Ακαδημία Εμπορικού Ναυτικού Ασπροπύργου
Σχολή Μηχανικών

Πτυχιακή εργασία

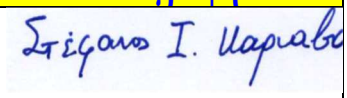
Εξόρυξη δεδομένων και εφαρμογές της

Σπουδαστής: Ντάϊκο Γεώργιος (ΑΜ 9273)

Επιβλέπων Καθηγητής: Στέφανος Ι. Καρναβάς, Μαθηματικός (M.Ed., Ph.D.)
Επίκουρος Καθηγητής

Ακαδημαϊκό έτος: 2025–2026

Ημερομηνία ανάθεσης: 15.11.2024
Ημερομηνία κατάθεσης: ….01.2026
Ημερομηνία εξέτασης: ….01.2026

Α/Α	Ονοματεπώνυμο	Χαρακτηρισμός	Υπογραφή
1	Στέφανος Ι. Καρναβάς Μαθηματικός (M.Ed., Ph.D.) Επίκουρος Καθηγητής	Άριστα 10	
2	Τσαγκανός Γεώργιος Αυτοματιστής (M.sc.) ΕΔΙΠ		
3	Πετεινάτος Ηλίας Ηλεκτρονικός (M.sc.) ΕΔΙΠ		
Τελικός χαρακτηρισμός			

Περίληψη

Οι σύγχρονες κοινωνίες κατακλύζονται από δεδομένα, η ποσότητα των οποίων διαρκώς αυξάνεται δυσανάλογα με την κατανόηση – ερμηνεία τους, διότι στα δεδομένα περιέχονται σημαντικές πληροφορίες που σπάνια αποσαφηνίζονται (π.χ. σε εμπορικό περιβάλλον, σε ανταγωνιστικά πλεονεκτήματα). Ενώ η ροή των δεδομένων βαίνει διαρκώς διογκούμενη, οι σύγχρονες μηχανές αναζήτησης, που ευρύτατα χρησιμοποιούνται από τους πολίτες, προσφέρουν αυξανόμενες ευκαιρίες για εξόρυξη γνώσεων από τα δεδομένα (Data mining) ή όπως συχνά λέγεται ανακάλυψη γνώσης σε βάσεις δεδομένων (Knowledge Discovery in Databases–KDD) και ενδεικτικά περιλαμβάνει την ανακάλυψη – συγκομιδή των πληροφοριών, την εξερευνητική ανάλυση των δεδομένων, τη μη επιβλεπόμενη αναγνώριση προτύπου. Η έκφραση KDD σημαίνει ανακάλυψη γνώσης σε βάσεις δεδομένων, δηλαδή είναι η διαδικασία της εύρεσης των έγκυρων πληροφοριών και της αναγνώρισης των χρήσιμων και κατανοητών προτύπων, σε μεγάλα σύνολα δεδομένων. Ένας ακριβής ορισμός της KDD ο οποίος δόθηκε το 1991 από τους Frawley, Piatesky–Shapiro, Matheus είναι: «*Ανακάλυψη γνώσης σε βάσεις δεδομένων, είναι η ντετερμινιστική διαδικασία της αναγνώρισης έγκυρων, καινοτόμων, ενδεχομένως χρήσιμών και εν τέλει κατανοητών προτύπων στα δεδομένα*».

Η έκφραση Data Mining, σημαίνει τη χρήση αλγορίθμων για την εξαγωγή των πληροφοριών και των προτύπων που παράγονται με τη διαδικασία KDD.

Λέξεις κλειδιά

Εξόρυξη δεδομένων, εξόρυξη γνώσης από βάσεις δεδομένων, εξαγωγή γνώσης, ανάλυση (δεδομένων/προτύπων, χρονοσειρών), αρχαιολογία δεδομένων, βυθοκόρηση δεδομένων, βάσεις δεδομένων (χωρικές, χρονοσειρών, κειμένου, πολυμέσων, object–oriented, object–relational), ανακάλυψη γνώσης από βάσεις δεδομένων, CRISP–DM (Cross–Industry Standard Process for Data Mining), κατηγοριοποίηση, συσταδοποίηση (διαιρετική, ασαφής, μη ασαφής, ιεραρχική, εννοιολογική, αριθμητικών δεδομένων, βασισμένη σε πλέγμα, βασισμένη στην πυκνότητα, PAM, K–means), μάθηση (μηχανική, μη εποπτευόμενη, εποπτευόμενη, ημιεποπτευόμενη, ανταγωνιστική, μη ανταγωνιστική), αλγόριθμοι (παράλληλοι, ιεραρχικοί, διαμεριστικοί, συσταδοποίησης), αλγόριθμοι αναζήτησης (παραμέτρων, μοντέλων), κανόνες συσχέτισης, μοντελοποίηση, προετοιμασία δεδομένων, εφαρμογές της εξόρυξης δεδομένων, μοτίβο (συχνό, δομημένο, διαδοχικό), υπερπροσαρμογή, εκπαίδευση του μοντέλου, σύνολο (εκπαίδευσης, δεδομένων), διάγραμμα ανύψωσης, πίνακας σύγχυσης (confusion matrix), πίνακας ταξινόμησης (classification matrix), διάγραμμα μπανάνας (banana chart), data (cleaning, integration, transformation, discretization, reduction), επαγωγική στατιστική, data warehouse, SPSSS, SAS, cloud computing, Bayesian θεωρία, βασικό κυβοειδές, κύβος δεδομένων, OnLine Analytical Processing (OLAP), παλινδρόμηση (regression), πρόβλεψη (prediction), summarization, κανόνες συσχέτισης (association rules), medoid.

Περιεχόμενα

Εισαγωγή

Μια σημαντική φάση της τεχνολογικής καινοτομίας που συνδέεται με την ταχεία ανάπτυξη των υπολογιστών, ξεκίνησε πριν από λίγες δεκαετίες και ονομάζεται εξόρυξη δεδομένων. Έφερε επανάσταση στον τρόπο που εργάζονται οι άνθρωποι, στις επιστήμες, στις επιχειρήσεις και στην καθημερινή ζωή. Επί αρκετά έτη αναπτύχθηκε μια πτυχή της τεχνολογικής καινοτομίας (όχι ανεξάρτητη από την ανάπτυξη των υπολογιστών), της δόθηκε η δική της αυτονομία, μεγάλες (μερικές φορές τεράστιες) μάζες πληροφοριών για ένα ευρύ φάσμα θεμάτων, έγιναν ξαφνικά διαθέσιμες απλά και φθηνά. Αυτό, οφείλονταν πρώτα στην ανάπτυξη των αυτόματων μεθόδων συλλογής των δεδομένων και στη συνέχεια στις βελτιώσεις στα ηλεκτρονικά συστήματα αποθήκευσης των πληροφοριών και στις σημαντικές μειώσεις του κόστους τους.

Αυτή η εξέλιξη, δε σχετιζόταν με μία εφεύρεση, αλλά ήταν συνέπεια πολλών καινοτόμων στοιχείων που συνέβαλαν από κοινού, στη δημιουργία αυτού που ονομάζεται κοινωνία της πληροφορίας. Σε αυτό το πλαίσιο, έχουν ανοιχτεί νέοι δρόμοι ευκαιριών και τρόποι εργασίας που είναι πολύ διαφορετικοί από αυτούς που χρησιμοποιούνταν στο παρελθόν.

1. Ο όρος εξόρυξη δεδομένων

Η εξόρυξη δεδομένων αναφέρεται στην εξαγωγή ή «εξόρυξη» γνώσης από μεγάλες ποσότητες δεδομένων. Ο όρος, είναι στην πραγματικότητα λανθασμένος και αυτό για να το αντιληφθούμε αρκεί να αναλογιστούμε ότι η εξόρυξη χρυσού από βράχους ή από άμμο αναφέρεται ως εξόρυξη χρυσού και όχι ως εξόρυξη βράχου ή άμμου. Έτσι, ο όρος «εξόρυξη δεδομένων» θα έπρεπε να είχε ονομαστεί καταλληλότερα «**εξόρυξη γνώσης από δεδομένα**», ο οποίος δυστυχώς είναι κάπως μακρύς. Ο όρος εξόρυξη γνώσης, ένας συντομότερος όρος, μπορεί να μην αντικατοπτρίζει την έμφαση στην εξόρυξη από μεγάλες ποσότητες δεδομένων. Παρ' όλα αυτά, η εξόρυξη είναι ένας έντονος όρος που χαρακτηρίζει τη διαδικασία, που βρίσκει ένα μικρό σύνολο πολύτιμου ψήγματος από μεγάλη ποσότητα πρώτης ύλης.

Έτσι, ένας τέτοιος λανθασμένος όρος, που περιέχει τόσο δεδομένα όσο και εξόρυξη, έγινε μια δημοφιλής επιλογή. Υπάρχουν πολλοί άλλοι όροι που έχουν παρόμοια ή ελαφρώς διαφορετική σημασία από την εξόρυξη δεδομένων, όπως η εξόρυξη γνώσης από βάσεις δεδομένων, η εξαγωγή γνώσης, η ανάλυση δεδομένων/προτύπων, η αρχαιολογία δεδομένων και η βυθοκόρηση δεδομένων.

Πολλοί αντιμετωπίζουν την εξόρυξη δεδομένων ως συνώνυμο ενός άλλου ευρέως χρησιμοποιούμενου όρου, του «ανακάλυψη γνώσης σε βάσεις δεδομένων» ή KDD, την οποία θα αναλύσουμε παρακάτω. Εναλλακτικά, άλλοι θεωρούν την εξόρυξη δεδομένων απλώς ένα βήμα σε ολόκληρη τη διαδικασία ανακάλυψης γνώσης σε βάσεις δεδομένων, αν και ουσιώδες, καθώς αποκαλύπτει κρυμμένα μοτίβα για αξιολόγηση. Το βέβαιο είναι πως η εξόρυξη δεδομένων αποτελεί μια διαδικασία ανακάλυψης γνώσης.

2. Η σημασία της εξόρυξης δεδομένων

Ο κύριος λόγος που η εξόρυξη δεδομένων έχει προσελκύσει μεγάλη προσοχή τα τελευταία έτη, οφείλεται στην ευρεία διαθεσιμότητα τεράστιων ποσοτήτων δεδομένων και στην άμεση ανάγκη μετατροπής τους σε χρήσιμες πληροφορίες και γνώσεις, που βρίσκουν εφαρμογές στη διαχείριση των επιχειρήσεων, στον έλεγχο της παραγωγής, στην ανάλυση της αγοράς, στον μηχανικό σχεδιασμό και στην επιστημονική εξερεύνηση. Η εξόρυξη δεδομένων, μπορεί να θεωρηθεί ως αποτέλεσμα της φυσικής εξέλιξης της τεχνολογίας πληροφοριών. Μια εξελικτική πορεία έχει

παρατηρηθεί στη βιομηχανία των βάσεων δεδομένων, στην ανάπτυξη των ακόλουθων λειτουργιών:

- συλλογή των δεδομένων και δημιουργία των βάσεων δεδομένων,
- διαχείριση των δεδομένων (συμπεριλαμβανομένης της αποθήκευσης και της ανάκτησης τους και της επεξεργασίας των βάσεων δεδομένων),
- ανάλυση και κατανόηση των δεδομένων (που περιλαμβάνει την αποθήκευση και την εξόρυξη τους).

Η σταθερή και εκπληκτική πρόοδος της τεχνολογίας του υλικού των υπολογιστών τις τελευταίες δεκαετίες, έχει οδηγήσει σε ισχυρές, οικονομικά προσιτές και μεγάλες προμήθειες υπολογιστών, εξοπλισμού συλλογής των δεδομένων και μέσων αποθήκευσης τους. Αυτή η τεχνολογία, παρέχει μεγάλη ώθηση στη βιομηχανία των βάσεων δεδομένων και των πληροφοριών και καθιστά διαθέσιμο έναν τεράστιο αριθμό βάσεων δεδομένων και αποθετηρίων πληροφοριών, για τη διαχείριση των συναλλαγών, την ανάκτηση των πληροφοριών και την ανάλυση των δεδομένων.

Η αφθονία των δεδομένων, σε συνδυασμό με την ανάγκη για ισχυρά εργαλεία ανάλυσης τους, έχει περιγραφεί ως μια κατάσταση «πλούσια σε δεδομένα, αλλά φτωχή σε πληροφορίες». Η ταχέως αναπτυσσόμενη, τεράστια ποσότητα δεδομένων, που συλλέγονται και αποθηκεύονται σε μεγάλες και πολυάριθμες βάσεις δεδομένων, έχει ξεπεράσει κατά πολύ την ανθρώπινη ικανότητά για κατανόηση, χωρίς ισχυρά εργαλεία. Ως αποτέλεσμα, τα δεδομένα που συλλέγονται σε μεγάλες βάσεις δεδομένων γίνονται «τάφοι δεδομένων», δηλαδή αρχεία δεδομένων που σπάνια επισκέπτονται ξανά.

Άρα, οι σημαντικές αποφάσεις, συχνά λαμβάνονται όχι με βάση τα πλούσια σε πληροφορίες δεδομένα που είναι αποθηκευμένα σε βάσεις δεδομένων, αλλά με βάση τη διαίσθηση ενός υπευθύνου λήψης αποφάσεων, απλώς και μόνο επειδή αυτός δε διαθέτει τα εργαλεία για να εξαγάγει την πολύτιμη γνώση που ενσωματώνεται στις τεράστιες ποσότητες δεδομένων. Επιπλέον, τρέχουσες τεχνολογίες συστημάτων εμπειρογνωμόνων, συνήθως βασίζονται σε χρήστες ή σε ειδικούς στον τομέα για την χειροκίνητη εισαγωγή γνώσης στις βάσεις γνώσεων. Δυστυχώς, αυτή η διαδικασία είναι εξαιρετικά χρονοβόρα, δαπανηρή, επιρρεπής σε προκαταλήψεις και σφάλματα. Τα εργαλεία εξόρυξης δεδομένων που εκτελούν ανάλυση των δεδομένων, μπορούν να αποκαλύψουν σημαντικά πρότυπα δεδομένων, συμβάλλοντας σημαντικά στις επιχειρηματικές στρατηγικές, στις βάσεις γνώσεων, στην επιστημονική και στην ιατρική έρευνα. Το διαρκώς διευρυνόμενο χάσμα μεταξύ των δεδομένων και των πληροφοριών, απαιτεί μια συστηματική ανάπτυξη εργαλείων εξόρυξης δεδομένων που θα μετατρέψουν τους τάφους δεδομένων, σε «χρυσά ψήγματα» γνώσης.

Άλλοι λόγοι που οδήγησαν στην ανάπτυξη της εξόρυξης δεδομένων, πέραν της τεράστιας αύξησης των δεδομένων, της μείωσης που έχει επέλθει στο κόστος της επεξεργασίας και της αύξηση της χωρητικότητας αποθήκευσης των δεδομένων αφορούν στη διαθεσιμότητα λογισμικού data mining, καθώς πολλές είναι εταιρείες οι οποίες έχουν αναπτύξει χρήσιμο λογισμικό εξόρυξης δεδομένων τα τελευταία έτη αλλά και το ανταγωνιστικό επιχειρηματικό περιβάλλον, που έχει επιφέρει στις περισσότερες χώρες η αυξημένη παγκοσμιοποίηση του εμπορίου, όπου για να ευδοκιμήσει στο σημερινό περιβάλλον μία επιχείρηση (ή ένας οργανισμός), είναι ζωτικής σημασίας να αποκτήσει ένα πλεονέκτημα, ώστε να μπορεί να αναπτυχθεί με τον πιο αποτελεσματικό τρόπο. Για να επιτευχθεί αυτό, ζωτικής σημασίας είναι η αξιοποίηση της τεχνολογίας, συμπεριλαμβανομένης της τεχνολογίας εξόρυξης δεδομένων. Μία από τις κεντρικές βάσεις για την επίτευξη πλεονεκτήματος, αποτελεί η οργανωσιακή ικανότητα να δημιουργεί νέα γνώση και να τη μεταφέρει σε διάφορα επίπεδα και τμήματα της επιχείρησης (ή του οργανισμού). Επειδή η γνώση είναι κεντρικής σημασίας για τη διαμόρφωση και για την εφαρμογή της στρατηγικής, η διαχείριση της γνώσης έχει γίνει

ένα βασικό στρατηγικό έργο που αντιμετωπίζουν οι διευθυντές για την επίτευξη επιτυχίας, στο σημερινό πολύπλοκο και δυναμικό περιβάλλον. Άρα, τα πιο πολύτιμα περιουσιακά στοιχεία των επιχειρήσεων του 21ου αιώνα είναι η γνώση και οι εργαζόμενοι στη γνώση.

3. Προβλήματα στην εξόρυξη δεδομένων

Η εξόρυξη δεδομένων, απαιτεί κατάλληλα εργαλεία για την αξιοποίηση των μαζικών στοιχείων της πληροφορίας, δηλαδή των δεδομένων. Με την πρώτη ματιά, αυτό μπορεί να φαίνεται παράδοξο, αλλά στην πραγματικότητα, τις περισσότερες φορές, σημαίνει ότι δε μπορούμε να λάβουμε σημαντικές πληροφορίες από μια τέτοια πληθώρα δεδομένων. Πρακτικά, η εξέταση των δεδομένων 2 χαρακτηριστικών 100 ατόμων, είναι πολύ διαφορετική από την εξέταση των αποτελεσμάτων 100 χαρακτηριστικών 1.000.000 ατόμων. Στην **1η περίπτωση**, τα απλά εργαλεία ανάλυσης δεδομένων μπορεί να οδηγήσουν σε σημαντικές πληροφορίες στο τέλος της διαδικασίας: συχνά ένα στοιχειώδες διάγραμμα διασποράς μπορεί να προσφέρει χρήσιμες ενδείξεις, αν και η επίσημη ανάλυση μπορεί να είναι πολύ πιο εξελιγμένη.

Στη **2η περίπτωση**, η εικόνα αλλάζει δραματικά, πολλά από τα απλά εργαλεία που χρησιμοποιήθηκαν στην προηγούμενη περίπτωση χάνουν την αποτελεσματικότητά τους. Για παράδειγμα, το διάγραμμα διασποράς 1.000.000 σημείων μπορεί να γίνει μια ενιαία άμορφη κηλίδα μελανιού και 100 χαρακτηριστικά μπορεί να παράγουν 100 x 99/2 από αυτές τις μορφές, οι οποίες είναι ταυτόχρονα πάρα πολλές και άχρηστες. Αυτό το απλό παράδειγμα, αναδεικνύει δύο πτυχές που περιπλέκουν την ανάλυση δεδομένων του αναφερόμενου τύπου. Η μία, αφορά το μέγεθος των δεδομένων, δηλαδή τον αριθμό των περιπτώσεων (ή των στατιστικών μονάδων) από τις οποίες αντλούνται πληροφορίες. Η άλλη, αφορά τις διαστάσεις των δεδομένων, δηλαδή τον αριθμό των χαρακτηριστικών (ή των μεταβλητών) των δεδομένων που συλλέγονται σε μια συγκεκριμένη μονάδα. Οι επιπτώσεις αυτών των συνιστωσών στην πολυπλοκότητα του προβλήματος είναι πολύ διαφορετικές μεταξύ τους, αλλά δεν είναι εντελώς ανεξάρτητες. Με απλοποίηση, που μπορεί να θεωρηθεί χονδροειδής αλλά βοηθά στην κατανόηση του προβλήματος, μπορούμε να πούμε ότι το μέγεθος επιφέρει αύξηση κυρίως στις υπολογιστικές πτυχές, ενώ το μέγεθος των διαστάσεων έχει μια σύνθετη επίδραση, που περιλαμβάνει μια υπολογιστική αύξηση παρόμοια με αυτή του μεγέθους και μια ταχεία αύξηση στην εννοιολογική πολυπλοκότητα των μοντέλων που χρησιμοποιούνται και κατά συνέπεια στην ερμηνεία και στη λειτουργική τους χρήση.

Δεν είναι δυνατόν όλα τα προβλήματα που ανακύπτουν σε αυτό το πλαίσιο να περιγραφούν μέσα σε ένα σχήμα, όπου ορίζονται εύκολα το μέγεθος και ακόμη δυσκολότερα, οι διαστάσεις των δεδομένων. Πρέπει επίσης να λάβουμε υπόψη την πιθανότητα τα δεδομένα να έχουν «άπειρο» μέγεθος, με την έννοια ότι μερικές φορές έχουμε μια συνεχή ροή δεδομένων. Η εξόρυξη δεδομένων αντιπροσωπεύει ακριβώς την εργασία της επεξεργασίας, μεγάλων ποσοτήτων (ή συνεχών ροών) δεδομένων, με στόχο την εξαγωγή χρήσιμων πληροφοριών σε όσους τις κατέχουν. Η έκφραση χρήσιμες πληροφορίες είναι σκόπιμα γενική: σε πολλές περιπτώσεις, το σημείο ενδιαφέροντος δεν προσδιορίζεται καθόλου εκ των προτέρων και συχνά το αναζητούμε εξορύσσοντας τα δεδομένα. Αυτή η πτυχή, διακρίνει την εξόρυξη δεδομένων από άλλες αναζητήσεις που σχετίζονται με την ανάλυση δεδομένων. Το τι μπορεί να αποτελεί χρήσιμη πληροφορία, ποικίλλει σημαντικά και εξαρτάται από το πλαίσιο στο οποίο λειτουργούμε και από τους στόχους που θέτουμε. Αυτή η παρατήρηση ισχύει και σε πολλά άλλα πλαίσια, αλλά στον τομέα της εξόρυξης δεδομένων έχει πρόσθετη αξία.

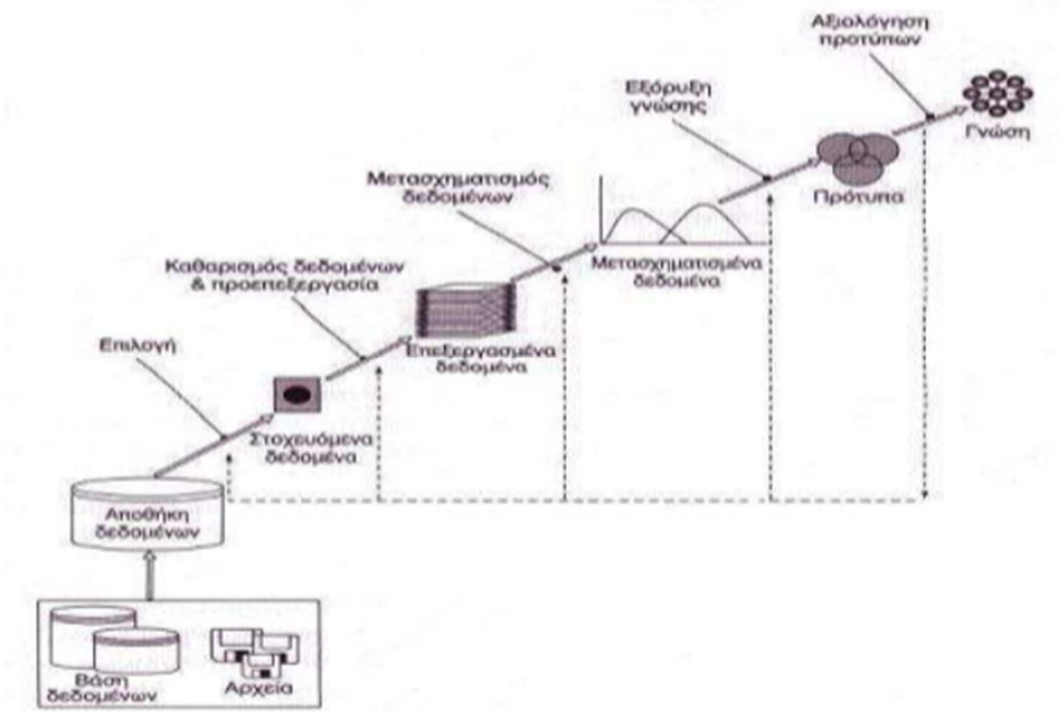
Έπειτα, το υπολογιστικό κόστος που συνδέεται με μεγάλα μεγέθη και με διαστάσεις δεδομένων, έχει προφανώς επιπτώσεις στη μέθοδο εργασίας με αυτά τα δεδομένα: καθώς αυξάνονται, οι μέθοδοι με υψηλό υπολογιστικό κόστος, καθίστανται λιγότερο εφικτές. Σαφώς, σε τέτοιες περιπτώσεις, δε μπορούμε να προσδιορίσουμε έναν ακριβή κανόνα, επειδή διάφοροι παράγοντες εκτός από αυτούς που έχουν ήδη αναφερθεί παίζουν ρόλο, όπως οι διαθέσιμοι πόροι για τον υπολογισμό και ο χρόνος που απαιτείται για τα αποτελέσματα. Ωστόσο, το αποτέλεσμα αναμφισβήτητα υπάρχει και εμποδίζει τη χρήση ορισμένων εργαλείων ή τουλάχιστον τα καθιστά λιγότερο πρακτικά, ενώ ευνοεί άλλα με χαμηλότερο υπολογιστικό κόστος. Υπάρχουν καταστάσεις στις οποίες αυτές οι πτυχές έχουν μόνο οριακή σημασία, επειδή η ποσότητα των δεδομένων δεν είναι αρκετή για να επηρεάσει το υπολογιστικό στοιχείο. Αυτό, οφείλεται εν μέρει στην τεράστια αύξηση της ισχύος των υπολογιστών.

Συχνά βλέπουμε αυτή την κατάσταση με ένα πρόβλημα μεγάλης κλίμακας, αν μπορεί να αναλυθεί σε υποπροβλήματα, τα οποία καθιστούν τμήματα των δεδομένων πιο διαχειρίσιμα. Οι πιο παραδοσιακές μέθοδοι παλαιότερης ηλικίας δεν έχουν ακόμη τεθεί σε παύση. Αντιθέτως, πολλά από αυτά, τα οποία αναπτύχθηκαν σε μια περίοδο περιορισμένων υπολογιστικών πόρων, είναι πολύ λιγότερο απαιτητικά από άποψη υπολογιστικής προσπάθειας και εξακολουθούν να ισχύουν αν εφαρμοστούν κατάλληλα. Έχει επανειλημμένα αναφερθεί η μεγάλη διαθεσιμότητα δεδομένων, τα οποία πλέον συλλέγονται με όλο και πιο συστηματικό και διεξοδικό τρόπο, ως σημείο εκκίνησης για την επεξεργασία. Ωστόσο, η μετατροπή των ακατέργαστων δεδομένων σε καθαρά δεδομένα είναι χρονοβόρα και μερικές φορές πολύ απαιτητική, με την προεπεξεργασία αυτών να παίζει πολύ σημαντικό ρόλο.

4. Διαδικασία της ανακάλυψης γνώσης από τις βάσεις δεδομένων–KDD

Η KDD είναι μία διαλογική και επαναληπτική διαδικασία, με είσοδο τα δεδομένα και με έξοδο τις χρήσιμες πληροφορίες, που αποτελείται από τα ακόλουθα στάδια:

- Ανάπτυξη και κατανόηση της περιοχής εφαρμογής και των στόχων του τελικού χρήστη.
- Επιλογή των δεδομένων, από πολλαπλές πηγές (βάσεις δεδομένων, αρχεία, μη ηλεκτρονικές πηγές) και καθορισμός του συνόλου στο οποίο θα εφαρμοσθεί η διαδικασία της εξόρυξης.
- Καθαρισμός (**Data cleaning**) και προεπεξεργασία των δεδομένων για την αφαίρεση outliers και θορύβου, για τη μέτρηση του θορύβου, για τη χάραξη στρατηγικών διαχείρισης των ελλειπόντων πεδίων στα δεδομένα, των διπλοτυπιών, των αντιφάσεων, για την απαλοιφή πλεονασμού. Ισχύει ότι για να έχουμε ποιοτικά αποτελέσματα από την εξόρυξη γνώσης, χρειαζόμαστε ποιοτικά δεδομένα. No quality data, no quality mining results.
- Ενοποίηση των δεδομένων (**Data integration**). Δημιουργία ενός συνόλου δεδομένων (μεταβλητές, δείγματα δεδομένων) επί των οποίων πρόκειται να εκτελεστεί η διαδικασία της εξόρυξης, με ενοποίηση των πολλαπλών βάσεων δεδομένων, των κύβων δεδομένων και των αρχείων.
- Μετασχηματισμός των δεδομένων (**Data transformation**) και διακριτοποίηση τους (**Data discretization**) σε μορφές κατάλληλες για την εφαρμογή της διαδικασίας εξόρυξης, με διαγραφή ή διόρθωση των εσφαλμένων δεδομένων και τη συγκέντρωση – αξιολόγηση των ελλιπών δεδομένων με χρήση μεθόδων μείωσης των διαστάσεων ή μετασχηματισμού, για μείωση των υπό εξέταση μεταβλητών ή εύρεση της κατάλληλης αντιπροσώπευσης των δεδομένων χωρίς μεταβλητές.



Γράφημα 1. Η βασική ροή των βημάτων της διαδικασίας KDD

• Μείωση των δεδομένων (**Data reduction**), διότι οι μεγάλες αποθήκες τους μπορεί να έχουν terabytes δεδομένων, συνεπώς η εξόρυξη γνώσης μπορεί να απαιτήσει χρόνο. Προκειμένου να βελτιωθεί η ποιότητα των αποτελεσμάτων, αφαιρούνται εκείνες οι ακραίες τιμές που εμφανίζονται σπάνια, μειώνεται η μεταβλητότητα των τιμών των δεδομένων και οι διαστάσεις του πλήθους των γνωρισμάτων. Συνηθίζεται να χρησιμοποιείται ως συνάρτηση μετασχηματισμού, ο λογάριθμος της τιμής αντί της ίδιας της τιμής. Στρατηγικές που ακολουθούνται:

- μείωση των διαστάσεων (dimension reduction),
- μείωση της πολυαριθμίας (numerosity reduction),
- data cube aggregation,
- instance selection,
- value discretization,
- συμπίεση των δεδομένων.

• Επιλογή των στόχων και των αλγορίθμων εξόρυξης δεδομένων.

• Εξόρυξη των δεδομένων, με εφαρμογές ευφυϊών μεθόδων (δέντρα, παλινδρόμηση, συσταδοποίηση–clustering, κανόνες κατηγοριοποίησης– classification rules), εστιάζοντας κυρίως στις μεθοδολογίες και στις τεχνικές εξαγωγής προτύπων δεδομένων ή στις περιγραφές δεδομένων από μεγάλες αποθήκες δεδομένων. Το ανωτέρω στάδιο, περιλαμβάνει την επιλογή της κωδικοποίησης προτύπων, της προεπεξεργασίας, της δειγματοληψίας και του μετασχηματισμού των δεδομένων πριν από το βήμα της εξόρυξης δεδομένων.

• Αξιολόγηση των προτύπων, προκειμένου αναγνωρισθούν τα ενδιαφέροντα.

• Σταθεροποίηση (ενσωμάτωση της εξορυγμένης γνώσης, στο σύστημα) και παρουσίαση των αποτελεσμάτων με χρήση τεχνικών οπτικοποίησης (visualization), σε δύο ή τρεις διαστάσεις, που μπορεί να είναι:

- Γραφικές (ραβδογράμματα, ιστογράμματα, πίττες)
- Γεωμετρικές (θηκογράμματα, διαγράμματα διασποράς)

➢ Ιεραρχικές (διαίρεση οθόνης – χώρου παρουσίασης σε περιοχές με κριτήριο τις τιμές των δεδομένων)

➢ Βασισμένες σε εικονίδια (με χρήση χρωμάτων και σχημάτων) ή εικονοστοιχεία (κάθε τιμή που αντιστοιχεί σε ένα δεδομένο, παρουσιάζεται ως ένα εικονοστοιχείο χρωματιστό με διαφορετικό χρώμα)

➢ Υβριδικές (συνδυασμός των ανωτέρω αναφερόμενων τεχνικών σε μία ενιαία παρουσίαση)

Η διαδικασία KDD είναι επαναληπτική και θα μπορούσε να περιέχει βρόχους μεταξύ οιονδήποτε εκ των ανωτέρω σταδίων.

5. Διαδικασία της εξόρυξης δεδομένων–Data Mining

Περιλαμβάνει τα μοντέλα συναρμολογήσεων των υπό μελέτη δεδομένων ή την εξαγωγή προτύπων από αυτά. Διατίθεται μεγάλο πλήθος αλγορίθμων εξόρυξης δεδομένων, αρκετοί εκ των οποίων χρησιμοποιούν τεχνικές από τομείς όπως η στατιστική, η μηχανική μάθηση, η αναγνώριση προτύπων, οι βάσεις δεδομένων. Το ουσιαστικότερο χαρακτηριστικό των αλγορίθμων εξόρυξης δεδομένων, που διαφοροποιεί τους περισσοτέρους εξ' αυτών από παρόμοιες τεχνικές, που ακολουθούνται στη στατιστική και στη μηχανική μάθηση, είναι ότι έχουν σχεδιασθεί με έμφαση στην επιλεξιμότητα εις ότι αφορά το μέγεθος του συνόλου των υπό εξέταση δεδομένων. Οι αλγόριθμοι εξόρυξης δεδομένων, προκύπτουν ως σύνθεση των ακόλουθων τριών παραγόντων:

• Της περιγραφής του μοντέλου ως προς:

➢ τη λειτουργία του. Καθορισμός των βασικών στόχων κατά τη διαδικασία της εξόρυξης δεδομένων (π.χ. classification ή clustering)

➢ την παραστατική του μορφή. Αν στα δεδομένα ταιριάζουν καλύτερα πολύπλοκα μοντέλα (δέντρα, νευρωνικά δίκτυα, γραφικά μοντέλα, κανόνες αποφάσεων, συγγενικά μοντέλα, δίκτυα Bayes), θα είναι δυσκολότερο να γίνουν κατανοητά και να ανταποκριθούν σε πραγματικές συνθήκες.

