ΚΕΦΑΛΑΙΟ 1

Εισαγωγή

Ο όγκος των δεδομένων που φυλάσσονται στα αρχεία και στις βάσεις δεδομένων αυξάνεται με έναν εκπληκτικό ρυθμό. Την ίδια στιγμή, οι χρήστες αυτών των δεδομένων επιζητούν από αυτά πιο εξειδικευμένες πληροφορίες. Ένας διευθυντής πωλήσεων δεν είναι πια ικανοποιημένος με μία απλή λίστα από στοιχεία πελατών αλλά θέλει λεπτομερείς πληροφορίες σχετικά με τις προηγούμενες αγορές των πελατών καθώς επίσης και με τις προβλέψεις για τις μελλοντικές αγορές τους. Απλές ερωτήσεις, που μπορούν να εκφραστούν σε μία δομημένη γλώσσα ερωτήσεων (SQL), δεν αρκούν για να υποστηρίξουν αυτές τις αυξανόμενες απαιτήσεις για πληροφορίες. Η εξόρυξη γνώσης από δεδομένα παρεμβαίνει προκειμένου να ικανοποιήσει αυτές τις ανάγκες. Η *εξόρυξη γνώσης από δεδομένα* (data mining)[1] συχνά ορίζεται σαν η εύρεση πληροφοριών που είναι κρυμμένες σε μια βάση δεδομένων. Εναλλακτικά, η εξόρυξη γνώσης από δεδομένα ονομάστηκε εξερευνητική ανάλυση δεδομένων, ανακάλυψη καθοδηγούμενη από δεδομένα και συμπερασματική μάθηση.

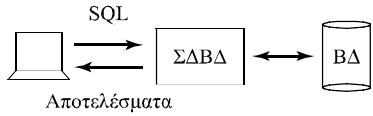
Οι παραδοσιακές ερωτήσεις σε βάσεις δεδομένων (Σχήμα 1.1), έχουν πρόσβαση σε μία βάση δεδομένων χρησιμοποιώντας μία καλά ορισμένη ερώτηση, η οποία εκφράζεται σε μία γλώσσα όπως είναι η SQL. Το αποτέλεσμα της ερώτησης αποτελείται από δεδομένα που προέρχονται από τις βάσεις δεδομένων και που ικανοποιούν την ερώτηση. Η έξοδος είναι συνήθως ένα υποσύνολο της βάσης των δεδομένων, αλλά μπορεί επίσης να είναι και μία εξαγόμενη όψη ή να περιέχει συναθροίσεις. Η προσπέλαση σε μία βάση δεδομένων, μέσω της εξόρυξης γνώσης από δεδομένα, διαφέρει από την παραδοσιακή προσπέλαση σε πολλά σημεία:

[1] ΣτΜ: Όπως σχολιάζουμε και στον Πρόλογο της Ελληνικής Έκδοσης, έχουμε επιλέξει τον όρο “εξόρυξη γνώσης από δεδομένα” ως μετάφραση του Αγγλικού όρου “data mining” αντί του διαδεδομένου “εξόρυξη δεδομένων” που εμφανίζεται συνήθως στην Ελληνική βιβλιογραφία.

- **Ερώτηση:** Η ερώτηση ίσως να μην είναι καλά σχηματισμένη ή να μην είναι με ακρίβεια ορισμένη. Εκείνος ο οποίος εξορύσσει τα δεδομένα, ίσως να μην είναι καν απόλυτα σίγουρος για αυτό που θέλει να βρει.

- **Δεδομένα:** Τα δεδομένα που προσπελαύνονται αποτελούν συνήθως μία διαφορετική έκδοση από εκείνα της αρχικής (επιχειρησιακής) βάσης δεδομένων. Τα δεδομένα έχουν καθαριστεί και τροποποιηθεί για να υποστηρίζουν καλύτερα τη διαδικασία της εξόρυξης.

- **Έξοδος:** Η έξοδος μιας ερώτησης εξόρυξης γνώσης από δεδομένα πιθανώς να μην αποτελεί ένα υποσύνολο της βάσης των δεδομένων. Αντί γι'αυτό, μπορεί να είναι η έξοδος από κάποιες αναλύσεις των περιεχομένων της βάσης δεδομένων.



ΣΧΗΜΑ 1.1: Προσπέλαση σε βάση δεδομένων.

Η τρέχουσα κατάσταση στην εξόρυξη γνώσης από δεδομένα είναι παρόμοια με αυτή που επικρατούσε στην επεξεργασία ερωτήσεων σε βάσεις δεδομένων στα τέλη της δεκαετίας του '60 και στις αρχές της δεκαετίας του '70. Κατά την επόμενη δεκαετία αναμφίβολα θα σημειωθεί μεγάλη πρόοδος, όσον αφορά στην εξόρυξη γνώσης. Πιθανώς θα δούμε την ανάπτυξη των μοντέλων «επεξεργασίας ερωτήσεων», των βιομηχανικών προτύπων (standards) και των αλγορίθμων που στοχεύουν στις εφαρμογές εξόρυξης γνώσης. Επίσης θα δούμε, πιθανότατα, νέες δομές δεδομένων σχεδιασμένες για την αποθήκευση βάσεων δεδομένων που χρησιμοποιούνται για εφαρμογές εξόρυξης γνώσης. Παρόλο που η περιοχή αυτή βρίσκεται ακόμα στην αρχή, την τελευταία δεκαετία έχουμε παρακολουθήσει μία εξάπλωση των αλγορίθμων, των εφαρμογών και των τεχνικών, σχετικών με εξόρυξη γνώσης. Το Παράδειγμα 1.1 παρουσιάζει μία τέτοια εφαρμογή.

ΠΑΡΑΔΕΙΓΜΑ 1.1

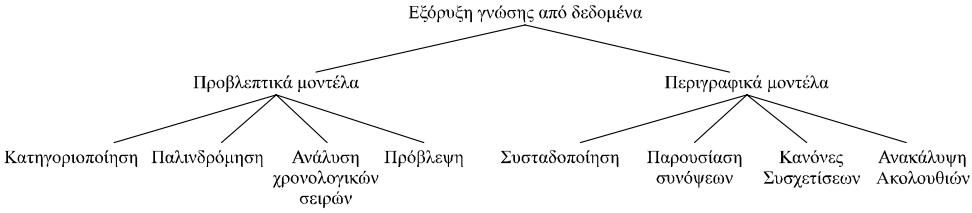
Οι εταιρείες πιστωτικών καρτών πρέπει να καθορίζουν, εάν θα εγκρίνουν αγορές μέσω πιστωτικών καρτών. Ας υποθέσουμε ότι με βάση το αγοραστικό ιστορικό ενός πελάτη, κάθε αγορά τοποθετείται σε μία από τέσσερις κατηγορίες: (1) να εγκριθεί, (2) να ζητηθούν επιπλέον στοιχεία ταυτότητας πριν από την έγκριση, (3) να μην εγκριθεί και (4) να μην εγκριθεί και να ενημερωθεί η αστυνομία. Οι λειτουργίες της εξόρυξης γνώσης από δεδομένα εξυπηρετούν δύο σκοπούς. Κατά πρώτον, τα δεδομένα του ιστορικού των πελατών πρέπει να εξεταστούν για να καθοριστεί πως ταιριάζουν στις τέσσερις κατηγορίες. Κατά δεύτερον, το πρόβλημα είναι πώς θα εφαρμοστεί αυτό το μοντέλο σε κάθε μία από τις νέες αγορές. Εάν και μπορεί να θεωρηθεί ότι το δεύτερο μέρος είναι πραγματικά μία απλή ερώτηση βάσης δεδομένων, το πρώτο μέρος δεν μπορεί να θεωρηθεί σαν τέτοια.

Η εξόρυξη γνώσης από δεδομένα περιλαμβάνει πολλούς διαφορετικούς αλγορίθμους για να εκπληρωθούν διαφορετικές εργασίες. Όλοι αυτοί οι αλγόριθμοι επιχειρούν να ταιριάξουν ένα μοντέλο στα δεδομένα. Οι αλγόριθμοι εξετάζουν τα δεδομένα και καθορίζουν ένα μοντέλο που να είναι το πλησιέστερο στα χαρακτηριστικά των δεδομένων που εξετάζονται. Οι αλγόριθμοι εξόρυξης γνώσης μπορεί να θεωρηθεί ότι αποτελούνται από τρία μέρη:

- *Μοντέλο*: Ο σκοπός του αλγορίθμου είναι να ταιριάξει το μοντέλο στα δεδομένα.
- *Προτίμηση*: Πρέπει να χρησιμοποιούνται κάποια κριτήρια για να ταιριάξει ένα μοντέλο έναντι ενός άλλου.
- *Αναζήτηση*: Όλοι οι αλγόριθμοι απαιτούν μία τεχνική για να κάνουν αναζήτηση στα δεδομένα.

Στο Παράδειγμα 1.1 τα δεδομένα μοντελοποιούνται με το διαχωρισμό τους σε τέσσερις κατηγορίες. Η αναζήτηση προϋποθέτει την εξέταση των δεδομένων που είναι σχετικά με προηγούμενες αγορές μέσω πιστωτικής κάρτας και των αποτελεσμάτων της εξέτασης, για να καθορίσει τα κριτήρια που πρέπει να χρησιμοποιηθούν, ώστε να οριστεί η δομή της κατηγορίας. Προτίμηση δίνεται στα κριτήρια εκείνα που φαίνεται να ταιριάζουν καλύτερα στα δεδομένα. Για παράδειγμα, πιθανώς να θέλαμε να εγκρίνουμε μία αγορά με πιστωτική κάρτα μικρού χρηματικού ποσού, όταν η πιστωτική κάρτα ανήκει σε έναν τακτικό πελάτη. Αντιστρόφως, δεν θα θέλαμε να εγκρίνουμε τη χρήση μίας πιστωτικής κάρτας για οποιαδήποτε αγορά σε περίπτωση που η κάρτα φέρεται ως κλεμμένη. Η διαδικασία αναζήτησης απαιτεί να είναι κατάλληλα ορισμένα τα κριτήρια που χρειάζονται, για να ταιριάξουν τα δεδομένα στις κατηγορίες.

Όπως φαίνεται στο Σχήμα 1.2, το μοντέλο που δημιουργείται μπορεί να είναι είτε προβλεπτικό είτε περιγραφικό μοντέλο. Σε αυτό το σχήμα, δείχνουμε κάτω από κάθε τύπο μοντέλου μερικές από τις πιο συνηθισμένες εργασίες εξόρυξης γνώσης από δεδομένα που χρησιμοποιούν αυτό το είδος του μοντέλου.



ΣΧΗΜΑ 1.2: Μοντέλα και εργασίες στην εξόρυξη γνώσης από δεδομένα.

Ένα *προβλεπτικό μοντέλο* (predictive model) κάνει μία πρόβλεψη για τις τιμές των δεδομένων, χρησιμοποιώντας γνωστά αποτελέσματα που έχει βρει από άλλα δεδομένα. Η μοντελοποίηση πρόβλεψης μπορεί να γίνει με βάση τη χρήση ιστορικών δεδομένων. Για παράδειγμα, η χρήση μιας πιστωτικής κάρτας μπορεί να μη γίνει δεκτή, όχι λόγω του πιστωτικού ιστορικού του πελάτη αλλά λόγω του ότι η τωρινή αγορά είναι σχετική με προηγούμενες αγορές οι οποίες διαδοχικά βρέθηκαν να έγιναν με κλεμμένες κάρτες. Το Παράδειγμα 1.1 χρησιμοποιεί μοντελοποίηση πρόβλεψης για να προβλέψει το πιστωτικό ρίσκο. Οι εργασίες εξόρυξης γνώσης από δεδομένα για το χτίσιμο ενός προβλεπτικού μο-

ντέλου περιλαμβάνουν κατηγοριοποίηση, παλινδρόμηση, ανάλυση χρονολογικών σειρών και πρόβλεψη. Η πρόβλεψη μπορεί να χρησιμοποιηθεί επίσης για να υποδηλώσει ένα συγκεκριμένο τύπο λειτουργίας εξόρυξης γνώσης από δεδομένα, όπως εξηγείται στη Ενότητα 1.1.4.

