004). *Data mining techniques*. 2nd ed. Indianapolis: Wiley.
4. **Dean, J.** (2014). *Big Data, Data Mining and Machine Learning*. Hoboken, New Jersey: Wiley.
5. **El.wikipedia.org.** (2018). *Αλγόριθμος ID3*. [online] Available at: https://el.wikipedia.org/wiki/%CE%91%CE%BB%CE%B3%CF%8C%CF%81%CE%B9%CE%B8%CE%BC%CE%BF%CF%82_ID3 [Accessed 27 Nov. 2018].
6. **El.wikipedia.org.** (2018). *Εξόρυξη δεδομένων*. [online] Available at: https://el.wikipedia.org/wiki/%CE%95%CE%BE%CF%8C%CF%81%CF%85%CE%BE%CE%B7_%CE%B4%CE%B5%CE%B4%CE%BF%CE%BC%CE%AD%CE%BD%CF%89%CE%BD [Accessed 27 Nov. 2018].
7. **El.wikipedia.org.** (2018). *Στατιστική ταξινόμηση*. [online] Available at: https://el.wikipedia.org/wiki/%CE%A3%CF%84%CE%B1%CF%84%CE%B9%CF%83%CF%84%CE%B9%CE%BA%CE%AE_%CF%84%CE%B1%CE%BE%CE%B9%CE%BD%CF%8C%CE%BC%CE%B7%CF%83%CE%B7 [Accessed 27 Nov. 2018].
8. **En.wikipedia.org.** (2019). *Educational data mining*. [online] Available at: https://en.wikipedia.org/wiki/Educational_data_mining [Accessed 11 Jan. 2019].
9. **En.wikipedia.org.** (2018). *Support vector machine*. [online] Available at: https://en.wikipedia.org/wiki/Support_vector_machine [Accessed 27 Nov. 2018].
10. **GEORGE, Gerard, HAAS, Martine R., and PENTLAND, Alex**. Big Data and Management: From the Editors. (2014). Academy of Management Journal. 57, (2), 321-326. Research Collection Lee Kong Chian School Of Business. Available at: http://ink.library.smu.edu.sg/lkcsb_research/4621
11. **Han, J. and Kamber, M.** (2006). *Data mining*. 2nd ed. Amsterdam: Elsevier, Morgan Kaufmann.
12. **Hand, D., Mannila, H. and Smyth, P.** (2001). *Principles of data mining*. Cambridge, Mass.: MIT Press.
13. **Kantardzic, M.** (2011). *Data Mining: Concepts, Models, Methods, and Algorithms, 2nd Edition*. 2nd ed. Hoboken, New Jersey: John Wiley & Sons.

14. **Kdnuggets.com.** (2018). *History of Data Mining*. [online] Available at: https://www.kdnuggets.com/2016/06/rayli-history-data-mining.html [Accessed 9 Dec. 2018].
15. **Korting, T. S.** (2014). 'C4.5 algorithm and Multivariate Decision Trees', *ResearchGate.* Available at: https://www.researchgate.net/profile/Thales_Koerting/publication/267945462_C45_algorithm_and_Multivariate_Decision_Trees/links/5475b99b0cf29afed612b236.pdf
16. **N. O. Sadiku, M., G. Eze, K. and M. Musa, S.** (2018). Data Mining in Healthcare. *International Journal of Advances in Scientific Research and Engineering*, 4(9), pp.90-92.
17. **Raste, K. S.** (2014). *BIG DATA ANALYTICS – HADOOP PERFORMANCE ANALYSIS.* San Diego State University.
18. **Russom, P.** (2011). *BIG DATA ANALYTICS*. TDWI BEST PRACTISES REPORTS.
19. **Witten, I. and Frank, E.** (2005). *Data mining*. 2nd ed. San Francisco, Calif.: Morgan Kaufmann.
20. **Καρασιώτου, Β.** (2010). *Υλοποίηση Αποθήκης Μεταναστευτικών Δεδομένων - OLAP Ανάλυση - Data mining μοντέλα.* Πανεπιστήμιο Πειραιώς.
21. **Μουζάκη, Δ.** (2006). *ΑΛΓΟΡΙΘΜΟΙ ΟΜΑΔΟΠΟΙΗΣΗΣ-ΤΑΞΙΝΟΜΗΣΗΣ.* Τ.Ε.Ι. Μεσολογγίου.
22. **Παγουρόπουλος, Α.** (2006). *Data Mining στην Χρηματοοικονομική Ανάλυση.* Πανεπιστήμιο Πατρών.
23. **Σταυλιώτης, Γ. Ε.** (2008). *ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ (DATA MINING) ΚΑΙ ΚΑΤΗΓΟΡΙΚΑ ΔΕΔΟΜΕΝΑ.* Πανεπιστήμιο Πειραιώς.
24. **Τσικριτέας, Χ.** (2015). *ΤΑ ΜΕΓΑΛΑ ΔΕΔΟΜΕΝΑ - ΙΔΙΩΤΙΚΟΤΗΤΑ ΚΑΙ ΡΥΘΜΙΣΤΙΚΕΣ ΑΡΧΕΣ.* Τ.Ε.Ι. Ηπείρου.