• Της αξιολόγησης του μοντέλου ως προς την ακρίβεια, τη χρησιμότητα τη δυνατότητα κατανόησης και την εγκυρότητα των προτύπων.

• Των αλγορίθμων αναζήτησης που είναι δυο τύπων:

➢ Αναζήτησης παραμέτρων (αναζητούν παραμέτρους που βελτιστοποιούν ένα κριτήριο αξιολόγησης για το μοντέλο).

➢ Αναζήτησης μοντέλων (εκτελούν μία επαναληπτική διαδικασία αναζήτησης για την αντιπροσώπευση των δεδομένων και αξιολογούνται τα αποτελέσματα).

Γράφημα 2. Οι «ρίζες» της εξόρυξης δεδομένων

6. Ιστορική εξέλιξη της εξόρυξης δεδομένων

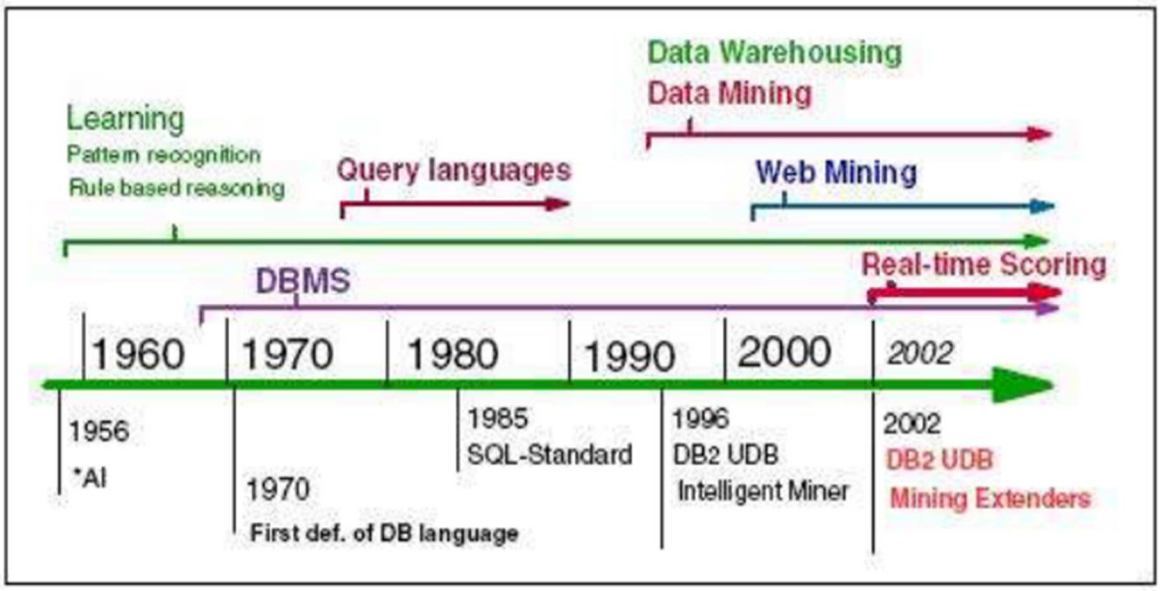
Στον πίνακα 1 που ακολουθεί, παρουσιάζεται εν συντομία η ιστορική εξέλιξη της εξόρυξής των δεδομένων, ενώ στον πίνακα 2 παρατίθεται χρονολογικά οι εξελίξεις στις περιοχές της τεχνητής νοημοσύνης, της ανάκτησης πληροφοριών, των βάσεων δεδομένων και της στατιστικής που τώρα κυριαρχούν στην εξόρυξη γνώσης από δεδομένα. Επιπλέον, στο γράφημα 3 παρουσιάζεται ένα χρονοδιάγραμμα της εξόρυξης των δεδομένων.

Αλγόριθμοι	**Βάσεις δεδομένων**	**Ανάκτηση πληροφοριών**	**Στατιστική**	**Μηχανική μάθηση**
⇩	⇩	⇩	⇩	⇩
Εξόρυξη γνώσης				

Πίνακας 1. Ιστορική άποψη της εξόρυξης γνώσης

Χρόνος	**Περιοχή**	**Συνεισφορά**
Τέλη 1700	Στατιστική	Θεώρημα πιθανοτήτων του Bayes
Αρχές 1900	Τεχνητή νοημοσύνη	Ανάλυση με παλινδρόμηση
Αρχές 1920	Στατιστική	Εκτίμηση μέγιστης πιθανοφάνειας
Αρχές 1940	Τεχνητή νοημοσύνη	Νευρωνικά δίκτυα
Αρχές 1950	Στατιστική	Πλησιέστερος γείτονας, απλός σύνδεσμος
Τέλη 1950	Τεχνητή νοημοσύνη	Perception
	Στατιστική	Επαναδειγματοληψία, μείωση μεροληψίας, εκτιμήτρια Jackknife
Αρχές 1960	Τεχνητή νοημοσύνη	Έναρξη μηχανικής μάθησης
	Βάσεις δεδομένων	Μαζικές αναφορές
Μέσα 1960	Στατιστική	Γραμμικά μοντέλα κατηγοριοποίησης, εξερευνητική ανάλυση δεδομένων (EDA)
	Ανάκτηση πληροφοριών	Μέτρα ομοιότητας, συσταδοποίηση
Τέλη 1960	Βάσεις δεδομένων	Σχεσιακό μοντέλο δεδομένων
Αρχές 1970	Ανάκτηση πληροφοριών	Έξυπνα συστήματα ανάκτησης πληροφοριών
Μέσα 1970	Τεχνητή νοημοσύνη	Γενετικοί αλγόριθμοι
Τέλη 1970	Στατιστική	Συσταδοποίηση K–means, Εκτίμηση με μη πλήρη δεδομένα (EM αλγόριθμος)
Αρχές 1980	Τεχνητή νοημοσύνη	Αυτό–οργανωμένα δίκτυα Kohonen
Μέσα 1980	Τεχνητή νοημοσύνη	Αλγόριθμοι δέντρων αποφάσεων
Αρχές 1990	Βάσεις δεδομένων	Αλγόριθμοι κανόνων συσχέτισης, παγκόσμιος ιστός, μηχανές αναζήτησης
1990	Βάσεις δεδομένων	Αποθήκες δεδομένων, άμεση αναλυτική επεξεργασία (OLAP)

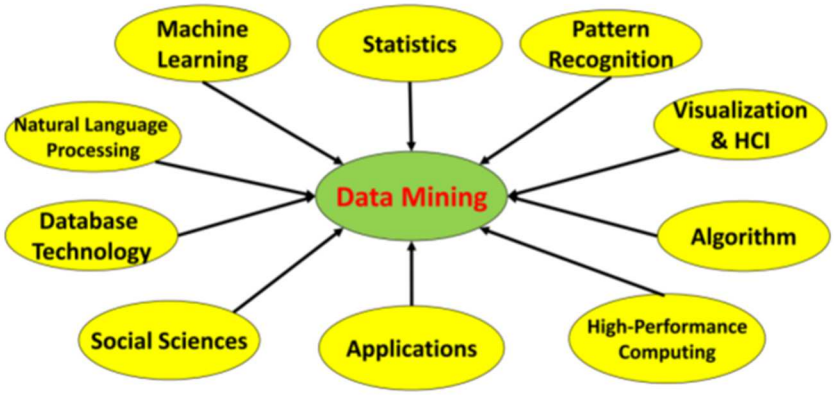
Πίνακας 2. Ιστορική εξέλιξη της εξόρυξης δεδομένων



Γράφημα 3. Χρονοδιάγραμμα της εξόρυξης δεδομένων

7. Η συμβολή της επιστήμης στην εξόρυξη δεδομένων

Η εξόρυξη δεδομένων, αποτελεί ένα σημείο σύγκλισης πολλών επιστημονικών πεδίων, όπως η μηχανική μάθηση, η στατιστική, η αναγνώριση προτύπων, η επεξεργασία της φυσικής γλώσσας, η τεχνολογία των βάσεων δεδομένων, η οπτικοποίηση και η αλληλεπίδραση ανθρώπου–υπολογιστή, οι αλγόριθμοι, η υπολογιστική υψηλής απόδοσης, οι κοινωνικές επιστήμες (γράφημα 4). Η διεπιστημονική φύση της έρευνας και της ανάπτυξης της εξόρυξης δεδομένων, συμβάλλει σημαντικά στην επιτυχία της και στις εκτεταμένες εφαρμογές της.



Γράφημα 4. Εξόρυξη δεδομένων: συμβολή πολλαπλών επιστημονικών κλάδων

7.1 Στατιστική και εξόρυξη δεδομένων

Η στατιστική, μελετά τη συλλογή, την ανάλυση, την ερμηνεία και την παρουσίαση των δεδομένων. Η εξόρυξη δεδομένων, έχει μια εγγενή σύνδεση με τη στατιστική. Ένα στατιστικό μοντέλο, είναι ένα σύνολο μαθηματικών συναρτήσεων που περιγράφουν τη συμπεριφορά των αντικειμένων σε μια κλάση–στόχο με βάση τυχαίες μεταβλητές και τις σχετικές κατανομές πιθανοτήτων. Τα στατιστικά μοντέλα, χρησιμοποιούνται ευρέως για τη μοντελοποίηση των δεδομένων και των κλάσεων δεδομένων. Π.χ. σε εργασίες εξόρυξης δεδομένων, όπως ο χαρακτηρισμός και η ταξινόμηση των δεδομένων, μπορούν να κατασκευαστούν στατιστικά μοντέλα

κλάσεων–στόχων. Με άλλα λόγια, τέτοια στατιστικά μοντέλα μπορούν να είναι το αποτέλεσμα μιας εργασίας εξόρυξης δεδομένων. Εναλλακτικά, οι εργασίες εξόρυξης δεδομένων μπορούν να κατασκευαστούν πάνω σε στατιστικά μοντέλα. Π.χ. μπορούμε να χρησιμοποιήσουμε στατιστικά στοιχεία για να μοντελοποιήσουμε τον θόρυβο και τις τιμές των ελλειπουσών δεδομένων. Στη συνέχεια, κατά την εξόρυξη μοτίβων σε ένα μεγάλο σύνολο δεδομένων, η διαδικασία εξόρυξης δεδομένων μπορεί να χρησιμοποιήσει το μοντέλο για να βοηθήσει στον εντοπισμό και στη διαχείριση των θορυβωδών ή ελλειπουσών τιμών στα δεδομένα.

Η στατιστική έρευνα, αναπτύσσει εργαλεία για την πρόβλεψη και για τη χρήση δεδομένων και στατιστικών μοντέλων. Στατιστικές μέθοδοι μπορούν να χρησιμοποιηθούν για τη σύνοψη ή για την περιγραφή μιας συλλογής δεδομένων. Η στατιστική, είναι χρήσιμη για την εξόρυξη διαφόρων μοτίβων από δεδομένα και για την κατανόηση των υποκείμενων μηχανισμών που δημιουργούν και επηρεάζουν τα μοτίβα. Η **επαγωγική στατιστική** (ή προγνωστική στατιστική), μοντελοποιεί τα δεδομένα με τρόπο που λαμβάνει υπόψη την τυχαιότητα και την αβεβαιότητα στις παρατηρήσεις και χρησιμοποιείται για την εξαγωγή συμπερασμάτων, σχετικά με τη διαδικασία ή τον πληθυσμό που ερευνάται. Οι στατιστικές μέθοδοι, μπορούν να χρησιμοποιηθούν για την επαλήθευση των αποτελεσμάτων της εξόρυξης δεδομένων.

Π.χ. μετά από την εξόρυξη ενός μοντέλου ταξινόμησης ή πρόβλεψης, το μοντέλο πρέπει να επαληθευτεί με στατιστικό έλεγχο υποθέσεων. Ένας έλεγχος στατιστικής υπόθεσης (μερικές φορές ονομάζεται επιβεβαιωτική ανάλυση δεδομένων), λαμβάνει στατιστικές αποφάσεις, χρησιμοποιώντας πειραματικά δεδομένα. Ένα αποτέλεσμα, ονομάζεται στατιστικά σημαντικό αν είναι απίθανο να έχει προκύψει τυχαία. Αν το μοντέλο ταξινόμησης ή πρόβλεψης ισχύει, τότε η περιγραφική στατιστική του μοντέλου αυξάνει την αξιοπιστία του. Η εφαρμογή στατιστικών μεθόδων στην εξόρυξη δεδομένων, δεν είναι καθόλου ασήμαντη. Συχνά, μια σοβαρή πρόκληση είναι ο τρόπος κλιμάκωσης μιας στατιστικής μεθόδου, σε ένα μεγάλο σύνολο δεδομένων. Πολλές στατιστικές μέθοδοι, έχουν υψηλή πολυπλοκότητα στον υπολογισμό. Όταν τέτοιες μέθοδοι εφαρμόζονται σε μεγάλα σύνολα δεδομένων που είναι επίσης κατανεμημένα σε πολλαπλές λογικές ή φυσικές τοποθεσίες, οι αλγόριθμοι πρέπει να σχεδιάζονται και να ρυθμίζονται προσεκτικά για να μειώνουν το υπολογιστικό κόστος. Αυτή η πρόκληση γίνεται ακόμη πιο δύσκολη για τις διαδικτυακές εφαρμογές, όπως οι προτάσεις διαδικτυακών ερωτημάτων στις μηχανές αναζήτησης, όπου η εξόρυξη δεδομένων απαιτείται για τη συνεχή διαχείριση των γρήγορων ροών δεδομένων, σε πραγματικό χρόνο. Η έρευνα για την εξόρυξη δεδομένων, έχει αναπτύξει πολλές επεκτάσιμες και αποτελεσματικές λύσεις για την ανάλυση των μαζικών συνόλων δεδομένων και των ροών δεδομένων. Επιπλέον, διαφορετικά είδη συνόλων δεδομένων και διαφορετικές εφαρμογές μπορεί να απαιτούν διαφορετικές μεθόδους ανάλυσης. Έχουν προταθεί και δοκιμαστεί αποτελεσματικές λύσεις, οι οποίες οδηγούν σε πολλές νέες, επεκτάσιμες μεθόδους στατιστικής ανάλυσης που βασίζονται στην εξόρυξη δεδομένων.

7.2 Μηχανική μάθηση και εξόρυξη δεδομένων

Η μηχανική μάθηση, διερευνά το πώς οι υπολογιστές μαθαίνουν (ή το πώς βελτιώνουν την απόδοσή τους) με βάση δεδομένα. Η μηχανική μάθηση, είναι ένας ταχέως αναπτυσσόμενος κλάδος, με πολλές νέες μεθοδολογίες και εφαρμογές που έχουν αναπτυχθεί τα τελευταία έτη. Γενικά, η μηχανική μάθηση ασχολείται με δύο κλασικά προβλήματα: την εποπτευόμενη μάθηση και τη μη εποπτευόμενη μάθηση.

7.2.1 Εποπτευόμενη μάθηση

Ένα κλασικό παράδειγμα εποπτευόμενης μάθησης είναι η ταξινόμηση. Η εποπτεία στη μάθηση, προέρχεται από τα παραδείγματα με ετικέτες στο σύνολο των δεδομένων εκπαίδευσης. Π.χ. για την αυτόματη αναγνώριση των χειρόγραφων ταχυδρομικών κωδίκων σε επιστολές του ταχυδρομείου, το σύστημα μάθησης λαμβάνει ένα σύνολο χειρόγραφων εικόνων των ταχυδρομικών κωδίκων και τις αντίστοιχες μηχανικά αναγνώσιμες μεταφράσεις τους ως παραδείγματα εκπαίδευσης και μαθαίνει (δηλαδή, υπολογίζει) ένα μοντέλο ταξινόμησης.

7.2.2 Μη εποπτευόμενη μάθηση

Ένα κλασικό παράδειγμα μη εποπτευόμενης μάθησης είναι η ομαδοποίηση. Η διαδικασία μάθησης είναι μη εποπτευόμενη, καθώς τα παραδείγματα εισόδου δε φέρουν ετικέτες κλάσης. Συνήθως, μπορούμε να χρησιμοποιήσουμε ομαδοποίηση για να ανακαλύψουμε ομάδες μέσα στα δεδομένα. Π.χ. μια μέθοδος μάθησης χωρίς επίβλεψη μπορεί να λάβει ως είσοδο, ένα σύνολο εικόνων χειρόγραφων ψηφίων.

Έστω ότι βρίσκει 10 συστάδες δεδομένων. Αυτές οι συστάδες μπορεί να αντιστοιχούν στα 10 διακριτά ψηφία από το 0 ως το 9, αντίστοιχα. Ωστόσο, επειδή τα δεδομένα εκπαίδευσης δε φέρουν ετικέτες, το μοντέλο μάθησης δε μπορεί να μας πει τη σημασιολογική σημασία των συστάδων που βρέθηκαν. Όσον αφορά αυτά τα δύο βασικά προβλήματα, η εξόρυξη δεδομένων και η μηχανική μάθηση μοιράζονται πολλές ομοιότητες.

7.2.3 Διαφορές της μηχανικής μάθησης και της εξόρυξης δεδομένων

Η εξόρυξη δεδομένων, διαφέρει από τη μηχανική μάθηση σε αρκετές σημαντικές πτυχές.

Πρώτον, ακόμη και σε παρόμοιες εργασίες όπως είναι η ταξινόμηση και η ομαδοποίηση, η εξόρυξη δεδομένων συχνά λειτουργεί σε πολύ μεγάλα σύνολα δεδομένων και σε άπειρες ροές δεδομένων. Η επεκτασιμότητα, μπορεί να αποτελέσει σημαντικό μέλημα και πολλοί αποτελεσματικοί και εξαιρετικά επεκτάσιμοι αλγόριθμοι εξόρυξης δεδομένων ή αλγόριθμοι εξόρυξης ροών πρέπει να αναπτυχθούν για την ολοκλήρωση τέτοιων εργασιών.

Δεύτερον, σε πολλά προβλήματα εξόρυξης δεδομένων, τα σύνολα δεδομένων είναι συνήθως μεγάλα, αλλά τα δεδομένα εκπαίδευσης μπορεί να είναι μάλλον μικρά, καθώς είναι ακριβό για τους ειδικούς να παρέχουν ετικέτες ποιότητας για πολλά παραδείγματα. Επομένως, η εξόρυξη δεδομένων πρέπει να καταβάλει μεγάλη προσπάθεια στην ανάπτυξη μεθόδων με ασθενή επίβλεψη. Αυτές, περιλαμβάνουν μεθοδολογίες όπως η:

- ημιεποπτευόμενη μάθηση με ένα μικρό σύνολο δεδομένων με ετικέτες, ένα μεγάλο σύνολο δεδομένων χωρίς ετικέτες,
- ενσωμάτωση ή το σύνολο πολλαπλών αδύναμων μοντέλων που λαμβάνονται από μη ειδικούς (π.χ. αυτά που λαμβάνονται μέσω crowdsourcing),
- εξ αποστάσεως εποπτεία, όπως η χρήση ευρέως διαθέσιμων και γενικών (αλλά εξ αποστάσεως σχετικών με το προς επίλυση πρόβλημα) βάσεων γνώσης (π.χ. Wikipedia, DBPedia),
- ενεργητική μάθηση, επιλέγοντας προσεκτικά παραδείγματα για να ερωτηθούν ειδικοί
- μεταφορά μάθησης, ενσωματώνοντας μοντέλα που έχουν αποκτηθεί από παρόμοιους τομείς προβλημάτων.

Η εξόρυξη δεδομένων, έχει επεκτείνει τέτοιες ασθενώς εποπτευόμενες μεθόδους για την κατασκευή μοντέλων ποιοτικής ταξινόμησης σε μεγάλα σύνολα

δεδομένων, με ένα πολύ περιορισμένο σύνολο δεδομένων εκπαίδευσης υψηλής ποιότητας.

Τρίτον, οι μέθοδοι μηχανικής μάθησης ενδέχεται να μην είναι σε θέση να χειριστούν πολλά είδη προβλημάτων ανακάλυψης γνώσης σε μεγάλα δεδομένα. Από την άλλη πλευρά, η εξόρυξη δεδομένων, η ανάπτυξη αποτελεσματικών λύσεων για συγκεκριμένα προβλήματα εφαρμογών, εμβαθύνει στον τομέα του προβλήματος και επεκτείνεται πολύ πέρα από το πεδίο εφαρμογής που καλύπτεται από τη μηχανική μάθηση. Π.χ. πολλά προβλήματα εφαρμογών, όπως η ανάλυση δεδομένων των επιχειρηματικών συναλλαγών, η ανάλυση της ακολουθίας εκτέλεσης των προγραμμάτων λογισμικού και η χημική και βιολογική δομική ανάλυση, χρειάζονται αποτελεσματικές μεθόδους για την εξόρυξη συχνών μοτίβων, διαδοχικών μοτίβων και δομημένων μοτίβων. Η έρευνα για την εξόρυξη δεδομένων, έχει δημιουργήσει πολλές κλιμακούμενες, αποτελεσματικές και ποικίλες μεθόδους εξόρυξης, για τέτοιες εργασίες. Π.χ. η ανάλυση κοινωνικών και πληροφοριακών δικτύων μεγάλης κλίμακας, θέτει πολλά απαιτητικά προβλήματα που μπορεί να μην ταιριάζουν στο τυπικό πεδίο εφαρμογής πολλών μεθόδων μηχανικής μάθησης, λόγω της αλληλεπίδρασης των πληροφοριών μεταξύ των συνδέσμων και των κόμβων σε τέτοια δίκτυα.

Η εξόρυξη δεδομένων, έχει αναπτύξει πολλές ενδιαφέρουσες λύσεις σε τέτοια προβλήματα. Από αυτή την άποψη, η εξόρυξη δεδομένων και η μηχανική μάθηση είναι δύο διαφορετικοί, αλλά στενά συνδεδεμένοι κλάδοι. Η εξόρυξη δεδομένων, εμβαθύνει σε συγκεκριμένους, απαιτητικούς από δεδομένα τομείς εφαρμογών, δεν περιορίζεται σε μία μόνο μεθοδολογία επίλυσης προβλημάτων και αναπτύσσει συγκεκριμένες (μερικές φορές μάλλον καινοτόμες), αποτελεσματικές και κλιμακούμενες λύσεις για πολλά απαιτητικά προβλήματα εφαρμογών. Είναι ένας νέος, ευρύς και πολλά υποσχόμενος ερευνητικός κλάδος για ερευνητές και επαγγελματίες, να μελετήσουν και να εργαστούν πάνω σε αυτόν.

7.3 Βάσεις δεδομένων και εξόρυξη δεδομένων

Η έρευνα των συστημάτων βάσεων δεδομένων, επικεντρώνεται στη δημιουργία, στη συντήρηση και στη χρήση των βάσεων δεδομένων για οργανισμούς και τελικούς χρήστες. Συγκεκριμένα, οι ερευνητές των συστημάτων βάσεων δεδομένων έχουν καθιερώσει καλά αναγνωρισμένες αρχές σε μοντέλα δεδομένων, γλώσσες ερωτημάτων, επεξεργασία και βελτιστοποίηση ερωτημάτων, αποθήκευση δεδομένων και μεθόδους ευρετηρίασης. Η τεχνολογία των βάσεων δεδομένων, είναι γνωστή για την επεκτασιμότητά της στην επεξεργασία πολύ μεγάλων, σχετικά δομημένων συνόλων δεδομένων. Πολλές εργασίες εξόρυξης δεδομένων, πρέπει να χειρίζονται μεγάλα σύνολα δεδομένων ή και δεδομένα σε πραγματικό χρόνο, γρήγορης ροής. Η εξόρυξη δεδομένων, μπορεί να αξιοποιήσει αποτελεσματικά τις επεκτάσιμες τεχνολογίες των βάσεων δεδομένων για να επιτύχει υψηλή απόδοση και επεκτασιμότητα, σε μεγάλα σύνολα δεδομένων. Επιπλέον, οι εργασίες εξόρυξης δεδομένων μπορούν να χρησιμοποιηθούν για την επέκταση της ικανότητας των υπαρχόντων συστημάτων βάσεων δεδομένων, ώστε να ικανοποιήσουν τις εξελιγμένες απαιτήσεις ανάλυσης των δεδομένων των χρηστών.

Πρόσφατα συστήματα βάσεων δεδομένων, έχουν αναπτύξει συστηματικές δυνατότητες ανάλυσης δεδομένων πάνω στα δεδομένα των βάσεων, χρησιμοποιώντας υποδομές αποθηκών δεδομένων (data warehouse) και εξόρυξης δεδομένων.

7.4 Εξόρυξη δεδομένων και επιστήμη δεδομένων

Με την τεράστια ποσότητα δεδομένων σε σχεδόν κάθε κλάδο και σε διάφορα είδη εφαρμογών, τα μεγάλα δεδομένα και η επιστήμη των δεδομένων, έχουν γίνει

λέξεις κλειδιά τα τελευταία έτη. Τα μεγάλα δεδομένα, αναφέρονται γενικά σε τεράστιες ποσότητες δομημένων και αδόμητων δεδομένων διαφόρων μορφών και η επιστήμη των δεδομένων είναι ένας διεπιστημονικός τομέας που χρησιμοποιεί επιστημονικές μεθόδους, διαδικασίες, αλγόριθμους και συστήματα για την εξαγωγή γνώσεων και πληροφοριών, από τεράστια δεδομένα διαφόρων μορφών. Σαφώς, η εξόρυξη δεδομένων παίζει ουσιαστικό ρόλο στην επιστήμη των δεδομένων. Για τους περισσότερους, η επιστήμη των δεδομένων είναι μια έννοια που ενοποιεί τη στατιστική, τη μηχανική μάθηση, την εξόρυξη δεδομένων και τις σχετικές μεθόδους τους, προκειμένου να κατανοήσουν και να αναλύσουν τεράστια δεδομένα.

Χρησιμοποιεί τεχνικές και θεωρίες που προέρχονται από πολλούς τομείς στο πλαίσιο των μαθηματικών, της στατιστικής, της επιστήμης των πληροφοριών και της επιστήμης των υπολογιστών. Για πολλούς ειδικούς του κλάδου, ο όρος επιστήμη των δεδομένων, αναφέρεται συχνά στην επιχειρηματική ανάλυση, στην επιχειρηματική ευφυΐα, στην προγνωστική μοντελοποίηση ή σε οποιαδήποτε ουσιαστική χρήση των δεδομένων και θεωρείται ως ένας λαμπρός όρος για την αναδιαμόρφωση της στατιστικής, της εξόρυξης δεδομένων, της μηχανικής μάθησης ή οποιουδήποτε είδους ανάλυσης δεδομένων. Μέχρι στιγμής, δεν υπάρχει συναίνεση σχετικά με έναν ορισμό ή κατάλληλο περιεχόμενο στα προγράμματα σπουδών της επιστήμης των δεδομένων σε πολλά πανεπιστήμια. Παρόλα αυτά, τα περισσότερα πανεπιστήμια λαμβάνουν τις βασικές γνώσεις που παράγονται στη στατιστική, στη μηχανική μάθηση, στην εξόρυξη δεδομένων, στις βάσεις δεδομένων και στην αλληλεπίδραση ανθρώπου – υπολογιστή ως το βασικό πρόγραμμα σπουδών για την εκπαίδευση στην επιστήμη των δεδομένων.

Δεν είναι περίεργο που η επιστήμη των δεδομένων, τα μεγάλα δεδομένα και η εξόρυξη δεδομένων είναι στενά συνδεδεμένες και αντιπροσωπεύουν μια αναπόφευκτη τάση στις εξελίξεις της επιστήμης και της τεχνολογίας.

7.5 Εξόρυξη δεδομένων και άλλοι κλάδοι των επιστημών

Εκτός από τη στατιστική, τη μηχανική μάθηση και την τεχνολογία των βάσεων δεδομένων, η εξόρυξη δεδομένων έχει στενές σχέσεις με πολλούς άλλους κλάδους.

Η πλειονότητα των δεδομένων του πραγματικού κόσμου είναι μη δομημένα, με τη μορφή κειμένου φυσικής γλώσσας, εικόνων ή δεδομένων ήχου – βίντεο. Άρα, η επεξεργασία της φυσικής γλώσσας, η υπολογιστική όραση, η αναγνώριση προτύπων, η επεξεργασία σήματος ήχου–βίντεο και η ανάκτηση πληροφοριών, θα προσφέρουν κρίσιμη βοήθεια στον χειρισμό τέτοιων δεδομένων. Στην πραγματικότητα, ο χειρισμός οποιωνδήποτε ειδών δεδομένων απαιτεί πολλή γνώση του τομέα για να ενσωματωθεί στον σχεδιασμό των αλγορίθμων εξόρυξης δεδομένων. Π.χ. η εξόρυξη βιοϊατρικών δεδομένων θα χρειαστεί την ενσωμάτωση γνώσεων από τις βιολογικές επιστήμες, τις ιατρικές επιστήμες και τη βιοπληροφορική. Η εξόρυξη γεωχωρικών δεδομένων θα χρειαστεί πολλή γνώση και τεχνικές από τη γεωγραφία και από τις επιστήμες των γεωχωρικών δεδομένων. Η εξόρυξη σφαλμάτων λογισμικού σε μεγάλα προγράμματα λογισμικού, θα χρειαστεί να ενσωματώσει τη μηχανική λογισμικού με την εξόρυξη δεδομένων. Η εξόρυξη σε μέσα κοινωνικής δικτύωσης και κοινωνικών δικτύων, θα χρειαστεί γνώσεις και δεξιότητες από τις κοινωνικές επιστήμες και από τις επιστήμες δικτύων. Τέτοια παραδείγματα μπορούν να συνεχιστούν επ' αόριστον, καθώς η εξόρυξη δεδομένων θα διεισδύσει σχεδόν σε κάθε τομέα. Μία σημαντική πρόκληση στην εξόρυξη δεδομένων είναι η αποτελεσματικότητα και η επεκτασιμότητα, καθώς συχνά πρέπει να χειριζόμαστε τεράστιες ποσότητες δεδομένων με κρίσιμους χρονικούς περιορισμούς. Η εξόρυξη δεδομένων, συνδέεται κρίσιμα με τον αποτελεσματικό σχεδιασμό αλγορίθμων, όπως αλγόριθμους χαμηλής πολυπλοκότητας, σταδιακούς και συνεχούς εξόρυξης δεδομένων. Συχνά, χρειάζεται να διερευνήσει υπολογισμούς

υψηλής απόδοσης, παράλληλους υπολογισμούς και κατανεμημένους υπολογισμούς, με προηγμένο υλικό και περιβάλλον Cloud computing ή cluster computing. Η εξόρυξη δεδομένων, είναι επίσης στενά συνδεδεμένη με την αλληλεπίδραση ανθρώπου–υπολογιστή. Οι χρήστες, πρέπει να αλληλοεπιδρούν με ένα σύστημα ή μια διαδικασία εξόρυξης δεδομένων με αποτελεσματικό τρόπο, λέγοντας στο σύστημα τι να εξορύξει, πώς να ενσωματώσει τις βασικές γνώσεις, πώς να εξορύξει και πώς να παρουσιάσει τα αποτελέσματα της εξόρυξης με έναν εύκολο στην κατανόηση (π.χ. μέσω ερμηνείας και οπτικοποίησης) και εύκολο στην αλληλεπίδραση τρόπο (π.χ. με φιλικό γραφικό περιβάλλον χρήστη και διαδραστική εξόρυξη).

Στην πραγματικότητα, στις μέρες μας, δεν υπάρχουν μόνο πολλά διαδραστικά συστήματα εξόρυξης δεδομένων, αλλά και πολλές λειτουργίες εξόρυξης δεδομένων κρυμμένες σε διάφορα είδη προγραμμάτων εφαρμογών. Είναι μη ρεαλιστικό να περιμένουμε από όλους στην κοινωνία να κατανοήσουν και να κατακτήσουν τις τεχνικές εξόρυξης δεδομένων. Απαγορεύεται στις βιομηχανίες να εκθέτουν τα μεγάλα σύνολα δεδομένων τους. Πολλά συστήματα, έχουν ενσωματωμένες λειτουργίες εξόρυξης δεδομένων, έτσι ώστε οι άνθρωποι να μπορούν να εκτελούν εξόρυξη δεδομένων ή να χρησιμοποιούν αποτελέσματα της εξόρυξης δεδομένων απλώς κάνοντας κλικ με το ποντίκι. Π.χ. οι έξυπνες μηχανές αναζήτησης και τα ηλεκτρονικά καταστήματα λιανικής πώλησης, εκτελούν τέτοια αόρατη εξόρυξη δεδομένων συλλέγοντας τα δεδομένα τους και το ιστορικό αναζήτησης ή αγορών των χρηστών, ενσωματώνοντας την εξόρυξη δεδομένων στα στοιχεία τους, για να βελτιώσουν την απόδοση, τη λειτουργικότητα και την ικανοποίηση των χρηστών.