Ένα *περιγραφικό μοντέλο* (descriptive model) αναγνωρίζει πρότυπα ή συσχετίσεις στα δεδομένα. Αντίθετα από το προβλεπτικό, το περιγρφικό μοντέλο λειτουργεί σαν ένα μέσο που διερευνά τις ιδιότητες των δεδομένων που εξετάζονται, όχι να προβλέπει νέες ιδιότητες. Η συσταδοποίηση, η παρουσίαση συνόψεων, οι κανόνες συσχετίσεων και η ανακάλυψη ακολουθιών συνήθως θεωρούνται σαν περιγραφικές εργασίες από τη φύση τους.

1.1 ΒΑΣΙΚΕΣ ΕΡΓΑΣΙΕΣ ΕΞΟΡΥΞΗΣ ΓΝΩΣΗΣ ΑΠΟ ΔΕΔΟΜΕΝΑ

Στις επόμενες παραγράφους μελετάμε σε συντομία μερικές από τις λειτουργίες της εξόρυξης γνώσης. Ακολουθούμε τα βασικά σημεία των εργασιών που φαίνονται στο Σχήμα 1.2. Αυτή η καταγραφή δεν είναι πλήρης αλλά προσπαθεί να είναι αρκετά επεξηγηματική. Φυσικά, αυτές οι μεμονωμένες εργασίες μπορούν να συνδυαστούν για να πάρουμε πιο εξειδικευμένες εφαρμογές της εξόρυξης γνώσης από δεδομένα.

1.1.1 Κατηγοριοποίηση

Η *κατηγοριοποίηση* (classification) απεικονίζει τα δεδομένα σε προκαθορισμένες ομάδες ή *κατηγορίες – κλάσεις* (classes). Αναφέρεται συχνά σαν εποπτευόμενη μάθηση, επειδή οι κατηγορίες – κλάσεις καθορίζονται πριν ακόμη εξεταστούν τα δεδομένα. Δύο παραδείγματα εφαρμογών κατηγοριοποίησης είναι ο καθορισμός, εάν θα δοθεί ένα τραπεζικό δάνειο και ο προσδιορισμός του πιστωτικού ρίσκου. Οι αλγόριθμοι κατηγοριοποίησης απαιτούν οι κατηγορίες να ορίζονται με βάση τις τιμές των γνωρισμάτων των δεδομένων. Συχνά περιγράφουν αυτές τις κατηγορίες κοιτάζοντας τα χαρακτηριστικά δεδομένων που είναι ήδη γνωστό ότι ανήκουν στις κατηγορίες. Η *αναγνώριση προτύπου* (pattern recognition) αποτελεί ένα είδος κατηγοριοποίησης, όπου ένα πρότυπο εισόδου κατηγοριοποιείται σε μία από διάφορες κατηγορίες, με βάση την εγγύτητά του ως προς αυτές τις προκαθορισμένες κατηγορίες. Το Παράδειγμα 1.1 παρουσιάζει ένα γενικό πρόβλημα κατηγοριοποίησης. Το Παράδειγμα 1.2 δείχνει ένα απλό παράδειγμα αναγνώρισης προτύπου.

ΠΑΡΑΔΕΙΓΜΑ 1.2

Ένας σταθμός ελέγχου ασφάλειας αεροδρομίου χρησιμοποιείται για να καθοριστεί, εάν οι επιβάτες είναι πιθανοί τρομοκράτες ή εγκληματίες. Για να γίνει αυτό, σαρώνεται με ειδικό σαρωτή το πρόσωπο κάθε επιβάτη και αναγνωρίζεται το βασικό του πρότυπο (απόσταση μεταξύ των ματιών, μέγεθος και σχήμα στόματος, σχήμα κεφαλιού, κ.λπ.). Αυτό το πρότυπο συγκρίνεται με τις καταχωρήσεις μιας βάσης δεδομένων για να διαπιστωθεί, εάν ταιριάζει με κάποια πρότυπα που συσχετίζονται με γνωστοποιημένους παραβάτες του νόμου.

1.1.2 Παλινδρόμηση

Η *παλινδρόμηση* (regression) χρησιμοποιείται για να απεικονιστεί ένα στοιχειώδες δεδομένο σε μία πραγματική μεταβλητή πρόβλεψης. Στην πραγματικότητα, η παλιν-

δρόμηση περιλαμβάνει την εκμάθηση της συνάρτησης που κάνει αυτή την απεικόνιση. Η παλινδρόμηση προϋποθέτει ότι τα σχετικά δεδομένα ταιριάζουν με μερικά γνωστά είδη συνάρτησης (π.χ. γραμμική, λογαριθμική κλπ.) και μετά καθορίζει την καλύτερη συνάρτηση αυτού του είδους που μοντελοποιεί τα δεδομένα που έχουν δοθεί. Ένα είδος ανάλυσης σφάλματος χρησιμοποιείται για να καθορίσει ποια συνάρτηση είναι «η καλύτερη». Η τυπική γραμμική παλινδρόμηση που περιγράφεται στο Παράδειγμα 1.3 αποτελεί ένα απλό παράδειγμα παλινδρόμησης.

ΠΑΡΑΔΕΙΓΜΑ 1.3

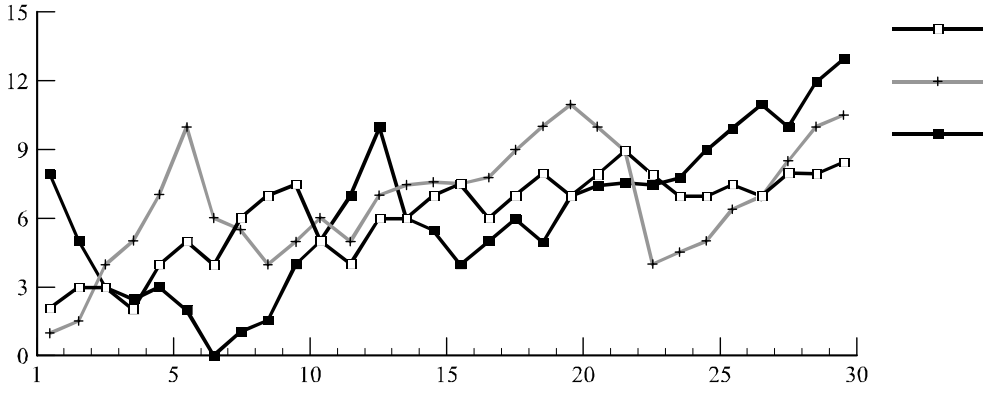
Μία καθηγήτρια πανεπιστημίου επιθυμεί οι αποταμιεύσεις της να φτάσουν σε ένα ορισμένο επίπεδο πριν από τη συνταξιοδότησή της. Περιοδικά, προβλέπει ποιες θα είναι οι αποταμιεύσεις της κατά τη συνταξιοδότησή της βασιζόμενη στην τρέχουσα τιμή τους και σε προηγούμενες τιμές. Χρησιμοποιεί έναν απλό γραμμικό τύπο παλινδρόμησης για να προβλέψει αυτήν την τιμή ταιριάζοντας προηγούμενες συμπεριφορές σε μία γραμμική συνάρτηση και στη συνέχεια χρησιμοποιεί αυτή τη συνάρτηση για να προβλέψει τις τιμές σε κάποιες στιγμές στο μέλλον. Βασιζόμενη σε αυτές τις τιμές, στη συνέχεια τροποποιεί το χαρτοφυλάκιο των επενδύσεών της.

1.1.3 Ανάλυση Χρονοσειρών

Με την *ανάλυση χρονολογικών σειρών* ή *χρονοσειρών* (time series analysis), μελετάται η τιμή ενός γνωρίσματος καθώς μεταβάλλεται στο χρόνο. Οι τιμές συνήθως λαμβάνονται σε ίσα χρονικά διαστήματα (ημερήσια, εβδομαδιαία, ωριαία, κοκ.). Για να παρασταθούν οπτικά οι χρονοσειρές χρησιμοποιείται ένα διάγραμμα χρονοσειρών (Σχήμα 1.3). Σε αυτό το σχήμα μπορεί κανείς εύκολα να δει ότι οι γραφικές παραστάσεις των Y και Z έχουν παρόμοια συμπεριφορά, ενώ του X φαίνεται να έχει λιγότερη αστάθεια. Υπάρχουν τρεις βασικές λειτουργίες που πραγματοποιούνται στην ανάλυση χρονοσειρών. Στη μία περίπτωση, χρησιμοποιούνται μονάδες μέτρησης απόστασης για να καθορίσουν την ομοιότητα ανάμεσα σε διαφορετικές χρονοσειρές. Στη δεύτερη περίπτωση, εξετάζεται η δομή της χρονοσειράς για να καθορίσει (και ίσως να κατηγοριοποιήσει) τη συμπεριφορά της. Μία τρίτη εφαρμογή θα μπορούσε να είναι η χρήση διαγραμμάτων χρονοσειρών για την πρόβλεψη μελλοντικών τιμών. Ένα παράδειγμα χρονοσειρών δίνεται στο Παράδειγμα 1.4.

ΠΑΡΑΔΕΙΓΜΑ 1.4

Ο κύριος Smith προσπαθεί να προσδιορίσει, εάν θα αγοράσει μετοχές από τις εταιρείες X, Y ή Z. Για τη χρονική περίοδο ενός μηνός παριστάνει γραφικά την ημερήσια τιμή της μετοχής κάθε εταιρείας. Το Σχήμα 1.3 δείχνει το διάγραμμα χρονοσειρών που δημιούργησε ο κύριος Smith. Χρησιμοποιώντας αυτό το διάγραμμα καθώς και σχετικές πληροφορίες που του παρέχει ο χρηματιστής του, ο κύριος Smith αποφασίζει να αγοράσει τη μετοχή X που είναι λιγότερο ασταθής και συνολικά παρουσιάζει ένα ελαφρά μεγαλύτερο σχετικό ποσό αύξησης από τις άλλες δύο μετοχές. Στην πραγματικότητα, οι μετοχές Y και Z έχουν παρόμοια συμπεριφορά. Η συμπεριφορά της Y, ανάμεσα στις μέρες 6 και 20, είναι πανομοιότυπη με αυτή της Z, ανάμεσα στις ημέρες 13 και 27.



ΣΧΗΜΑ 1.3: Διαγράμματα χρονοσειρών.

1.1.4 Πρόβλεψη

Πολλές από τις πρακτικές εφαρμογές εξόρυξης γνώσης μπορούν να θεωρηθούν σαν πρόβλεψη μελλοντικών καταστάσεων με γνώση των προηγούμενων και των σημερινών δεδομένων. Η *πρόβλεψη* (prediction) μπορεί να θεωρηθεί σαν ένα είδος κατηγοριοποίησης. (Σημείωση: Αυτή η εργασία εξόρυξης γνώσης είναι διαφορετική από το μοντέλο πρόβλεψης, παρόλο που η διαδικασία πρόβλεψης αποτελεί έναν τύπο μοντέλου πρόβλεψης.) Η διαφορά είναι ότι ως πρόβλεψη θεωρείται περισσότερο το να δίνεται τιμή σε μία μελλοντική κατάσταση παρά σε μία τρέχουσα. Εδώ αναφερόμαστε σε ένα είδος εφαρμογής παρά σε μια προσέγγιση μοντελοποίησης, όπως συζητήθηκε προηγουμένως. Οι εφαρμογές πρόβλεψης περιλαμβάνουν πρόγνωση πλημμύρων, αναγνώριση ομιλίας, μηχανική μάθηση και αναγνώριση προτύπου. Εάν και μπορούν να προβλεφθούν οι μελλοντικές τιμές με τεχνικές ανάλυσης χρονοσειρών ή παλινδρόμησης, μπορούν να χρησιμοποιηθούν επίσης και άλλες προσεγγίσεις. Το Παράδειγμα 1.5 επεξηγεί τη διαδικασία.

ΠΑΡΑΔΕΙΓΜΑ 1.5

Η πρόβλεψη μιας πλημμύρας είναι δύσκολο πρόβλημα. Μία προσέγγιση περιλαμβάνει τη χρήση οργάνων παρακολούθησης και ελέγχου που έχουν τοποθετηθεί σε διάφορα σημεία του ποταμού. Αυτά τα όργανα συλλέγουν δεδομένα σχετικά με την πρόβλεψη της πλημμύρας: ύψος της στάθμης του νερού, ποσότητα βροχής, χρόνος, υγρασία, κοκ. Στη συνέχεια μπορεί να προβλεφθεί το ύψος της στάθμης του νερού σε ένα σημείο του ποταμού στο οποίο είναι πιθανό να δημιουργηθεί πλημμύρα, βάσει των δεδομένων που συλλέχθηκαν από αισθητήρες που βρίσκονται στον ποταμό πάνω από το σημείο αυτό. Η πρόβλεψη πρέπει να γίνει σε σχέση με το χρόνο που συλλέχθηκαν τα δεδομένα.

1.1.5 Συσταδοποίηση

Η *συσταδοποίηση* (clustering) είναι παρόμοια με την κατηγοριοποίηση εκτός από το ότι οι συστάδες – ομάδες δεδομένων – δεν είναι προκαθορισμένες αλλά ορίζονται κυρίως

από τα ίδια τα δεδομένα. Η συσταδοποίηση αναφέρεται εναλλακτικά και σαν μη εποπτευόμενη μάθηση ή τμηματοποίηση. Μπορεί να θεωρηθεί σαν μια διαμέριση ή τμηματοποίηση των δεδομένων σε ομάδες που μπορεί να είναι ή να μην είναι διακριτές μεταξύ τους. Η συσταδοποίηση συνήθως επιτυγχάνεται με τον καθορισμό της ομοιότητας, ως προς προκαθορισμένα γνωρίσματα, ανάμεσα στα δεδομένα. Τα πιο σχετικά δεδομένα ομαδοποιούνται στις ίδιες ομάδες. Το Παράδειγμα 1.6 δίνει ένα απλό παράδειγμα συσταδοποίησης. Αφού οι ομάδες δεν είναι προκαθορισμένες χρειάζεται ένας ειδικός του πεδίου για να ερμηνεύσει τη σημασία των συστάδων που δημιουργούνται.

ΠΑΡΑΔΕΙΓΜΑ 1.6

Μία αλυσίδα πολυκαταστημάτων δημιουργεί ειδικούς καταλόγους, που στοχεύουν σε διάφορες δημογραφικές ομάδες, με βάση γνωρίσματα όπως το εισόδημα, ο τόπος διαμονής και τα φυσικά χαρακτηριστικά των δυνητικών πελατών (ηλικία, ύψος, βάρος κ.λπ.). Προκειμένου να καθορίσει σε ποιους από τους πελάτες των διαφόρων καταλόγων θα σταλεί ταχυδρομικά διαφημιστικό υλικό και προκειμένου να δημιουργηθούν καινούργιοι και πιο συγκεκριμένοι κατάλογοι, η εταιρεία κάνει συσταδοποίηση των πιθανών πελατών βασιζόμενη στις προκαθορισμένες τιμές γνωρισμάτων. Τα αποτελέσματα της συσταδοποίησης χρησιμοποιούνται στη συνέχεια από τη διεύθυνση προκειμένου να δημιουργηθούν ειδικοί κατάλογοι που θα διανεμηθούν στο πιο κατάλληλο τμήμα του πληθυσμού, βάσει της ομάδας που αντιστοιχεί σε αυτόν τον κατάλογο.

Μία ειδική κατηγορία συσταδοποίησης ονομάζεται *κατάτμηση* (segmentation). Με την κατάτμηση, μια βάση δεδομένων χωρίζεται σε διακριτές ομάδες παρόμοιων εγγραφών που ονομάζονται *τμήματα* (segments). Η κατάτμηση συχνά θεωρείται πανομοιότυπη με την συσταδοποίηση. Κατά άλλους, η κατάτμηση θεωρείται σαν ένας ειδικός τύπος συσταδοποίησης που εφαρμόζεται στην ίδια τη βάση δεδομένων. Σε αυτό το βιβλίο, οι δύο όροι, ομαδοποίηση και κατάτμηση, χρησιμοποιούνται εναλλακτικά.

1.1.6 Παρουσίαση Συνόψεων

Η *παρουσίαση συνόψεων* (summarization) απεικονίζει τα δεδομένα σε υποσύνολά τους με συνοδευτικές απλές περιγραφές. Η σύνοψη των δεδομένων ονομάζεται επίσης και *χαρακτηρισμός* (characterization) ή *γενίκευση* (generalization). Εξάγει ή παράγει αντιπροσωπευτικές πληροφορίες σχετικά με τις βάσεις δεδομένων. Αυτό γίνεται ανακτώντας, στην πραγματικότητα, τμήματα από τα δεδομένα. Εναλλακτικά, μπορούν να εξαχθούν από τα δεδομένα συνοπτικές πληροφορίες (όπως είναι ο μέσος όρος κάποιου αριθμητικού γνωρίσματος). Εν ολίγοις, η παρουσίαση συνόψεων χαρακτηρίζει τα περιεχόμενα της βάσης δεδομένων. Το Παράδειγμα 1.7 παρουσιάζει αυτήν τη διαδικασία.

ΠΑΡΑΔΕΙΓΜΑ 1.7

΄Ενα από τα πολλά κριτήρια που χρησιμοποιεί το *U.S. News & World Report* για να συγκρίνει τα πανεπιστήμια, είναι η μέση βαθμολογία SAT ή ACT [GM99]. Αυτό είναι

μία συνοπτική παρουσίαση των δεδομένων που χρησιμοποιείται για να αξιολογηθεί ο τύπος και το μορφωτικό επίπεδο των φοιτητών.

1.1.7 Κανόνες Συσχέτισης

Η *ανάλυση συνδέσμων* (link analysis), που εναλλακτικά αναφέρεται και σαν *ανάλυση συγγένειας* (affinity analysis) ή *συσχέτιση* (association), αναφέρεται στη διαδικασία εκείνη της εξόρυξης γνώσης που αποκαλύπτει συσχετίσεις μεταξύ των δεδομένων. Το καλύτερο παράδειγμα αυτού του είδους της εφαρμογής είναι ο προσδιορισμός κανόνων συσχετίσεων. Ένας *κανόνας συσχέτισης* (association rule) είναι ένα μοντέλο που αναγνωρίζει ειδικούς τύπους συσχέτισης μεταξύ δεδομένων. Αυτές οι συσχετίσεις συχνά χρησιμοποιούνται στις λιανικές πωλήσεις για να αναγνωριστούν προϊόντα που συχνά αγοράζονται μαζί. Το Παράδειγμα 1.8 δείχνει τη χρήση των κανόνων συσχετίσεων στην *"ανάλυση καλαθιού αγορών"* (market basket analysis). Εδώ τα δεδομένα που αναλύονται αποτελούνται από πληροφορίες σχετικά με τα προϊόντα που αγοράζει ένας πελάτης. Συσχετίσεις χρησιμοποιούνται επίσης σε πολλές άλλες εφαρμογές, όπως είναι η πρόβλεψη της αποτυχίας λειτουργίας των τηλεπικοινωνιακών διακοπτών.

ΠΑΡΑΔΕΙΓΜΑ 1.8

Ένα κατάστημα λιανικής πώλησης τροφίμων προσπαθεί να αποφασίσει εάν θα βάλει το ψωμί σε έκπτωση. Προκειμένου να βοηθηθεί ο πωλητής να καθορίσει τον αντίκτυπο αυτής της απόφασης, δημιουργεί κανόνες συσχέτισης που δείχνουν ποια άλλα προϊόντα αγοράζονται συχνά μαζί με το ψωμί. Βρίσκει ότι στο 60% των περιπτώσεων που πωλείται ψωμί πωλούνται και κουλουράκια και ότι στο 70% των περιπτώσεων πωλούνται επίσης και ζελεδάκια. Βασιζόμενος σε αυτά τα δεδομένα προσπαθεί να εκμεταλλευτεί τη συσχέτιση ανάμεσα στο ψωμί, τα κουλουράκια και τα ζελεδάκια βάζοντας μερικά κουλουράκια και μερικά ζελεδάκια στο τέλος του διαδρόμου εκεί που είναι τοποθετημένο το ψωμί. Επιπλέον αποφασίζει να μη βάλει αυτά τα προϊόντα ταυτόχρονα σε έκπτωση.

Η χρήση των κανόνων συσχετίσεων για τις όποιες αποφάσεις πρέπει να γίνεται πολύ προσεκτικά επειδή υπάρχει ο κίνδυνος αυτές οι συσχετίσεις να είναι τυχαίες. Οι συσχετίσεις αυτές μπορεί να μην αντιπροσωπεύουν καμία έμφυτη σχέση ανάμεσα στα δεδομένα (κάτι που ισχύει για παράδειγμα στις συναρτησιακές εξαρτήσεις). Πιθανώς να μην υπάρχει καμία συσχέτιση ανάμεσα στο ψωμί και στα κουλουράκια, η οποία να προκαλεί τα δύο προϊόντα να αγοράζονται μαζί. Ούτε υπάρχει καμία εγγύηση ότι αυτή η συσχέτιση θα εμφανίζεται και στο μέλλον. Ωστόσο, οι κανόνες συσχετίσεων μπορούν να χρησιμοποιηθούν για να βοηθήσουν τη διοίκηση των καταστημάτων λιανικής πώλησης στην αποτελεσματική διαφήμιση, στο μάρκετινγκ και στον έλεγχο της αποθήκης.

1.1.8 Ανακάλυψη Ακολουθιών

Η *ακολουθιακή ανάλυση* (sequential analysis) ή αλλιώς *ανακάλυψη ακολουθιών* (sequence discovery) χρησιμοποιείται για να καθοριστούν σειριακά πρότυπα στα δεδομένα. Αυτά τα πρότυπα βασίζονται σε μία χρονική ακολουθία ενεργειών. Αυτά τα πρότυπα είναι παρό-

μοια με τις συσχετίσεις στο ότι συσχετίζονται τα δεδομένα (ή τα γεγονότα) που εξάγονται, με τη διαφορά ότι η συσχέτισή τους αυτή βασίζεται στο χρόνο. Αντίθετα με την ανάλυση καλαθιού αγορών, που προϋποθέτει να γνωρίζουμε ποια προϊόντα αγοράστηκαν ταυτόχρονα, στην ανακάλυψη ακολουθιών τα προϊόντα αγοράζονται με κάποια σειρά κατά τη διάρκεια μιας περιόδου. Το Παράδειγμα 1.9 επεξηγεί την ανακάλυψη μερικών απλών προτύπων. Ένας παρόμοιος τύπος ανακάλυψης μπορεί να βρεθεί μέσα στην ακολουθία των προϊόντων που αγοράζονται. Για παράδειγμα, οι περισσότεροι άνθρωποι που αγοράζουν CD players ίσως και να αγοράζουν μέσα σε μία εβδομάδα και CDs. Όπως θα δούμε, οι χρονικοί κανόνες συσχέτισης πράγματι εμπίπτουν σε αυτήν την κατηγορία.