8. Σχεδιασμός και υλοποίηση ενός έργου εξόρυξης δεδομένων

Υπάρχουν διαφορετικές προσεγγίσεις σε ότι αφορά τον σχεδιασμό και την υλοποίηση ενός έργου εξόρυξης δεδομένων (data mining project). Μία προσέγγιση, αποτελεί μια προσαρμογή της γνωστής διαδικασίας ανάπτυξης λογισμικού και είναι πιθανό να περιλαμβάνει τα ακόλουθα έξι βήματα:

- Ανάλυση των απαιτήσεων
- Επιλογή και συλλογή των δεδομένων
- Καθαρισμός και προετοιμασία των δεδομένων
- Εξερεύνηση και επικύρωση της εξόρυξης των δεδομένων
- Υλοποίηση, αξιολόγηση και παρακολούθηση
- Οπτικοποίηση των αποτελεσμάτων

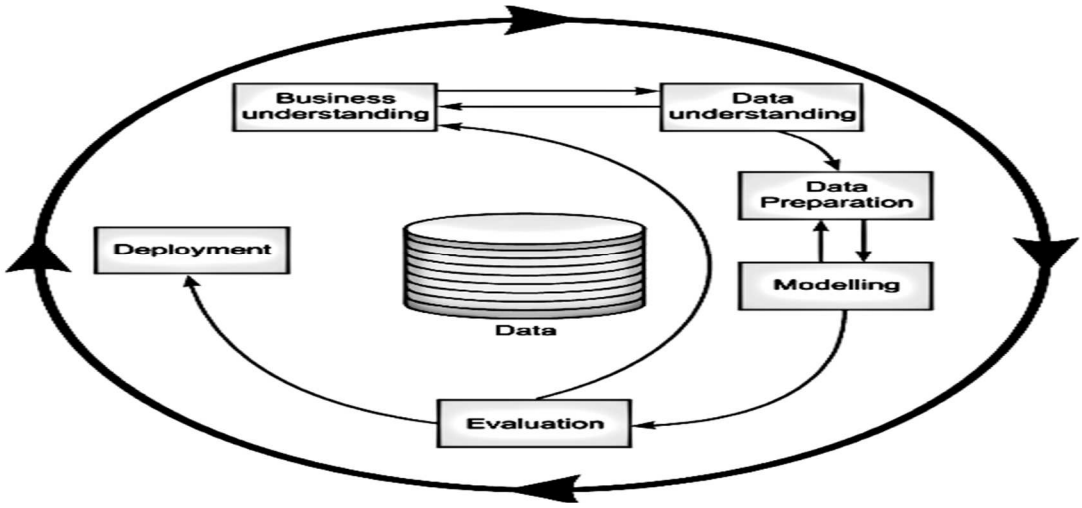
Πέρα από την παραπάνω προσέγγιση, έχουν προταθεί και άλλες, συχνά παρόμοιες μεθοδολογίες όπως π.χ. το SPSS (Statistical Package for the Social Sciences) που συνέστησε την προσέγγιση 5A η οποία αποτελείται από 5 βήματα:

- Αξιολόγηση (assess),
- Πρόσβαση (access),
- Ανάλυση (analyze),
- Δράση (action) και
- Αυτοματοποίηση (automate),

ή το SAS (Statistical Analysis System) που χρησιμοποιεί μια διαφορετική προσέγγιση 5 βημάτων η οποία αποτελείται πάλι από 5 βήματα:

- Δειγματοληψία (sample),
- Εξερεύνηση (explore),
- Τροποποίηση (modify),
- Μοντελοποίηση (model),
- Αξιολόγηση (assess)

με τις δύο αυτές προσεγγίσεις να μη φαίνεται να δίνουν έμφαση στον καθαρισμό και στην προετοιμασία των δεδομένων, αν και είναι πολύ πιθανό αυτά τα βήματα να περιλαμβάνονται στα προτεινόμενα βήματα. Μια άλλη προσέγγιση είναι η CRISP–DM (Cross–Industry Standard Process for Data Mining) (γράφημα 5), η οποία προτάθηκε ως μια επιπλέον προσέγγιση το 1998 από μια κοινοπραξία προμηθευτών και χρηστών (συμπεριλαμβανομένων των Daimler Chrysler, SPSS και NCR) και θεωρείται πιο πρακτική, επιτυχημένη και ευρέως υιοθετημένη.



Γράφημα 5. Το μοντέλο εξόρυξης δεδομένων CRISP

9. Η προσέγγιση CRISP–DM

Η προσέγγιση CRISP–DM, αποτελείται από 6 βήματα:

- Επιχειρηματική κατανόηση (business understanding),
- Κατανόηση των δεδομένων (data understanding),
- Προετοιμασία των δεδομένων (data preparation),
- Μοντελοποίηση (modeling),
- Αξιολόγηση (evaluation),
- Ανάπτυξη (deployment),

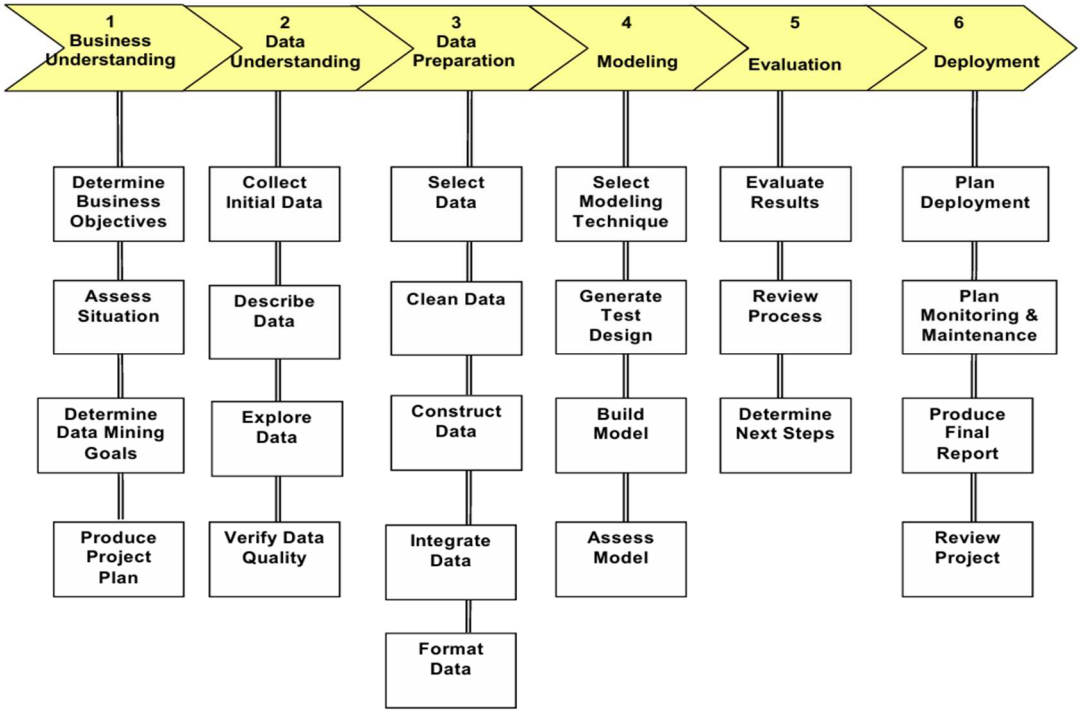
τα οποία παρουσιάζονται στο γράφημα 6 και θα αναλυθούν στη συνέχεια.

9.1 Επιχειρηματική κατανόηση

Αυτή η αρχική φάση, επικεντρώνεται στην κατανόηση των στόχων και των απαιτήσεων του έργου από επιχειρηματική οπτική γωνία και στη συνέχεια στη μετατροπή αυτής της γνώσης σε έναν ορισμό προβλήματος εξόρυξης δεδομένων και σε ένα προκαταρκτικό σχέδιο, που έχει σχεδιαστεί για την επίτευξη των στόχων.

9.1.1 Καθορισμός των επιχειρηματικών στόχων

Ο 1ος στόχος του αναλυτή δεδομένων είναι να κατανοήσει πλήρως, από επιχειρηματική οπτική γωνία, τι πραγματικά θέλει να πετύχει ο πελάτης. Συχνά, ο πελάτης έχει πολλούς ανταγωνιστικούς στόχους και περιορισμούς που πρέπει να εξισορροπηθούν σωστά. Στόχος του αναλυτή είναι να αποκαλύψει σημαντικούς παράγοντες, στην αρχή, που μπορούν να επηρεάσουν το αποτέλεσμα του έργου. Μια πιθανή συνέπεια της παραμέλησης αυτού του βήματος, είναι η καταβολή μεγάλης προσπάθειας για την παραγωγή των σωστών απαντήσεων, σε λάθος ερωτήσεις.



Γράφημα 6. Μεθοδολογία της διαδικασίας εξόρυξης δεδομένων CRISP–DM

9.1.2 Αξιολόγηση της κατάστασης

Περιλαμβάνει πιο λεπτομερή διερεύνηση όλων των πόρων, των περιορισμών, των υποθέσεων και των άλλων παραγόντων που πρέπει να ληφθούν υπόψη κατά τον καθορισμό του στόχου της ανάλυσης δεδομένων και του σχεδίου του έργου. Ο αναλυτής, θέλει να διευκρινίσει τις λεπτομέρειες.

9.1.3 Προσδιορισμός των στόχων στην εξόρυξη δεδομένων

Ένας επιχειρηματικός στόχος δηλώνει τους στόχους, με επιχειρηματικούς όρους. Ένας στόχος εξόρυξης δεδομένων, δηλώνει τους στόχους του έργου με τεχνικούς όρους. Π.χ. ο επιχειρηματικός στόχος μπορεί να είναι «Αύξηση των πωλήσεων του καταλόγου σε υπάρχοντες πελάτες». Ένας στόχος εξόρυξης δεδομένων μπορεί να είναι «Πρόβλεψη πόσα widgets θα αγοράσει ένας πελάτης, δεδομένων των αγορών του τα τελευταία τρία έτη, δημογραφικών πληροφοριών (π.χ. ηλικία, μισθός, πόλη, ταχυδρομικός κώδικας) και της τιμής του αντικειμένου».

9.1.4 Δημιουργία του σχεδίου του έργου

Γίνεται περιγραφή του προβλεπόμενου σχεδίου για την επίτευξη των στόχων της εξόρυξης δεδομένων και ως εκ τούτου, την επίτευξη των επιχειρηματικών στόχων. Το σχέδιο, πρέπει να καθορίζει το αναμενόμενο σύνολο των βημάτων που θα εκτελεστούν κατά τη διάρκεια του υπόλοιπου έργου, συμπεριλαμβανομένης μιας αρχικής επιλογής εργαλείων και τεχνικών.

9.2 Κατανόηση των δεδομένων

Η φάση της κατανόησης των δεδομένων, ξεκινά με μια αρχική συλλογή δεδομένων και προχωρά με δραστηριότητες για

- την εξοικείωση με τα δεδομένα,
- τον εντοπισμό των προβλημάτων της ποιότητας των δεδομένων,
- την ανακάλυψη των πρώτων πληροφοριών, σχετικά με τα δεδομένα,

• την ανίχνευση ενδιαφερόντων υποσυνόλων, για τη διαμόρφωση υποθέσεων για κρυφές πληροφορίες.

9.2.1 Συλλογή των αρχικών δεδομένων

Αποκτούνται, εντός του έργου, τα δεδομένα (ή η πρόσβαση στα δεδομένα) που αναφέρονται στους πόρους του έργου. Αυτή η αρχική συλλογή, περιλαμβάνει τη φόρτωση δεδομένων, αν είναι απαραίτητο για την κατανόηση τους. Π.χ. αν εφαρμόζεται ένα συγκεκριμένο εργαλείο για την κατανόηση των δεδομένων, είναι απολύτως λογικό να φορτωθούν τα δεδομένα σε αυτό το εργαλείο. Αυτή η προσπάθεια, μπορεί να οδηγήσει σε αρχικά βήματα προετοιμασίας των δεδομένων.

9.2.2. Περιγραφή των δεδομένων

Γίνεται εξέταση των «ακαθάριστων» ή «επιφανειακών» ιδιοτήτων των δεδομένων που αποκτήθηκαν και αναφορά των αποτελεσμάτων.

9.2.3 Εξερεύνηση των δεδομένων

Αντιμετωπίζει τα ερωτήματα της εξόρυξης δεδομένων που μπορούν να αντιμετωπιστούν, χρησιμοποιώντας την υποβολή ερωτημάτων, την οπτικοποίηση και την αναφορά. Αυτές οι αναλύσεις, μπορούν να αντιμετωπίσουν άμεσα τους στόχους της εξόρυξης δεδομένων. Μπορούν επίσης να συμβάλουν ή να βελτιώσουν την περιγραφή των δεδομένων, τις αναφορές ποιότητας και να τροφοδοτήσουν τον μετασχηματισμό και άλλες προετοιμασίες δεδομένων που απαιτούνται για περαιτέρω ανάλυση.

9.2.4 Επαλήθευση της ποιότητας των δεδομένων

Πραγματοποιείται εξέταση της ποιότητας των δεδομένων, απαντώντας σε ερωτήματα όπως π.χ. «*Είναι τα δεδομένα πλήρη; Είναι σωστά; Είναι αυτές οι τιμές που λείπουν; Αν ναι, πώς αναπαρίστανται, που εμφανίζονται και πόσο συχνές είναι;*».

9.3 Προετοιμασία των δεδομένων

Η φάση της προετοιμασίας των δεδομένων, καλύπτει όλες τις δραστηριότητες για την κατασκευή του τελικού συνόλου δεδομένων (δεδομένα που θα τροφοδοτηθούν στα εργαλεία μοντελοποίησης), από τα αρχικά ακατέργαστα δεδομένα. Οι εργασίες προετοιμασίας των δεδομένων, είναι πιθανό να εκτελεστούν πολλές φορές και όχι με κάποια προκαθορισμένη σειρά. Οι εργασίες, περιλαμβάνουν την επιλογή πίνακα εγγραφής και χαρακτηριστικών, τον μετασχηματισμό και τον καθαρισμό των δεδομένων για τα εργαλεία μοντελοποίησης.

9.3.1 Επιλογή των δεδομένων

Σε αυτό το στάδιο, παίρνεται η απόφαση σχετικά με τα δεδομένα που θα χρησιμοποιηθούν για ανάλυση. Τα κριτήρια, περιλαμβάνουν τη συνάφεια με τους στόχους της εξόρυξης δεδομένων, την ποιότητα και τους τεχνικούς περιορισμούς, όπως τα όρια στον όγκο ή τους τύπους δεδομένων. Η επιλογή των δεδομένων, καλύπτει την επιλογή χαρακτηριστικών (στήλες) και την επιλογή εγγραφών (γραμμές), σε έναν πίνακα.

9.3.2 Καθαρισμός των δεδομένων

Εδώ επιτελείται η αύξηση της ποιότητας των δεδομένων, στο επίπεδο που απαιτείται, από τις επιλεγμένες τεχνικές ανάλυσης. Προβλήματα που μπορούν να προκύψουν με τα «βρώμικα δεδομένα» περιλαμβάνουν ελλείποντα δεδομένα, κενές

τιμές, ανύπαρκτες τιμές και ελλιπή δεδομένα. Ο καθαρισμός των δεδομένων μπορεί να περιλαμβάνει την επιλογή καθαρών υποσυνόλων των δεδομένων, την εισαγωγή κατάλληλων προεπιλογών ή πιο φιλόδοξων τεχνικών, όπως η αντικατάσταση των βρώμικων δεδομένων με παράγωγες τιμές ή η δημιουργία ξεχωριστών μοντέλων για εκείνες τις οντότητες που διαθέτουν βρώμικα δεδομένα. Ωστόσο, αυτές οι προσεγγίσεις μπορούν να εισαγάγουν πρόσθετα προβλήματα. Συγκεκριμένα, το φιλτράρισμα των προβληματικών δεδομένων μπορεί να εισαγάγει μεροληψία του δείγματος στα δεδομένα και η χρήση επικαλύψεων των δεδομένων θα μπορούσε να εισαγάγει ελλείπουσες τιμές.

9.3.3 Κατασκευή των δεδομένων

Περιλαμβάνει εποικοδομητικές λειτουργίες προετοιμασίας των δεδομένων, όπως η παραγωγή παράγωγων χαρακτηριστικών, η ολόκληρων νέων εγγραφών ή μετασχηματισμένων τιμών για υπάρχοντα χαρακτηριστικά.

9.3.4 Ενσωμάτωση των δεδομένων

Δύο μέθοδοι που χρησιμοποιούνται για την ενσωμάτωση των δεδομένων είναι η συγχώνευση των δεδομένων και η δημιουργία συγκεντρωτικών τιμών. Σε αυτές τις μεθόδους, οι πληροφορίες συνδυάζονται από πολλαπλούς πίνακες ή από άλλες πηγές πληροφοριών για τη δημιουργία νέων εγγραφών ή τιμών. Π.χ. η συγχώνευση πινάκων, αναφέρεται στη σύνδεση δύο ή περισσότερων πινάκων που έχουν διαφορετικές πληροφορίες για τα ίδια αντικείμενα. Η δημιουργία συγκεντρωτικών τιμών, αναφέρεται στον υπολογισμό νέων τιμών που υπολογίζονται με τη σύνοψη πληροφοριών από πολλαπλές εγγραφές, πίνακες ή άλλες πηγές πληροφοριών.

9.3.5 Μορφοποίηση των δεδομένων

Οι μετασχηματισμοί μορφοποίησης, αναφέρονται κυρίως σε συντακτικές τροποποιήσεις που γίνονται στα δεδομένα, οι οποίες δεν αλλάζουν τη σημασία τους, αλλά ενδέχεται να απαιτούνται από το εργαλείο μοντελοποίησης.

9.4 Μοντελοποίηση

Επιλέγονται και εφαρμόζονται διάφορες τεχνικές μοντελοποίησης και οι παράμετροί τους βαθμονομούνται σε βέλτιστες τιμές. Συνήθως, αρκετές τεχνικές μπορούν να εφαρμοστούν στον ίδιο τύπο προβλήματος εξόρυξης δεδομένων. Ορισμένες τεχνικές, απαιτούν μια συγκεκριμένη μορφή δεδομένων. Άρα, συχνά απαιτείται επιστροφή στη φάση της προετοιμασίας των δεδομένων.

9.4.1 Επιλογή της τεχνικής μοντελοποίησης

Ως $1^{ο}$ βήμα στη μοντελοποίηση, πραγματοποιείται η επιλογή της πραγματικής τεχνικής μοντελοποίησης που θα χρησιμοποιηθεί. Αν επιλέχθηκε ένα εργαλείο στην επιχειρηματική κατανόηση, αυτή η εργασία αναφέρεται στην επιλογή της συγκεκριμένης τεχνικής μοντελοποίησης. Π.χ. κατασκευή δέντρων αποφάσεων ή δημιουργία ενός νευρωνικού δικτύου.

9.4.2 Δημιουργία του σχεδιασμού δοκιμής

Πριν από τη δημιουργία ενός μοντέλου, πρέπει να οριστεί μια διαδικασία για να ελεγχθεί η ποιότητα και η εγκυρότητα του. Π.χ. σε εργασίες εποπτευόμενης εξόρυξης των δεδομένων, όπως η ταξινόμηση, είναι σύνηθες να χρησιμοποιούνται ποσοστά σφάλματος ως μέτρα ποιότητας για τα μοντέλα εξόρυξης δεδομένων. Άρα, αν ο σχεδιασμός δοκιμής καθορίζει ότι το σύνολο των δεδομένων πρέπει να χωριστεί σε

σύνολα εκπαίδευσης και δοκιμών, το μοντέλο βασίζεται στο σύνολο εκπαίδευσης και η ποιότητά του εκτιμάται στο σύνολο δοκιμών.

9.4.3 Δημιουργία του μοντέλου

Ο σκοπός της δημιουργίας μοντέλων είναι η χρήση των προβλέψεων, για τη λήψη πιο εμπεριστατωμένων επιχειρηματικών αποφάσεων. Ο πιο σημαντικός στόχος κατά τη δημιουργία ενός μοντέλου είναι η σταθερότητα, που σημαίνει ότι το μοντέλο πρέπει να κάνει προβλέψεις που θα ισχύουν, όταν εφαρμόζεται σε δεδομένα που δεν έχουν ληφθεί ακόμη. Ανεξάρτητα από την τεχνική της εξόρυξης δεδομένων που χρησιμοποιείται, τα βασικά βήματα που χρησιμοποιούνται για τη δημιουργία προγνωστικών μοντέλων είναι τα ίδια. Το σύνολο μοντέλων, πρέπει πρώτα να χωριστεί σε τρία στοιχεία:

- το σύνολο εκπαίδευσης,
- το σύνολο δοκιμών,
- το σύνολο αξιολόγησης.

Ένα τέταρτο σύνολο δεδομένων, το σύνολο βαθμολογιών, δεν αποτελεί μέρος του συνόλου μοντέλων. Κάθε ένα από αυτά τα στοιχεία, πρέπει να είναι εντελώς ξεχωριστό. Δηλαδή, δεν θα πρέπει να έχουν κοινές εγγραφές, καθώς κάθε σύνολο εκτελεί έναν ξεχωριστό σκοπό. Τα μοντέλα δημιουργούνται χρησιμοποιώντας δεδομένα από το παρελθόν, προκειμένου το μοντέλο να κάνει προβλέψεις για το μέλλον. Αυτή η διαδικασία, ονομάζεται εκπαίδευση του μοντέλου. Σε αυτό το βήμα, οι αλγόριθμοι της εξόρυξης δεδομένων βρίσκουν μοτίβα που έχουν προγνωστική αξία.

Στη συνέχεια, το μοντέλο βελτιώνεται χρησιμοποιώντας το σύνολο δοκιμών. Το μοντέλο, πρέπει να βελτιωθεί για να αποφευχθεί η απομνημόνευση του συνόλου εκπαίδευσης. Αυτό το βήμα, διασφαλίζει ότι το μοντέλο είναι πιο γενικό (δηλαδή σταθερό) και θα έχει καλή απόδοση σε μη ορατά δεδομένα. Στη συνέχεια, η απόδοση του μοντέλου εκτιμάται χρησιμοποιώντας το σύνολο αξιολόγησης. Το σύνολο αξιολόγησης είναι εντελώς ξεχωριστό και διακριτό από τα σύνολα εκπαίδευσης και δοκιμής. Το σύνολο αξιολόγησης, χρησιμοποιείται για την αξιολόγηση της αναμενόμενης ακρίβειας του μοντέλου, όταν εφαρμόζεται σε δεδομένα εκτός του συνόλου μοντέλων. Τέλος, το μοντέλο εφαρμόζεται στο σύνολο των βαθμολογιών.

Το σύνολο των βαθμολογιών δεν είναι προταξινομημένο και δεν αποτελεί μέρος του συνόλου των μοντέλων που χρησιμοποιείται για τη δημιουργία του μοντέλου των δεδομένων. Τα αποτελέσματα για το σύνολο των βαθμολογιών, δεν είναι γνωστά εκ των προτέρων. Το τελικό μοντέλο, εφαρμόζεται στο σύνολο των βαθμολογιών για να γίνουν προβλέψεις. Οι προγνωστικές πλάκες, πιθανώς θα χρησιμοποιηθούν για τη λήψη πιο ενημερωμένων επιχειρηματικών αποφάσεων. Η υπερπροσαρμογή (ένα πρόβλημα που μπορεί να προκύψει) είναι ότι το μοντέλο που δημιουργείται μπορεί να υπερπροσαρμοστεί στα δεδομένα. Η υπερπροσαρμογή, σημαίνει ότι η προδιαγραφή ενός μοντέλου είναι σε μεγάλο βαθμό ένα τεχνούργημα των ιδιομορφιών του συνόλου των δεδομένων που χρησιμοποιείται για την κατασκευή του (δηλαδή, του συνόλου εκπαίδευσης). Η υπερπροσαρμογή, προκύπτει όταν ένα μοντέλο ουσιαστικά απομνημονεύει τα δεδομένα στα οποία κατασκευάστηκε. Το μοντέλο, πρέπει να μάθει τα μοτίβα για να τα αναγνωρίσει σε μελλοντικά σύνολα δεδομένων που δεν είναι ορατά, αλλά το μοντέλο δεν πρέπει να απομνημονεύει τα μοτίβα. Το πρόβλημα με το μοντέλο που απομνημονεύει το σύνολο εκπαίδευσης, είναι ότι όταν το μοντέλο βαθμολογεί μια άγνωστη εγγραφή, θα χρησιμοποιήσει τα αποτελέσματα από το σύνολο μοντέλων αν υπάρχει αντιστοιχία και αν όχι, θα παράγει μια τυχαία εικασία.

Σε αυτήν την περίπτωση, το μοντέλο είναι εντελώς ασταθές, δηλαδή δεν θα τα πάει καλύτερα από τυχαία στοιχεία στο σύνολο βαθμολογιών.

9.4.4 Αξιολόγηση του μοντέλου

Το μοντέλο πρέπει τώρα να αξιολογηθεί, ώστε να διασφαλιστεί ότι πληροί τα κριτήρια επιτυχίας της εξόρυξης δεδομένων και περνά τα επιθυμητά κριτήρια δοκιμής. Αυτό το βήμα, είναι μια καθαρά τεχνική αξιολόγηση που βασίζεται στο αποτέλεσμα των εργασιών μοντελοποίησης. Δύο εργαλεία που χρησιμοποιούνται συνήθως για την αξιολόγηση της απόδοσης διαφορετικών μοντέλων, είναι το διάγραμμα ανύψωσης και ο πίνακας σύγχυσης. Ένα διάγραμμα ανύψωσης (lift chart), που μερικές φορές ονομάζεται διάγραμμα αθροιστικών κερδών (cumulative gains chart) ή διάγραμμα μπανάνας (banana chart), είναι ένα μέτρο της απόδοσης του μοντέλου. Δείχνει πώς οι απαντήσεις (π.χ. σε μια άμεση αλληλογραφία ή σε μια χειρουργική θεραπεία), αλλάζουν με την εφαρμογή του μοντέλου. Αυτός ο λόγος αλλαγής, που ελπίζουμε ότι είναι η αύξηση του ποσοστού απόκρισης, ονομάζεται ανύψωση. Ένα διάγραμμα ανύψωσης, υποδεικνύει ποιο υποσύνολο του συνόλου των δεδομένων περιέχει το μεγαλύτερο δυνατό ποσοστό θετικών απαντήσεων. Όσο υψηλότερη είναι η καμπύλη ανύψωσης από την αρχική τιμή, τόσο καλύτερη είναι η απόδοση του μοντέλου, καθώς η αρχική τιμή αντιπροσωπεύει το μηδενικό μοντέλο, το οποίο δεν είναι καθόλου μοντέλο. Το καλύτερο μοντέλο, δεν είναι αυτό με την υψηλότερη ανύψωση κατά την κατασκευή του. Είναι το μοντέλο που αποδίδει καλύτερα σε μη ορατά, μελλοντικά δεδομένα.

Ένας πίνακας σύγχυσης (confusion matrix), που μερικές φορές ονομάζεται πίνακας ταξινόμησης (classification matrix), χρησιμοποιείται για την αξιολόγηση της ακρίβειας πρόβλεψης ενός μοντέλου. Μετρά αν ένα μοντέλο είναι συγκεχυμένο ή όχι, δηλαδή αν κάνει λάθη στις προβλέψεις του. Διάφοροι κανόνες ταξινόμησης χρησιμοποιούνται για τη δημιουργία ενός πίνακα σύγχυσης. Οι κανόνες ταξινόμησης που ενσωματώνουν προηγούμενες πιθανότητες, μεταγενέστερες πιθανότητες και κόστος λανθασμένης ταξινόμησης, βασίζονται στη στατιστική Bayesian θεωρία αποφάσεων. Η Bayesian θεωρία, ουσιαστικά αναθεωρεί τις προηγούμενες πιθανότητες με βάση πρόσθετες διαθέσιμες πληροφορίες. Στο τέλος των διαδικασιών δημιουργίας και αξιολόγησης του μοντέλου, το καταλληλότερο μοντέλο θα είναι αυτό που πληροί τους επιχειρηματικούς στόχους.

9.5 Αξιολόγηση

Τα προηγούμενα βήματα αξιολόγησης, ασχολήθηκαν με παράγοντες όπως η ακρίβεια και η γενικότητα του μοντέλου. Αυτό το βήμα, αξιολογεί τον βαθμό στον οποίο το μοντέλο πληροί τους επιχειρηματικούς στόχους και επιδιώκει να προσδιορίσει αν υπάρχει κάποιος επιχειρηματικός λόγος για τον οποίο το μοντέλο είναι ελλιπές.

Συγκρίνει τα αποτελέσματα, με τα κριτήρια αξιολόγησης που ορίζονται στην αρχή του έργου. Ένας καλός τρόπος για να ορίσουμε τα συνολικά αποτελέσματα ενός έργου εξόρυξης δεδομένων είναι να χρησιμοποιήσουμε την εξίσωση:
results= f(models, findings) Στην εξίσωση, ορίζουμε το συνολικό αποτέλεσμα του έργου εξόρυξης δεδομένων όχι μόνο ως τα μοντέλα, αλλά και ως τα ευρήματα που μπορούν να οριστούν σαν οτιδήποτε (εκτός από το μοντέλο) που είναι σημαντικό για την επίτευξη των στόχων της επιχείρησης.

9.5.1 Αξιολόγηση των αποτελεσμάτων

Μια άλλη επιλογή αξιολόγησης είναι η δοκιμή του/των μοντέλου/ων σε δοκιμαστικές εφαρμογές στην πραγματική εφαρμογή, αν το επιτρέπει ο χρόνος και ο προϋπολογισμός.

9.5.2 Διαδικασία της αναθεώρησης

Σε αυτό το σημείο, το μοντέλο που προκύπτει φαίνεται ικανοποιητικό και μοιάζει να πληρεί τις επιχειρηματικές ανάγκες. Είναι πλέον σκόπιμο να γίνει μια πιο εμπεριστατωμένη ανασκόπηση του έργου εξόρυξης δεδομένων, προκειμένου να διαπιστωθεί αν υπάρχει κάποιος σημαντικός παράγοντας ή εργασία που έχει παραλειφθεί με κάποιον τρόπο. Σε αυτό το στάδιο της εξόρυξης δεδομένων, η διαδικασία αναθεώρησης λαμβάνει τη μορφή αναθεώρησης διασφάλισης ποιότητας.

9.5.3 Καθορισμός των επόμενων βημάτων

Σύμφωνα με τα αποτελέσματα της αξιολόγησης και με την ανασκόπηση της διαδικασίας, ο αναλυτής αποφασίζει πώς θα προχωρήσει σε αυτό το στάδιο. Ο αναλυτής πρέπει να αποφασίσει αν θα ολοκληρώσει το έργο και θα προχωρήσει στην ανάπτυξη ή αν θα ξεκινήσει περαιτέρω επαναλήψεις ή αν θα δημιουργήσει νέα έργα εξόρυξης δεδομένων.

9.6 Ανάπτυξη

Η δημιουργία του μοντέλου γενικά, δεν είναι το τέλος του έργου. Ακόμα κι αν ο σκοπός του μοντέλου είναι η αύξηση της γνώσης των δεδομένων, η γνώση που αποκτάται πρέπει να οργανωθεί και να παρουσιαστεί με τρόπο που ο πελάτης μπορεί να χρησιμοποιήσει. Ανάλογα με τις απαιτήσεις, η φάση ανάπτυξης μπορεί να είναι τόσο απλή όσο η δημιουργία μιας αναφοράς ή τόσο περίπλοκη όσο η εφαρμογή μιας επαναλήψιμης διαδικασίας εξόρυξης δεδομένων. Σε πολλές περιπτώσεις, ο πελάτης, όχι ο αναλυτής δεδομένων, είναι αυτός που θα εκτελέσει τα βήματα ανάπτυξης.

Ωστόσο, ακόμη και αν ο αναλυτής δεν εκτελέσει την προσπάθεια ανάπτυξης, είναι σημαντικό για τον πελάτη να κατανοήσει εκ των προτέρων, ποιες ενέργειες πρέπει να εκτελεστούν για να αξιοποιηθούν τα μοντέλα που δημιουργήθηκαν.

9.6.1 Σχεδιασμός της ανάπτυξης

Για την ανάπτυξη του αποτελέσματος εξόρυξης δεδομένων στην επιχείρηση, αυτή η εργασία λαμβάνει τα αποτελέσματα της αξιολόγησης και αναπτύσσει μια στρατηγική για την ανάπτυξη. Αν έχει προσδιοριστεί μια γενική διαδικασία για τη δημιουργία του σχετικού μοντέλου, αυτή η διαδικασία τεκμηριώνεται εδώ για μεταγενέστερη ανάπτυξη.

9.6.2 Σχεδιασμός της παρακολούθησης και της συντήρησης

Η παρακολούθηση και η συντήρηση, είναι σημαντικά ζητήματα αν το αποτέλεσμα της εξόρυξης δεδομένων γίνει μέρος της καθημερινής επιχείρησης και του περιβάλλοντος της. Η προσεκτική προετοιμασία μιας στρατηγικής συντήρησης, βοηθά στην αποφυγή άσκοπα μεγάλων περιόδων λανθασμένης χρήσης των αποτελεσμάτων της εξόρυξης δεδομένων. Για την παρακολούθηση της ανάπτυξης του αποτελέσματος της εξόρυξης δεδομένων, το έργο χρειάζεται ένα λεπτομερές σχέδιο για τη διαδικασία παρακολούθησης. Αυτό το σχέδιο, λαμβάνει υπόψη τον συγκεκριμένο τύπο ανάπτυξης.

9.6.3 Σύνταξη της τελικής έκθεσης

Στο τέλος του έργου, ο επικεφαλής του έργου και η ομάδα συντάσσουν την τελική έκθεση. Ανάλογα με το σχέδιο ανάπτυξης, αυτή η έκθεση μπορεί να είναι μόνο μια σύνοψη του έργου και των εμπειριών του (αν δεν έχουν ήδη καταγραφεί ως συνεχιζόμενη δραστηριότητα) ή μπορεί να είναι μια τελική και ολοκληρωμένη παρουσίαση του αποτελέσματος της εξόρυξης δεδομένων.

9.6.4 Αναθεώρηση του έργου

Γίνεται αξιολόγηση τι πήγε σωστά, τι πήγε στραβά, τι έγινε καλά και τι χρειάζεται βελτίωση.