ΠΑΡΑΔΕΙΓΜΑ 1.9

Ο webmaster της εταιρείας XYZ περιοδικά αναλύει τα δεδομένα καταγραφών στο Web για να προσδιορίσει τον τρόπο που οι χρήστες έχουν πρόσβαση στις ιστοσελίδες της εταιρείας. Συγκεκριμένα, ενδιαφέρεται να προσδιορίσει τις ακολουθίες ιστοσελίδων που προσπελαύνονται συχνότερα. Ανακαλύπτει ότι το 70% των ανθρώπων που επισκέπτονται τη σελίδα A ακολουθούν ένα από τα ακόλουθα πρότυπα συμπεριφοράς: A, B, C ή A, D, B, C ή A, E, B, C. Στη συνέχεια, αποφασίζει να προσθέσει έναν απευθείας σύνδεσμο από τη σελίδα A στη σελίδα C.

1.2 Η ΕΞΟΡΥΞΗ ΩΣ ΣΤΑΔΙΟ ΤΗΣ ΑΝΑΚΑΛΥΨΗΣ ΓΝΩΣΗΣ ΣΕ ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ

Οι όροι *ανακάλυψη γνώσης σε βάσεις δεδομένων* (Knowledge Discovery in Databases – σε συντομία, KDD) και *εξόρυξη γνώσης από δεδομένα* (data mining) συχνά χρησιμοποιούνται εναλλακτικά για την ίδια έννοια. Στην πραγματικότητα, έχουν δοθεί πολλές διαφορετικές ονομασίες σε αυτήν τη διαδικασία ανακάλυψης χρήσιμων (κρυμμένων) προτύπων από τα δεδομένα: εξαγωγή γνώσης, ανακάλυψη πληροφοριών, εξερευνητική ανάλυση δεδομένων, συγκομιδή πληροφοριών, μη επιβλεπόμενη αναγνώριση προτύπου. Στα τελευταία χρόνια, ο όρος KDD έχει χρησιμοποιηθεί για να εκφράσει μια διαδικασία που αποτελείται από πολλά βήματα, ένα από τα οποία είναι η εξόρυξη γνώσης από δεδομένα. Αυτή είναι και η προσέγγιση που υιοθετείται σε αυτό το βιβλίο. Οι ακόλουθοι ορισμοί είναι τροποποιήσεις των αρχικών που δίνονται στα [FPSS96c, FPSS96a].

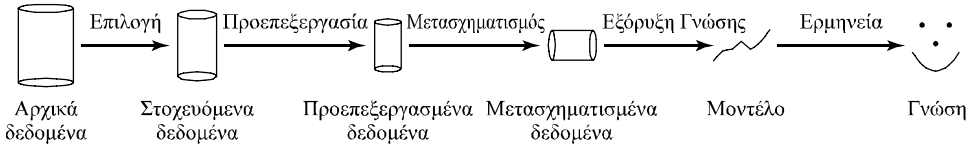
> **ΟΡΙΣΜΟΣ 1.1.** Η **ανακάλυψη γνώσης σε βάσεις δεδομένων (KDD)** είναι η διαδικασία εύρεσης χρήσιμων πληροφοριών και προτύπων στα δεδομένα.

> **ΟΡΙΣΜΟΣ 1.2.** Η **εξόρυξη γνώσης από δεδομένα** είναι η χρήση αλγορίθμων για την εξαγωγή των πληροφοριών και προτύπων που παράγονται με τη διαδικασία KDD.

Η διαδικασία KDD συχνά θεωρείται πολύπλοκη, ωστόσο εμείς τη θεωρούμε στην πιο γενική της μορφή που περιλαμβάνει συγκεκριμένα απλούστερα βήματα. Μία τυπική ερώτηση SQL σε μία βάση δεδομένων μπορεί να θεωρηθεί σαν το τμήμα της εξόρυξης γνώσης από δεδομένα μιας KDD διαδικασίας. Πράγματι, αυτό μπορεί να θεωρηθεί σαν κάτι απλό και συνηθισμένο. Όμως, δεν ίσχυε κάτι τέτοιο πριν από 30 χρόνια. Εάν ήταν

να μεταφερθούμε 30 χρόνια μπροστά στο μέλλον, τις διαδικασίες που σήμερα βρί-σκουμε δύσκολες και πολύπλοκες θα τις θεωρούσαμε εξίσου απλές. Ο ορισμός της KDD περιλαμβάνει τη λέξη-κλειδί *χρήσιμο*. Εάν και μερικοί ορισμοί περιλαμβάνουν τον όρο «εν δυνάμει χρήσιμο», πιστεύουμε ότι, εάν οι πληροφορίες που βρίσκονται σε αυτή τη διαδικασία δεν είναι χρήσιμες, τότε δεν είναι στην πραγματικότητα πληροφορίες. Φυσικά το αν κάτι είναι χρήσιμο ή όχι, είναι σχετική έννοια και εξαρτάται από τα άτομα που εμπλέκονται.

Η KDD είναι μία διαδικασία που περιλαμβάνει πολλά διαφορετικά βήματα. Η είσοδος σε αυτή τη διαδικασία είναι τα δεδομένα, και οι χρήσιμες πληροφορίες που επιθυμούν οι χρήστες είναι η έξοδος. Όμως, ο αντικειμενικός σκοπός δεν είναι από την αρχή ξεκάθαρος. Η διαδικασία από μόνη της είναι διαδραστική και συνήθως απαιτείται πολύς χρόνος για την ολοκλήρωσή της. Για να διασφαλιστεί η χρησιμότητα και η ακρίβεια των αποτελεσμάτων αυτής της διαδικασίας, συνήθως χρειάζεται η συνεργασία ειδικών του πεδίου εφαρμογής με ειδικούς της διαδικασίας KDD καθ' όλη τη διάρκεια της διαδικασίας αυτής. Το Σχήμα 1.4 (τροποποιημένο από το [FPSS96c]) επεξηγεί τη συνολική διαδικασία της ανακάλυψης γνώσης σε βάσεις δεδομένων.



ΣΧΗΜΑ 1.4: Διαδικασία KDD (τροποποιημένη από [FPSS96c]).

Η KDD διαδικασία αποτελείται από τα επόμενα πέντε βήματα [FPSS96c]:

- **Επιλογή:** Τα δεδομένα που χρειάζονται για τη διαδικασία της ανακάλυψης γνώσης μπορούν να προέλθουν από πολλές διαφορετικές και ετερογενείς πηγές δεδομένων. Σε αυτό το πρώτο βήμα συλλέγονται δεδομένα από διάφορες βάσεις δεδομένων, αρχεία και μη ηλεκτρονικές πηγές.

- **Προεπεξεργασία:** Τα δεδομένα που πρόκειται να χρησιμοποιηθούν κατά την διαδικασία, ίσως να είναι λανθασμένα ή ελλιπή. Ίσως υπάρχουν ανώμαλα δεδομένα από πολλαπλές πηγές που περιλαμβάνουν διαφορετικούς τύπους δεδομένων και διαφορετικές μονάδες μέτρησης. Σε αυτό το βήμα μπορούν να πραγματοποιηθούν πολλές και διαφορετικές δραστηριότητες. Τα λανθασμένα δεδομένα μπορεί να διορθωθούν ή να αφαιρεθούν, ενώ τα ελλιπή δεδομένα πρέπει να συλλεχθούν ή να εκτιμηθούν (συχνά χρησιμοποιώντας εργαλεία εξόρυξης γνώσης από δεδομένα).

- **Μετασχηματισμός:** Τα δεδομένα που προέρχονται από διαφορετικές πηγές χρειάζεται να μετατραπούν σε ένα κοινό σχήμα για την περαιτέρω επεξεργασία τους. Μερικά δεδομένα ίσως απαιτείται να κωδικοποιηθούν ή να μετασχηματιστούν σε πιο χρήσιμα σχήματα. Μπορεί να μειωθούν τα δεδομένα για να ελαττωθεί ο αριθμός των πιθανών τιμών των δεδομένων που θα ληφθούν υπόψη.

- **Εξόρυξη γνώσης από δεδομένα:** Με βάση το είδος της εξόρυξης που είναι να

εκτελεστεί, σε αυτό το βήμα εφαρμόζονται αλγόριθμοι στα τροποποιημένα δεδομένα για να προκύψουν τα επιθυμητά αποτελέσματα.

- **Ερμηνεία / αξιολόγηση:** Είναι πολύ σημαντικό το πώς θα παρουσιαστούν στους χρήστες τα αποτελέσματα της εξόρυξης γνώσης, επειδή η χρησιμότητα ή μη των αποτελεσμάτων μπορεί να εξαρτάται ακριβώς από αυτήν την παρουσίαση. Σε αυτό το τελευταίο βήμα χρησιμοποιούνται διάφορες στρατηγικές οπτικοποίησης και γραφικές διεπαφές χρήστη (GUI).

Για να προετοιμαστούν τα δεδομένα για εξόρυξη γνώσης και να παραχθούν αποτελέσματα με περισσότερο νόημα χρησιμοποιούνται τεχνικές μετασχηματισμού. Για να διευκολυνθεί η χρήση αυτών των τεχνικών που απαιτούν ειδικούς τύπους κατανομής δεδομένων μπορεί να τροποποιηθεί η πραγματική κατανομή των δεδομένων. Μπορούν να συνδυαστούν τιμές γνωρισμάτων για να δώσουν νέες τιμές, μειώνοντας έτσι την πολυπλοκότητα των δεδομένων. Για παράδειγμα, η σημερινή ημερομηνία και η ημερομηνία γέννησης, θα μπορούσαν να αντικατασταθούν από την ηλικία. Ένα γνώρισμα θα μπορούσε να αντικατασταθεί από ένα άλλο. Ένα παράδειγμα θα ήταν η αντικατάσταση μίας ακολουθίας που περιέχει τις πραγματικές τιμές ενός γνωρίσματος, με τις διαφορές μεταξύ των διαδοχικών τιμών. Μπορούμε να χειριστούμε ευκολότερα τις τιμές των γνωρισμάτων διαμερίζοντάς τις σε διαστήματα και χρησιμοποιώντας αυτά τα διακριτά διαστήματα τιμών. Μερικές τιμές δεδομένων μπορούν και να αφαιρεθούν. Οι ακραίες τιμές, που εμφανίζονται σπάνια, μπορούν να αφαιρεθούν. Αν εφαρμοστεί μια μεταβλητή στις τιμές μπορούν να τροποποιηθούν τα δεδομένα. Μία συνηθισμένη συνάρτηση μετασχηματισμού είναι η χρήση του λογάριθμου της τιμής παρά της ίδιας της τιμής. Αυτές οι τεχνικές κάνουν την διαδικασία της εξόρυξης γνώσης από δεδομένα ευκολότερη με τη μείωση των διαστάσεων (του πλήθους των γνωρισμάτων) ή τη μείωση της μεταβλητότητας των τιμών των δεδομένων. Η αφαίρεση των ακραίων τιμών μπορεί πραγματικά να βελτιώσει την ποιότητα των αποτελεσμάτων. Όμως, η τροποποίηση δεδομένων πρέπει να γίνει με προσοχή, όπως με προσοχή πρέπει να γίνουν και όλα τα άλλα βήματα της διαδικασίας KDD. Εάν η τροποποίηση γίνει με λανθασμένο τρόπο τότε θα αλλάξουν ριζικά τα δεδομένα και τα αποτελέσματα από την εξόρυξη γνώσης από δεδομένα θα είναι ανακριβή.