10. Δεδομένα που πραγματοποιείται εξόρυξη δεδομένων

Εξόρυξη δεδομένων μπορεί να πραγματοποιηθεί σε μια σειρά από διαφορετικές αποθήκες δεδομένων. Κατ' αρχήν, η εξόρυξη δεδομένων πρέπει να εφαρμόζεται σε κάθε είδους αποθετήριο πληροφοριών. Αυτό περιλαμβάνει σχεσιακές βάσεις δεδομένων, αποθήκες δεδομένων, βάσεις δεδομένων συναλλαγών, προηγμένα συστήματα βάσεων δεδομένων, επίπεδα αρχεία και τον παγκόσμιο ιστό. Τα προηγμένα συστήματα βάσεων δεδομένων, περιλαμβάνουν αντικειμενοστρεφείς (object–oriented databases) και αντικειμενοσχεσιακές βάσεις δεδομένων (object–relational databases) και συγκεκριμένες βάσεις δεδομένων προσανατολισμένες σε εφαρμογές, όπως χωρικές βάσεις δεδομένων, βάσεις δεδομένων χρονοσειρών, βάσεις δεδομένων κειμένου και βάσεις δεδομένων πολυμέσων. Οι προκλήσεις και οι τεχνικές της εξόρυξης, μπορεί να διαφέρουν για κάθε ένα από τα συστήματα αποθετηρίων.

10.1 Σχεσιακές βάσεις δεδομένων

Ένα σύστημα βάσεων δεδομένων, που ονομάζεται επίσης σύστημα διαχείρισης βάσεων δεδομένων (DBMS–database management system), αποτελείται από μια συλλογή αλληλένδετων δεδομένων, γνωστή ως βάση δεδομένων και από ένα σύνολο προγραμμάτων λογισμικού για τη διαχείριση και για την πρόσβαση στα δεδομένα. Τα προγράμματα λογισμικού, περιλαμβάνουν μηχανισμούς για τον ορισμό δομών βάσεων δεδομένων, για την αποθήκευση δεδομένων, για την ταυτόχρονη, κοινόχρηστη ή κατανεμημένη πρόσβαση σε δεδομένα και για τη διασφάλιση της συνέπειας και της ασφάλειας των αποθηκευμένων πληροφοριών, παρά τις διακοπές λειτουργίας του συστήματος ή τις προσπάθειες μη εξουσιοδοτημένης πρόσβασης. Μια σχεσιακή βάση δεδομένων είναι μια συλλογή πινάκων, σε κάθε έναν από τους οποίους έχει εκχωρηθεί ένα μοναδικό όνομα. Κάθε πίνακας, αποτελείται από ένα σύνολο χαρακτηριστικών (στήλες ή πεδία) και συνήθως αποθηκεύει έναν μεγάλο αριθμό πλειάδων (εγγραφές ή γραμμές). Κάθε πλειάδα σε έναν σχεσιακό πίνακα, αντιπροσωπεύει ένα αντικείμενο που αναγνωρίζεται από ένα μοναδικό κλειδί και περιγράφεται από ένα σύνολο τιμών χαρακτηριστικών.

10.2 Αποθήκες δεδομένων

Μια αποθήκη δεδομένων (data warehouse) είναι ένα αποθετήριο πληροφοριών που συλλέγονται από πολλαπλές πηγές, αποθηκεύονται σε ένα ενοποιημένο σχήμα και συνήθως βρίσκονται σε μία μόνο τοποθεσία. Οι αποθήκες δεδομένων, κατασκευάζονται μέσω μιας διαδικασίας καθαρισμού, μετασχηματισμού, ενσωμάτωσης, φόρτωσης και περιοδικής ανανέωσης των δεδομένων. Μια αποθήκη δεδομένων, συνήθως μοντελοποιείται από μια πολυδιάστατη δομή βάσης δεδομένων, όπου κάθε διάσταση αντιστοιχεί σε ένα χαρακτηριστικό ή σε ένα σύνολο χαρακτηριστικών στο σχήμα και κάθε κελί αποθηκεύει την τιμή κάποιου συγκεντρωτικού μέτρου, όπως ο αριθμός ή το ποσό πωλήσεων. Η πραγματική φυσική δομή μιας αποθήκης δεδομένων, μπορεί να είναι μια σχεσιακή αποθήκη δεδομένων ή ένας πολυδιάστατος κύβος δεδομένων.

Παρέχει μια πολυδιάστατη προβολή των δεδομένων και επιτρέπει τον προϋπολογισμό και τη γρήγορη πρόσβαση σε συνοπτικά δεδομένα. Στην ερευνητική βιβλιογραφία σχετικά με τις αποθήκες δεδομένων, η δομή κύβου δεδομένων που αποθηκεύει το πρωτόγονο ή το χαμηλότερο επίπεδο πληροφοριών, ονομάζεται βασικό

κυβοειδές. Οι αντίστοιχες πολυδιάστατες (κύβοι) δομές υψηλότερου επιπέδου, ονομάζονται (μη βασικά) κυβοειδή. Ένα βασικό κυβοειδές μαζί με όλα τα αντίστοιχα κυβοειδή υψηλότερου επιπέδου, σχηματίζουν έναν κύβο δεδομένων. Παρέχοντας πολυδιάστατες προβολές δεδομένων και τον προϋπολογισμό συνοπτικών δεδομένων, τα συστήματα αποθήκης δεδομένων είναι κατάλληλα για OnLine Analytical Processing (OLAP). Οι λειτουργίες OLAP, χρησιμοποιούν γνώσεις υποβάθρου σχετικά με τον τομέα των δεδομένων που μελετώνται, προκειμένου να επιτρέψουν την παρουσίαση δεδομένων σε διαφορετικά επίπεδα αφαίρεσης. Τέτοιες λειτουργίες, προσαρμόζονται σε διαφορετικές οπτικές γωνίες χρηστών. Παρόλο που τα εργαλεία αποθήκης δεδομένων βοηθούν στην υποστήριξη της ανάλυσης δεδομένων, απαιτούνται πρόσθετα εργαλεία για την εξόρυξη δεδομένων, για να επιτρέψουν μια πιο σε βάθος και αυτοματοποιημένη ανάλυση.

10.3 Συναλλακτικές βάσεις δεδομένων

Γενικά, μια συναλλακτική βάση δεδομένων αποτελείται από ένα αρχείο όπου κάθε εγγραφή αντιπροσωπεύει μια συναλλαγή. Μια συναλλαγή, συνήθως περιλαμβάνει έναν μοναδικό αριθμό αναγνώρισης συναλλαγής και μια λίστα με τα στοιχεία που αποτελούν τη συναλλαγή (όπως είδη που αγοράστηκαν σε ένα κατάστημα). Η βάση δεδομένων συναλλαγών, μπορεί να έχει πρόσθετους πίνακες που σχετίζονται με αυτήν, οι οποίοι περιέχουν άλλες πληροφορίες σχετικά με την πώληση, όπως την ημερομηνία της συναλλαγής, τον αριθμό αναγνώρισης πελάτη, τον αριθμό αναγνώρισης του πωλητή και του υποκαταστήματος στο οποίο πραγματοποιήθηκε η πώληση.

10.4. Προηγμένα συστήματα βάσεων δεδομένων και προηγμένες εφαρμογές βάσεων δεδομένων

Τα σχεσιακά συστήματα βάσεων δεδομένων, έχουν χρησιμοποιηθεί ευρέως σε επιχειρηματικές εφαρμογές. Με την πρόοδο της τεχνολογίας των βάσεων δεδομένων, έχουν εμφανιστεί και αναπτύσσονται διάφορα είδη προηγμένων συστημάτων βάσεων δεδομένων για να καλύψουν τις απαιτήσεις των νέων εφαρμογών των βάσεων δεδομένων. Οι νέες εφαρμογές των βάσεων δεδομένων, περιλαμβάνουν τη διαχείριση χωρικών δεδομένων (χάρτες), δεδομένων μηχανικού σχεδιασμού (σχεδιασμός κτιρίων, στοιχείων συστήματος ή ολοκληρωμένων κυκλωμάτων), δεδομένων υπερκειμένου και πολυμέσων (συμπεριλαμβανομένων δεδομένων κειμένου, εικόνας, βίντεο και ήχου), δεδομένων που σχετίζονται με τον χρόνο (ιστορικά αρχεία ή δεδομένα χρηματιστηρίου) και τον παγκόσμιο ιστό (ένα τεράστιο, ευρέως κατανεμημένο αποθετήριο πληροφοριών που διατίθεται μέσω του διαδικτύου).

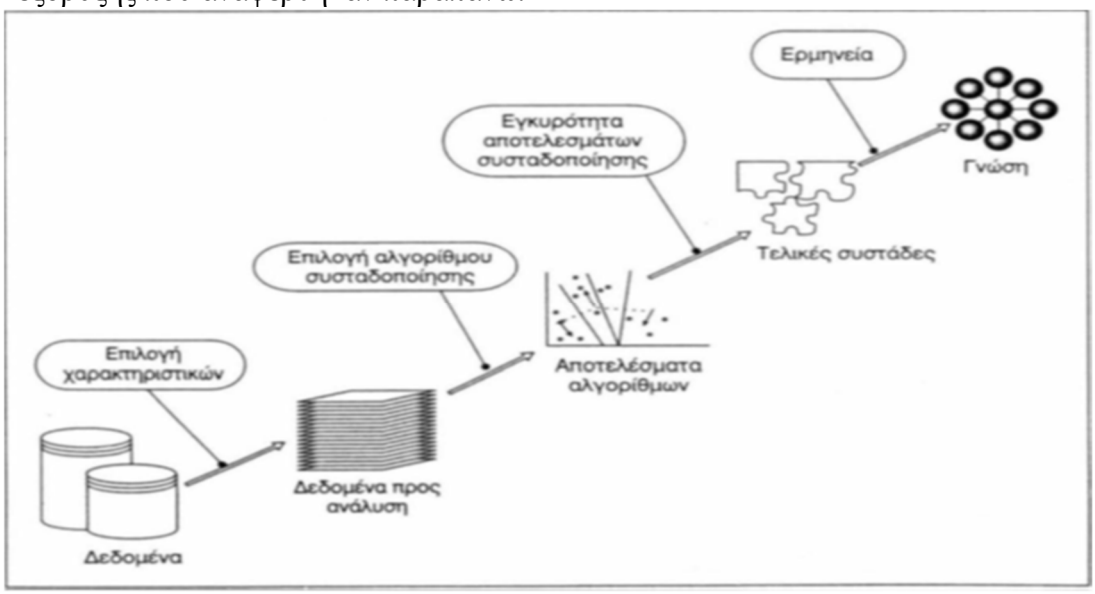
Αυτές οι εφαρμογές, απαιτούν αποτελεσματικές δομές δεδομένων και κλιμακούμενες μεθόδους για τη διαχείριση των σύνθετων δομών των αντικειμένων, των αρχείων μεταβλητού μήκους, των ημιδομημένων ή των μη δομημένων δεδομένων, των δεδομένων κειμένου και πολυμέσων και των σχημάτων βάσεων δεδομένων με σύνθετες δομές και δυναμικές αλλαγές. Αντιμετωπίζοντας αυτές τις ανάγκες, έχουν αναπτυχθεί προηγμένα συστήματα βάσεων δεδομένων και ειδικά συστήματα βάσεων δεδομένων, προσανατολισμένα στις εφαρμογές. Αυτά περιλαμβάνουν αντικειμενοστρεφή και αντικειμενοσχεσιακά συστήματα βάσεων δεδομένων, χωρικά συστήματα βάσεων δεδομένων, χρονικά και χρονοσειριακά συστήματα βάσεων δεδομένων, συστήματα βάσεων δεδομένων κειμένου και πολυμέσων, ετερογενή και παλαιότερα συστήματα βάσεων δεδομένων και τα παγκόσμια συστήματα πληροφοριών που βασίζονται στο Web. Ενώ τέτοιες βάσεις δεδομένων ή αποθετήρια πληροφοριών απαιτούν εξελιγμένες εγκαταστάσεις για την αποτελεσματική

αποθήκευση, ανάκτηση και ενημέρωση μεγάλων ποσοτήτων σύνθετων δεδομένων, παρέχουν επίσης γόνιμο έδαφος και εγείρουν πολλά απαιτητικά ζητήματα έρευνας και εφαρμογής για την εξόρυξη δεδομένων.

11. Βασικές εργασίες της εξόρυξης γνώσης από δεδομένα

- **Κατηγοριοποίηση (Classification).** Απεικονίζει τα δεδομένα σε προκαθορισμένες ομάδες – κλάσεις (classes). Αναφέρεται και ως εποπτευόμενη μάθηση, διότι οι κλάσεις καθορίζονται πριν από την εξέταση των δεδομένων. Η αναγνώριση προτύπου (pattern recognition), αποτελεί ένα είδος κατηγοριοποίησης.
- **Παλινδρόμηση (Regression).** Χρησιμοποιείται προκειμένου να απεικονιστεί ένα στοιχειώδες δεδομένο σε μία πραγματική μεταβλητή πρόβλεψης. Περιλαμβάνει την εκμάθηση της συνάρτησης που κάνει την απεικόνιση, υπό την προϋπόθεση ότι τα σχετικά δεδομένα ταιριάζουν με γνωστά είδη συναρτήσεων (γραμμική, λογαριθμική) και ακολούθως καθορίζει την καλύτερη συνάρτηση που μοντελοποιεί τα δοθέντα δεδομένα. Η συνάρτηση μεταφοράς μπορεί να είναι γραμμική (linear), ή μη γραμμική (non–linear), ή στοχαστική (stochastic), ή βηματική (step by step).
- **Ανάλυση χρονοσειρών (Time series analysis).** Εξετάζεται η τιμή ενός γνωρίσματος καθώς μεταβάλλεται με τον χρόνο, με τιμές που λαμβάνονται σε ίσα χρονικά διαστήματα (ωριαία, ημερήσια, εβδομαδιαία). Στην ανάλυση των χρονοσειρών, πραγματοποιούνται τρεις βασικές λειτουργίες.
  - Χρησιμοποίηση μονάδων μέτρησης της απόστασης, για τον καθορισμό της ομοιότητας μεταξύ των χρονοσειρών.
  - Μελέτη της δομής της χρονοσειράς, προκειμένου να καθορισθεί (ή και να κατηγοριοποιηθεί) η συμπεριφορά της.
  - Χρήση διαγραμμάτων χρονοσειρών, για την πρόβλεψη των μελλοντικών τιμών.
- **Πρόβλεψη (prediction).** Θεωρείται το να δίνεται τιμή σε μία μελλοντική κατάσταση παρά σε μία τρέχουσα, βρίσκοντας εφαρμογή σε πρόγνωση πλημμυρών, αναγνώριση ομιλίας, μηχανική μάθηση, αναγνώριση προτύπου. Συνεπώς, μπορεί να θεωρηθεί ως ένα είδος κατηγοριοποίησης.
- **Συσταδοποίηση (clustering).** Αναφέρεται και ως μη εποπτευόμενη μάθηση ή τμηματοποίηση και μπορεί να θεωρηθεί ως μία διαμέριση των δεδομένων σε ομάδες, που μπορεί να είναι (ή να μην είναι) διακριτές μεταξύ τους. Επιτυγχάνεται με τον καθορισμό της ομοιότητας, ανάμεσα στα δεδομένα, ως προς προκαθορισμένα χαρακτηριστικά. Μοιάζει με την κατηγοριοποίηση, αλλά διαφέρει στο ότι οι ομάδες δεδομένων (συστάδες) δεν είναι προκαθορισμένες, αλλά ορίζονται κυρίως από τα δεδομένα, με το να ομαδοποιούνται στις ίδιες ομάδες τα πιο σχετικά δεδομένα.
- **Παρουσίαση συνόψεων (summarization)** ή χαρακτηρισμός (characterization) ή γενίκευση (generalization). Απεικονίζει τα δεδομένα σε υποσύνολα τους, με συνοδευτικές απλές αντιπροσωπευτικές περιγραφές ή συνοπτικές πληροφορίες (μέσος όρος ενός χαρακτηριστικού) σχετικά με τις βάσεις δεδομένων.
- **Κανόνες συσχέτισης (association rules)** ή ανάλυση συνδέσμων (link analysis) ή ανάλυση συγγένειας (affinity analysis). Πρόκειται για μοντέλα που αναγνωρίζουν ειδικούς τύπους συσχέτισης μεταξύ των δεδομένων και εφαρμόζονται στις πωλήσεις λιανικής, προκειμένου αναγνωρισθούν προϊόντα που συχνά αγοράζονται μαζί (ανάλυση καλαθιού αγορών – market basket analysis).
- **Ανακάλυψη ακολουθιών (sequence discovery)** ή **ακολουθιακή ανάλυση (sequential analysis)**. Χρησιμοποιείται προκειμένου να καθορισθούν στα δεδομένα, σειριακά πρότυπα, που η συσχέτιση τους βασίζεται σε μία χρονική ακολουθία ενεργειών. Ενώ η ανάλυση του καλαθιού των αγορών προϋποθέτει γνώση του ποια

προϊόντα αγοράσθηκαν ταυτόχρονα, στην ανακάλυψη ακολουθιών τα προϊόντα αγοράζονται με κάποια σειρά στη διάρκεια μίας περιόδου.
Παρακάτω παρουσιάζονται αναλυτικότερα ορισμένες εκ των βασικότερων τεχνικών εξόρυξης που αναφέρθηκαν παραπάνω.



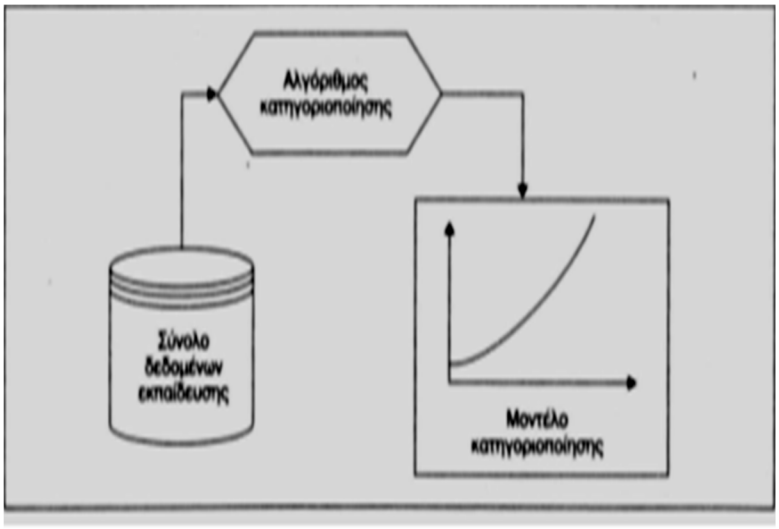
Γράφημα 7. Βήματα της διαδικασίας της συσταδοποίησης

11.1. Κατηγοριοποίηση

Χαρακτηρίζεται ως μία από τις βασικές τεχνικές εξόρυξης γνώσης και αποσκοπεί στην ανάθεση ενός στοιχείου σε ένα προκαθορισμένο σύνολο κατηγοριών (classes), δηλαδή κατηγοριοποιεί ένα στοιχείο σε μία από τις διαφορετικές κατηγορίες που έχουν προκαθορισθεί, με στόχο τη δημιουργία ενός μοντέλου που θα μπορεί να χρησιμοποιηθεί για κατηγοριοποίηση μελλοντικών δεδομένων των οποίων η κατηγοριοποίηση είναι άγνωστη. Εφαρμόζεται στην ιατρική διάγνωση, στην έγκριση δανείων, στην ανίχνευση σφαλμάτων σε βιομηχανικές εφαρμογές και στην κατηγοριοποίηση τάσεων στην οικονομία. Προβλήματα κατηγοριοποίησης έχουν μελετηθεί εκτενώς στη στατιστική, στην αναγνώριση προτύπων (patterns) και στη μηχανική μάθηση (machine learning). Η διαδικασία κατηγοριοποίησης δεδομένων, περιγράφεται από δύο βήματα.

1ο βήμα. Εκμάθηση (Learning)

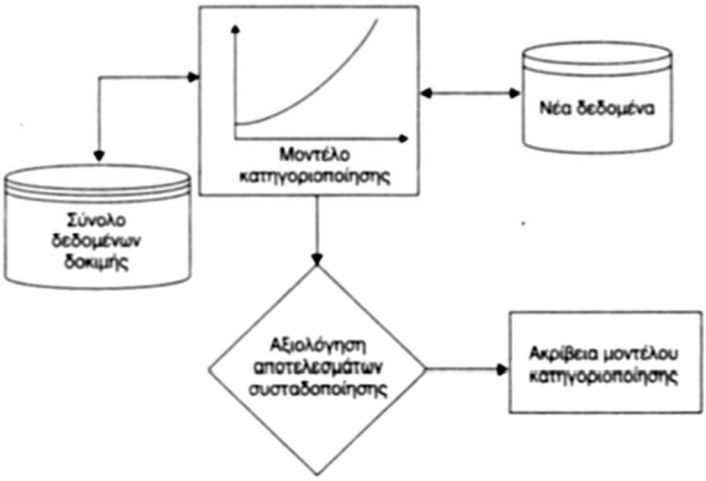
Ένας αλγόριθμος κατηγοριοποίησης αναλύει τα δεδομένα εκπαίδευσης (training data) προκειμένου να κατασκευάσουν εν συνεχεία το μοντέλο (model) το οποίο είναι γνωστό ως κατηγοριοποιητής (classifier) και αναπαρίσταται με τη μορφή κανόνων κατηγοριοποίησης (classification rules), δέντρων αποφάσεων (decision trees) ή μαθηματικών τύπων (mathematical formulas). Επειδή η κατηγορία των δειγμάτων εκπαίδευσης είναι γνωστή, αυτό το βήμα ονομάζεται και εποπτευόμενη μάθηση (supervised learning).



Γράφημα 8. Διαδικασία της εκμάθησης

2ο βήμα. Κατηγοριοποίηση (Classification)

Γίνεται χρήση δοκιμαστικών μοντέλων (test data), προκειμένου υπολογισθεί η ακρίβεια (accuracy) του μοντέλου, η οποία ορίζεται ως το ποσοστό των δειγμάτων δοκιμής (training samples) που κατηγοριοποιήθηκαν σωστά από το υπό εκπαίδευση μοντέλο, σε ένα καθορισμένο σύνολο δοκιμών επαφής. Εφόσον γίνει αποδεκτή η ακρίβεια του μοντέλου, επιτρέπεται πλέον αυτό να χρησιμοποιηθεί για κατηγοριοποίηση μελλοντικών δειγμάτων δεδομένων, των οποίων η κατηγοριοποίηση είναι άγνωστη.



Γράφημα 9. Διαδικασία της κατηγοριοποίησης

Οι στατιστικοί αλγόριθμοι, βασίζονται άμεσα, στη χρήση της στατιστικής πληροφορίας. Οι αλγόριθμοι που βασίζονται στην απόσταση, χρησιμοποιούν μέτρα ομοιότητας ή απόστασης, προκειμένου εκτελέσουν την κατηγοριοποίηση. Τεχνικές δέντρων αποφάσεων και νευρωνικών δικτύων (Neural networks) χρησιμοποιούν αυτές τις δομές, προκειμένου να εκτελέσουν την κατηγοριοποίηση. Οι αλγόριθμοι κατηγοριοποίησης που βασίζονται σε κανόνες, δημιουργούν if–then κανόνες για να εκτελέσουν την κατηγοριοποίηση.

Κατηγοριοποίηση				
↓	↓	↓	↓	↓
Στατιστική	**Απόσταση**	**Κανόνες**	**Νευρωνικά δίκτυα**	**Δέντρα αποφάσεων**

Πίνακας 3. Κατηγορίες αλγορίθμων κατηγοριοποίησης

Προς επίλυση του προβλήματος της κατηγοριοποίησης, έχουν αναπτυχθεί τρείς μέθοδοι (καθορισμού των ορίων, χρήσης των κατανομών πιθανότητας και χρήση εκ των υστέρων πιθανοτήτων). Για τον χειρισμό των ελλιπών τιμών, υπάρχουν οι ακόλουθες προσεγγίσεις:

- Αγνοούνται τα ελλιπή δεδομένα
- Στα ελλιπή δεδομένα, δίδεται υποθετικά κάποια τιμή.
- Τα ελλιπή δεδομένα, θεωρούνται όλα ότι έχουν μία συγκεκριμένη τιμή από μόνα τους.

11.2. Συσταδοποίηση

Είναι μία από τις πλέον χρήσιμες διεργασίες της διαδικασίας εξόρυξης γνώσης για την ανακάλυψη συστάδων και για τον προσδιορισμό κατανομών ή προτύπων, που παρουσιάζουν ενδιαφέρον στα υπό μελέτη δεδομένα. Εφαρμόζεται στη μη εποπτευόμενη μάθηση (unsupervised learning), στην αναγνώριση προτύπων, στην αριθμητική ταξινομία (numerical taxonomy), στη βιολογία, στην οικολογία, στις κοινωνικές επιστήμες και στη θεωρία γράφων.

11.2.1 Διαδικασία της συσταδοποίησης

Περιλαμβάνει τα ακόλουθα βήματα:

- Επιλογή των χαρακτηριστικών γνωρισμάτων,
- Αλγόριθμο της συσταδοποίησης,
- Μέτρο της γειτνίασης (proximity measure),
- Κριτήριο της συσταδοποίησης,
- Επικύρωση των αποτελεσμάτων,
- Ερμηνεία των αποτελεσμάτων.

11.2.2 Εφαρμογές της συσταδοποίησης

Ενδεικτικά, αναφέρονται τα ακόλουθα πεδία εφαρμογής:

- Μείωση των δεδομένων,
- Παραγωγή της υπόθεσης,
- Έλεγχος της υπόθεσης,
- Πρόβλεψη βασισμένη σε συστάδες.

11.2.3 Μέθοδοι της συσταδοποίησης

Οι αλγόριθμοι συσταδοποίησης, ταξινομούνται σύμφωνα με τον τύπο των δεδομένων που εισάγονται στον αλγόριθμο, τη μέθοδο που καθορίζει τη συσταδοποίηση του συνόλου δεδομένων, τη θεωρία και τις θεμελιώδεις έννοιες στις οποίες είναι βασισμένες οι τεχνικές ανάλυσης συστάδας.

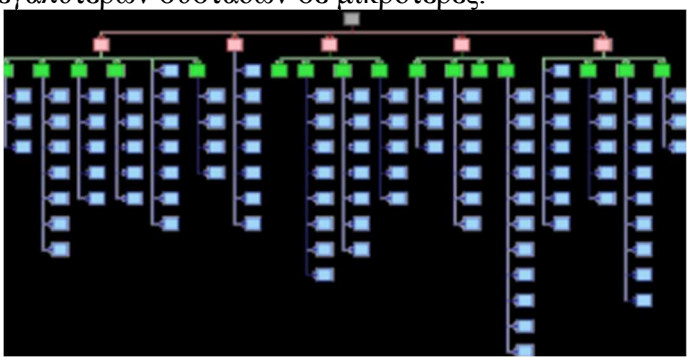
11.2.4 Κατηγοριοποίηση των αλγορίθμων με βάση τη μέθοδο συσταδοποίησης

Οι αλγόριθμοι, ταξινομούνται στους παρακάτω τύπους:

- Διαιρετική συσταδοποίηση (**partitional clustering**). Βασίζεται στην άμεση αποσύνθεση του συνόλου των δεδομένων, σε ένα σύνολο μη σχετιζόμενων συστάδων.
- Η ασαφής συσταδοποίηση (**fuzzy clustering**). Χρησιμοποιεί τεχνικές ασαφούς λογικής, προκειμένου να ομαδοποιήσει δεδομένα και θεωρεί ότι ένα αντικείμενο μπορεί να ταξινομηθεί σε περισσότερες από μία συστάδες.
- Η μη ασαφής συσταδοποίηση (**crisp clustering**). Θεωρεί μη επικαλυπτόμενα χωρίσματα, σημαίνοντας ότι ένα στοιχείο του συνόλου των δεδομένων ανήκει ή δεν ανήκει σε μία κατηγορία.

• Συσταδοποίηση βασισμένη στα δίκτυα Kohonen (**Kohonen net clustering**), η οποία είναι βασισμένη στις έννοιες των νευρωνικών δικτύων.

• Ιεραρχική συσταδοποίηση (**Hierarchical clustering**). Αυτοί οι αλγόριθμοι, βασίζονται στη διαδοχική σύνδεση των μικρότερων συστάδων σε μεγαλύτερες ή σε διάσπαση των μεγαλύτερων συστάδων σε μικρότερες.



Γράφημα 10. Δενδρογράφημα

• Συσταδοποίηση βασισμένη στην πυκνότητα (**Density–based clustering**).

• Συσταδοποίηση βασισμένη σε πλέγμα (**Grid–based clustering**). Οι αλγόριθμοι αυτού του τύπου, χωρίζουν τον χώρο σε έναν πεπερασμένο αριθμό κελιών και ακολούθως κάνουν όλες τις διαδικασίες στον κβαντοποιημένο χώρο.

• Συσταδοποίηση (**Subspace clustering**). Οι αλγόριθμοι αυτοί, προσπαθούν να βρουν τα υποσύνολα του αρχικού χώρου όπου τα αποτελέσματα είναι καλύτερα.

11.2.5 Κατηγοριοποίηση αλγορίθμων με βάση τον τύπο των δεδομένων

Υπάρχει η συσταδοποίηση των αριθμητικών δεδομένων και η εννοιολογική συσταδοποίηση.

11.2.6 Ιεραρχικοί αλγόριθμοι

Δημιουργούν στην πραγματικότητα σύνολα συστάδων. Διακρίνονται σύμφωνα με τη μέθοδο που παράγουν συστάδες, σε συσωρευτικούς ιεραρχικούς αλγορίθμους (Agglomerative) και σε διαιρετικούς ιεραρχικούς αλγορίθμους (Divisive). Οι συσωρευτικοί αλγόριθμοι, διαφοροποιούνται μεταξύ τους ως προς το πώς συγχωνεύονται οι συστάδες σε κάθε επίπεδο, με τις ακόλουθες τεχνικές:

• Τεχνική απλού συνδέσμου (Single link technique)

• Αλγόριθμος πλήρους συνδέσμου (Complete link algorithm)

• Αλγόριθμος μέσου συνδέσμου (Average link algorithm)

Οι διαιρετικοί ιεραρχικοί αλγόριθμοι συσταδοποίησης (Divisive clustering), τοποθετούν αρχικά όλα τα στοιχεία σε μία συστάδα η οποία στη συνέχεια διασπάται σε δύο επιμέρους συστάδες και η διαδικασία αυτή επαναλαμβάνεται μέχρι που το κάθε στοιχείο να ανήκει στη δική του συστάδα.

11.2.7 Διαμεριστικοί αλγόριθμοι (Partitional algorithms)

Οι συστάδες, δημιουργούνται σε ένα βήμα. Για τον προσδιορισμό της καταλληλότητας των προτεινομένων λύσεων, χρησιμοποιούνται μέτρα ποιότητας (μετρικές, συναρτήσεις κριτηρίων). Ένα συνηθισμένο μέτρο, είναι μία μετρική τετραγωνικού σφάλματος (squared error), που μετρά την τετραγωνική απόσταση των σημείων της συστάδας από το κέντρο της. Ορισμένοι αντιπροσωπευτικοί αλγόριθμοι αυτής της κατηγορίας είναι:

• Συσταδοποίηση με γενετικούς αλγορίθμους, οι οποίοι εκτελούν καθολική αντί τοπικής αναζήτηση των ενδεχομένων λύσεων.

• Δέντρο ελάχιστης ζεύξης

• Συσταδοποίηση Partitioning around medoids–PAM. Αναπτύχθηκε από τους Kaufmann και Rousseeuw, αποτελεί μία από τις γνωστότερες k–medoids μεθόδους συσταδοποίησης, η ποιότητα της οποίας μετράτε με βάση τη μέση διαφοροποίηση ανάμεσα σε ένα αντικείμενο και στο medoid της συστάδας που ανήκει. Medoid ονομάζονται τα αντικείμενα–αντιπρόσωποι και είναι εκείνα που βρίσκονται εγγύτερα στα κέντρα των συστάδων.

• Συσταδοποίηση K–means. Επαναληπτικός αλγόριθμος, στον οποίο τα στοιχεία μετακινούνται μεταξύ των διαφόρων συνόλων συστάδων, μέχρι να επιτευχθεί το επιθυμητό σύνολο συστάδων. Βασίζεται στην άμεση αποσύνθεση του συνόλου των δεδομένων σε ένα σύνολο ασυσχέτιστων συστάδων και χρησιμοποιεί σταθερό και εξ΄ αρχής δεδομένο αριθμό συστάδων που θα δημιουργηθούν (όσα και τα κέντρα). Διάφορα στατιστικά πακέτα (SPSS, SAS, BMPD) που χρησιμοποιούν τον αλγόριθμο K–means, υιοθετούν το καθένα τη δική του έκδοση.

Πλεονεκτήματα	Μειονεκτήματα
Απλός	Δε δουλεύει με μη αριθμητικά δεδομένα
Κατανοητός	Μη ντετερμινιστικός
Ταχύτητα σύγκλισης	Πρέπει να ορισθεί ο αριθμός των clusters
Τα αντικείμενα ανατίθεται αυτόματα σε κάποιο cluster	Όλα τα αντικείμενα πρέπει υποχρεωτικά να ανήκουν σε κάποιο cluster

Πίνακας 4. Πλεονεκτήματα και μειονεκτήματα του αλγορίθμου K–means.

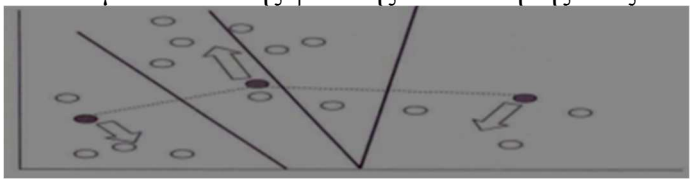
Ορισμένες παραλλαγές του αλγορίθμου K–means είναι:

➢ Ο αλγόριθμος ISODATA, που περιλαμβάνει μία διαδικασία αναζήτησης του καλύτερου αριθμού συστάδων, με βάση κάποιο κόστος εκτέλεσης.