Η *οπτικοποίηση* (visualization) αναφέρεται ως η οπτική παρουσίαση των δεδομένων. Η κλασική έκφραση που λέει ότι «μία εικόνα αξίζει όσο χίλιες λέξεις» βεβαίως και είναι σωστή όταν εξετάζουμε τη δομή των δεδομένων. Για παράδειγμα, μία γραφική παράσταση που δείχνει την κατανομή μιας μεταβλητής δεδομένων είναι πιο κατανοητή και ίσως πιο κατατοπιστική από έναν τύπο για την αντίστοιχη κατανομή. Η χρήση των τεχνικών οπτικοποίησης επιτρέπει στους χρήστες να συνοψίζουν, να εξάγουν και να αντιλαμβάνονται πιο πολύπλοκα αποτελέσματα από αυτά που τους επιτρέπουν να αντιλαμβάνονται οι πιο μαθηματικοί και πιο περιγραφικοί τρόποι παρουσίασης των αποτελεσμάτων. Οι τεχνικές οπτικοποίησης μπορεί να είναι:

- **Γραφικές:** Μπορούν να χρησιμοποιηθούν οι παραδοσιακές γραφικές παραστάσεις, όπως τα ραβδογράμματα, οι πίτες, τα ιστογράμματα και τα γραμμογράμματα.

- **Γεωμετρικές:** Οι γεωμετρικές τεχνικές περιλαμβάνουν θηκογράμματα και διαγράμματα διασποράς.

- **Βασισμένες σε εικονίδια:** Χρησιμοποιώντας σχήματα, χρώματα, ή εικονίδια μπορούμε να βελτιώσουμε την παρουσίαση των αποτελεσμάτων.

- **Βασισμένες σε εικονοστοιχεία:** Με αυτές τις τεχνικές, κάθε τιμή που αντιστοιχεί σε δεδομένο παρουσιάζεται σαν ένα εικονοστοιχείο χρωματισμένο με μοναδικό τρόπο.

- **Ιεραρχικές:** Αυτές οι τεχνικές διαιρούν ιεραρχικά το χώρο παρουσίασης (οθόνη) σε περιοχές, βασιζόμενες στις τιμές των δεδομένων.

- **Υβριδικές:** Οι προηγούμενες τεχνικές μπορούν να συνδυαστούν σε μία ενιαία παρουσίαση.

Οποιαδήποτε από τις παραπάνω προσεγγίσεις μπορεί να είναι 2-διάστατη ή 3-διάστατη. Μπορούν να χρησιμοποιηθούν εργαλεία οπτικοποίησης, για να συνοψίσουν τα δεδομένα, όπως θα έκανε από μόνη της μία τεχνική εξόρυξης γνώσης. Επιπρόσθετα, μπορεί να χρησιμοποιηθεί η οπτικοποίηση για να εμφανίσει τα πολύπλοκα αποτελέσματα των εργασιών της εξόρυξης γνώσης από δεδομένα.

Η διαδικασία εξόρυξης γνώσης είναι από μόνη της πολύπλοκη. Όπως θα δούμε σε επόμενα κεφάλαια, υπάρχουν πολλοί αλγόριθμοι και πολλές εφαρμογές της εξόρυξης γνώσης από δεδομένα. Αυτοί οι αλγόριθμοι πρέπει να εφαρμοστούν προσεκτικά για να είναι αποτελεσματικοί. Τα πρότυπα που ανακαλύπτονται πρέπει να ερμηνεύονται και να αξιολογούνται σωστά για να προκύπτουν πληροφορίες που να είναι ακριβείς και να έχουν κάποια ιδιαίτερη σημασία.

1.2.1 Η Ανάπτυξη της Εξόρυξης γνώσης από δεδομένα

Η σημερινή εξέλιξη στις λειτουργίες και στα προϊόντα της εξόρυξης γνώσης από δεδομένα είναι αποτέλεσμα πολλών χρόνων επιρροής από πολλούς επιστημονικούς κλάδους, όπως είναι οι βάσεις δεδομένων, η ανάκτηση πληροφοριών, η στατιστική, οι αλγόριθμοι, και η μηχανική μάθηση (Σχήμα 1.5).

ΣΧΗΜΑ 1.5: Ιστορική άποψη της εξόρυξης γνώσης από δεδομένα.

Μία άλλη περιοχή της πληροφορικής, που επηρέασε σημαντικά τη διαδικασία KDD είναι η περιοχή των πολυμέσων και των γραφικών. Ένας βασικός στόχος είναι να μπορέσει να δοθεί μία περιγραφή με νόημα στα αποτελέσματα της διαδικασίας KDD. Επειδή προκύπτουν συχνά πολλά διαφορετικά αποτελέσματα, είναι πολύπλοκο πρόβλημα να δοθεί μία τέτοια περιγραφή. Οι τεχνικές οπτικοποίησης συχνά περιλαμβάνουν εξειδικευμένα πολυμέσα και γραφικές παρουσιάσεις. Επιπλέον, οι τεχνικές εξόρυξης γνώσης από δεδομένα μπορούν να εφαρμοστούν σε εφαρμογές πολυμέσων.

Αντίθετα με τη μέχρι τώρα έρευνα σε αυτές τις διαφορετικές περιοχές, μία μεγάλη τάση στην περιοχή των βάσεων δεδομένων θέλει να συνδυάζονται τα αποτελέσματα από αυτούς τους, διαφορετικούς κατά τα φαινόμενα, επιστημονικούς κλάδους σε μία ενοποιημένη προσέγγιση με βάση τα δεδομένα ή τους αλγορίθμους. Αν και η εξέλιξη αυτή βρίσκεται σε νηπιακό στάδιο, ο τελικός της στόχος είναι να δημιουργήσει μία σφαιρική εικόνα της περιοχής η οποία θα διευκολύνει την ολοκλήρωση, των διάφορων τύπων των εφαρμογών σε υπάρχοντα πεδία για το χρήστη.

Ο Πίνακας 1.1 δείχνει τις εξελίξεις στις περιοχές της Τεχνητής Νοημοσύνης (ΤΝ), της Ανάκτησης Πληροφοριών (ΑΠ), των Βάσεων Δεδομένων (ΒΔ), και της Στατιστικής που κυριαρχούν στη σύγχρονη εικόνα της εξόρυξης γνώσης από δεδομένα. Αυτές οι διαφορετικές επιρροές από το παρελθόν, οι οποίες οδήγησαν στην ανάπτυξη της περιοχής της εξόρυξης γνώσης από δεδομένα, συντέλεσαν στη δημιουργία διαφορετικών απόψεων για το τι είναι στην πραγματικότητα οι λειτουργίες της εξόρυξης γνώσης [RG99]:

- Η ***επαγωγή*** χρησιμοποιείται για να οδηγηθούμε από μία πολύ εξειδικευμένη γνώση σε πιο γενικές πληροφορίες. Αυτό το είδος της τεχνικής συχνά υπάρχει στις εφαρμογές της ΤΝ.

- Επειδή ο πρωταρχικός αντικειμενικός στόχος της εξόρυξης γνώσης από δεδομένα είναι να περιγράψει μερικά χαρακτηριστικά ενός συνόλου δεδομένων από ένα γενικό μοντέλο, αυτή η προσέγγιση μπορεί να θεωρηθεί σαν ένα είδος ***συμπίεσης.*** Εδώ, τα λεπτομερή δεδομένα της βάσης δεδομένων «αφαιρούνται» και συμπιέζονται σε μία μικρότερη περιγραφή των χαρακτηριστικών των δεδομένων που βρίσκονται στο μοντέλο.

- Όπως διατυπώθηκε προηγουμένως, η διαδικασία της εξόρυξης γνώσης από δεδομένα μπορεί να θεωρηθεί από μόνη της σαν ένας τύπος διαδικασίας υποβολής ***ερωτήσεων*** στη σχετική βάση δεδομένων. Πράγματι, η έρευνα στην εξόρυξη γνώσης από δεδομένα τείνει προς την κατεύθυνση εκείνη όπου αναζητείται ο τρόπος ορισμού μιας ερώτησης εξόρυξης γνώσης και το κατά πόσο μπορεί να αναπτυχθεί μία γλώσσα ερωτήσεων (σαν την SQL) που να περιλαμβάνει τόσους πολλούς διαφορετικούς τύπους επερωτήσεων εξόρυξης γνώσης.

- Η περιγραφή μιας μεγάλης βάσης δεδομένων μπορεί να θεωρηθεί σαν να χρησιμοποιούμε ***προσέγγιση*** προκειμένου να αποκαλυφθούν κρυμμένες πληροφορίες σχετικές με τα δεδομένα.

- Όταν εργαζόμαστε με μεγάλες βάσεις δεδομένων, η επίδραση του μεγέθους και η ικανότητα ανάπτυξης ενός αφηρημένου μοντέλου μπορούν να θεωρηθούν σαν ένας τύπος προβλήματος ***αναζήτησης***.

ΠΙΝΑΚΑΣ 1.1: Χρονοδιάγραμμα της εξέλιξης της εξόρυξης γνώσης από δεδομένα

Χρόνος	Περιοχή	Συνεισφορά	Αναφορά
Τέλη του 1700	Στατ.	Θεώρημα των πιθανοτήτων του Bayes	[Bay63]
Αρχές του 1900	ΤΝ	Ανάλυση με παλινδρόμηση	
Αρχές του 1920	Στατ.	Εκτιμήτρια μέγιστης πιθανοφάνειας	[Fis21]
Αρχές του 1940	ΤΝ	Νευρωνικά δίκτυα	[MP43]
Αρχές του 1950		Πλησιέστερος γείτονας	[FJ51]
Αρχές του 1950		Απλός σύνδεσμος	[FLP+51]
Τέλη του 1950	ΤΝ	Perceptron	Ros58]
Τέλη του 1950	Στατ.	Επαναδειγματοληψία, μείωση μεροληψίας, Jackknife εκτιμήτρια	
Αρχές του 1960	ΤΝ	Έναρξη μηχανικής μάθησης	[FF63]
Αρχές του 1960	ΒΔ	Μαζικές αναφορές	
Μέσα του 1960		Δένδρα αποφάσεων	[HMS66]
Μέσα του 1960	Στατ.	Γραμμικά μοντέλα κατηγοριοποίησης	[Nil65]
	ΑΠ	Μέτρα ομοιότητας	
	ΑΠ	Συσταδοποίηση	
	Στατ.	Εξερευνητική ανάλυση δεδομένων (EDA)	
Τέλη του 1960	ΒΔ	Σχεσιακό μοντέλο δεδομένων	[Cod70]
Αρχές του 1970	ΑΠ	Έξυπνα συστήματα ΑΠ	[Sal71]
Μέσα του 1970	ΤΝ	Γενετικοί αλγόριθμοι	[Hol75]
Τέλη του 1970	Στατ.	Εκτίμηση με μη πλήρη δεδομένα (EM αλγόριθμος)	[DLR77]
Τέλη του 1970	Στατ.	Συσταδοποίηση K-means	
Αρχές του 1980	ΤΝ	Αυτο-οργανωμένα δίκτυα Kohonen	[Koh82]
Μέσα του 1980	ΤΝ	Αλγόριθμοι δένδρων αποφάσεων	[Qui86]
Αρχές του 1990	ΒΔ	Αλγόριθμοι κανόνων συσχετίσεων Παγκόσμιος ιστός και μηχανές αναζήτησης	
1990	ΒΔ	Αποθήκες δεδομένων	
1990	ΒΔ	Άμεση αναλυτική επεξεργασία (OLAP)	