➢ Ο fuzzy C–means, που επεκτείνει τον κλασσικό αλγόριθμο K–means με χρήση της ασαφούς λογικής.

➢ Ο SAS PROC FACTULUS, που ελέγχει τη διαδικασία συσταδοποίησης υιοθετώντας δυο επιπλέον παραμέτρους, την max_rad (ελέγχει τον ελάχιστο αριθμό στοιχείων που μπορεί να έχει κάθε συστάδα) και την min_rad (καθορίζει ότι η απόσταση κάθε στοιχείου μίας συστάδας από το κέντρο αυτής, δε θα είναι μεγαλύτερη του max_rad).

• Αλγόριθμος ενέργειας δεσμού (Bond energy algorithm–BEA). Χρησιμοποιείται στη σχεδίαση βάσεων δεδομένων για τον καθορισμό του τρόπου ομαδοποίησης των δεδομένων και της φυσικής τοποθέτησης τους στον δίσκο.



Γράφημα 11. Αρχικοποίηση K–means

Γράφημα 12. Μήτρα συγγένειας για τον αλγόριθμο BEA

• CLARA (Clustering large applications). Βασίζεται στη δειγματοποίηση (samling) και (σε αντίθεση με τον PAM) δε βρίσκει αντικείμενα αντιπροσώπους για ολόκληρο το σύνολο δεδομένων, αλλά λαμβάνει με τυχαίο τρόπο ένα δείγμα του συνόλου δεδομένων, εφαρμόζει μόνο στο δείγμα τον PAM και βρίσκει τα medoids (αντικείμενα αντιπρόσωποι) του δείγματος.

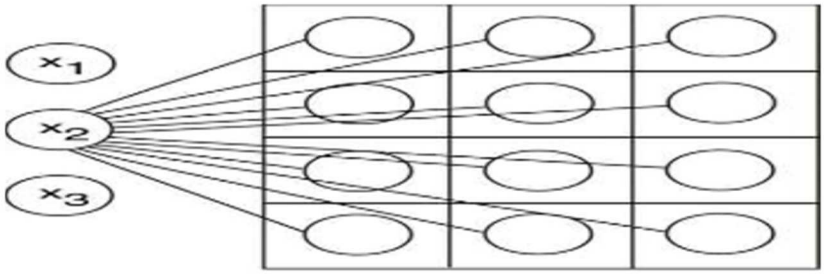
• CLARANS (Clustering large applications based on randomized search). Προσπαθεί να συνδυάσει τους αλγορίθμους PAM και CLARA εκτελώντας κάθε φορά αναζήτηση, μόνο σε ένα υποσύνολο του συνόλου των δεδομένων, ενώ δεν περιορίζεται σε κάποιο δείγμα μία δεδομένη στιγμή. Ο αριθμός των γειτόνων που μπορούν να δοκιμασθούν τυχαία, περιορίζεται από την παράμετρο maxneighbor (μέγιστος αριθμός γειτόνων που μπορούν να εξετασθούν).

• Συσταδοποίηση με νευρωνικά δίκτυα. Χρησιμοποιούν μη επιβλεπόμενη μάθηση, προσπαθώντας να βρουν χαρακτηριστικά, στα δεδομένα που χαρακτηρίζουν την επιθυμητή έξοδο (ψάχνουν για συστάδες παρόμοιων δεδομένων). Υπάρχουν δυο βασικοί τύποι η:

➢ μη ανταγωνιστική μάθηση (noncompetitive learning) στην οποία το βάρος μεταξύ δύο κόμβων αλλάζει, προκειμένου να είναι ανάλογο και των δυο τιμών εξόδου.

➢ ανταγωνιστική μάθηση (competitive learning) στην οποία οι κόμβοι επιτρέπεται να ανταγωνίζονται μεταξύ τους και «ο νικητής τα παίρνει όλα». Στη διάρκεια της εκπαίδευσης, οι κόμβοι του επιπέδου εξόδου συσχετίζονται με συγκεκριμένες πλειάδες του συνόλου των δεδομένων εισόδου, γεγονός που οδηγεί στη συσταδοποίηση αυτών των πλειάδων σε μία συστάδα.

Εδώ ανήκει και ο αυτοοργανωσιακός χάρτης (SOFM) (self–organizing feature map) η λειτουργία του οποίου βασίζεται στον τρόπο λειτουργίας των τεχνητών νευρωνικών δικτύων, δηλαδή η διέγερση των νευρώνων επηρεάζει και τη διέγερση άλλων νευρώνων που βρίσκονται κοντά τους. Οι νευρώνες που βρίσκονται σε μεγάλες μεταξύ τους αποστάσεις, φαίνεται να αλληλοαναχαιτίζονται και να έχουν συγκεκριμένες διακριτές μεταξύ τους λειτουργίες. Ο δημοφιλέστερος SOFM είναι ο αυτοοργανωσιακός χάρτης Kohonen, που χρησιμοποιείται ευρέως στα εμπορικά προϊόντα εξόρυξης γνώσης για την εκτέλεση συσταδοποίησης.



Γράφημα 13. Δίκτυο Kohonen

Οι αλγόριθμοι μηχανικής μάθησης, κατηγοριοποιούνται ως εξής:

• **Επιβλεπόμενη μάθηση** (supervised learning), όπου ο αλγόριθμος κατασκευάζει μία συνάρτηση που απεικονίζει δεδομένες εισόδους, σε γνωστές επιθυμητές εξόδους (σύνολο εκπαίδευσης) με απώτερο στόχο τη γενίκευση αυτής της συνάρτησης και για εισόδους με άγνωστη έξοδο (σύνολο ελέγχου). Π.χ. κατηγοριοποίηση

• **Μη επιβλεπόμενη μάθηση** (unsupervised learning), όπου ο αλγόριθμος κατασκευάζει ένα μοντέλο για κάποιο σύνολο εισόδων χωρίς να γνωρίζει τις επιθυμητές εξόδους για το σύνολο της εκπαίδευσης. Π.χ. συσταδοποίηση

Η ανάλυση των αλγορίθμων μηχανικής μάθησης, είναι ένας κλάδος της στατιστικής που ονομάζεται θεωρία μάθησης.

11.2.8 Συσταδοποίηση σε μεγάλες βάσεις δεδομένων

Οι περισσότεροι αλγόριθμοι που έχουν μέχρι τώρα παρουσιασθεί, ίσως είναι ακατάλληλοι για μεγάλες βάσεις δεδομένων, διότι θεωρούν πως όλα τα στοιχεία υπάρχουν ταυτόχρονα και ότι όλες οι παραδοχές είναι ρεαλιστικές, γεγονός που αντίκειται στις δυναμικές βάσεις δεδομένων. Οι αλγόριθμοι που θεωρούνται κατάλληλοι για συσταδοποίηση σε μεγάλες βάσεις δεδομένων, εξετάζουν από ένα θέμα σχετιζόμενο με την εκτέλεση της συσταδοποίησης, σε ένα περιβάλλον μάθησης και προκειμένου να είναι αποδοτικοί πρέπει να:

- μην απαιτούν περισσότερο από μία σάρωση της βάσης δεδομένων
- μπορούν να διακόπτονται προσωρινά, να σταματούν οριστικά και να συνεχίζουν την εκτέλεση τους μετά από προσωρινή διακοπή
- είναι online, δηλαδή να είναι ικανοί, κατά τη διάρκεια εκτέλεσης τους, να παράσχουν πληροφόρηση σχετικά με την κατάσταση και τη βελτίωση τους, κάθε χρονική στιγμή
- επεξεργάζονται κάθε πλειάδα, μία μόνο φορά
- δουλεύουν με περιορισμένη κύρια μνήμη
- μπορούν, καθώς προσθαφαιρούνται δεδομένα από τη βάση δεδομένων, να ενημερώνουν αυξητικά τα αποτελέσματα τους.
- είναι ικανοί να εκτελούν διαφορετικές τεχνικές για σάρωση της βάσης δεδομένων.

Αλγόριθμος Cure. Είναι αλγόριθμος συσταδοποίησης που συνδυάζει τεχνικές τυχαίας δειγματοποίησης (sampling) και τμηματοποίησης (partitioning) με βασικά χαρακτηριστικά:

- είναι εύρωστος στην παρουσία outliers
- αναγνωρίζει συστάδες αυθαιρέτων (arbitrary) σχημάτων
- οι απαιτήσεις του σε χώρο αποθήκευσης είναι γραμμική συνάρτηση του αριθμού των στοιχείων εισόδου.

Όσο το μέγεθος του συνόλου των δεδομένων εισόδου αυξάνεται, τόσο αυξάνεται και η υπολογιστική πολυπλοκότητα του αλγορίθμου Cure.

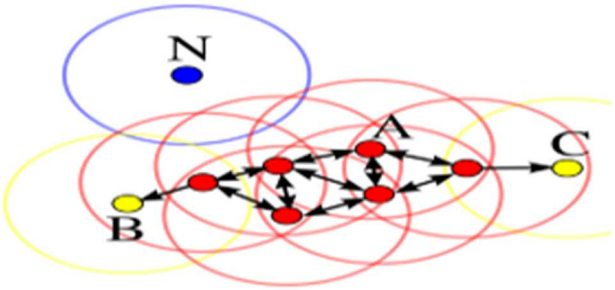
Αλγόριθμος BIRCH (Balanced iterative reducing and clustering using hierarchies). Κατασκευάσθηκε για τη συσταδοποίηση μεγάλου πλήθους μετρικών δεδομένων, είναι αυξητικός, ιεραρχικός, γραμμικός ως προς τον χώρο και ως προς τον χρόνο, εφαρμόζεται μόνο σε αριθμητικά δεδομένα και κύριο γνώρισμα του είναι η χρήση του χαρακτηριστικού συσταδοποίησης (clustering feature–CF), μίας τριάδας που περιέχει πληροφορία για τη συστάδα.

11.2.9 Συσταδοποίηση βασισμένη στην πυκνότητα

Οι αλγόριθμοι που βασίζονται στην πυκνότητα (Density–based), θεωρούν τις συστάδες σαν πυκνές περιοχές αντικειμένων που χωρίζονται από περιοχές χαμηλής πυκνότητας. Στην κατηγορία αυτή, ανήκουν οι παρακάτω αλγόριθμοι.

- **Αλγόριθμος DBSCAN (Density based spatial clustering of applications with noise).** Στηρίζεται στο ότι η περιοχή που εκτείνεται γύρω από το αντικείμενο μίας συστάδας (γειτονιά αντικειμένου) σε συγκεκριμένη ακτίνα (Eps), πρέπει να περιέχει έναν ελάχιστο αριθμό αντικειμένων–σημείων (MinPts). Τα σημεία κατηγοριοποιούνται σε κεντρικά (core points), προσεγγίσιμα ή πυκνά–προσεγγίσιμα

(density–reachable points) και ακραία (outliers). Δεν απαιτεί τον εκ των προτέρων προσδιορισμό του αριθμού των συστάδων, έχει καλή ευαισθησία στον θόρυβο, δεν επηρεάζεται από ακραίες τιμές, μπορεί να καταλήξει σε αυθαίρετα σχήματα συστάδων και να εντοπίσει ακόμα και μια συστάδα ευρισκόμενη γύρω από κάποια άλλη.



Γράφημα 14. Δημιουργία συστάδων με χρήση DBSCAN για MinPts=3

• **Αλγόριθμος DENCLUE (Density based clustering).** Διαχειρίζεται καλά, σύνολα δεδομένων που περιέχουν θόρυβο, επιτρέπει την ανακάλυψη συστάδων με περίεργες γεωμετρίες (arbitrary shape clusters) σε πολυδιάστατα σύνολα δεδομένων και μοντελοποιεί τη συνολική πυκνότητα των σημείων, με αναλυτικό τρόπο, ως το άθροισμα των συναρτήσεων επιρροής (influence functions), του συνόλου δεδομένων. Η συνάρτηση επιρροής, περιγράφει την επίδραση ενός σημείου, από το σύνολο των δεδομένων, στη γειτονιά του και density attractors είναι τα τοπικά μέγιστα της συνολικής συνάρτησης πυκνότητας.

11.2.10 Συσταδοποίηση υποχώρων (Subspace clustering)

Εξετάζει προβλήματα που προκύπτουν από τις υψηλές διαστάσεις (High dimensionality) δεδομένων, εξαιτίας των οποίων και του θορύβου, σχεδόν κάθε περιοχή στον χώρο έχει χαμηλή πυκνότητα σημείων. Στην κατηγορία αυτή, ανήκουν οι παρακάτω αλγόριθμοι.

• **Αλγόριθμος CLIQUE (Clustering in quest).** Προχωρά από χαμηλότερης σε υψηλότερης διάστασης υποχώρους και ανακαλύπτει σε κάθε υπόχωρο (subspace) τις πυκνές περιοχές.

• **Αλγόριθμος Proclus.** Αναζητά τέτοια υποσύνολα διαστάσεων ώστε, τα σημεία δεδομένων να είναι πολύ πυκνά ομαδοποιημένα, στους αντίστοιχους υποχώρους. Ο χρήστης, καθορίζει τον αριθμό των συστάδων και τον μέσο αριθμό διαστάσεων, ανά συστάδα. Στις ασύνδετες συστάδες, δεν υπολογίζει τμηματοποίηση των στοιχείων.

11.2.11 Αλγόριθμοι συσταδοποίησης για σύνολα δεδομένων με λεκτικές τιμές

Οι περισσότεροι από τους κλασσικούς αλγορίθμους συσταδοποίησης, εφαρμόζονται σε σύνολα δεδομένων με αριθμητικές τιμές. Όμως, οι εφαρμογές εξόρυξης γνώσης περιλαμβάνουν και μη αριθμητικά δεδομένα (categorical data), των οποίων η μετατροπή σε αριθμητικές τιμές δεν είναι πάντα αποτελεσματική. Οι παρακάτω αλγόριθμοι, εφαρμόζονται αποτελεσματικά για τη συσταδοποίηση μη αριθμητικών δεδομένων.

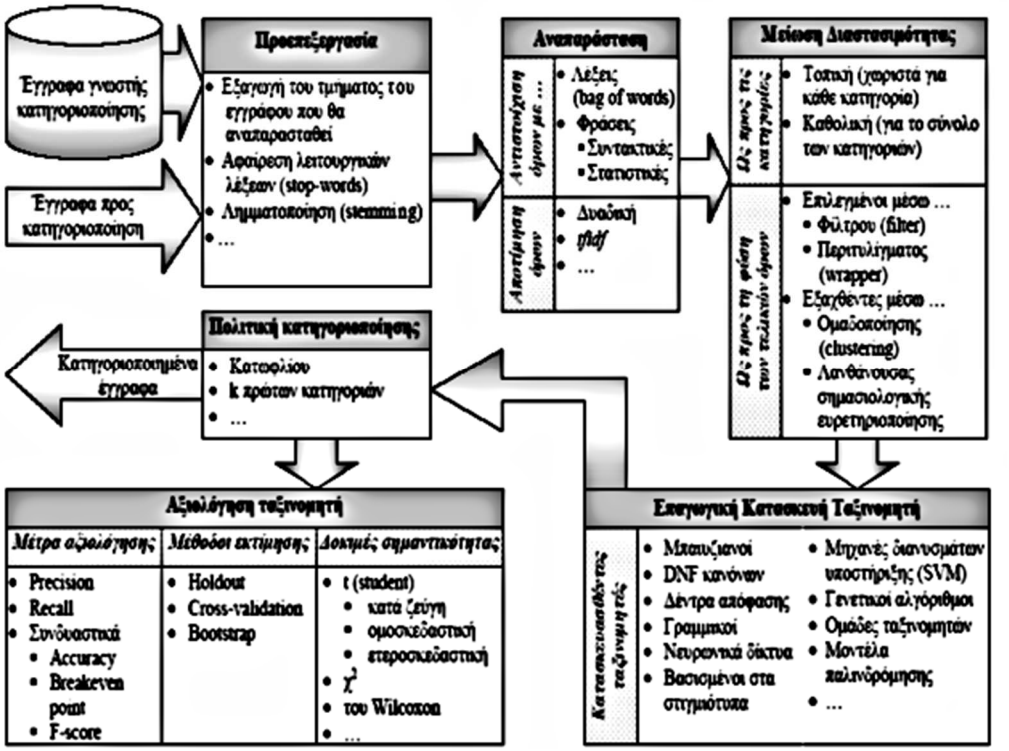
• **Αλγόριθμος ROCK (Robust clustering algorithm for categorical attributes).** Ιεραρχικός αλγόριθμος συσταδοποίησης που μπορεί να χειριστεί αποτελεσματικά Boolean και κατηγορικά δεδομένα, εισάγοντας δυο έννοιες, του γείτονα (neighbor) και των δεσμών (links).

• **Αλγόριθμοι βασισμένοι στον K–Means για λεκτικά δεδομένα.** Διαιρετικοί (Partitional) αλγόριθμοι συσταδοποίησης με κυριότερους τους K–prototypes και K–

mode, που σχεδιάστηκαν από τον Huang. Είναι αποτελεσματικότεροι για μεγάλα σύνολα λεκτικών δεδομένων, σε σχέση με άλλους ιεραρχικούς αλγορίθμους οι οποίοι λόγω της πολυπλοκότητας τους καθίστανται μη αποδοτικοί, για μεγάλα σύνολα δεδομένων.

• **Αλγόριθμος K–prototypes.** Σχεδιάστηκε για συσταδοποίηση μεγάλων συνόλων βάσεων δεδομένων, με αριθμητικές και λεκτικές τιμές. Ορίζει ένα μέτρο ανομοιότητας, το οποίο λαμβάνει υπόψη του γνωρίσματα με αριθμητικές και με λεκτικές τιμές.

• **Αλγόριθμος K–mode.** Αποτελεί απλούστευση του ανωτέρω αλγορίθμου, καθώς λαμβάνει υπόψη του μόνο δεδομένα με λεκτικές τιμές. Βασίζεται στον K–Means στον οποίο έχουν γίνει τρεις τροποποιήσεις.



Γράφημα 15. Σχεδιασμός συστήματος αυτόματης κατηγοριοποίησης κειμένου

11.2.12 Ιεραρχική συσταδοποίηση βασισμένη σε γράφους

Αυτοί οι αλγόριθμοι, στοχεύουν στον συνδυασμό των πλεονεκτημάτων της ιεραρχικής συσταδοποίησης και της συσταδοποίησης που βασίζεται σε γράφους. Στην κατηγορία αυτή, ανήκουν οι παρακάτω αλγόριθμοι.

• **CHAMELEON.** Συσσωρευτικός (agglomerative), ιεραρχικός, αλγόριθμος που βρίσκει τις συστάδες του συνόλου δεδομένων, χρησιμοποιώντας έναν αλγόριθμο δυο φάσεων και μετρά την ομοιότητα δυο συστάδων, που βασίζονται σε ένα δυναμικό μοντέλο.

• **C^2P.** Πρόσφατος αλγόριθμος συσταδοποίησης, αποτελούμενος από δύο βασικές φάσεις, που συνδυάζει τα χαρακτηριστικά των ιεραρχικών αλγορίθμων και της θεωρίας γράφων, εκμεταλλευόμενος τις δομές ευρετηρίων και την επεξεργασία των ερωτήσεων του πιο κοντινού ζευγαριού (Closest pair queries–CPQ) στις χωρικές βάσεις δεδομένων.

11.2.13 Αποδοτικότητα της συσταδοποίησης

Η αποδοτικότητα της διαδικασίας συσταδοποίησης, επηρεάζεται από την κλιμάκωση (scaling) και από τη διαβάθμιση (weighting). Η κλιμάκωση, έχει να κάνει με διαφορετικές μεταβλητές που μετρούνται σε διαφορετικές μονάδες μέτρησης. Η κλιμάκωση, γίνεται μετατρέποντας όλες τις μεταβλητές (γνωρίσματα) και τις τιμές τους 0 έως 1 ή –1 έως 1. Η στάθμιση, είναι το διαφορετικό ενδιαφέρον που μπορεί να έχουν κάποιες μεταβλητές σε σχέση με άλλες.

11.3. Κανόνες συσχέτισης

Συνήθως, οι κανόνες συσχέτισης χειρίζονται κατηγορικά δεδομένα, μπορούν όμως να εφαρμοσθούν και σε πεδία που δεν είναι κατηγορικά. Ενδιαφέρον δεν παρουσιάζουν όλες οι συσχετίσεις, αλλά μόνο όσες είναι σημαντικές. Η σημαντικότητα μετριέται από δυο χαρακτηριστικά, την **εμπιστοσύνη** (μετρά την ισχύ του κανόνα) και την **υποστήριξη** (μετρά πόσο συχνά εμφανίζεται στη βάση δεδομένων). Η πιο κοινή προσέγγιση για την εύρεση των κανόνων συσχέτισης είναι η διάσπαση του προβλήματος σε δύο μέρη:

- Εύρεση συχνών στοιχειοσυνόλων (frequent item set)
- Δημιουργία κανόνων από τα συχνά στοιχειοσύνολα.

11.3.1 Ταξινόμηση αλγορίθμων

Οι αλγόριθμοι εύρεσης των κανόνων συσχέτισης, ταξινομούνται ως προς τον στόχο, τον τύπο δεδομένων, την πηγή δεδομένων, την τεχνική, τη στρατηγική των σημειοσυνόλων, τη στρατηγική συναλλαγών, τη δομή των στοιχειοσυνόλων, τις δομές των δεδομένων συναλλαγών, τη βελτιστοποίηση, την αρχιτεκτονική και τη στρατηγική παραλληλισμού. Οι σημαντικότεροι αλγόριθμοι που χρησιμοποιούνται για την εύρεση κανόνων συσχέτισης είναι:

- Ο **αλγόριθμος Apriori**. Είναι ο γνωστότερος αλγόριθμος που χρησιμοποιείται στα περισσότερα εμπορικά προϊόντα και η βασική του ιδέα είναι η δημιουργία υποψηφίων στοιχειοσυνόλων, ενός συγκεκριμένου μεγέθους και στη συνέχεια, η σάρωση της βάσης δεδομένων για να μετρηθούν και παρουσιασθούν, αν αυτά είναι συχνά.
- Ο **αλγόριθμος AprioriTID.** Επειδή έχει καλύτερα αποτελέσματα στα τελευταία περάσματα από ότι στα πρώτα, συνηθίζεται να χρησιμοποιείται ο Apriori για τις πρώτες επαναλήψεις. Ο αλγόριθμος **AprioriHybrid**, συνδυάζει τα πλεονεκτήματα των δυο προαναφερθέντων αλγορίθμων.
- **Ο αλγόριθμος δειγματοληψίας.** Μειώνει τον αριθμό των περασμάτων της βάσης σε ένα (στην καλύτερη περίπτωση) ή δύο (στη χειρότερη περίπτωση). Το δείγμα της βάσης δεδομένων, επιλέγεται έτσι ώστε να μπορεί να χωρέσει στη μνήμη.

11.3.2 Παράλληλοι και κατανεμημένοι αλγόριθμοι

Οι περισσότεροι αλγόριθμοι, επιδιώκουν τον παραλληλισμό των δεδομένων (Data parallelism) ή τον παραλληλισμό των εργασιών (Task parallelism). Με τον παραλληλισμό των εργασιών, οι υποψήφιοι (και η βάση δεδομένων) διαμερίζονται και καταμετρώνται ξεχωριστά σε κάθε επεξεργαστή. Ο αλγόριθμος DDA (Data distribution algorithm) είναι παραλληλισμού εργασιών. Ο αλγόριθμος CDA (Count distribution algorithm) είναι παραλληλισμού δεδομένων.

12. Κατηγορίες των μεθόδων εξόρυξης δεδομένων

Υπάρχουν δυο βασικοί στόχοι, η πρόβλεψη σε μεγάλα σύνολα δεδομένων (π.χ. μελλοντική αξία) και η εφαρμογή τεχνικών περιγραφής.

Η **πρόβλεψη,** έχει ως στόχο την εκτίμηση της συμπεριφοράς κάποιων μεταβλητών που παρουσιάζουν ενδιαφέρον και βασίζονται – επηρεάζονται από τη συμπεριφορά άλλων μεταβλητών. Ένα προβλεπτικό μοντέλο (predictive model), χρησιμοποιώντας γνωστά αποτελέσματα που έχει βρει από τα άλλα δεδομένα, επιχειρεί πρόβλεψη για τις τιμές των άλλων δεδομένων.

Η **περιγραφή,** αναπαριστά τα δεδομένα μίας πολύπλοκης βάσης δεδομένων, με κατανοητό και αξιοποιήσιμο στόχο. Σε ότι αφορά την εξόρυξη γνώσης, η περιγραφή τείνει να είναι σημαντικότερη από την πρόβλεψη. Αντιθέτως, σε ότι αφορά την αναγνώριση προτύπων και την εφαρμογή της μηχανικής μάθησης, η πρόβλεψη είναι σημαντικότερη. Ένα περιγραφικό μοντέλο (descriptive model) όπως η συσταδοποίηση, η παρουσίαση συνόψεων, οι κανόνες συσχετίσεων και η ανακάλυψη ακολουθιών, αναγνωρίζει πρότυπα ή συσχετίσεις στα δεδομένα, λειτουργώντας ως το μέσο που διερευνά τις ιδιότητες των υπό εξέταση δεδομένων, αλλά δεν προβλέπει νέες ιδιότητες.

Εξόρυξη γνώσης	
Προβλεπτικά μοντέλα	**Περιγραφικά μοντέλα**
Κατηγοριοποίηση (Classification)	Συσταδοποίηση (Clustering)
Παλινδρόμηση (Regression)	Παρουσίαση συνόψεων (Summarization)
Ανάλυση χρονοσειρών (Time series analysis)	Κανόνες συσχέτισης (Association rules)
Πρόβλεψη (Prediction)	Ανακάλυψη ακολουθιών (Sequence discovery)

Πίνακας 5. Μοντέλα εξόρυξης δεδομένων

13. Απαιτήσεις της εξόρυξης δεδομένων

Αρχικά, προσδιορίζεται το είδος των χαρακτηριστικών που αναμένεται να έχει ένα σύστημα εξόρυξης και ακολούθως οι απαιτήσεις που πρέπει να ληφθούν υπόψη, στην ανάπτυξη των τεχνικών εξόρυξης δεδομένων. Οι σημαντικότερες απαιτήσεις είναι:

- **Χειρισμός των διαφορετικών τύπων δεδομένων.** Παρόλο που υπάρχουν διαφορετικοί τύποι και βάσεις δεδομένων, οι βάσεις είναι στη συντριπτική τους πλειοψηφία συγγενείς μεταξύ τους, άρα ένα σύστημα εξόρυξης δεδομένων πρέπει να εφαρμόζεται αποτελεσματικά σε διαφορετικούς τύπους δεδομένων (περιέχουν σύνθετους τύπους δεδομένων όπως δομές δεδομένων, σύνθετα αντικείμενα, υπερκείμενο, στοιχεία πολυμέσων, χωροχρονικά στοιχεία).
- **Απόδοση και εξελιξιμότητα των αλγορίθμων εξόρυξης δεδομένων.** Οι αλγόριθμοι πρέπει να προσαρμοσθούν κατάλληλα, στα μεγάλα σύνολα δεδομένων, ώστε να είναι αποτελεσματική η εξόρυξη γνώσης και αποδεκτός – αναμενόμενος ο χρόνος εκτέλεσης τους. Αλγόριθμοι με εκθετική και με πολυωνυμική πολυπλοκότητα, δε θεωρούνται κατάλληλοι.
- **Χρησιμότητα, βεβαιότητα και εκφραστικότητα των αποτελεσμάτων της εξόρυξης δεδομένων**. Η εξορυγμένη γνώση, πρέπει να παρουσιάζει με ακριβή τρόπο τα περιεχόμενα των βάσεων δεδομένων. Ο θόρυβος και οι outliers που αντιπροσωπεύουν τις εξαιρέσεις, πρέπει να αντιμετωπίζονται αποτελεσματικά.
- **Διαλογική ανακάλυψη γνώσης στα πολυεννοιολογικά επίπεδα.** Επιτρέπει στον χρήστη να αλληλοεπιδράσει με ένα σύστημα, καθορίζοντας τις ερωτήσεις της εξόρυξης δεδομένων, προκειμένου αλλάζοντας την εστίαση των δεδομένων να κατευθύνει τη διαδικασία εξόρυξης σε πολλαπλά επίπεδα.
- **Εξόρυξη γνώσης από διαφορετικές πηγές δεδομένων.** Το τεράστιο πλήθος δεδομένων, η υψηλή κατανομή τους και η υπολογιστική τους πολυπλοκότητα, οδηγούν στην ανάπτυξη παραλλήλων και κατανεμημένων αλγορίθμων εξόρυξης δεδομένων.

14. Λειτουργίες της εξόρυξης δεδομένων

Οι λειτουργίες της εξόρυξης δεδομένων είναι:

- **Επαγωγή.** Εμφανίζεται συχνά σε εφαρμογές τεχνητής νοημοσύνης και χρησιμοποιείται προκειμένου μία πολύ εξειδικευμένη γνώση, να οδηγήσει σε γενικότερες πληροφορίες.
- **Συμπίεση**, διότι στοχεύει στην περιγραφή ορισμένων μόνο χαρακτηριστικών ενός συνόλου δεδομένων, από ένα γενικό μοντέλο.
- Η περιγραφή μίας μεγάλης βάσης δεδομένων, μπορεί να εκληφθεί ως μία προσπάθεια αποκάλυψης κρυμμένων πληροφοριών που περιέχονται στα δεδομένα.
- Η επίδραση του μεγέθους μίας μεγάλης βάσης δεδομένων και η ικανότητα ανάπτυξης ενός αφηρημένου μοντέλου, δύναται να θεωρηθούν ως ένας τύπος προβλήματος αναζήτησης, που περιλαμβάνει έναν τρόπο ορισμού της διαδικασίας υποβολής των ερωτήσεων στη βάση δεδομένων, με απώτερο στόχο την ανάπτυξη μίας γλώσσας ερωτήσεων (όπως στην SQL) η οποία να περιλαμβάνει πολλούς διαφορετικούς τύπους ερωτήσεων.

15. Υλοποίηση της εξόρυξης γνώσης από δεδομένα

Οι παρακάτω παράγοντες, σχετίζονται άμεσα με την υλοποίηση της εξόρυξης γνώσης από δεδομένα.

- **Ανθρώπινη αλληλεπίδραση.** Είναι απαραίτητη η αλληλεπίδραση ανάμεσα στους ειδικούς του πεδίου εφαρμογής (ταυτοποιούν τα δεδομένα εκπαίδευσης και ορίζουν τα επιθυμητά αποτελέσματα) και στους ειδικούς της συγκεκριμένης τεχνικής εξόρυξης γνώσης (μορφοποιούν τις ερωτήσεις και βοηθούν στην ερμηνεία των αποτελεσμάτων).
- **Υπερπροσαρμογή (over–fitting).** Εμφανίζεται όταν το μοντέλο δεν ταιριάζει σε μελλοντικές καταστάσεις της βάσης δεδομένων.
- **Ακραίες τιμές (outliers).** Στις πολύ μεγάλες βάσεις δεδομένων, υπάρχουν πολλές καταχωρήσεις που δεν ταιριάζουν με το αναπτυχθέν μοντέλο.
- **Ερμηνεία αποτελεσμάτων.** Πρέπει να γίνεται από ειδικούς, άλλως δε θα είναι κατανοητή από τον μέσο χρήστη.
- **Οπτικοποίηση αποτελεσμάτων.** Είναι ιδιαίτερα χρήσιμη για την ευκολότερη κατανόηση των αποτελεσμάτων.
- **Μεγάλα σύνολα δεδομένων.** Όταν ένας αλγόριθμος εξόρυξης γνώσης που έχει σχεδιασθεί για μικρά σύνολα δεδομένων, εφαρμοσθεί σε πολύ μεγάλα σύνολα δεδομένων, τότε εμφανίζονται προβλήματα, καθιστώντας τον αναποτελεσματικό. Προς αντιμετώπιση του ανωτέρω προβλήματος, χρησιμοποιούνται η δειγματοληψία και ο παραλληλισμός.
- **Υψηλές διαστάσεις.** Πρόκειται για πρόβλημα γνωστό και ως κατάρα των υψηλών διαστάσεων (dimensionality curse), εννοώντας ότι από τα πολλά γνωρίσματα που εμπλέκονται, είναι δύσκολο να καθορισθεί επακριβώς ποια πρέπει να χρησιμοποιηθούν. Προς αντιμετώπιση του ανωτέρω προβλήματος, χρησιμοποιείται η μείωση των υψηλών διαστάσεων (dimensionality reduction) δηλαδή η μείωση των γνωρισμάτων, χωρίς να είναι πάντα αυτό εύκολο.
- **Δεδομένα πολυμέσων.** Επειδή οι περισσότεροι αλγόριθμοι που χρησιμοποιούνται στοχεύουν στα παραδοσιακά είδη δεδομένων (αριθμητικά, χαρακτήρες, κείμενο), η χρήση των δεδομένων πολυμέσων τους περιπλέκει ή τους καθιστά ακατάλληλους.
- **Ελλιπή δεδομένα.** Αν συμπληρωθούν κατ΄ εκτίμηση, ενδέχεται να οδηγήσουν σε λάθος αποτελέσματα κατά την εξόρυξη γνώσης από τα δεδομένα.

• **Άσχετα δεδομένα.** Πιθανόν ορισμένα από τα γνωρίσματα που περιλαμβάνονται στη βάση δεδομένων, να μην έχουν ενδιαφέρον για την εξόρυξη γνώσης που πραγματοποιείται.