Έχει ενδιαφέρον να σκεφτούμε τα διαφορετικά προβλήματα εξόρυξης γνώσης από δεδομένα και πώς αυτά μπορούν να ειδωθούν από διαφορετικές σκοπιές ανάλογα με την οπτική γωνία και το επιστημονικό υπόβαθρο του ερευνητή ή του σχεδιαστή. Αναφέρουμε αυτές τις διαφορετικές απόψεις μόνο και μόνο για να δώσουμε στον αναγνώστη την «πλήρη εικόνα» της εξόρυξης γνώσης. Συχνά, λόγω διαφορετικού επιστημονικού υπόβαθρου, μπορούμε να βρούμε τα ίδια προβλήματα (και ίσως ακόμα και τις ίδιες λύσεις) να περιγράφονται με διαφορετικό τρόπο. Πράγματι, οι διαφορετικές ορολογίες μπορούν να οδηγήσουν σε παρανοήσεις και δυσαρέσκεια μεταξύ των εμπλεκομένων. Μπορούμε να δούμε στατιστικούς να εκφράζουν τις ανησυχίες τους όταν γενικεύονται εκτιμήσεις (προσεγγίσεις), ενώ δεν θα έπρεπε να γενικεύονται. Οι ερευνητές των βάσεων δεδομένων εκφράζουν την ανησυχία τους για τη μη αποδοτικότητα πολλών από τους προτεινόμενους αλγορίθμους ΤΝ, ιδίως όταν οι τελευταίοι χρησιμοποιούνται σε πολύ μεγάλες βάσεις δεδομένων. Η ΑΠ και όσοι ενδιαφέρονται για την εξόρυξη γνώσης από δεδομένα - κείμενα ανησυχούν για το ότι πολλοί αλγόριθμοι στοχεύουν μόνο σε αριθμητικά δεδομένα. Η προσέγγιση που υιοθετείται σε αυτό το βιβλίο είναι να εξεταστούν οι συνεισφορές όλων αυτών των επιστημονικών κλάδων στην εξόρυξη γνώσης από δεδομένα.

Υπάρχουν τουλάχιστον δύο θέματα που χαρακτηρίζουν μία προσέγγιση εξέτασης των εννοιών της εξόρυξης γνώσης από τη σκοπιά των βάσεων δεδομένων: η αποτελεσματικότητα και η κλιμάκωση. Όλες οι λύσεις των προβλημάτων πρέπει να είναι ικανές να εφαρμόζονται στις βάσεις δεδομένων του πραγματικού κόσμου. Όσον αφορά στην αποτελεσματικότητα, ενδιαφερόμαστε για τους αλγορίθμους και τις δομές δεδομένων που χρησιμοποιούνται. Θα μπορούσε να χρησιμοποιηθεί παραλληλισμός για να βελτιωθεί η αποτελεσματικότητα. Επιπλέον, είναι σημαντικό πώς συμπεριφέρονται οι προτεινόμενοι αλγόριθμοι καθώς τροποποιείται η βάση δεδομένων. Πολλοί αλγόριθμοι εξόρυξης γνώσης που έχουν προταθεί μπορούν να δουλέψουν καλά σε μία στατική βάση δεδομένων, αλλά είναι ιδιαίτερα αναποτελεσματικοί όταν γίνονται αλλαγές στη βάση δεδομένων. Ενδιαφερόμαστε κυρίως για το πώς αποδίδουν οι αλγόριθμοι σε πολύ μεγάλες βάσεις δεδομένων παρά για το πώς λειτουργούν σε απλοϊκά προβλήματα. Επίσης υποθέτουμε ότι τα δεδομένα αποθηκεύονται στο δίσκο και ότι ενδεχομένως είναι κατανεμημένα.

1.3 ΘΕΜΑΤΑ ΕΞΟΡΥΞΗΣ ΓΝΩΣΗΣ ΑΠΟ ΔΕΔΟΜΕΝΑ

Υπάρχουν πολλά σημαντικά θέματα υλοποίησης που σχετίζονται με την εξόρυξη γνώσης από δεδομένα:

1. **Ανθρώπινη αλληλεπίδραση:** Αφού τα προβλήματα της εξόρυξης γνώσης από δεδομένα συνήθως δεν ορίζονται με ακρίβεια, μπορεί να είναι αναγκαία μια αλληλεπίδραση μεταξύ των ειδικών του πεδίου εφαρμογής με τους ειδικούς της συγκεκριμένης τεχνικής εξόρυξης γνώσης. Οι δεύτεροι χρησιμοποιούνται προκειμένου να μορφοποιήσουν τις ερωτήσεις και να βοηθήσουν στην ερμηνεία των αποτελεσμάτων. Οι πρώτοι είναι απαραίτητοι για να ταυτοποιήσουν τα δεδομένα εκπαίδευσης και να ορίσουν τα επιθυμητά αποτελέσματα.
2. **Υπερπροσαρμογή:** Όταν προκύπτει ένα μοντέλο που συσχετίζεται με μία δεδομένη κατάσταση μίας βάσης δεδομένων, είναι επιθυμητό αυτό το μοντέλο να ταιριάζει επίσης και σε μελλοντικές καταστάσεις της βάσης δεδομένων. Η *υπερπροσαρμογή (overfitting)* εμφανίζεται όταν το μοντέλο δεν ταιριάζει σε μελλοντικές καταστάσεις. Αυτό μπορεί να συμβαίνει εξαιτίας υποθέσεων που γίνονται για τα δεδομένα ή απλά μπορεί να συμβαίνει εξαιτίας του μικρού μεγέθους των δεδομένων εκπαίδευσης. Έστω, για παράδειγμα, ένα μοντέλο κατηγοριοποίησης που κατατάσσει τους υπαλλήλους σε 'κοντούς', 'μέτριους' ή 'ψηλούς', σε μια βάση δεδομένων που αφορά εργαζομένους. Εάν τα δεδομένα εκπαίδευσης είναι αρκετά λίγα, το μοντέλο ίσως λανθασμένα δείξει ότι κάθε άτομο με ύψος κάτω από 1.80 είναι 'κοντό' επειδή στη βάση με τα δεδομένα εκπαίδευσης υπάρχει μόνο μία καταχώριση για ύψος κάτω από 1.80. Σε αυτήν την περίπτωση, πολλοί υπάλληλοι λανθασμένα θα καταχωρηθούν σαν 'κοντοί'. Η υπερπροσαρμογή μπορεί επίσης να εμφανιστεί και σε άλλες περιπτώσεις, ακόμα και όταν δεν αλλάζουν τα δεδομένα.
3. **Ακραίες τιμές:** Υπάρχουν συχνά πολλές καταχωρήσεις δεδομένων που δεν ταιριάζουν σωστά στο μοντέλο που έχει αναπτυχθεί. Αυτό συμβαίνει συχνά στις πολύ μεγάλες βάσεις δεδομένων. Εάν το μοντέλο που θα δημιουργηθεί περιλαμβάνει αυτές τις *ακραίες τιμές* (outliers), τότε ίσως να μη συμπεριφέρεται σωστά για τα μη ακραία δεδομένα.
4. **Ερμηνεία των αποτελεσμάτων:** Με τα σημερινά δεδομένα, τα αποτελέσματα από

την εξόρυξη γνώσης πρέπει να ερμηνεύονται από ειδικούς του πεδίου, αλλιώς θα είναι χωρίς νόημα για το μέσο χρήστη.

5. **Οπτικοποίηση των αποτελεσμάτων:** Η οπτικοποίηση των αποτελεσμάτων των αλγορίθμων εξόρυξης γνώσης είναι χρήσιμη για να δούμε και να κατανοήσουμε ευκολότερα τα αποτελέσματα αυτά.
6. **Μεγάλα σύνολα δεδομένων:** Τα ογκώδη σύνολα δεδομένων δημιουργούν προβλήματα όταν εφαρμόζονται αλγόριθμοι εξόρυξης γνώσης που έχουν σχεδιαστεί για μικρά σύνολα δεδομένων. Πολλές εφαρμογές μοντελοποίησης αυξάνονται εκθετικά στον αριθμό των δεδομένων και γι'αυτόν το λόγο οι εφαρμογές αυτές είναι αναποτελεσματικές στα μεγαλύτερα σύνολα δεδομένων. Αποτελεσματικά εργαλεία για να αντιμετωπιστεί το πρόβλημα της κλιμάκωσης είναι η δειγματοληψία και ο παραλληλισμός.
7. **Υψηλές διαστάσεις:** Το σχήμα μίας συμβατικής βάσης δεδομένων μπορεί να αποτελείται από πολλά διαφορετικά γνωρίσματα. Το πρόβλημα εδώ είναι ότι ίσως δεν χρειάζονται όλα τα γνωρίσματα για να λυθεί ένα συγκεκριμένο πρόβλημα εξόρυξης γνώσης. Στην πράξη, αν χρησιμοποιήσουμε κάποια γνωρίσματα μπορεί να εμποδίσουμε τη σωστή ολοκλήρωση μίας εργασίας. Η χρήση άλλων γνωρισμάτων μπορεί απλά να αυξήσει τη συνολική πολυπλοκότητα και να μειώσει την απόδοση ενός αλγορίθμου. Αυτό το πρόβλημα μερικές φορές αναφέρεται σαν *η κατάρα των υψηλών διαστάσεων* (dimensionality curse), εννοώντας ότι υπάρχουν πολλά γνωρίσματα (διαστάσεις) που εμπλέκονται και είναι δύσκολο να καθοριστεί ποια γνωρίσματα πρέπει να χρησιμοποιηθούν. Μία λύση στο πρόβλημα των υψηλών διαστάσεων είναι να μειωθούν τα γνωρίσματα, κάτι που αναφέρεται ως *μείωση των υψηλών διαστάσεων* (dimensionality reduction). Όμως, δεν είναι πάντα εύκολο να προσδιοριστούν τα γνωρίσματα που δεν χρειάζονται.
8. **Δεδομένα πολυμέσων:** Οι περισσότεροι από τους αλγορίθμους που έχουν προταθεί κατά καιρούς στοχεύουν στα παραδοσιακά είδη δεδομένων (αριθμητικά, χαρακτήρες, κείμενο, κ.λπ.). Η χρήση των δεδομένων πολυμέσων, σαν και αυτά που βρίσκουμε στις γεωγραφικές βάσεις δεδομένων, περιπλέκει ή καθιστά ακατάλληλους πολλούς από τους αλγορίθμους αυτούς.
9. **Ελλιπή δεδομένα:** Κατά τη διάρκεια της φάσης της προεπεξεργασίας στη διαδικασία KDD, τα δεδομένα που λείπουν μπορούν να συμπληρωθούν με κατ' εκτίμηση τιμές. Αυτή η προσέγγιση, καθώς και άλλες προσεγγίσεις που αντιμετωπίζουν το πρόβλημα των ελλιπών δεδομένων, ενδεχομένως οδηγούν σε λανθασμένα αποτελέσματα κατά την εξόρυξη γνώσης από δεδομένα.
10. **Άσχετα δεδομένα:** Μερικά γνωρίσματα στη βάση δεδομένων ίσως να μην έχουν ενδιαφέρον όσον αφορά στη συγκεκριμένη εργασία εξόρυξης γνώσης που πραγματοποιείται.
11. **Δεδομένα με θόρυβο:** Μερικές τιμές των γνωρισμάτων μπορεί να είναι άκυρες ή λανθασμένες. Αυτές οι τιμές συνήθως διορθώνονται πριν τρέξουμε την εφαρμογή της εξόρυξης γνώσης από δεδομένα.
12. **Δεδομένα που αλλάζουν:** Οι βάσεις δεδομένων δεν μπορεί να θεωρηθούν ότι είναι στατικές. Όμως, οι περισσότεροι αλγόριθμοι εξόρυξης γνώσης υποθέτουν ότι η βάση δεδομένων είναι στατική. Αυτό απαιτεί ο αλγόριθμος να ξανατρέχει από την αρχή κάθε φορά που αλλάζει η βάση δεδομένων.

13. **Ολοκλήρωση:** Η διαδικασία KDD σήμερα δεν αποτελεί μέρος των συνηθισμένων εργασιών επεξεργασίας των δεδομένων. Οι απαιτήσεις της KDD μπορεί να αντιμετωπίζονται σαν ιδιαίτερες, ασυνήθιστες, ή σαν απαιτήσεις της «μιας φοράς». Οι απαιτήσεις αυτές γίνονται άρα αναποτελεσματικές και όχι αρκετά γενικές για να χρησιμοποιούνται σε συνεχή βάση. Φυσικά ένας επιθυμητός στόχος είναι η ενσωμάτωση των λειτουργιών της εξόρυξης γνώσης σε παραδοσιακά συστήματα διαχείρισης βάσεων δεδομένων.
14. **Εφαρμογή:** Αποτελεί πρόκληση το να προσδιοριστεί η ενδεικνυόμενη χρήση για μια πληροφορία που προήλθε από τη λειτουργία της εξόρυξης γνώσης. Πράγματι, η αποτελεσματική ερμηνεία των αποτελεσμάτων θεωρείται μερικές φορές, από τα στελέχη μίας επιχείρησης, πιο δύσκολο έργο από το τρέξιμο ενός αλγορίθμου. Επειδή τα δεδομένα είναι πληροφορίες που δεν ήταν γνωστές στο παρελθόν, οι τεχνικές των επιχειρήσεων πρέπει να τροποποιηθούν για να καθορίσουν τον τρόπο με τον οποίο θα χρησιμοποιήσουν τις κρυμμένες πληροφορίες.

Αυτά τα θέματα πρέπει να αντιμετωπιστούν από τους αλγόριθμους και τα προϊόντα της εξόρυξης γνώσης από δεδομένα.

1.4 ΜΕΤΡΑ ΑΞΙΟΛΟΓΗΣΗΣ ΤΗΣ ΕΞΟΡΥΞΗΣ ΓΝΩΣΗΣ ΑΠΟ ΔΕΔΟΜΕΝΑ

Η μέτρηση της αποτελεσματικότητας και της χρησιμότητας μιας τεχνικής εξόρυξης γνώσης από δεδομένα δεν είναι πάντα απλή διαδικασία. Στην πράξη, μπορούν να χρησιμοποιηθούν διαφορετικά μέτρα αξιολόγησης για διαφορετικές τεχνικές και επίσης ανάλογα με το επίπεδο ενδιαφέροντος. Για να αξιολογηθεί συνολικά μία επιχείρηση ή να αξιολογηθεί η χρησιμότητα της τεχνικής μπορεί να χρησιμοποιηθεί σαν μέτρο αξιολόγησης η *απόδοση της επένδυσης* (return on investment - ROI). Το μέτρο ROI εξετάζει τη διαφορά ανάμεσα στο κόστος εφαρμογής της τεχνικής από τη μία και στην εξοικονόμηση ή στα κέρδη από την άλλη που προκύπτουν από τη χρήση της τεχνικής αυτής. Φυσικά, η διαφορά αυτή ίσως είναι κάτι δύσκολα μετρήσιμο αφού η απόδοση δύσκολα ποσοτικοποιείται. Η διαφορά αυτή θα μπορούσε να μετρηθεί σαν αύξηση στις πωλήσεις, σαν μείωση στις δαπάνες διαφήμισης, ή σαν το άθροισμα αυτών των δύο. Σε μια συγκεκριμένη διαφημιστική καμπάνια, η οποία υλοποιείται μέσω διαφημιστικών καταλόγων που θα σταλούν ταχυδρομικά, το ποσοστό των ατόμων που θα πάρουν τον κατάλογο και ο αριθμός των αγορών ανά άτομο θα μπορούσε να παρέχει ένα μέτρο υπολογισμού της αποτελεσματικότητας της ταχυδρομικής αποστολής των καταλόγων.

Σε αυτό το βιβλίο, χρησιμοποιούμε μία προσέγγιση περισσότερο σχετική με την επιστήμη των υπολογιστών και τις βάσεις δεδομένων για να αξιολογήσουμε διάφορες τεχνικές εξόρυξης γνώσης από δεδομένα. Υποθέτουμε ότι η διοίκηση της επιχείρησης έχει καθορίσει ότι θα γίνει μία συγκεκριμένη εφαρμογή εξόρυξης γνώσης από δεδομένα. Στη συνέχεια θα καθορίσει τη συνολική αποτελεσματικότητα της προσέγγισης χρησιμοποιώντας το ROI ή ένα ανάλογο μέτρο. Ο αντικειμενικός μας σκοπός είναι να συγκρίνουμε τις εναλλακτικές λύσεις που υπάρχουν για την υλοποίηση μιας εργασίας εξόρυξης γνώσης. Τα μέτρα αξιολόγησης που χρησιμοποιούνται περιλαμβάνουν τα τυπικά μέτρα αξιολόγησης ως προς χώρο και ως προς χρόνο, βάσει της ανάλυσης πολυπλοκότητας. Σε μερικές περιπτώσεις, όπως στην περίπτωση της ακρίβειας στην κατηγοριοποίηση, χρησιμοποιούνται πιο ειδικά μέτρα για την αξιολόγηση της εργασίας εξόρυξης γνώσης.

1.5 ΚΟΙΝΩΝΙΚΕΣ ΕΠΙΠΤΩΣΕΙΣ ΤΗΣ ΕΞΟΡΥΞΗΣ ΓΝΩΣΗΣ

Η ενσωμάτωση των τεχνικών εξόρυξης γνώσης από στις καθημερινές δραστηριότητες αποτελεί πια συνηθισμένη δραστηριότητα. Καθημερινά ερχόμαστε αντιμέτωποι με διαφημίσεις, και οι επιχειρήσεις έχουν γίνει πιο αποτελεσματικές στο να μειώσουν τα έξοδά τους με χρήση της διαδικασίας KDD. Όμως, οι «εχθροί» της εξόρυξης γνώσης από δεδομένα ανησυχούν ότι αυτές οι πληροφορίες παρέχονται με κόστος την καταπάτηση της ιδιωτικής ζωής. Οι εφαρμογές εξόρυξης γνώσης μπορούν να εξάγουν πολλές δημογραφικές πληροφορίες που αφορούν πελάτες, οι οποίες ήταν πριν άγνωστες ή κρυμμένες στα δεδομένα. Η μη εξουσιοδοτημένη χρήση αυτών των δεδομένων θα μπορούσε να οδηγήσει στην αποκάλυψη πληροφοριών που θεωρούνται εμπιστευτικές.

Πρόσφατα έχουμε παρατηρήσει ένα αυξανόμενο ενδιαφέρον στις τεχνικές εξόρυξης γνώσης από δεδομένα που χρησιμοποιούνται σε εφαρμογές όπως είναι η ανίχνευση απάτης, η αναγνώριση υπόπτων για εγκλήματα και η πρόβλεψη των πιθανών τρομοκρατών. Αυτά μπορούν να θεωρηθούν σαν τύποι προβλημάτων κατηγοριοποίησης. Η προσέγγιση που συχνά χρησιμοποιείται εδώ είναι η δημιουργία ενός «προφίλ», με μια τυπική συμπεριφορά και τα κατάλληλα χαρακτηριστικά. Πράγματι, πολλές τεχνικές κατηγοριοποίησης λειτουργούν αναγνωρίζοντας τις τιμές των γνωρισμάτων που εμφανίζονται συχνά για την υπό εξέταση κατηγορία – κλάση. Στη συνέχεια, κατηγοριοποιούνται οι καταγραφές με βάση αυτές τις τιμές των γνωρισμάτων. Ας μην ξεχνάμε ότι αυτές οι προσεγγίσεις της κατηγοριοποίησης *δεν είναι τέλειες*. Μπορεί να γίνουν λάθη. Το ότι κάποιος αγοράζει με πιστωτική κάρτα μια σειρά από προϊόντα που συνήθως αγοράζονται όταν η πιστωτική κάρτα είναι κλεμμένη, δεν σημαίνει ότι η κάρτα του είναι κλεμμένη ή ότι ο συγκεκριμένος καταναλωτής είναι εγκληματίας.

Οι χρήστες των τεχνικών εξόρυξης γνώσης πρέπει να είναι ευαισθητοποιημένοι σε αυτά τα θέματα και δεν θα πρέπει να παραβιάζουν κατευθύνσεις ή οδηγίες σχετικές με θέματα προστασίας προσωπικών δεδομένων.

1.6 Η ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ ΑΠΟ ΤΗΝ ΣΚΟΠΙΑ ΤΩΝ ΒΑΣΕΩΝ ΔΕΔΟΜΕΝΩΝ

Η εξόρυξη γνώσης από δεδομένα μπορεί να μελετηθεί από πολλές διαφορετικές σκοπιές. Ένας ερευνητής ΑΠ ίσως να επικεντρωνόταν στη χρήση των τεχνικών εξόρυξης γνώσης από δεδομένα για να έχει πρόσβαση σε κείμενα. Ένας στατιστικός ίσως κοιτούσε τεχνικές, όπως η ανάλυση χρονοσειρών, η δοκιμή υποθέσεων και εφαρμογές του θεωρήματος Bayes. Ένας ειδικός στη μηχανική μάθηση ίσως να ενδιαφερόταν για τους αλγορίθμους της εξόρυξης γνώσης που μαθαίνουν, και ένας ερευνητής αλγορίθμων θα ενδιαφερόταν να μελετήσει και να συγκρίνει τους αλγορίθμους έχοντας σα βάση τον τύπο του αλγορίθμου και την πολυπλοκότητά του.

Η μελέτη της εξόρυξης γνώσης από δεδομένα, από την σκοπιά των βάσεων δεδομένων, περιλαμβάνει την εξέταση όλων των ειδών των εφαρμογών και των τεχνικών της εξόρυξης γνώσης από δεδομένα. Όμως, ενδιαφερόμαστε κυρίως για αυτές που έχουν πρακτικό ενδιαφέρον. Αφού δεν περιορίζεται το ενδιαφέρον μας σε κανένα συγκεκριμένο είδος αλγορίθμου ή προσέγγισης, εξετάζουμε τα επόμενα θέματα:

- **Κλιμάκωση:** Οι αλγόριθμοι που δεν αποδίδουν καλά όταν υπάρχει κλιμάκωση των δεδομένων, όπως πράγματι συμβαίνει στις πραγματικές ογκώδεις βάσεις δεδομένων, είναι περιορισμένης χρήσης. Με αυτό συσχετίζεται το γεγονός ότι οι τεχνικές πρέπει να λειτουργούν ανεξάρτητα από το μέγεθος της διαθέσιμης κύριας μνήμης.

- **Πραγματικά δεδομένα:** Τα πραγματικά δεδομένα έχουν θόρυβο και πολλές ελλιπείς τιμές γνωρισμάτων. Οι αλγόριθμοι θα πρέπει να μπορούν να δουλεύουν ακόμα και παρουσία αυτών των προβλημάτων.