• **Δεδομένα με θόρυβο.** Αν μερικές τιμές των γνωρισμάτων είναι άκυρες ή λανθασμένες, τότε πρέπει να διορθωθούν, πριν τρέξει ο αλγόριθμος εξόρυξης γνώσης από δεδομένα.

• **Δεδομένα που αλλάζουν.** Οι βάσεις δεδομένων, δεν είναι στατικές όπως τις θεωρούν οι αλγόριθμοι εξόρυξης γνώσης. Άρα, κάθε φορά που η βάση δεδομένων επικαιροποιήται, απαιτείται να ξανατρέξει ο αλγόριθμος.

16. Μέτρα αξιολόγησης

Το μέτρο ROI (Return on investment–ROI), εξετάζει τη διαφορά ανάμεσα στο κόστος εφαρμογής της τεχνικής και στην εξοικονόμηση (ή στα κέρδη) που προκύπτει από τη χρήση αυτής της τεχνικής. Η διαφορά, μπορεί να μετρηθεί ως αύξηση των πωλήσεων ή ως μείωση της διαφημιστικής δαπάνης ή ως άθροισμα και των δύο.

17. Εξόρυξη γνώσης από τη σκοπιά των βάσεων δεδομένων

Περιλαμβάνει την εξέταση όλων των ειδών των εφαρμογών και των τεχνικών εξόρυξης γνώσης από δεδομένα, με ιδιαίτερη έμφαση στα παρακάτω θέματα:

• Κλιμάκωση
• Πραγματικά δεδομένα
• Ενημέρωση
• Ευχρηστία

Οι περισσότερες εργασίες εξόρυξης γνώσης από δεδομένα τη σημερινή εποχή, βασίζονται σε συγκεκριμένους αλγορίθμους που πραγματοποιούν ξεχωριστά κάθε πράξη. Στις αρχές της δεκαετίας του 1970 εμφανίστηκαν τα συστήματα διαχείρισης βάσεων δεδομένων–ΣΔΒΔ (Database management systems–DBMS), η επιτυχία των οποίων οφείλονταν εν μέρει, στον αφηρημένο ορισμό των δεδομένων και στους βασικούς κανόνες προσπέλασης για έναν μικρό πυρήνα λειτουργικών απαιτήσεων.

Ένας από τους λόγους που οι σχεσιακές βάσεις δεδομένων είναι δημοφιλείς στην εποχή μας, είναι η ανάπτυξη της SQL η οποία είναι εύκολη στη χρήση και έχει γίνει βιομηχανικό πρότυπο γλώσσας, που υλοποιείται από όλους τους κατασκευαστές ΣΔΒΔ.

18. Λογισμικά εξόρυξης δεδομένων

Συνοπτικά, παρουσιάζονται τα σημαντικότερα προγράμματα που εφαρμόζονται στις τεχνικές εξόρυξης δεδομένων.

• **Clementine.** Θεωρείτε μία από τις πλέον αξιόλογες πλατφόρμες, αποτελεί προϊόν της SPSS, περιλαμβάνει πολλά εργαλεία εξόρυξης δεδομένων, δίνει έμφαση στη μοντελοποίηση της πρόβλεψης, περιλαμβάνει εξόρυξη κειμένου (Text mining) και εξόρυξη από τον παγκόσμιο ιστό (Web mining).

• **R**. Είναι ελεύθερη στατιστική γλώσσα προγραμματισμού, με πάρα πολλές δυνατότητες, στηρίζεται στην ανάπτυξη προγραμμάτων μέσω πακέτων (packages), τα οποία διατίθεται επίσης ελεύθερα.

• **SAS**. Από τα σημαντικότερα προγράμματα πρακτικής εφαρμογής της εξόρυξης δεδομένων (και κειμένου). Προσφέρει λύσεις επιχειρηματικής ευφυΐας (business intelligence) και προβλεπτικής ανάλυσης (predictive analysis).

• **Orange**. Βιβλιοθήκη αντικειμένων πυρήνα και ρουτινών της C++, περιέχει ρουτίνες για εισαγωγή και για χειρισμό δεδομένων και αποτελεί συλλογή από υποπρογράμματα (modules) python.

• **Rapid Miner**. Είναι το πρώην YALE (Yet another learning environment), δηλαδή ένα περιβάλλον για μηχανική εκμάθηση, για εξόρυξη δεδομένων, για προβλεπτική και επιχειρησιακή ανάλυση που χρησιμοποιείται στην έρευνα, στην εκπαίδευση, στην ανάπτυξη εφαρμογών στη γρήγορη προτυποποίηση και σε βιομηχανικές εφαρμογές (τραπεζικό και ασφαλιστικό κλάδο, έρευνα αγοράς, τηλεπικοινωνίες, αεροπλοΐα, αυτοκινητιστική βιομηχανία, τομείς ενέργειας/ φαρμάκου/ πληροφορικής/ εμπορίου).

• **WEKA (Waikato environment for knowledge analysis)**. Προϊόν ελεύθερου λογισμικού, διατίθεται δωρεάν με άδεια χρήσης GPL, ευρύτατα χρησιμοποιούμενο στην εξόρυξη δεδομένων. Δημιουργήθηκε το 1992 εξαιτίας της αισθητής ανάγκης για ένα πρόγραμμα που θα επέτρεπε στους ερευνητές να χρησιμοποιούν τεχνολογικά εξελιγμένες τεχνικές μάθησης. Το 1997 δημιουργήθηκε η $1^{η}$ έκδοση του, πλήρως βασισμένη σε Java.

•**Tanagra**. Δωρεάν λογισμικό εξόρυξης δεδομένων γραμμένο σε Java και C++, που διαθέτει αλγορίθμους εξόρυξης δεδομένων, διερευνητικής ανάλυσης δεδομένων, μηχανικής εκμάθησης και στατιστικής εκμάθησης.

• **Rattle**. Πακέτο λογισμικού ανοικτού κώδικα, σχεδιάστηκε ειδικά για να διευκολύνει τη μετάβαση από την απλή και βασική εξόρυξη δεδομένων, στην εξελιγμένη ανάλυση δεδομένων, χρησιμοποιώντας μία ισχυρή προγραμματιστική στατιστική γλώσσα, την R.

• **Carrot²**. Ανοικτού κώδικα μηχανή αναζήτησης για συσταδοποίηση αποτελεσμάτων, γραμμένο σε Java, μπορεί αυτόματα να συσταδοποιήσει μικρές συλλογές εγγράφων.

• **XL-MINER**. Είναι μία add–in εφαρμογή του Excel. Παρέχει ένα σύνολο σαφών τεχνικών ανάλυσης, στηριζόμενων σε στατιστικές μεθόδους μηχανικής εκμάθησης. Είναι ικανό να χειριστεί πολύ μεγάλα σύνολα δεδομένων, που μπορεί να ξεπερνούν τη χωρητικότητα του excel.

19. Πλάνες της εξόρυξης δεδομένων

Παρακάτω περιγράφονται κάποιες εκ των κυριότερων λανθασμένων αντιλήψεων που υπάρχουν γύρω από την εξόρυξη των δεδομένων:

Πλάνη 1. Υπάρχουν εργαλεία εξόρυξης δεδομένων που μπορούμε να απελευθερώσουμε στα αποθετήρια δεδομένων μας και να βρούμε απαντήσεις στα προβλήματά μας.

Πραγματικότητα. Δεν υπάρχουν αυτόματα εργαλεία εξόρυξης δεδομένων που θα λύσουν μηχανικά τα προβλήματά μας ενώ περιμένουμε. Αντίθετα, η εξόρυξη δεδομένων είναι μια διαδικασία που ενσωματώνεται στο συνολικό επιχειρηματικό ή ερευνητικό σχέδιο δράσης.

Πλάνη 2. Η διαδικασία εξόρυξης δεδομένων είναι αυτόνομη, απαιτώντας ελάχιστη ή καθόλου ανθρώπινη εποπτεία.

Πραγματικότητα. Η εξόρυξη δεδομένων δεν είναι μαγεία. Χωρίς εξειδικευμένη ανθρώπινη εποπτεία, η τυφλή χρήση λογισμικού εξόρυξης δεδομένων θα δώσει μόνο λάθος απάντηση, σε λάθος ερώτηση, που εφαρμόζεται σε λάθος τύπο δεδομένων. Επιπλέον, η λάθος ανάλυση είναι χειρότερη από την έλλειψη ανάλυσης, καθώς οδηγεί σε συστάσεις πολιτικής που πιθανότατα θα αποδειχθούν δαπανηρές αποτυχίες. Ακόμα και μετά από την ανάπτυξη του μοντέλου, η εισαγωγή νέων δεδομένων συχνά απαιτεί ενημέρωση του μοντέλου. Η συνεχής παρακολούθηση της ποιότητας και άλλα μέτρα αξιολόγησης, πρέπει να αξιολογούνται από ανθρώπινους αναλυτές.

Πλάνη 3. Η εξόρυξη δεδομένων αποδίδει αρκετά γρήγορα.

Πραγματικότητα. Τα ποσοστά απόδοσης ποικίλλουν, ανάλογα με το κόστος εκκίνησης, το κόστος προσωπικού ανάλυσης, το κόστος προετοιμασίας αποθήκευσης των δεδομένων και ούτω καθεξής.

Πλάνη 4. Τα πακέτα λογισμικού εξόρυξης δεδομένων, είναι διαισθητικά και εύχρηστα.

Πραγματικότητα. Και πάλι, η ευκολία χρήσης ποικίλλει. Ωστόσο, ανεξάρτητα από το τι ισχυρίζονται ορισμένες διαφημίσεις προμηθευτών λογισμικού, δεν μπορούμε απλώς να αγοράσουμε κάποιο λογισμικό εξόρυξης δεδομένων, να το εγκαταστήσουμε, να καθίσουμε και να το παρακολουθήσουμε να λύνει όλα τα προβλήματά μας. Π.χ. οι αλγόριθμοι απαιτούν συγκεκριμένες μορφές δεδομένων, οι οποίες μπορεί να απαιτούν σημαντική προεπεξεργασία. Οι αναλυτές δεδομένων, πρέπει να συνδυάζουν τη γνώση του αντικειμένου με την αναλυτική σκέψη και την εξοικείωση με το συνολικό επιχειρηματικό ή ερευνητικό μοντέλο.

Πλάνη 5. Η εξόρυξη δεδομένων, θα εντοπίσει τις αιτίες των επιχειρηματικών ή ερευνητικών μας προβλημάτων.

Πραγματικότητα. Η διαδικασία ανακάλυψης γνώσης θα βοηθήσει να αποκαλύψουμε πρότυπα συμπεριφοράς. Και πάλι, εναπόκειται σε εμάς στους ανθρώπους να εντοπίσουμε τις αιτίες.

Πλάνη 6. Η εξόρυξη δεδομένων, θα καθαρίσει αυτόματα την ακατάστατη βάση δεδομένων μας.

Πραγματικότητα. Λοιπόν, όχι αυτόματα. Ως προκαταρκτική φάση στη διαδικασία εξόρυξης δεδομένων, η προετοιμασία δεδομένων συχνά ασχολείται με δεδομένα που δεν έχουν εξεταστεί ή χρησιμοποιηθεί εδώ και χρόνια. Άρα, οι οργανισμοί που ξεκινούν μια νέα επιχείρηση εξόρυξης δεδομένων, συχνά αντιμετωπίζουν το πρόβλημα δεδομένων που βρίσκονται εδώ και χρόνια, είναι ξεπερασμένα και χρειάζονται σημαντική ενημέρωση.

Πλάνη 7. Η εξόρυξη δεδομένων, παρέχει πάντα θετικά αποτελέσματα.

Πραγματικότητα. Δεν υπάρχει εγγύηση θετικών αποτελεσμάτων κατά την εξόρυξη δεδομένων για αξιοποιήσιμη γνώση. Η εξόρυξη δεδομένων, δεν αποτελεί πανάκεια για την επίλυση επιχειρηματικών προβλημάτων. Αλλά, αν χρησιμοποιηθεί σωστά, από άτομα που κατανοούν τα εμπλεκόμενα μοντέλα, τις απαιτήσεις δεδομένων και τους συνολικούς στόχους του έργου, μπορεί πράγματι να προσφέρει αξιοποιήσιμα και εξαιρετικά κερδοφόρα αποτελέσματα.

20. Πλεονεκτήματα της εξόρυξης δεδομένων

Η εξόρυξη δεδομένων, έχει γίνει η κορυφαία προτεραιότητα πολλών διαχειριστών πληροφορικής. Ποια είναι τα πλεονεκτήματα που έχει υποσχεθεί να προσφέρει; Η απάντηση σε αυτό το ερώτημα δεν θα δοθεί μέχρι να διερευνηθούν διεξοδικά όλες οι πιθανές εφαρμογές της. Σε αυτό το πρώιμο στάδιο, τα ακόλουθα πλεονεκτήματα της εξόρυξης δεδομένων είναι ευρέως αναγνωρισμένα:

- **Παροχή καλύτερων πληροφοριών για την επίτευξη ανταγωνιστικού πλεονεκτήματος.** Αυτό το πλεονέκτημα, είναι το κύριο κίνητρο για την εξόρυξη δεδομένων και έχει αναφερθεί επανειλημμένα σε διάφορα άρθρα. Η εξόρυξη δεδομένων, έχει μια ισχυρή αναλυτική ικανότητα να παράγει πληροφορίες, οι οποίες επιτρέπουν σε έναν οργανισμό να κατανοήσει καλύτερα τον εαυτό του, τους πελάτες του και την αγορά στην οποία ανταγωνίζεται. Όταν χρησιμοποιείται ως εργαλείο μάρκετινγκ, συχνά οδηγεί σε ισχυρότερο ανταγωνιστικό πλεονέκτημα, μια προσέγγιση πωλήσεων βασισμένη σε τεκμήρια, ένα σχέδιο μάρκετινγκ προσανατολισμένο στον πελάτη, μικρότερους κύκλους πωλήσεων και μειωμένο λειτουργικό κόστος.

• **Προσθήκη αξίας σε μια αποθήκη δεδομένων.** Μια αποθήκη δεδομένων, από μόνη της είναι ένα μεγάλο αποθετήριο μη δομημένων δεδομένων και η εξόρυξη δεδομένων είναι η διαδικασία ανάλυσης των δεδομένων και μετατροπής τους σε χρήσιμες πληροφορίες. Οι οργανισμοί, έχουν βιώσει μια απόσβεση της επένδυσής τους σε αποθήκες δεδομένων από 10 ως 70 φορές, μετά από την προσθήκη στοιχείων εξόρυξης δεδομένων.

• **Επίλυση προβλημάτων έρευνας.** Σε πολλές κοινωνικές επιστήμες και επιχειρηματικές καταστάσεις, η διεξαγωγή πραγματικών πειραμάτων είναι σχεδόν αδύνατη. Η εξόρυξη δεδομένων, είναι σε θέση να παρέχει σε αυτές τις ερευνητικές ατζέντες ένα πιο περιορισμένο σύνολο υποθέσεων εργασίας, για περαιτέρω διερεύνηση με βάση μεγάλα, αδόμητα σύνολα δεδομένων.

• **Αύξηση της λειτουργικής αποδοτικότητας.** Η ικανότητα της εξόρυξης δεδομένων να οργανώνει και να αναλύει γρήγορα μια μεγάλη ομάδα δεδομένων, έχει αυξήσει δραματικά την αποδοτικότητα του χώρου εργασίας. Επιτρέπει στους χρήστες να δημιουργούν σύνθετες οικονομικές καταστάσεις σε λίγα λεπτά, σε σύγκριση με εβδομάδες με τις παραδοσιακές μεθόδους.

• **Παροχή ευελιξίας στη χρήση δεδομένων.** Με την εξόρυξη δεδομένων, οι χρήστες έχουν αποκτήσει τον έλεγχο των δεδομένων. Αντί να αφήνουν το σύστημα να προωθεί τα δεδομένα, είναι πλέον σε θέση να αντλούν τα δεδομένα που χρειάζονται. Μπορούν να αφήσουν τη φαντασία τους να καλπάσει και να χειραγωγήσουν τα δεδομένα με διάφορους τρόπους, για να απαντήσουν στις ερωτήσεις τους. Η νέα εύκολη διεπαφή των εργαλείων εξόρυξης δεδομένων και η τεχνολογία πελάτη/διακομιστή, έχουν κάνει τις πληροφορίες άμεσα προσβάσιμες από μεμονωμένους χρήστες.

• **Μείωση του λειτουργικού κόστους**. Τα σύγχρονα εργαλεία εξόρυξης δεδομένων, είναι κατασκευασμένα από εξαιρετικά εξελιγμένα στοιχεία υλικού και λογισμικού. Επιτρέπουν σε αυτά τα εργαλεία να αναλύουν αποτελεσματικά, τεράστια σύνολα δεδομένων, με μειωμένο λειτουργικό κόστος. Η περίπτωση της Bank of America, έδειξε ότι μετά από την εφαρμογή της τεχνολογίας εξόρυξης δεδομένων, το κόστος ανά ερώτημα μειώθηκε από 2.430 $ σε μόλις 24 $.

• **Έτοιμο προς χρήση.** Σε αντίθεση με τις παραδοσιακές μεθόδους ανάλυσης δεδομένων, η εξόρυξη δεδομένων δεν απαιτεί σχεδόν καθόλου προεπεξεργασία των δεδομένων πριν από την ανάλυση. Μπορεί να χρησιμοποιήσει ένα μείγμα αριθμητικών, κατηγορικών και ημερολογιακών δεδομένων και μπορεί να ανεχθεί δεδομένα που λείπουν και είναι θορυβώδη. Τα αποτελέσματα, έχουν τη μορφή έτοιμων προς χρήση επιχειρηματικών κανόνων, χωρίς σχεδόν καμία στατιστική εμπειρογνωμοσύνη και εικασίες.

21. Μειονεκτήματα της εξόρυξης δεδομένων

Παρά τα πολλά πλεονεκτήματα, η εξόρυξη δεδομένων δεν είναι χωρίς μειονεκτήματα. Οι οργανισμοί θα αντιμετωπίσουν αυτά τα προβλήματα, κατά την ανάπτυξη της τεχνολογίας εξόρυξης δεδομένων. Η κατανόησή τους, βοηθά τους διαχειριστές των πληροφοριακών συστημάτων να έχουν ρεαλιστικές προσδοκίες και να προετοιμαστούν για πιθανά αρνητικά αποτελέσματα.

• **Δεν υπάρχει οριστική απάντηση.** Η εξόρυξη δεδομένων, αποφέρει χρήσιμες γνώσεις και ενδείξεις, αλλά όχι οριστικές απαντήσεις. Οι οριστικές απαντήσεις, πρέπει να επιτευχθούν μέσω πολύ πιο αυστηρών επιστημονικών πειραματισμών. Οι εμπειρίες από τη Wall Street, έχουν δείξει ότι αυτή η τεχνολογία μπορεί να μην ξεπεράσει τις παραδοσιακές μεθόδους. Άρα, οι χρήστες πρέπει να έχουν ρεαλιστικές προσδοκίες για τα αποτελέσματα της εξόρυξης δεδομένων.

• **Υψηλό κόστος.** Το κόστος εφαρμογής της εξόρυξης δεδομένων είναι πολύ υψηλό. Άρα, ίσως να μην είναι κατάλληλο σε ορισμένα επιχειρηματικά περιβάλλοντα.

• **Πολύπλοκο και χρονοβόρο έργο.** Η εμπειρία από τους προγραμματιστές των συστημάτων εξόρυξης δεδομένων, έχει δείξει ότι χρειάζεται πολύς χρόνος για να γίνει σωστά το έργο. Οι προγραμματιστές, προτείνουν να επικεντρωθούμε στη σταδιακή ανάπτυξη και στα οφέλη.

• **Απόρρητο.** Τα λεπτομερή δεδομένα σχετικά με τα άτομα που χρησιμοποιούνται στην εξόρυξη δεδομένων, ενδέχεται να συνεπάγονται παραβίαση της ιδιωτικότητας. Αυτό το πρόβλημα επιδεινώνεται, όταν εμπλέκεται ο παγκόσμιος ιστός, επειδή οι λεπτομερείς προσωπικές πληροφορίες είναι εύκολα προσβάσιμες και μπορούν να πέσουν σε λάθος χέρια.

• **Υψηλές απαιτήσεις γνώσης από τον χρήστη.** Παρά την ολοένα και πιο απλή διεπαφή και την αυτοματοποίηση των διαδικασιών σκέψης, η εξόρυξη δεδομένων είναι πιο κατάλληλη για άτομα με υπόβαθρο στη στατιστική, στην επιχειρησιακή έρευνα και στις διοικητικές επιστήμες. Η ευκολία χρήσης, γίνεται ένας κρίσιμος παράγοντας για την προσέλκυση περισσότερων επιχειρήσεων να επενδύσουν σε αυτήν την τεχνολογία.

• **Μη διαχειρίσιμη βάση δεδομένων.** Πολλοί συγγραφείς, έχουν προτείνει ότι οι οργανισμοί πρέπει να αυξήσουν τρομερά το μέγεθος των βάσεων δεδομένων τους ώστε να κάνουν εξόρυξη δεδομένων. Ωστόσο, ορισμένοι ανησυχούν ότι αυτό θα οδηγήσει σε μη διαχειρίσιμες και περιττές βάσεις δεδομένων.

• **Λανθασμένες πληροφορίες από σφάλματα στα δεδομένα.** Τα τεράστια δεδομένα που χρησιμοποιούνται στην εξόρυξη δεδομένων, περιέχουν αναπόφευκτα ανθρώπινα λάθη. Οι πληροφορίες που παράγονται, πρέπει να χρησιμοποιούνται με προσοχή για την αποφυγή αγωγών σε τομείς όπως οι προσλήψεις. Οι ειδικοί, προτείνουν τη χρήση μόνο σχετικών πληροφοριών για την εξόρυξη προς μείωση τέτοιων κινδύνων.

22. Εφαρμογές της εξόρυξης δεδομένων

Όπου υπάρχουν δεδομένα, υπάρχουν και εφαρμογές εξόρυξης τους. Πριν αναφερθούμε σε διαφόρους τομείς που εφαρμόζεται η εξόρυξη των δεδομένων, θα κάνουμε μια ομαδοποίηση των εφαρμογών στις ακόλουθες 6 σχετικές και όχι ασύνδετες ομάδες:

• **Πρόβλεψη και περιγραφή**. Είναι σημαντικό η κάθε επιχείρηση να μπορεί να προβλέπει ορισμένες πτυχές της ζήτησης (μελλοντικές τάσεις της αγοράς), ώστε να βοηθήσει στον σχεδιασμό του μέλλοντος.

• **Μάρκετινγκ σχέσεων**, καθώς για πολλές επιχειρήσεις το μάρκετινγκ έχει μεγάλη σημασία. Σήμερα, πολλές σύγχρονες επιχειρήσεις επικεντρώνονται στις σχέσεις με τους πελάτες ως το κεντρικό επιχειρηματικό ζήτημα. Αυτές οι επιχειρήσεις, ασχολούνται με την αύξηση της αξίας για τους πελάτες, μέσω της ανάλυσης του κύκλου ζωής του πελάτη. Ένας νέος όρος που χρησιμοποιείται στις επιχειρήσεις, ονομάζεται διαχείριση των σχέσεων με τους πελάτες (CRM–customer relationship management) και παρέχει νέες ευκαιρίες στις επιχειρήσεις για να ενεργήσουν με βάση τις έννοιες του μάρκετινγκ σχέσεων. Συνήθως, οι πελάτες έχουν μια αξία ζωής, όχι μόνο την αξία μιας μόνο πώλησης. Η σχέση με τους πελάτες, ασχολείται με την οικοδόμηση μιας μακροπρόθεσμης σχέσης μεταξύ ενός πελάτη και μιας επιχείρησης. Αυτό μπορεί να περιλαμβάνει την αναγνώριση των πελατών, την αξία τους, τη διατήρηση τους και την ανάπτυξη τους. Η εξόρυξη δεδομένων μπορεί να βοηθήσει στην ανάλυση των προφίλ των πελατών, στην ανακάλυψη εναυσμάτων πωλήσεων, στον εντοπισμό των κρίσιμων ζητημάτων που καθορίζουν την αφοσίωση των πελατών και στη βελτίωση της διατήρησης τους. Αυτό, περιλαμβάνει την ανάλυση των προφίλ

των πελατών και τη βελτίωση των σχεδίων άμεσου μάρκετινγκ. Στο μάρκετινγκ σχέσεων, μπορούν να χρησιμοποιηθούν ορισμένες τεχνικές εξόρυξης δεδομένων. Π.χ. μπορεί να είναι δυνατή η χρήση της ανάλυσης των συστάδων για τον εντοπισμό πελατών κατάλληλων για διασταυρούμενες πωλήσεις, άλλων προϊόντων.

• **Δημιουργία του προφίλ των πελατών**, η οποία μπορεί να φανεί πολύ χρήσιμη για την πλειοψηφία των επιχειρήσεων. Η δημιουργία του προφίλ των πελατών, αποτελεί τη διαδικασία χρήσης σχετικών και διαθέσιμων πληροφοριών για την περιγραφή των χαρακτηριστικών μιας ομάδας πελατών και για τον εντοπισμό των διακριτικών παραγόντων τους, από άλλους πελάτες και οδηγούς για τις αγοραστικές τους αποφάσεις. Η δημιουργία προφίλ, μπορεί να βοηθήσει μια επιχείρηση να εντοπίσει τους πιο πολύτιμους πελάτες της, ώστε να μπορεί να διαφοροποιήσει τις ανάγκες και τις αξίες τους. Μπορεί να βοηθήσει στην αύξηση της αξίας ζωής ορισμένων πελατών. Τα προφίλ των πελατών, μπορεί να περιλαμβάνουν πληροφορίες σχετικά με το πώς οι πελάτες ξοδεύουν χρήματα, που και τι τείνουν να αγοράζουν, ποιοι είναι οι πιο κερδοφόροι πελάτες και πού θα μπορούσε η επιχείρηση να βρει περισσότερους υποψήφιους πελάτες, παρόμοιους με τους πιο πολύτιμους.

• **Τμηματοποίηση των πελατών**, η οποία μπορεί να είναι χρήσιμη σε πολλές περιπτώσεις και σχετίζεται με το μάρκετινγκ σχέσεων και τη δημιουργία του προφίλ των πελατών που αναφέρθηκαν παραπάνω. Ουσιαστικά, είναι μια διαδικασία εύρεσης υποομάδων παρόμοιων ανθρώπων, μέσα σε ένα σύνολο δεδομένων και μπορεί να είναι χρήσιμη στο μάρκετινγκ. Μπορεί να υπάρχουν διάφοροι τύποι τμηματοποίησης (δημογραφικός, προσανατολιστικός, συμπεριφορικός). Αποτελεί έναν τρόπο αξιολόγησης και προβολής των ατόμων, με βάση την κατάσταση και τις ανάγκες τους.

Η εξόρυξη δεδομένων, μπορεί να χρησιμοποιηθεί για την τμηματοποίηση των πελατών, για την αύξηση της διατήρησης τους και για την προώθηση των διασταυρούμενων πωλήσεων υπηρεσιών. Μπορεί να χρησιμοποιηθεί για την τμηματοποίηση των υποκαταστημάτων και για την αξιολόγηση της απόδοσης διαφόρων τραπεζικών καναλιών (τηλεφωνική ή ηλεκτρονική τραπεζική). Επιπλέον, μπορεί να χρησιμοποιηθεί για την κατανόηση και για την πρόβλεψη της συμπεριφοράς και της κερδοφορίας των πελατών, για την ανάπτυξη νέων προϊόντων και υπηρεσιών και για την αποτελεσματική προώθηση νέων προσφορών.

• **Αναγνώριση και ανίχνευση ακραίων τιμών**, συμπεριλαμβανομένων των περιπτώσεων απάτης ή ασυνήθιστων περιπτώσεων, με τις τεχνικές εξόρυξης δεδομένων να χρησιμοποιούνται σε περιπτώσεις ανίχνευσης απάτης όπως:

- Απάτη με πιστωτικές κάρτες
- Απάτη κοινωνικής πρόνοιας
- Απάτη ασφάλισης
- Ιατρική απάτη
- Φορολογική απάτη
- Τελωνειακή απάτη και λαθρεμπόριο
- Απάτη τηλεπικοινωνιών

• **Σχεδιασμός και προώθηση ιστοσελίδων**, καθώς η εξόρυξη δεδομένων από το διαδίκτυο μπορεί να χρησιμοποιηθεί για να ανακαλυφθεί ο τρόπος με τον οποίο οι χρήστες πλοηγούνται σε έναν ιστότοπο και τα αποτελέσματα μπορούν να βοηθήσουν στη βελτίωση του σχεδιασμού του ιστότοπου και στην καλύτερη προβολή του στο διαδίκτυο. Η εξόρυξη δεδομένων, μπορεί να χρησιμοποιηθεί σε διασταυρούμενες πωλήσεις, προτείνοντας σε έναν πελάτη του διαδικτύου είδη που μπορεί να τον ενδιαφέρουν, συσχετίζοντας ιδιότητες σχετικά με τον πελάτη ή είδη που έχει παραγγείλει το άτομο, με μια βάση δεδομένων ειδών που έχουν παραγγείλει άλλοι πελάτες στο παρελθόν. Η εξόρυξη δεδομένων, όντας ένας κλάδος που καθοδηγείται

έντονα από τις εφαρμογές, έχει σημειώσει μεγάλες επιτυχίες σε πολλές εφαρμογές και τομείς. Είναι αδύνατο να απαριθμηθούν όλες οι εφαρμογές και όλοι αυτοί οι τομείς όπου παίζει κρίσιμο ρόλο. Για να καταδείξουμε τη σημασία των εφαρμογών της, θα αναφερθούμε παρακάτω σε μερικούς τομείς που βρίσκει εφαρμογή.

22.1 Τραπεζικές και χρηματοοικονομικές υπηρεσίες

Οι τραπεζικές και χρηματοοικονομικές υπηρεσίες, είναι ένας ταχέως μεταβαλλόμενος ανταγωνιστικός κλάδος. Χρησιμοποιεί την εξόρυξη δεδομένων για μια ποικιλία εργασιών, όπως η δημιουργία του προφίλ των πελατών για την καλύτερη κατανόησή τους, η αναγνώριση της απάτης, η αξιολόγηση των κινδύνων στις ασφάλειες, στα προσωπικά και στεγαστικά δάνεια και για καλύτερη πρόβλεψη των τιμών των μετοχών, των επιτοκίων, των συναλλαγματικών ισοτιμιών και των τιμών των βασικών προϊόντων. Στον τομέα της αξιολόγησης της πιστοληπτικής ικανότητας, η εξόρυξη δεδομένων μπορεί να βοηθήσει στη δημιουργία ενός αυτοματοποιημένου συστήματος υποστήριξης αποφάσεων, που θα επέτρεπε στις εταιρείες παροχής πιστωτικών καρτών ή δανείων να αξιολογούν γρήγορα και με ακρίβεια τον κίνδυνο και να εγκρίνουν ή να απορρίπτουν μια αίτηση. Μια άλλη εφαρμογή της εξόρυξης δεδομένων στις τραπεζικές και χρηματοοικονομικές υπηρεσίες, είναι ο εντοπισμός πελατών υψηλής αξίας, καθώς δεν είναι όλοι οι πελάτες εξίσου κερδοφόροι.

Συχνά, μπορεί να ισχύει ένας κανόνας 80–20 ή παρόμοιος, δηλαδή το 20% των πελατών μπορεί να είναι υπεύθυνο για το 80% του κέρδους ή το 30% των πελατών μπορεί να αντιπροσωπεύει το 70% του κέρδους. Επίσης, η απώλεια πελατών αποτελεί τομέα σημαντικού ενδιαφέροντος στον τραπεζικό χώρο, όπως και αλλού (λιανικό εμπόριο, τηλεπικοινωνίες). Πολλές μελέτες εξόρυξης δεδομένων έχουν διερευνήσει την απώλεια πελατών. Μια εταιρεία, συνήθως έχει μια αίσθηση του πόσο αξίζει πραγματικά να διατηρήσει έναν πελάτη που είναι πιθανό να φύγει, επομένως και της κλίμακας της προσπάθειας που θα ήταν κατάλληλη για μια εκστρατεία διατήρησης.

22.2 Βιολογία, ιατρική επιστήμη και υγειονομική περίθαλψη

Η βιολογία, η ιατρική και η υγειονομική περίθαλψη, παράγουν τεράστια δεδομένα, σε εκθετική κλίμακα. Τα βιοϊατρικά δεδομένα, λαμβάνουν πολλές μορφές (απεικόνιση, κινητή υγεία, ηλεκτρονικά αρχεία υγείας). Με τη διαθεσιμότητα πιο αποτελεσματικών μεθόδων ψηφιακής συλλογής, οι βιοϊατρικοί επιστήμονες και οι κλινικοί ιατροί βρίσκονται αντιμέτωποι με όλο και μεγαλύτερα σύνολα δεδομένων και προσπαθούν να επινοήσουν δημιουργικούς τρόπους ώστε να εξετάσουν αυτό το βουνό δεδομένων και να τα κατανοήσουν. Πράγματι, τα δεδομένα που θεωρούνταν μεγάλα τώρα φαίνονται μικρά, καθώς η ποσότητα των δεδομένων που συλλέγονται σε μια μόνο ημέρα από έναν ερευνητή, μπορεί να ξεπεράσει αυτό που θα μπορούσε να είχε συλλεχθεί κατά τη διάρκεια της καριέρας του πριν από μια δεκαετία. Αυτός ο κατακλυσμός βιοϊατρικών πληροφοριών, απαιτεί νέα σκέψη σχετικά με τον τρόπο με τον οποίο τα δεδομένα μπορούν να διαχειριστούν, να αναλυθούν για την περαιτέρω επιστημονική κατανόηση και για τη βελτίωση της υγειονομικής περίθαλψης. Η εξόρυξη βιοϊατρικών δεδομένων, περιλαμβάνει πολλές απαιτητικές εργασίες εξόρυξης δεδομένων, όπως η εξόρυξη μαζικών γονιδιωματικών και πρωτεωμικών δεδομένων αλληλουχίας, η εξόρυξη συχνών υπογραφικών μοτίβων για την ταξινόμηση των βιολογικών δεδομένων, η εξόρυξη ρυθμιστικών δικτύων, ο χαρακτηρισμός και η πρόβλεψη αλληλεπιδράσεων πρωτεΐνης–πρωτεΐνης, η ταξινόμηση και η προγνωστική ανάλυση ιατρικών εικόνων, η εξόρυξη βιολογικού κειμένου, η κατασκευή δικτύου βιολογικών πληροφοριών από δεδομένα βιοκειμένου, η εξόρυξη ηλεκτρονικών αρχείων υγείας και η εξόρυξη βιοϊατρικών δικτύων. Η εξόρυξη δεδομένων, έχει

χρησιμοποιηθεί εντατικά και εκτενώς από πολλούς οργανισμούς υγειονομικής περίθαλψης και μπορεί να ωφελήσει σε μεγάλο βαθμό όλα τα εμπλεκόμενα μέρη. Π.χ. η εξόρυξη δεδομένων μπορεί να βοηθήσει τους ασφαλιστές της υγειονομικής περίθαλψης, να εντοπίσουν απάτες και καταχρήσεις, τους οργανισμούς υγειονομικής περίθαλψης να λάβουν αποφάσεις διαχείρισης των σχέσεων με τους πελάτες, τους ιατρούς να εντοπίσουν αποτελεσματικές θεραπείες και βέλτιστες πρακτικές και τους ασθενείς να λαμβάνουν καλύτερες και πιο προσιτές υπηρεσίες υγειονομικής περίθαλψης. Γενικά, οι εφαρμογές της εξόρυξης δεδομένων στις υπηρεσίες υγειονομικής περίθαλψης περιλαμβάνουν, μεταξύ άλλων, τα ακόλουθα:

- Μοντελοποίηση των αποτελεσμάτων της υγείας και πρόβλεψη των αποτελεσμάτων των ασθενών
- Μοντελοποίηση της κλινικής γνώσης των συστημάτων υποστήριξης των αποφάσεων
- Βιοπληροφορική
- Φαρμακευτική έρευνα
- Επιχειρηματική ευφυΐα όπως η διαχείριση της υγειονομικής περίθαλψης, η διαχείριση των σχέσεων με τους πελάτες, η ανίχνευση της απάτης και της κατάχρησης
- Έλεγχος των λοιμώξεων
- Κατάταξη των νοσοκομείων
- Αναγνώριση των ασθενών υψηλού κινδύνου
- Αξιολόγηση της αποτελεσματικότητας της θεραπείας

22.3 Ανίχνευση απάτης και πρόληψη εγκλήματος

Οι ακραίες τιμές, μπορούν να ανακαλυφθούν με τη χρήση της εξόρυξης δεδομένων από την τεράστια ποσότητα δεδομένων. Οι ακραίες τιμές, μπορούν να αναγνωριστούν ανακαλύπτοντας τα σπάνια μοτίβα στα δεδομένα. Τα σπάνια μοτίβα γενικά, ανήκουν σε δόλια/εγκληματική δραστηριότητα. Ως εκ τούτου, με τη βοήθεια της ανίχνευσης των ακραίων τιμών ή/και της εξόρυξης σπάνιων μοτίβων, οι πιθανές απάτες μπορούν να εντοπιστούν και να προβλεφθούν, ώστε να αποτραπεί η εμφάνιση εγκλημάτων. Έχουν δημοσιευτεί αρκετές μελέτες περιπτώσεων, σχετικά με τη χρήση τεχνικών εξόρυξης δεδομένων στην ανάλυση δεδομένων εγκληματικότητας.

Ένα εξαιρετικά αμφιλεγόμενο θέμα στην εγκληματολογία, είναι το κατά πόσον υπάρχουν τύποι εγκληματιών. Πολλοί εγκληματολόγοι είναι κατά της πιθανότητας των ξεχωριστών εγκληματικών τύπων, ενώ άλλοι υποστηρίζουν σθεναρά την ύπαρξή τους. Πιθανώς η εξόρυξη των δεδομένων να μπορεί να δώσει μια απάντηση στο δίλημμα αυτό. Οι υπηρεσίες πληροφοριών, συλλέγουν και αναλύουν πληροφορίες για τη διερεύνηση των τρομοκρατικών δραστηριοτήτων. Μια πρόκληση για τις υπηρεσίες επιβολής του νόμου και για τις υπηρεσίες πληροφοριών, είναι η δυσκολία ανάλυσης του μεγάλου όγκου των δεδομένων που εμπλέκονται σε εγκληματικές και τρομοκρατικές δραστηριότητες. Η εξόρυξη δεδομένων, καθιστά εύκολη, βολική και πρακτική την εξερεύνηση πολύ μεγάλων βάσεων δεδομένων για οργανισμούς.

Οι διαφορετικές τεχνικές εξόρυξης δεδομένων, χρησιμοποιούνται στην εξόρυξη δεδομένων εγκληματικότητας. Η εξαγωγή οντοτήτων, χρησιμοποιείται για την αυτόματη αναγνώριση προσώπου, διεύθυνσης, οχήματος, ναρκωτικών και προσωπικών περιουσιακών στοιχείων από αστυνομικές αφηγηματικές αναφορές. Τεχνικές συσταδοποίησης, χρησιμοποιούνται για την αυτόματη συσχέτιση διαφορετικών αντικειμένων, όπως πρόσωπα, οργανισμούς, οχήματα κ.λπ., στα αρχεία εγκλημάτων. Η ανίχνευση της απόκλισης, εφαρμόζεται στην ανίχνευση απάτης, στην ανίχνευση εισβολών δικτύου και σε άλλες αναλύσεις εγκλημάτων που περιλαμβάνουν την ανίχνευση μη φυσιολογικών δραστηριοτήτων. Η ταξινόμηση, χρησιμοποιείται για

την ανίχνευση ανεπιθύμητης αλληλογραφίας μέσω email και για την εύρεση των δημιουργών που στέλνουν ανεπιθύμητα emails. Ο συγκριτής συμβολοσειρών, χρησιμοποιείται για την ανίχνευση παραπλανητικών πληροφοριών στο ποινικό μητρώο. Η ανάλυση των κοινωνικών δικτύων, χρησιμοποιείται για την ανάλυση των ρόλων και των συσχετίσεων των εγκληματιών, σε ένα εγκληματικό δίκτυο.

22.4 Επιχειρηματική ευφυΐα

Είναι κρίσιμο για τις επιχειρήσεις, να αποκτήσουν καλύτερη κατανόηση του εμπορικού πλαισίου του οργανισμού τους, όπως οι πελάτες τους, η αγορά, η προσφορά, οι πόροι και οι ανταγωνιστές. Οι τεχνολογίες επιχειρηματικής ευφυΐας (BI–business intelligence) παρέχουν ιστορικές, τρέχουσες και προγνωστικές απόψεις των επιχειρηματικών δραστηριοτήτων. Παραδείγματα περιλαμβάνουν αναφορές, διαδικτυακή αναλυτική επεξεργασία, διαχείριση επιχειρηματικής απόδοσης, ανταγωνιστική ευφυΐα, συγκριτική αξιολόγηση και προγνωστική ανάλυση. Χωρίς την εξόρυξη δεδομένων, πολλές επιχειρήσεις ενδέχεται να μην είναι σε θέση να πραγματοποιήσουν αποτελεσματική ανάλυση της αγοράς, να συγκρίνουν τα σχόλια των πελατών για παρόμοια προϊόντα, να ανακαλύψουν τα δυνατά και τα αδύνατα σημεία των ανταγωνιστών τους, να διατηρήσουν πελάτες υψηλής αξίας και να λάβουν έξυπνες επιχειρηματικές αποφάσεις. Σαφώς, η εξόρυξη δεδομένων είναι ο πυρήνας της επιχειρηματικής ευφυΐας. Τα διαδικτυακά εργαλεία αναλυτικής επεξεργασίας στην επιχειρηματική ευφυΐα, βασίζονται στην αποθήκευση δεδομένων και στην πολυδιάστατη εξόρυξη τους. Οι τεχνικές ταξινόμησης και πρόβλεψης, είναι ο πυρήνας της προγνωστικής ανάλυσης στην επιχειρηματική ευφυΐα, για την οποία υπάρχουν πολλές εφαρμογές στην ανάλυση αγορών, προμηθειών και πωλήσεων. Η συσταδοποίηση, παίζει κεντρικό ρόλο στη διαχείριση των σχέσεων με τους πελάτες, η οποία τους ομαδοποιεί με βάση τα κοινά τους χαρακτηριστικά. Χρησιμοποιώντας τεχνικές πολυδιάστατης σύνοψης, μπορούμε να κατανοήσουμε καλύτερα τα χαρακτηριστικά της κάθε ομάδας πελατών και να αναπτύξουμε προσαρμοσμένα προγράμματα ανταμοιβής τους.

22.5 Μηχανές αναζήτησης ιστού

Μια μηχανή αναζήτησης ιστού, είναι ένας εξειδικευμένος διακομιστής υπολογιστών που αναζητά πληροφορίες στον ιστό. Τα αποτελέσματα της αναζήτησης ενός ερωτήματος χρήστη, συχνά επιστρέφονται ως λίστα (μερικές φορές ονομάζονται επιτυχίες). Οι επιτυχίες, μπορεί να αποτελούνται από ιστοσελίδες, από εικόνες και από άλλους τύπους αρχείων. Ορισμένες μηχανές αναζήτησης, αναζητούν και επιστρέφουν δεδομένα που είναι διαθέσιμα σε δημόσιες βάσεις δεδομένων ή σε ανοιχτούς καταλόγους. Οι μηχανές αναζήτησης, διαφέρουν από τους καταλόγους ιστού στο ότι οι κατάλογοι ιστού συντηρούνται από ανθρώπινους συντάκτες, ενώ οι μηχανές αναζήτησης λειτουργούν αλγοριθμικά ή με ένα μείγμα αλγοριθμικής και ανθρώπινης εισόδου. Οι μηχανές αναζήτησης, θέτουν μεγάλες προκλήσεις στην εξόρυξη δεδομένων.

Πρώτον, πρέπει να χειρίζονται μια τεράστια και συνεχώς αυξανόμενη ποσότητα δεδομένων. Συνήθως, τέτοια δεδομένα δε μπορούν να υποστούν επεξεργασία χρησιμοποιώντας μία ή λίγες μηχανές. Αντ' αυτού, οι μηχανές αναζήτησης συχνά χρειάζεται να χρησιμοποιούν υπολογιστικά σύννεφα, τα οποία αποτελούνται από χιλιάδες ή και εκατοντάδες χιλιάδες υπολογιστές που εξορύσσουν συλλογικά την τεράστια ποσότητα δεδομένων. Η κλιμάκωση των μεθόδων εξόρυξης δεδομένων, μέσω υπολογιστικών σύννεφων και μεγάλων κατανεμημένων συνόλων δεδομένων, είναι ένας τομέας ενεργής έρευνας και ανάπτυξης.

Δεύτερον, οι μηχανές αναζήτησης ιστού, συχνά πρέπει να χειρίζονται διαδικτυακά δεδομένα. Μια μηχανή αναζήτησης, μπορεί να είναι σε θέση να κατασκευάσει ένα μοντέλο εκτός σύνδεσης, σε τεράστια σύνολα δεδομένων. Για να το κάνει αυτό, μπορεί να κατασκευάσει έναν ταξινομητή ερωτημάτων που αντιστοιχίζει ένα ερώτημα αναζήτησης σε προκαθορισμένες κατηγορίες με βάση το θέμα του ερωτήματος (δηλαδή, αν το ερώτημα αναζήτησης «μήλο» προορίζεται για την ανάκτηση πληροφοριών σχετικά με ένα φρούτο ή με μια μάρκα υπολογιστών).

Ακόμα και αν ένα μοντέλο κατασκευάζεται εκτός σύνδεσης, η προσαρμογή του στο διαδίκτυο πρέπει να είναι αρκετά γρήγορη, ώστε να απαντά σε ερωτήματα χρηστών σε πραγματικό χρόνο. Μια άλλη πρόκληση είναι η διατήρηση και η σταδιακή ενημέρωση ενός μοντέλου, σε ταχέως αναπτυσσόμενες ροές δεδομένων.

Π.χ. ένας ταξινομητής ερωτημάτων μπορεί να χρειάζεται να συντηρείται, καθώς νέα ερωτήματα αναδύονται συνεχώς και η κατανομή των δεδομένων μπορεί να αλλάζει. Οι περισσότερες από τις υπάρχουσες μεθόδους εκπαίδευσης μοντέλων, είναι στατικές άρα, δε μπορούν να χρησιμοποιηθούν σε ένα τέτοιο σενάριο.

Τρίτον, οι μηχανές αναζήτησης στο Web, συχνά πρέπει να αντιμετωπίζουν ερωτήματα που τίθενται ελάχιστες φορές. Έστω ότι μια μηχανή αναζήτησης θέλει να παρέχει προτάσεις ερωτημάτων που βασίζονται στο περιβάλλον. Δηλαδή, όταν ένας χρήστης υποβάλλει ένα ερώτημα, τότε η μηχανή αναζήτησης προσπαθεί να συμπεράνει το περιβάλλον του ερωτήματος, χρησιμοποιώντας το προφίλ του χρήστη και το ιστορικό των ερωτημάτων του, προκειμένου να επιστρέψει πιο προσαρμοσμένες απαντήσεις μέσα σε ένα μικρό κλάσμα του δευτερολέπτου. Παρόλο που ο συνολικός αριθμός των ερωτημάτων που τίθενται μπορεί να είναι τεράστιος, πολλά ερωτήματα μπορεί να τεθούν μόνο μία ή λίγες φορές. Τέτοια σοβαρά ασύμμετρα δεδομένα, αποτελούν πρόκληση για πολλές μεθόδους εξόρυξης δεδομένων και μηχανικής μάθησης.

22.6 Κοινωνικά μέσα και κοινωνικά δίκτυα

Η επικράτηση των κοινωνικών μέσων και των κοινωνικών δικτύων, έχει αλλάξει ριζικά τη ζωή μας, τον τρόπο που ανταλλάσσουμε πληροφορίες και κοινωνικοποιούμαστε. Με τεράστιες ποσότητες δεδομένων κοινωνικών μέσων και δικτύων που είναι διαθέσιμες, είναι κρίσιμο να αναλύουμε τέτοια δεδομένα για να εξάγουμε εφαρμόσιμα μοτίβα και τάσεις. Η εξόρυξη δεδομένων από τα μέσα κοινωνικής δικτύωσης, είναι η διερεύνηση των τεράστιων ποσοτήτων δεδομένων των κοινωνικών μέσων, προκειμένου να διακρίνουμε μοτίβα και τάσεις (π.χ. σχετικά με τη χρήση των μέσων κοινωνικής δικτύωσης, τις διαδικτυακές κοινωνικές συμπεριφορές, τις συνδέσεις μεταξύ ατόμων, τη συμπεριφορά των ηλεκτρονικών αγορών, την ανταλλαγή περιεχομένου). Αυτά τα μοτίβα και οι τάσεις, έχουν χρησιμοποιηθεί για την ανίχνευση κοινωνικών συμβάντων, για την παρακολούθηση και επιτήρηση της δημόσιας υγείας, για την ανάλυση των συναισθημάτων στα μέσα κοινωνικής δικτύωσης, για τις συστάσεις στα μέσα κοινωνικής δικτύωσης, για την προέλευση των πληροφοριών, για την ανάλυση της αξιοπιστίας των μέσων κοινωνικής δικτύωσης και για την ανίχνευση των κοινωνικών αποστολέων της ανεπιθύμητης αλληλογραφίας.

Η εξόρυξη δεδομένων από τα μέσα κοινωνικής δικτύωσης, είναι η διερεύνηση των δομών των κοινωνικών δικτύων και των πληροφοριών που σχετίζονται με τέτοια δίκτυα, μέσω της χρήσης των δικτύων και των μεθόδων της θεωρίας γραφημάτων και της εξόρυξης δεδομένων. Οι δομές των κοινωνικών δικτύων, χαρακτηρίζονται με βάση τους κόμβους (μεμονωμένους δρώντες, άτομα ή πράγματα εντός του δικτύου) και τους δεσμούς, τις ακμές ή τους συνδέσμους (σχέσεις ή αλληλεπιδράσεις) που τους συνδέουν. Παραδείγματα κοινωνικών δομών που συνήθως απεικονίζονται μέσω της

ανάλυσης των κοινωνικών δικτύων, περιλαμβάνουν τα δίκτυα κοινωνικών μέσων, την εξάπλωση των memes, τα δίκτυα φιλίας και γνωριμιών, τα γραφήματα συνεργασίας, τη συγγένεια, τη μετάδοση ασθενειών και τις σεξουαλικές σχέσεις. Αυτά τα δίκτυα, συχνά απεικονίζονται μέσω κοινωνιογραμμάτων στα οποία οι κόμβοι αναπαρίστανται ως σημεία και οι δεσμοί ως γραμμές. Η εξόρυξη δεδομένων από τα κοινωνικά δίκτυα, έχει χρησιμοποιηθεί για την ανίχνευση κρυφών κοινοτήτων, την αποκάλυψη της εξέλιξης και της δυναμικής των κοινωνικών δικτύων, τον υπολογισμό μετρήσεων του δικτύου (π.χ. κεντρικότητα, μεταβατικότητα, αμοιβαιότητα, ισορροπία, κατάσταση και ομοιότητα), την ανάλυση του τρόπου με τον οποίο διαδίδονται οι πληροφορίες σε ιστότοπους κοινωνικής δικτύωσης, τη μέτρηση και τη μοντελοποίηση της επιρροής των κόμβων/υποδομών και τη διεξαγωγή ανάλυσης των κοινωνικών δικτύων βάσει της τοποθεσίας.

22.7 Λιανικό εμπόριο και υπηρεσίες

Το εμπόριο, οι επιχειρήσεις και η επιχειρηματικότητα, αποτελούν έναν πολύ σημαντικό τομέα της ανάπτυξης. Τα περισσότερα από τα δεδομένα που αφορούν επιχειρηματικές συναλλαγές, αποθηκεύονται σε αποθήκες δεδομένων και δεν έχουν ποτέ ξανά πρόσβαση για σκοπούς καθαρισμού ή ανάλυσης. Αν υποστούν κατάλληλη επεξεργασία από αναλυτές δεδομένων, αυτά τα δεδομένα θα μπορούσαν να δημιουργήσουν χρήσιμες σχέσεις, να προβλέψουν επερχόμενες συναλλαγές και να διευκολύνουν τους ιδιοκτήτες των επιχειρήσεων να ρυθμίζουν τις αγορές των πελατών τους. Η κατανόηση των αγοραστικών συνηθειών και των προτιμήσεων των πελατών, είναι απαραίτητη για τη στρατηγική των λιανοπωλητών. Η εξόρυξη δεδομένων, μπορεί να παρέχει αυτές τις πληροφορίες. Μια αποτελεσματική εφαρμογή της εξόρυξης δεδομένων στο περιβάλλον της λιανικής πώλησης, είναι η ανάλυση καλαθιού αγοράς (MBA–Market Basket Analysis) ή ανάλυση καλαθιού αγορών (SBA–Shopping Basket Analysis). Αναλύει τα χαρακτηριστικά του καλαθιού αγορών των πελατών, από δεδομένα ηλεκτρονικού σημείου πώλησης (EPOS–Electronic Point of Sale) και εφαρμόζει τα ευρήματα για την έναρξη αποτελεσματικών προωθητικών ενεργειών και διαφημίσεων. Μεγάλοι χρήστες της εξόρυξης δεδομένων στον κλάδο του λιανικού εμπορίου περιλαμβάνουν την Wal–Mart και την Victoria's Secret.

22.8 Τηλεμάρκετινγκ και άμεσο μάρκετινγκ

Οι εταιρείες τηλεμάρκετινγκ και άμεσου μάρκετινγκ, έχουν επιτύχει μεγάλες εξοικονομήσεις και είναι σε θέση να στοχεύουν τους πελάτες με μεγαλύτερη ακρίβεια, χρησιμοποιώντας την εξόρυξη δεδομένων. Οι άμεσοι έμποροι, διαμορφώνουν και αποστέλλουν ταχυδρομικώς τους καταλόγους των προϊόντων τους με βάση το ιστορικό των αγορών και τα δημογραφικά δεδομένα των πελατών. Οι τηλεπωλητές, είναι πλέον σε θέση να μειώσουν τον αριθμό των κλήσεων που πραγματοποιούνται και να αυξήσουν την αναλογία των επιτυχημένων κλήσεων. Ενώ το αποτέλεσμα διαφέρει από κλάδο σε κλάδο, τα τηλεφωνικά προγράμματα μεγάλων αποστάσεων άμεσου μάρκετινγκ έχουν αυξήσει το ποσοστό επιτυχίας από 10% ως 20% σε ένα εντυπωσιακό 30% ως 40%.

22.9 Ηλεκτρονικό εμπόριο

Είναι ο πιο πιθανός τομέας για εφαρμογή της εξόρυξης δεδομένων, διότι πολλά από τα συστατικά που απαιτούνται για την επιτυχημένη εξόρυξη δεδομένων είναι εύκολα διαθέσιμα: τα αρχεία δεδομένων είναι άφθονα, η ηλεκτρονική συλλογή παρέχει αξιόπιστα δεδομένα, η γνώση μπορεί εύκολα να μετατραπεί σε δράση και η απόδοση της επένδυσης μπορεί να μετρηθεί. Η ενσωμάτωση του ηλεκτρονικού εμπορίου και της

εξόρυξης δεδομένων, βελτιώνει σημαντικά τα αποτελέσματα και καθοδηγεί τους χρήστες στη δημιουργία γνώσης και στη λήψη σωστών επιχειρηματικών αποφάσεων.

Αυτή η ενσωμάτωση, επιλύει αποτελεσματικά πολλά σημαντικά προβλήματα που σχετίζονται με τα οριζόντια εργαλεία εξόρυξης δεδομένων, συμπεριλαμβανομένης της τεράστιας προσπάθειας που απαιτείται για την προεπεξεργασία των δεδομένων πριν αυτά χρησιμοποιηθούν για εξόρυξη και καθιστά τα αποτελέσματα της εξόρυξης αξιοποιήσιμα.

22.10 Τηλεπικοινωνίες

Η εξόρυξη δεδομένων, χρησιμοποιείται από παρόχους τηλεπικοινωνιακών/ κινητών υπηρεσιών για:

- τη διαμόρφωση και τον σχεδιασμό στρατηγικών,
- τον σχεδιασμό της καμπάνιας και για το μάρκετινγκ,
- τη διατήρηση των πελατών,
- τα πακέτα πελατών, με βάση την τμηματοποίηση τους,
- τη βέλτιστη αξιοποίηση της υποδομής των επικοινωνιών.

Χρησιμοποιώντας την ταξινόμηση και την ομαδοποίηση, οι πάροχοι υπηρεσιών κινητής τηλεφωνίας μπορούν να διαμορφώσουν στρατηγικές για την προώθηση του άμεσου μάρκετινγκ. Με τη βοήθεια της συσταδοποίησης ακολουθούμενης από την ταξινόμηση, οι πελάτες μπορούν να τμηματοποιηθούν σε διάφορες ομάδες για να προβλέψουν όσους μετακινούνται. Οι συγκεκριμένες στρατηγικές και τα πακέτα μάρκετινγκ, μπορούν να διαμορφωθούν και να σχεδιαστούν για την προσέλκυση πελατών, ώστε να διατηρηθούν από τον πάροχο των υπηρεσιών. Με βάση τις προσδιορισμένες ομάδες πελατών, τα συγκεκριμένα πακέτα μπορούν να διαμορφωθούν με βάση τις ανάγκες των διαφόρων ομάδων πελατών. Για τον σχεδιασμό πακέτων, μπορεί να χρησιμοποιηθεί η ανάλυση συσχέτισης. Το πρότυπο χρήσης δικτύου, μπορεί να αναλυθεί χρησιμοποιώντας εξόρυξη δεδομένων για τον εντοπισμό της υποχρησιμοποιούμενης και της υπερχρησιμοποιούμενης υποδομής του δικτύου, έτσι ώστε η συνολική υποδομή να μπορεί να χρησιμοποιηθεί και να βελτιωθεί, σύμφωνα με τις απαιτήσεις.

22.11 Εκπαίδευση

Μια επερχόμενη εφαρμογή της εξόρυξης δεδομένων αναλαμβάνεται από τα ανώτατα εκπαιδευτικά ιδρύματα, λόγω της σταδιακής αύξησης της ποσότητας των δεδομένων με την πάροδο των ετών. Η εξόρυξη δεδομένων σε αυτόν τον κλάδο, χρησιμοποιείται για την κατανόηση της συμπεριφοράς των φοιτητών, όπως οι τάσεις που θα υποδείκνυαν τη μεταφορά φοιτητών, την τάση των πιστωτικών μονάδων, τα σύνολα δεξιοτήτων των διαφόρων ομάδων φοιτητών και τα πλεονάζοντα χαρακτηριστικά τους. Διαφορετικοί λόγοι, για τους οποίους μπορεί να εφαρμοστεί η τεχνολογία εξόρυξης στην ανώτατη εκπαίδευση, είναι:

- η ανάλυση και η οπτικοποίηση των βαθμολογιών, με τη χρήση μαθηματικών και γραφημάτων,
- η πρόβλεψη της επίδοσης, με τεχνικές παλινδρόμησης και κανόνες ασαφούς συσχέτισης,
- η ανίχνευση των ακραίων τιμών κάθε ομάδας, μέσω εποπτευόμενης, μη εποπτευόμενης ή ημιεποπτευόμενης μάθησης,
- η ομαδοποίηση των φοιτητών και η διαχείριση των μαθημάτων ανάλογα με την ομαδοποίηση (όπως k–Means), βασισμένη σε μοντέλα και σε ιεραρχική συσσωμάτωση,

• ο σχεδιασμός και ο προγραμματισμός των ωραρίων και των ωρών μελέτης, χρησιμοποιώντας παλινδρόμηση, ομαδοποίηση και ταξινόμηση.

Τα δέντρα αποφάσεων και τα νευρωνικά δίκτυα συμπλεγμάτων οπισθοδιάδοσης, χρησιμοποιούνται σήμερα για τον σχεδιασμό των μαθημάτων. Η εξόρυξη δεδομένων στην εξ αποστάσεως εκπαίδευση, παράγει αυτόματα χρήσιμες πληροφορίες για την ενίσχυση της μαθησιακής διαδικασίας με βάση την τεράστια ποσότητα των δεδομένων που παράγονται από τους καθηγητές και από τις αλληλεπιδράσεις των φοιτητών με το διαδικτυακό περιβάλλον της εξ αποστάσεως εκπαίδευσης. Οι εφαρμογές της εξόρυξης δεδομένων, μεταφέρουν τα δεδομένα σε πληροφορίες και σε ανατροφοδότηση στο περιβάλλον της ηλεκτρονικής μάθησης. Αυτή η λύση, μετατρέπει μεγάλες ποσότητες άχρηστων δεδομένων σε ένα έξυπνο σύστημα παρακολούθησης και συστάσεων που εφαρμόζεται στη μαθησιακή διαδικασία. Στην εκπαίδευση μέσω του διαδικτύου, οι μέθοδοι εξόρυξης δεδομένων χρησιμοποιούνται για τη βελτίωση του εκπαιδευτικού υλικού.

Οι σχέσεις, ανακαλύπτονται μεταξύ των δεδομένων χρήσης που συλλέγονται, κατά τη διάρκεια των συνεδριών των φοιτητών. Αυτή η γνώση είναι πολύ χρήσιμη για τον καθηγητή, ο οποίος θα μπορούσε να αποφασίσει ποιες αλλαγές είναι οι καταλληλότερες για τη βελτίωση της αποτελεσματικότητας του μαθήματος. Οι μέθοδοι εξόρυξης δεδομένων, χρησιμοποιούνται για να παρέχουν στους εκπαιδευόμενους προσαρμοστική ανατροφοδότηση, σε πραγματικό χρόνο, σχετικά με τη φύση και με τα πρότυπα της διαδικτυακής επικοινωνίας τους, ενώ μαθαίνουν συνεργατικά. Αυτό, καθιστά δυνατή την αύξηση της ευαισθητοποίησης των εκπαιδευόμενων. Η εφαρμογή των μεθόδων εξόρυξης δεδομένων σε εκπαιδευτικές συνομιλίες, είναι εφικτή και μπορεί να φέρει βελτίωση στα μαθησιακά περιβάλλοντα.

22.12 Κατασκευαστές

Η εξόρυξη δεδομένων, χρησιμοποιείται ευρέως στις μεταποιητικές βιομηχανίες για τον έλεγχο και για τον προγραμματισμό των τεχνικών διαδικασιών παραγωγής. Π.χ. η LTV steel corporation, κατάφερε να μειώσει τα ελλαττωματικά της προϊόντα κατά 99% χρησιμοποιώντας την εξόρυξη δεδομένων, προς ανίχνευση πιθανών προβλημάτων ποιότητας.

22.13 Ασφαλιστικές εταιρείες

Ο ασφαλιστικός κλάδος, είναι εντατικός σε δεδομένα. Η εξόρυξη δεδομένων, έχει πρόσφατα παράσχει στις ασφαλιστικές εταιρείες έναν πλούτο χρήσιμων πληροφοριών, που εξάγονται από τεράστιες βάσεις δεδομένων προς λήψη αποφάσεων. Αυτές οι πληροφορίες, επιτρέπουν στις ασφαλιστικές εταιρείες να γνωρίζουν καλύτερα τους πελάτες τους και να εντοπίζουν την ασφαλιστική απάτη πιο αποτελεσματικά. Η Empire blue cross και η Blue shield, είναι μεταξύ των επιτυχημένων χρηστών της τεχνολογίας εξόρυξης δεδομένων.

22.14 Αθλητισμός

Ο αθλητισμός, είναι ιδανικός για την εφαρμογή εργαλείων και τεχνικών εξόρυξης δεδομένων, διότι συλλέγονται τεράστιες ποσότητες στατιστικών στοιχείων για κάθε παίκτη, ομάδα, παιχνίδι και σεζόν. Η εξόρυξη δεδομένων, μπορεί να χρησιμοποιηθεί από τους αθλητικούς οργανισμούς με τη μορφή της στατιστικής ανάλυσης, της ανακάλυψης προτύπων και της πρόβλεψης των αποτελεσμάτων.

Τα μοτίβα στα δεδομένα, είναι πολλές φορές χρήσιμα στην πρόβλεψη των μελλοντικών γεγονότων. Η εξόρυξη δεδομένων, μπορεί να χρησιμοποιηθεί για την ανίχνευση, για την πρόβλεψη της απόδοσης, για την επιλογή των παικτών, για την

προπονητική, για την εκπαίδευση και για την χάραξη της στρατηγικής. Οι τεχνικές εξόρυξης δεδομένων, χρησιμοποιούνται για τον προσδιορισμό της καλύτερης ομάδας σε ένα ομαδικό άθλημα σε μια σεζόν, περιοδεία ή παιχνίδι.

22.15 Εφορία

Το σύστημα εξόρυξης δεδομένων που εφαρμόζεται στην εφορία, για τον εντοπισμό ατόμων με υψηλό εισόδημα που εμπλέκονται σε καταχρηστικά φορολογικά καταφύγια, δίνει καλά αποτελέσματα. Οι κύριες γραμμές έρευνας, περιλαμβάνουν την οπτικοποίηση των σχέσεων και την εξόρυξη δεδομένων, για τον εντοπισμό και για την κατάταξη πιθανώς καταχρηστικών συναλλαγών φοροαποφυγής.

22.16 Ψηφιακή βιβλιοθήκη

Η ψηφιακή βιβλιοθήκη ανακτά, συλλέγει, αποθηκεύει και διατηρεί τα ψηφιακά δεδομένα. Η έλευση των ηλεκτρονικών πόρων και η αυξημένη χρήση τους στις βιβλιοθήκες, έχει επιφέρει σημαντικές αλλαγές. Τα δεδομένα και οι πληροφορίες, είναι διαθέσιμα σε διαφορετικές μορφές. Αυτές οι μορφές περιλαμβάνουν κείμενο, εικόνες, βίντεο, ήχο, εικόνα, χάρτες, άρα η ψηφιακή βιβλιοθήκη είναι ένας κατάλληλος τομέας για την εφαρμογή της εξόρυξης δεδομένων.

22.17 Φαρμακευτική βιομηχανία

Η φαρμακευτική βιομηχανία, είναι γνωστή για την εκτέλεση ποσοτικών αναλύσεων στην κλινική έρευνα και στην έρευνα αγοράς. Στα τμήματα μάρκετινγκ, οι εφαρμογές εξόρυξης δεδομένων χρησιμοποιούνται για τον σχεδιασμό των πωλήσεων και για το άμεσο μάρκετινγκ σε ιατρούς και ασθενείς. Οι τεχνικές εξόρυξης δεδομένων, έχουν χρησιμοποιηθεί αρκετά καλά σε μια ποικιλία κρίσιμων επιχειρηματικών αποφάσεων, στη φαρμακευτική βιομηχανία. Χρησιμοποιήθηκαν για πρόβλεψη στα χρονοδιαγράμματα παραγωγής σε εργοστάσια, για τον προσδιορισμό του δυναμικού της αγοράς, σε κρίσιμες αποφάσεις σχετικά με την έναρξη/μη έναρξη εργασιών, για την ανάπτυξη ενώσεων ή για την πραγματοποίηση οικονομικών προβλέψεων σε μετόχους και επενδυτές στη Wall Street.

22.18 Συστήματα συστάσεων

Τα συστήματα συστάσεων, παρέχουν στους ενδιαφερόμενους ποικίλες συστάσεις που μπορεί να ενδιαφέρουν τους χρήστες που χρησιμοποιούν την εξόρυξη δεδομένων. Τα συστήματα συστάσεων εξετάζουν τις συναλλαγές των χρηστών, τα προφίλ τους, τις λέξεις–κλειδιά, τα κοινά χαρακτηριστικά μεταξύ των στοιχείων και εκτιμούν ένα στοιχείο για τον χρήστη. Πολλές τεχνικές εξόρυξης δεδομένων (μηχανική μάθηση, στατιστική, ανάκτηση πληροφοριών), χρησιμοποιούνται στα συστήματα συστάσεων. Π.χ. στο μάρκετινγκ, το σύστημα συστάσεων μπορεί να προτείνει είδη που είναι ίδια με αυτά που ζήτησε ο χρήστης στο παρελθόν ή εξετάζοντας τις προτιμήσεις άλλων πελατών που έχουν παρόμοια συμπεριφορά με τον χρήστη, να του προτείνει κάτι παρόμοιο.

23. Εξόρυξη δεδομένων και κοινωνία

Καθώς η εξόρυξη δεδομένων διεισδύει στην καθημερινότητά μας, είναι σημαντικό να μελετήσουμε τον αντίκτυπο της στην κοινωνία. Πώς μπορούμε να την χρησιμοποιήσουμε προς όφελος της κοινωνίας; Πώς μπορούμε να προστατευτούμε από την κακή της χρήση; Η ακατάλληλη αποκάλυψη ή η χρήση δεδομένων και η πιθανή παραβίαση της ατομικής ιδιωτικότητας και των δικαιωμάτων προστασίας δεδομένων, είναι τομείς ανησυχίας που πρέπει να αντιμετωπιστούν. Η εξόρυξη δεδομένων θα

βοηθήσει στην επιστημονική ανακάλυψη, στη διαχείριση επιχειρήσεων, στην ανάκαμψη της οικονομίας και στην προστασία της ασφάλειας (π.χ. ανακάλυψη σε πραγματικό χρόνο των εισβολέων και των κυβερνοεπιθέσεων). Ωστόσο, ενέχει τον κίνδυνο της ακούσιας αποκάλυψης ορισμένων εμπιστευτικών επιχειρηματικών ή κυβερνητικών πληροφοριών και αποκάλυψης προσωπικών πληροφοριών ενός ατόμου.

Οι μελέτες σχετικά με την ασφάλεια των δεδομένων στην εξόρυξη τους, στη δημοσίευση τους και στην εξόρυξη τους με σεβασμό στην ιδιωτικότητα, αποτελούν ένα συνεχές σημαντικό ερευνητικό θέμα. Η φιλοσοφία είναι η παρατήρηση της ευαισθησίας των δεδομένων, η διατήρηση της ασφάλειας τους και του απορρήτου των ανθρώπων, κατά την εκτέλεση μίας επιτυχημένης εξόρυξης δεδομένων.

24. Εξόρυξη δεδομένων και ηθική

Η χρήση δεδομένων, ιδίως σχετικά με άτομα, για την εξόρυξη δεδομένων έχει σοβαρές ηθικές επιπτώσεις και οι επαγγελματίες των τεχνικών εξόρυξης δεδομένων πρέπει να ενεργούν υπεύθυνα, γνωρίζοντας τα ηθικά ζητήματα που περιβάλλουν τη συγκεκριμένη διαδικασία. Όταν εφαρμόζεται σε άτομα, η εξόρυξη δεδομένων χρησιμοποιείται συχνά για να διακρίνει ποιος λαμβάνει το δάνειο, ποιος λαμβάνει την ειδική προσφορά κ.ο.κ. Ορισμένες διακρίσεις (φυλετικές, σεξουαλικές, θρησκευτικές), είναι ανήθικες και παράνομες. Ωστόσο, η κατάσταση είναι περίπλοκη, όλα εξαρτώνται από την εφαρμογή. Η χρήση σεξουαλικών και φυλετικών πληροφοριών για ιατρική διάγνωση είναι σίγουρα ηθική, αλλά η χρήση των ίδιων πληροφοριών κατά την εξόρυξη συμπεριφοράς στην πληρωμή ενός δανείου δεν είναι. Ακόμα και όταν απορρίπτονται ευαίσθητες πληροφορίες, υπάρχει ο κίνδυνος να κατασκευαστούν μοντέλα που βασίζονται σε μεταβλητές που μπορούν να αποδειχθούν ότι υποκαθιστούν τα φυλετικά ή σεξουαλικά χαρακτηριστικά.

Π.χ. οι άνθρωποι ζουν συχνά σε περιοχές που συνδέονται με συγκεκριμένες εθνοτικές ταυτότητες, επομένως η χρήση ενός κωδικού περιοχής σε μια μελέτη εξόρυξης δεδομένων διατρέχει τον κίνδυνο να κατασκευαστούν μοντέλα που βασίζονται στη φυλή, παρόλο που οι φυλετικές πληροφορίες έχουν αποκλειστεί ρητά από τα δεδομένα. Είναι ευρέως αποδεκτό ότι, πριν οι άνθρωποι λάβουν την απόφαση να παράσχουν προσωπικές πληροφορίες, πρέπει να γνωρίζουν το πώς αυτές θα χρησιμοποιηθούν, για ποιόν σκοπό, ποια μέτρα θα ληφθούν για την προστασία της εμπιστευτικότητας και της ακεραιότητάς τους, ποιες είναι οι συνέπειες της παροχής ή της απόκρυψης των πληροφοριών και τυχόν δικαιώματα προσφυγής που μπορεί να έχουν. Κάθε φορά που συλλέγονται τέτοιες πληροφορίες, τα άτομα πρέπει να ενημερώνονται σχετικά, όχι με νομικά ψιλά γράμματα, αλλά με σαφήνεια και σε απλή γλώσσα που μπορούν να κατανοήσουν. Οι δυνατότητες των τεχνικών εξόρυξης δεδομένων, σημαίνουν ότι οι τρόποι με τους οποίους μπορεί να χρησιμοποιηθεί ένα αποθετήριο δεδομένων μπορεί να εκτείνονται πολύ πέρα από αυτό που είχε σχεδιαστεί όταν συλλέχθηκαν αρχικά τα δεδομένα. Αυτό, δημιουργεί ένα σοβαρό πρόβλημα: είναι απαραίτητο να προσδιοριστούν οι συνθήκες υπό τις οποίες συλλέχθηκαν τα δεδομένα και για ποιους σκοπούς μπορούν να χρησιμοποιηθούν. Παρέχει η κυριότητα των δεδομένων, το δικαίωμα χρήσης τους με τρόπους διαφορετικούς από αυτούς που υποτίθεται ότι καταγράφηκαν αρχικά; Προφανώς, στην περίπτωση των ρητά συλλεγόμενων προσωπικών δεδομένων, δεν το κάνει. Αλλά γενικά, η κατάσταση είναι περίπλοκη. Αναδύονται εκπληκτικά αποτελέσματα από την εξόρυξη δεδομένων.

Π.χ. έχει αναφερθεί ότι μία από τις κορυφαίες ομάδες καταναλωτών στη Γαλλία, διαπίστωσε ότι τα άτομα με κόκκινα αυτοκίνητα είναι πιο πιθανό να αθετήσουν τα δάνεια αυτοκινήτου τους. Ποια είναι η κατάσταση μιας τέτοιας ανακάλυψης; Σε ποιες πληροφορίες βασίζεται; Υπό ποιες συνθήκες συλλέχθηκαν αυτές οι πληροφορίες;

Με ποιους τρόπους είναι ηθικό να τις χρησιμοποιούμε; Σαφώς, οι ασφαλιστικές εταιρείες ασχολούνται με τις διακρίσεις μεταξύ των ανθρώπων με βάση στερεότυπα (οι νέοι άνδρες πληρώνουν ακριβά για την ασφάλεια αυτοκινήτου), αλλά αυτά δε βασίζονται αποκλειστικά σε στατιστικές συσχετίσεις. Περιλαμβάνουν επίσης, γνώση της κοινής λογικής για τον κόσμο. Το αν το προηγούμενο εύρημα λέει κάτι για το είδος του ατόμου που επιλέγει ένα κόκκινο αυτοκίνητο ή αν πρέπει να απορριφθεί ως άσχετο, είναι θέμα ανθρώπινης κρίσης που βασίζεται στη γνώση του κόσμου και όχι σε καθαρά στατιστικά κριτήρια. Όταν παρουσιάζονται δεδομένα, πρέπει να αναρωτηθούμε ποιος επιτρέπεται να έχει πρόσβαση σε αυτά, για ποιον σκοπό συλλέχθηκαν και τι είδους συμπεράσματα είναι θεμιτό να εξαχθούν από αυτά.

Η ηθική διάσταση, εγείρει δύσκολα ερωτήματα για όσους ασχολούνται με την πρακτική εξόρυξη δεδομένων. Είναι απαραίτητο να λάβουμε υπόψη τους κανόνες της κοινότητας, που έχει συνηθίσει να διαχειρίζεται το είδος των δεδομένων που εμπλέκονται, πρότυπα που μπορεί να έχουν εξελιχθεί κατά τη διάρκεια δεκαετιών ή αιώνων, αλλά που ο ειδικός της πληροφορίας μπορεί να μη γνωρίζει. Π.χ. στην κοινότητα των βιβλιοθηκών, θεωρείται δεδομένο ότι η ιδιωτικότητα των αναγνωστών είναι ένα δικαίωμα που προστατεύεται με ζήλο. Αν τηλεφωνήσει κάποιος στη βιβλιοθήκη του πανεπιστημίου και ρωτήσει ποιος έχει δανειστεί το τάδε εγχειρίδιο, δε θα του πουν. Όσοι δημιουργούν ψηφιακές βιβλιοθήκες, μπορεί να μην γνωρίζουν αυτές τις ευαισθησίες και μπορεί να ενσωματώσουν συστήματα εξόρυξης δεδομένων που αναλύουν και συγκρίνουν τις αναγνωστικές συνήθειες των ατόμων, ώστε να τους προτείνουν νέα βιβλία, ίσως να πουλήσουν τα αποτελέσματα σε εκδότες!

Εκτός από τα πρότυπα της κοινότητας για τη χρήση δεδομένων, πρέπει να τηρούνται λογικά και επιστημονικά πρότυπα, κατά την εξαγωγή συμπερασμάτων από αυτά. Αν καταλήξουμε σε συμπεράσματα (όπως ότι οι ιδιοκτήτες κόκκινων αυτοκινήτων ενέχουν μεγαλύτερο πιστωτικό κίνδυνο), πρέπει να τους επισυνάψουμε επιφυλάξεις και να τις υποστηρίξουμε με επιχειρήματα, πέρα από τα καθαρά στατιστικά δεδομένα. Το θέμα είναι ότι, η εξόρυξη δεδομένων είναι απλά ένα εργαλείο σε όλη τη διαδικασία: Οι άνθρωποι είναι αυτοί που λαμβάνουν τα αποτελέσματα, μαζί με άλλες γνώσεις και αποφασίζουν ποια ενέργεια θα εφαρμόσουν.

Η εξόρυξη δεδομένων εγείρει ένα άλλο ερώτημα, που είναι στην πραγματικότητα πολιτικό, σε τι χρησιμοποιούνται οι πόροι της κοινωνίας; Μια εφαρμογή της εξόρυξης δεδομένων είναι στην ανάλυση καλαθιών, όπου αναλύονται τα αρχεία των ταμείων του super market, για την ανίχνευση συσχετίσεων μεταξύ των ειδών που αγοράζουν οι άνθρωποι. Πώς θα πρέπει να χρησιμοποιηθούν οι πληροφορίες που προκύπτουν; Πρέπει ο διευθυντής του super market να τοποθετήσει την μπύρα και τα πατατάκια μαζί, για να διευκολύνει τους αγοραστές, ή να τα τοποθετήσει σε μεγαλύτερη απόσταση μεταξύ τους, μεγιστοποιώντας τον χρόνο παραμονής τους στο κατάστημα και ως εκ τούτου, αυξάνοντας την πιθανότητα να παρασυρθούν σε απρογραμμάτιστες περαιτέρω αγορές; Πρέπει ο διευθυντής να μετακινήσει τις πιο ακριβές, άρα πιο κερδοφόρες πάνες κοντά στην μπύρα, αυξάνοντας τις πωλήσεις σε βιαστικούς πατέρες, ενός είδους υψηλού περιθωρίου κέρδους και να προσθέσει περισσότερα πολυτελή προϊόντα για μωρά εκεί κοντά; Φυσικά, όσοι χρησιμοποιούν προηγμένες τεχνολογίες πρέπει να λάβουν υπόψη τη σοφία αυτού που κάνουν. Αν τα δεδομένα χαρακτηρίζονται ως καταγεγραμμένα γεγονότα, τότε οι πληροφορίες είναι το σύνολο των μοτίβων ή των προσδοκιών που αποτελούν τη βάση των δεδομένων. Έτσι, θα μπορούσε να οριστεί η γνώση ως η συσσώρευση του συνόλου των προσδοκιών και η σοφία ως η αξία που αποδίδεται στη γνώση. Όπως και να έχει, η εξόρυξη δεδομένων είναι μια τεχνολογία που πρέπει να λάβουμε σοβαρά υπόψη.

25. Η ανάγκη για ανθρώπινη κατεύθυνση στην εξόρυξη δεδομένων

Πολλοί προμηθευτές λογισμικού, προωθούν το αναλυτικό τους λογισμικό ως μια εφαρμογή plug and play έτοιμη προς χρήση, που θα παράσχει λύσεις σε διαφορετικά δυσεπίλυτα προβλήματα, χωρίς την ανάγκη ανθρώπινης επίβλεψης ή αλληλεπίδρασης. Μερικοί πρώιμοι ορισμοί της εξόρυξης δεδομένων, ακολούθησαν αυτή την εστίαση στον αυτοματισμό. Οι Berry και Linoff, στο «Τεχνικές εξόρυξης δεδομένων για μάρκετινγκ, πωλήσεις και υποστήριξη πελατών», έδωσαν τον ακόλουθο ορισμό για την εξόρυξη δεδομένων

«*Η εξόρυξη δεδομένων είναι η διαδικασία εξερεύνησης και ανάλυσης, με αυτόματα ή ημιαυτόματα μέσα, μεγάλων ποσοτήτων δεδομένων, προκειμένου να ανακαλυφθούν ουσιαστικά μοτίβα και κανόνες*».

Τρία έτη αργότερα, στο «Κατακτώντας την εξόρυξη δεδομένων», επανεξετάζουν τον ορισμό τους και αναφέρουν ότι

«*Αν υπάρχει κάτι για το οποίο μετανιώνουμε, είναι η φράση με αυτόματα ή ημιαυτόματα μέσα... επειδή πιστεύουμε ότι έχει δοθεί υπερβολική έμφαση στις αυτόματες τεχνικές και όχι αρκετή στην εξερεύνηση και στην ανάλυση*».

Αυτό, έχει παραπλανήσει πολλούς ώστε να πιστεύουν ότι η εξόρυξη δεδομένων είναι ένα προϊόν που μπορεί να αγοραστεί και όχι ένας κλάδος που πρέπει να κατακτηθεί. Ο αυτοματισμός, δεν υποκαθιστά την ανθρώπινη συμβολή. Οι άνθρωποι πρέπει να συμμετέχουν ενεργά, σε κάθε φάση της διαδικασίας εξόρυξης δεδομένων. Ο George Grinstein του Πανεπιστημίου της Μασαχουσέτης, το διατύπωσε ως εξής

«*Φανταστείτε ένα μαύρο κουτί, ικανό να απαντήσει σε οποιαδήποτε ερώτηση του τεθεί. Οποιαδήποτε ερώτηση. Θα εξαλείψει αυτό την ανάγκη για ανθρώπινη συμμετοχή όπως προτείνουν πολλοί; Ακριβώς το αντίθετο. Το θεμελιώδες πρόβλημα, εξακολουθεί να καταλήγει σε ένα ζήτημα ανθρώπινης διεπαφής. Πώς διατυπώνω σωστά την ερώτηση; Πώς μπορώ να ορίσω τις παραμέτρους ώστε να λάβω μια λύση, που να είναι εφαρμόσιμη στη συγκεκριμένη περίπτωση που με ενδιαφέρει; Πώς μπορώ να λάβω τα αποτελέσματα σε εύλογο χρόνο και σε μορφή που μπορώ να την καταλάβω;*»

Σημειώστε ότι όλες οι ερωτήσεις συνδέουν τη διαδικασία ανακάλυψης με εμένα, για ανθρώπινη κατανάλωση. Αντί να ρωτάμε πού εντάσσονται οι άνθρωποι στην εξόρυξη δεδομένων, πρέπει να αναρωτηθούμε πώς μπορούμε να την εντάξουμε στην ανθρώπινη διαδικασία επίλυσης προβλημάτων. Η δύναμη των αλγορίθμων εξόρυξης δεδομένων που είναι ενσωματωμένοι στο λογισμικό μαύρου κουτιού που είναι διαθέσιμο σήμερα, καθιστά την κακή χρήση τους αναλογικά πιο επικίνδυνη.

Όπως ακριβώς συμβαίνει με κάθε νέα τεχνολογία πληροφοριών, η εξόρυξη δεδομένων είναι εύκολο να γίνει επιβλαβής. Οι ερευνητές, μπορεί να εφαρμόσουν ακατάλληλη ανάλυση σε σύνολα δεδομένων που απαιτούν μια εντελώς διαφορετική προσέγγιση, ή να προκύψουν μοντέλα που βασίζονται σε εντελώς ψευδείς υποθέσεις. Άρα, απαιτείται κατανόηση των στατιστικών και μαθηματικών δομών των μοντέλων που διέπουν το λογισμικό.

Παράρτημα

Πίνακας 6. Τεχνικές εξόρυξης δεδομένων και οι εφαρμογές τους

Τεχνική / Κατηγορία	Περιγραφή	Εργαλεία / Υλοποίηση	Περιορισμοί	Εφαρμογές
Στατιστική	Γραφική και πινακοποιημένη αναπαράσταση δεδομένων	WebStat, AccessWatch, Analog	• Δεν αναλύει μεμονωμένα στοιχεία • Δεν λειτουργεί σε ετερογενή δεδομένα • Μικρά δείγματα είναι παραπλανητικά	• Μετρήσεις επισκέψεων ιστού • Γραφήματα & 3D απεικονίσεις
Εξόρυξη δεδομένων ιστού (Web mining)	Εξόρυξη δεδομένων σχετικών με ιστοσελίδες	Winautomation, import.io, CrawlMonster	• Παραβίαση ιδιωτικότητας • Άσχετα αποτελέσματα	•Δομή ιστοσελίδων • Περιεχόμενο • Χρήση ιστού
Ταξινόμηση (Classification)	Ομαδοποίηση αντικειμένων με παρόμοια χαρακτηριστικά	KNIME, RapidMiner, Weka, TreeView	• Δεν είναι χρήσιμο για ετερογενή δεδομένα	• Μείωση πολυπλοκότητας • Προγραμματισμός
Ομαδοποίηση (Clustering)	Ομαδοποίηση δεδομένων με ομοιότητες	Cluster 3.0, Java TreeView, PYCLUSTER	• Δεν υποστηρίζει κοινή αποθήκευση • Λειτουργικά σφάλματα	• Ανοχή σφαλμάτων • Συντήρηση συστημάτων
Διαδοχικά πρότυπα (Sequential patterns)	Εύρεση προτύπων με συγκεκριμένη σειρά	XAffinity(TM), SPMF, Mininggo	• Απαιτεί μεγάλη βάση δεδομένων	• Τοποθέτηση προϊόντων • Πρόβλεψη καταστροφών • Ανίχνευση ιατρικών ενδείξεων
Κανόνες συσχέτισης (Association rules)	Ανάλυση σχέσεων τύπου "αν–τότε"	FPM, Bart Goethals, FriDA, KNIME, Magnum Opus	• Πολύπλοκοι αλγόριθμοι • Πολλές παράμετροι	•LMS συστήματα •Χρηματιστηριακές συναλλαγές
Πρόβλεψη (Prediction)	Εκτίμηση μελλοντικών τιμών από προηγούμενες	EDM, business tools	• Αλλαγή τάσεων → λανθασμένες προβλέψεις	• Μετεωρολογία • Συμπεριφορά μαθητών • Επιχειρήσεις
Ανάλυση συσχέτισης (Correlation analysis)	Εντοπισμός προτύπων σε σήματα, ήχους, εικόνες	Google Trends, Google Flu Trends, Google Correlate	• Δεν αποδεικνύει αιτιότητα	• Έρευνα • Νευροεπιστήμες • Χρηματοοικονομικά

Τεχνική / Κατηγορία	Περιγραφή	Εργαλεία / Υλοποίηση	Περιορισμοί	Εφαρμογές
Αιτιώδης εξόρυξη (Casual data mining)	Ανάλυση σχέσεων στα δεδομένα	Weka, RapidMiner, KNIME, Rattle	• Ποιότητα, ασφάλεια, ιδιωτικότητα	• Υγεία • Οικονομικά • Εκπαίδευση
Ανίχνευση ανωμαλιών (Outlier detection)	Εντοπισμός ακραίων τιμών & αποκλίσεων	CMSR Data Miner	• Πολύπλοκα πιθανοθεωρητικά μοντέλα	• Βιομηχανική ζημιά • Απάτη • Εισβολές • Δημόσια υγεία
Εξόρυξη κειμένου (Text mining)	Εξαγωγή πληροφορίας από κείμενο	Carrot2, GATE, Gensim, OpenNLP, Orange, Stanbol, KNIME, PLOS	• Αδόμητα δεδομένα • Συντακτικά & σημασιολογικά λάθη	• Διαχείριση εγγράφων • Social media • Ασφάλεια • CRM • Εκπαίδευση
Ανάλυση κοινωνικών δικτύων (SNA)	Μελέτη κοινωνικών δομών μέσω δικτύων	Commetrix, Cytoscape, Cuttlefish, EgoNet, Gephi, GraphChi, Graphviz	• Ρίσκο απάτης • Σπατάλη χρόνου • Παραβίαση ιδιωτικότητας	• Συνδεσιμότητα • Κοινοποίηση πληροφοριών • Στοχευμένη διαφήμιση
Δέντρα αποφάσεων (Decision Trees)	Δενδροειδές μοντέλο αποφάσεων	SilverDecisions, GATree, Gambit, KNIME, RapidMiner, Smiles, YaDT	• Μικρές αλλαγές → αλλαγή δέντρου • Πολυπλοκότητα	• Μοντελοποίηση • Επιλογή χαρακτηριστικών • Ερμηνεία δεδομένων
Τεχνική k-Nearest Neighbor (k-NN)	Ταξινόμηση & παλινδρόμηση βάσει γειτόνων	Weka, Kaldi, MEKA, MODLEM, recognition systems	• Εύρεση k • Υπολογιστικά ακριβό	• Βιοπληροφορική • Συστάσεις • Υπολογιστική όραση • Μάρκετινγκ • Ψηφιακή αναζήτηση
Εξόρυξη διεργασιών (Process mining)	Ανάλυση διαδικασιών από log files	ProM, XESame, OpenXES, MXMLib	• Θόρυβος • Ελλιπή δεδομένα • Πολυπλοκότητα	• Ανακάλυψη διαδικασιών • Έλεγχος συμμόρφωσης

Πίνακας 7. Σχέση μεταξύ των εργασιών εξόρυξης δεδομένων και των τεχνικών εξόρυξης δεδομένων

	Εργασίες εξόρυξης δεδομένων						
Τεχνικές εξόρυξης δεδομένων	**Σύνοψη**	**Χαρακτηρισμός & διαχωρισμός**	**Ταξινόμηση**	**Ομαδοποίηση (Clustering)**	**Συσχέτιση**	**Ανάλυση ακραίων τιμών**	**Παλινδρόμηση & ανάλυση τάσεων**
Στατιστική	✓	✓	✓	✓	✓	✓	✓
Μηχανική μάθηση		✓	✓	✓	✓	✓	✓
Νευρωνικά δίκτυα		✓	✓	✓	✓	✓	✓
Συστήματα βάσεων δεδομένων	✓	✓			✓	✓	
Γενετικοί αλγόριθμοι			✓	✓	✓	✓	
Ασαφή σύνολα και λογική		✓	✓	✓	✓	✓	
Οπτικοποίηση		✓	✓	✓		✓	✓

Βιβλιογραφία

Ελληνόγλωσση βιβλιογραφία

• Γεωργούλη Α. (2015). *Τεχνητή νοημοσύνη: μια εισαγωγική προσέγγιση*, Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών.

• Διαμαντάρας Κ. (2007). *Τεχνητά νευρωνικά δίκτυα*, Κλειδάριθμος.

• Κεραυνού Ε. (2000). *Τεχνητή νοημοσύνη και έμπειρα συστήματα*, τόμος Α», ΕΑΠ.

• Λυκοθανάσης Σ. (2001). *Γενετικοί αλγόριθμοι και εφαρμογές*, ΕΑΠ, Σχολή θετικών επιστημών και τεχνολογίας, Πάτρα.

• Πεπινίδης Σ., Λαζάρου Α.–Μ., Χατζηδάκης Ι.–Π. (2015). *Λογισμικό εξόρυξης δεδομένων WEKA: Αναλυτικό εγχειρίδιο χρήσης και εφαρμογές*, Τμήμα διοίκησης επιχειρήσεων, ΤΕΙ δυτικής Ελλάδας, Πάτρα.

Ξενόγλωσση βιβλιογραφία

• Adhikari, A. and Adhikari, J. (2015). Advances in knowledge discovery in databases, Springer international publishing AG.

• Agrawal, R., Gehrke, J., Gunopulos, D. and Raghavan, P. (1998). Automatic subspace clustering of high dimensional data for data mining applications, Proceedings of the ACM SIGMOD Conference on Management of Data.• Ahlemeyer–Stubbe A. and Coleman S. A practical guide to data mining for business and industry, John Wiley & sons, Ltd.

• Alpaydin, E. (2010). Introduction to machine learning, 2nd edition, The MIT press, Cambridge.

• Azzalini, A., & Scarpa, B. (2012). Data analysis and data mining: An introduction. Oxford University Press.

• Bench–Capon, T.J.K. (1990). Knowledge representation: An approach to artificial intelligence, Academic press.

• Berson, A. and Smith, S. J. (1997). Data warehousing, data mining and OLAP, McGraw–Hill

• Bramer, M. (2013). Principles of data mining, Springer–Verlang, London.

• Cha, S. and T.C. (2009). A genetic algorithm for constructing compact binary decision trees, Journal of pattern recognition research, volume 1, pp. 1–13.

• Chen, Ming-Syan & Han, Jiawei & Yu, Philip. (1996). Data mining: An overview from a database perspective. Knowledge and Data Engineering, IEEE Transactions on. 8. 866 - 883.

•Chen, Lei-da & Sakaguchi, Toru & Frolick, Mark. (2000). Data Mining Methods, Applications, and Tools. IS Management. 17. 1-6.

• Chul Kwak, Oh–Wook Kwon (2008). Cardiac disorder classification based on extreme learning machine, World academy of science, engineering and technology, volume 48.

• Coremen, T., Leiserson C., et al. (2009). Algorithms, 3rd edition, MIT press.

• Dean, J. (2014). Big data, data mining and machine learning, SAS Institute Inc.

• Deshpande, Shrinivas & Thakare, V. M. & Mandal, H & India, Amravati. (2010). Data Mining System and Applications: A Review. International Journal of Distributed and Parallel systems. 1.

• Findler, N.V. (1979). Associative networks–representation and use of knowledge by computers, Academic press

• Gan, G. and Wu, J. (2007). Data clustering: theory, algorithms and applications, Society for industrial and applied mathematics, Philadelphia.

• Garcia, S., Herrera, F., Luengo, J. (2015). Data preprocessing in data mining, Springer international publishing AG.

• Chakrabarti, S., Cox, E., Frank, E., Güting, R. H., Han, J., Jiang, X., Kamber, M., Lightstone, S. S., Nadeau, T. P., Neapolitan, R. E., Pyle, D., Refaat, M., Schneider, M., Teorey, T. J., & Witten I. H. (2008). Data mining: Know it all. Morgan Kaufmann

• Ghosh, J. et. al (2007). Top 10 algorithms in data mining, pp. 14–37, Knowledge information system.

• Giarratano, J. and Riley, G. (1994). Expert systems: Principles and programming, 2nd edition, International Thomson publishing.

• Gupta, G. K. (2014). Introduction to data mining with case studies (3rd ed.). PHI Learning Pvt. Ltd.
• Gupta, Manoj & Chandra, Pravin. (2020). A comprehensive survey of data mining. International Journal of Information Technology. 1-15.
• Haijian, Shi (2007). Best–first decision tree learning, Hamilton, New Zeland.
• Han, J. and Kamber, M. (2006). Data mining concepts and techniques, 2nd edition, Amsterdam, Kaufmann publishers.
• Han, J., Pei, J., & Tong, H. (2022). Data mining: Concepts and techniques (4th ed.). Morgan Kaufmann.
• Hanson, R. et al. (1991). Artificial intelligence research branch. Bayesian classification theory, Citeseer.
• Hayden Wimmer, Loreen M. Powell (2015). A comparison of open source tools for data science, Proceedings of the conference on information systems applied research Wilmington, North Carolina, USA, ISSN: 2167–1508, volume 8, No 3651.
• Haykin, S. (2009). Neural networks and machine learning, 3rd edition, Pearson.
• Jackson, Joyce. (2002). Data Mining; A Conceptual Overview. Communications of the Association for Information Systems. 8. 267-296.
• Jackson, P. (1999). Introduction to expert systems, 3rd edition, Addison–Wesley.
• Johnson, L. and Karavnou, E. (1988). Expert systems architectures, International Thomson publishing.
• Kaplana Rangra, Dr. K.L. Bansal (2014). Comparative study of data mining tools, International journal of advanced research in computer science αvδ software engineering, volume 4, issue 6, ISSN: 2277 128X, www.ijarcsse.com
• Konar, A. (2000). Artificial intelligence and soft computing: Behavioral cognitive modeling of the human brain, CRC Press LLC.
• Kuncheva, L.I. (2004). Combining pattern classifiers: methods and algorithms, John Willey and Sons.
• Larose, D. T. (2005). Discovering knowledge in data: An introduction to data mining. John Wiley & Sons.
• Larose, D. T., & Larose, C. D. (2014). Discovering knowledge in data: An introduction to data mining (2nd ed.). John Wiley & Sons
• Luger, G.F. and Stubblefield, W.A. (1998). Artificial intelligence: structures and strategies for complex problem solving, Addison–Wesley.
• Madni, Hussain Ahmad & Anwar, Zahid & Shah, Munam. (2017). Data mining techniques and applications — A decade review. 1-7.
• Maimon, O. and Rokach, L. (2008). Soft computing for knowledge discovery and data mining, Springer science & business media Ltd.
• Maimon, O. and Rokach, L. (2010). Data mining and Knowledge discovery handbook, Springer.
• Maksood, F.Z., & Achuthan, G. (2016). Analysis of Data Mining Techniques and its Applications. International Journal of Computer Applications, 140, 6-14
Mikut, R. and Reischl, M. (2011), Data mining tools. WIREs Data Mining Knowl Discov, 1: 431-443.
• Mitchell, T.M. (2010). *Generative and discriminative classifiers: Naive bayes, logistic regression*.
• Mitchell, T.M. (1997). Machine learning, McGraw Hill.
• Musen et al. (2014). Clinical decision support systems in biomedical informatics, pp. 643–674, Springer London.
• Neha Chauhan, Nisha Gautam (2015). Parametric comparison of data mining tools, *International journal of advanced technology in engineering & science*, volume N° 3, issue 11, ISSN: 2348-7550. www.ijates.com
• Nilsson, N.J. (1980). *Principles of artificial intelligence*, Tioga publishing Co.
• Ogiela, M., Jain, L., (2012). Computational intelligence paradigms in advanced pattern classification, Springer–Verlag.

• Prasad et al. (2011). A comparative study of machine learning algorithms as expert systems in medical diagnosis (Asthma) in advances in computer science and information technology, pp. 570–576, Springer Berlin Heidelberg.
• Rich, E. and Knight, K. (1991). *Artificial intelligence*, 2nd edition, McGraw–Hill.
• Roiger, J.R. and Geatz, W.M. (2003). *Data mining: A tutorial–based primer*, Addison–Wesley.
• Simovici, D.A. (2011). Data mining of medical data: Opportunities and challenges in mining association rules, University of Massachusetts Boston.
• Suh, S.C. (2012). *Practical applications of data mining*, Jones & Bertlett learning.
• Tan, Pang–Ning, et al, (2005). Introduction to data mining, Addison–Wesley & Longman publishing Co, Inc., Boston, MA, USA.
• Vaughan A. (2015). *Adaptive machine learning for modeling & control of non–stationary, near chaotic combustion in real time*, University of Michigan.
• Vercellis, C. (2009). Business intelligence: Data mining and optimization for decision making, John Wiley and sons.
• Winston, P.H. (1992). *Artificial intelligence*, 3rd edition, Addison–Wesley.
• Witten, I., Hall, M., Eibe, F. (2011). Data mining: practical machine learning tools and techniques, Morgan Kaufmann publishers.
• Wu, J. (2012). *Advances in K–means clustering: A data mining thinking*, Springer publishing company.
• Yong Joo Chung. A classification approach for the heart sound signals using hidden Markov models.