- **Ενημέρωση:** Πολλοί αλγόριθμοι εξόρυξης γνώσης από δεδομένα δουλεύουν με στατικές βάσεις δεδομένων. Αυτό δεν μπορεί να θεωρηθεί ρεαλιστική υπόθεση.

- **Ευκολία στη χρήση:** Πολλοί αλγόριθμοι μπορεί μεν να δουλεύουν καλά αλλά να είναι δυσνόητοι και δύσχρηστοι, άρα μη αποδεκτοί από τους χρήστες.

Αυτά τα θέματα είναι κρίσιμα, εάν στοχεύουμε σε εφαρμογές που να γίνουν αποδεκτές και να χρησιμοποιηθούν στο χώρο εργασίας. Καθ'όλη τη διάρκεια του κειμένου θα αναφερόμαστε στο πώς οι τεχνικές αποδίδουν στα παραπάνω θέματα αλλά και σε άλλα θέματα υλοποίησης.

Η εξόρυξη γνώσης από δεδομένα σήμερα βρίσκεται σε μία κατάσταση παρόμοια με αυτή των βάσεων δεδομένων στις αρχές της δεκαετίας του 1960. Εκείνη την εποχή, κάθε εφαρμογή βάσης δεδομένων υλοποιούταν ανεξάρτητα, ακόμα και αν υπήρχαν πολλές ομοιότητες μεταξύ διαφορετικών εφαρμογών. Στα μέσα της δεκαετίας του 1960, παρουσιάστηκε μία πληθώρα εργαλείων που έμοιαζαν με συστήματα διαχείρισης βάσης δεδομένων (σαν τα συστήματα χρέωσης υλικών, συμπεριλαμβανομένων των DBOMP και CFMS). Παρόλο που με αυτά η ανάπτυξη των εφαρμογών έγινε ευκολότερη, παρέμεναν ακόμα να υπάρχουν διαφορετικά εργαλεία για διαφορετικές εφαρμογές. Η εμφάνιση των Συστημάτων Διαχείρισης Βάσεων Δεδομένων – ΣΔΒΔ (Database Management Systems – DBMS) έγινε στις αρχές της δεκαετίας του 1970. Η επιτυχία τους οφειλόταν, εν μέρει, στον αφηρημένο ορισμό των δεδομένων και στους βασικούς κανόνες προσπέλασης για ένα μικρό πυρήνα λειτουργικών απαιτήσεων. Τέτοια αφαιρετική διαδικασία δεν έχει ακόμα γίνει στις εργασίες της εξόρυξης γνώσης από δεδομένα. Κάθε εργασία αντιμετωπίζεται ξεχωριστά. Οι περισσότερες εργασίες εξόρυξης γνώσης από δεδομένα (σήμερα) βασίζονται σε συγκεκριμένους αλγόριθμους που θα πραγματοποιήσουν καθεμία πράξη ξεχωριστά. Δεν υπάρχει καμία γενικά αποδεκτή αφαίρεση σε ένα μικρό σύνολο βασικών αρχών. Ένας στόχος μερικών ερευνητών των βάσεων δεδομένων είναι η ανάπτυξη μίας τέτοιας αφαιρετικής διαδικασίας.

Ένα κρίσιμο σημείο της αφαίρεσης στις βάσεις δεδομένων είναι η υποστήριξη της επεξεργασίας ερωτήσεων. Ένας λόγος για τον οποίο οι σχεσιακές βάσεις δεδομένων είναι τόσο δημοφιλείς σήμερα, είναι η ανάπτυξη της SQL. Η SQL είναι εύκολη στη χρήση (τουλάχιστον σε σχέση με τις παλαιότερες γλώσσες ερωτήσεων, όπως ήταν η DBTG και η IMS DML) και έχει γίνει βιομηχανικό πρότυπο γλώσσας που υλοποιείται από όλους τους κατασκευαστές ΣΔΒΔ. Η SQL έχει επίσης καλά ορισμένες στρατηγικές βελτιστοποίησης. Εάν και σήμερα δεν υπάρχει κάποια γλώσσα που να ανταποκρίνεται στην εξόρυξη γνώσης από δεδομένα, στην περιοχή αυτή γίνεται συνεχώς μία προσπάθεια επέκτασης της SQL προκειμένου να υποστηρίξει τις εργασίες της εξόρυξης γνώσης από δεδομένα.

1.7 ΤΟ ΜΕΛΛΟΝ

Η έλευση του σχεσιακού μοντέλου και της SQL αποτέλεσαν ορόσημα στην εξέλιξη των συστημάτων βάσεων δεδομένων. Σήμερα, η εξόρυξη γνώσης από δεδομένα είναι κάτι

παραπάνω από ένα σύνολο από εργαλεία τα οποία μπορούν να χρησιμοποιηθούν για να ανακαλύψουν κρυμμένες πληροφορίες από τις βάσεις δεδομένων. Παρά την ύπαρξη πολλών εργαλείων που βοηθούν σε αυτή τη διαδικασία, δεν υπάρχει ένα μοντέλο ή μία προσέγγιση που να τα περιλαμβάνει όλα. Σύντομα στα επόμενα χρόνια, θα υπάρξουν όχι μόνο περισσότεροι αλγόριθμοι με καλύτερες διεπαφές, αλλά θα γίνουν και βήματα για την ανάπτυξη ενός μοντέλου εξόρυξης γνώσης από δεδομένα που θα τα περιέχει όλα. Εάν και δε θα μοιάζει με το σχεσιακό μοντέλο, πιθανότατα θα περιέχει παρόμοια στοιχεία: αλγόριθμους, μοντέλο δεδομένων και μέτρα αξιολόγησης (όπως στις κανονικές μορφές). Τα σημερινά εργαλεία της εξόρυξης γνώσης από δεδομένα απαιτούν υψηλή ανθρώπινη αλληλεπίδραση όχι μόνο για να οριστεί η απαίτηση αλλά επίσης και για να ερμηνευτούν τα αποτελέσματα. Καθώς τα εργαλεία γίνονται καλύτερα και πιο ολοκληρωμένα, αυτή η εκτεταμένη ανθρώπινη αλληλεπίδραση πιθανότατα θα μειωθεί. Οι εφαρμογές της εξόρυξης γνώσης από δεδομένα είναι διαφορετικών ειδών, με αποτέλεσμα να είναι επιθυμητή η δημιουργία ενός ολοκληρωμένου μοντέλου εξόρυξης γνώσης. Σημαντική ανάπτυξη θα ήταν η δημιουργία μίας εξειδικευμένης «γλώσσας ερωτήσεων» η οποία θα περιλαμβάνει τις παραδοσιακές SQL συναρτήσεις όπως επίσης και πιο πολύπλοκες ερωτήσεις σαν και αυτές που υπάρχουν στις εφαρμογές άμεσης αναλυτικής επεξεργασίας (Online Analytical Processing – OLAP) και της εξόρυξης γνώσης από δεδομένα.

Ήδη έχει προταθεί μια γλώσσα ερωτήσεων εξόρυξης γνώσης που ονομάζεται *DMQL* (data mining query language) και η οποία βασίζεται στην SQL. Αντίθετα με την SQL, όπου υποτίθεται ότι υπάρχει προσπέλαση μόνο σε σχεσιακές βάσεις δεδομένων, η DMQL επιτρέπει την προσπέλαση σε πληροφορίες όπως η ιεραρχία εννοιών. Μία άλλη διαφορά είναι ότι τα δεδομένα που ανακτώνται δεν χρειάζεται να αποτελούν ένα υποσύνολο ή μία συνάθροιση των δεδομένων των σχέσεων. Έτσι, μία DMQL δήλωση πρέπει να υποδείξει το είδος της γνώσης που πρόκειται να εξορυχτεί. Μία άλλη διαφορά είναι ότι μία DMQL δήλωση μπορεί να υποδηλώνει την απαραίτητη σημασία ή το κατώφλι που πρέπει να ικανοποιεί η πληροφορία που εξορύσσεται. Ακολουθεί μια BNF δήλωση της DMQL (από το [Zao99]):

```
DMQL ::=
   USE DATABASE ⟨database_name⟩
   {USE HIERARCHY ⟨hierarchy_name⟩ FOR ⟨attribute⟩}
   ⟨rule_spec⟩
   RELATED TO ⟨attr_or_agg_list⟩
   FROM ⟨relation(s)⟩
   [WHERE ⟨condition⟩]
   [ORDER BY ⟨order list⟩]
   {WITH [⟨kinds of⟩] THRESHOLD = ⟨threshold_value⟩
                                    [FOR ⟨attribute(s)⟩]}
```

Η καρδιά της DMQL δήλωσης είναι το τμήμα εκείνο με τον κανόνα προσδιορισμού. Σε αυτό το τμήμα δηλώνεται η απαίτηση της εξόρυξης γνώσης από δεδομένα, η οποία μπορεί να είναι μία από τις παρακάτω [HFW+96]:

- Μία ***γενικευμένη σχέση*** λαμβάνεται εάν γενικευτούν τα δεδομένα της εισόδου.

- Ένας **χαρακτηριστικός κανόνας** είναι μία συνθήκη που ικανοποιείται από όλες σχεδόν τις εγγραφές της κατηγορίας που μελετάται.
- Ένας **κανόνας διαχωρισμού** είναι μία συνθήκη η οποία ικανοποιείται από μία κατηγορία που μελετάται αλλά είναι διαφορετική από τις συνθήκες που ικανοποιούνται σε άλλες κατηγορίες.

- Ένας **κανόνας κατηγοριοποίησης** είναι ένα σύνολο από κανόνες που χρησιμοποιούνται για να κατηγοριοποιηθούν τα δεδομένα.

Έχει δημιουργηθεί ο όρος *Σύστημα Διαχείρισης Ανακάλυψης Γνώσης και Δεδομένων – ΣΔΑΓΔ* (Knowledge and Data Discovery Management System –KDDMS) για να περιγράψει τη μελλοντική γενιά των συστημάτων εξόρυξης γνώσης από δεδομένα τα οποία δεν θα περιλαμβάνουν μόνο εργαλεία εξόρυξης γνώσης αλλά επίσης και τεχνικές που θα χειρίζονται τα σχετικά δεδομένα, θα εξασφαλίζουν τη συνέπειά τους και θα προσφέρουν συνδρομικότητα και ανάκαμψη. Ένα ΣΔΑΓΔ θα παρέχει προσπέλαση μέσω ειδικών ερωτήσεων εξόρυξης γνώσης από δεδομένα οι οποίες θα έχουν βελτιστοποιηθεί για να γίνει η προσπέλαση αποτελεσματική.

Πρόσφατα παρουσιάστηκε ένα καινούργιο μοντέλο επεξεργασίας της διαδικασίας KDD, το επονομαζόμενο *CRISP-DM (CRoss-Industry Standard Process for Data Mining)*, με πολλές διαφορετικές εφαρμογές. Το μοντέλο απευθύνεται σε όλα τα βήματα της KDD, συμπεριλαμβανομένης και της συντήρησης των αποτελεσμάτων του βήματος της εξόρυξης γνώσης από δεδομένα. Ο κύκλος ζωής του CRISP-DM περιλαμβάνει τα επόμενα βήματα: κατανόηση του είδους της επιχείρησης, κατανόηση των δεδομένων, προετοιμασία των δεδομένων, μοντελοποίηση, ανάπτυξη. Τα βήματα που περιλαμβάνονται στο μοντέλο CRISP-DM μπορούν συνοπτικά να ονομαστούν σαν «τα 5Α»: assess, access, analyze, act, automate – προσδιορίζω, προσπελαύνω, αναλύω, ενεργώ, αυτοματοποιώ.

1.8 ΑΣΚΗΣΕΙΣ

1. Προσδιορίστε και περιγράψτε τις φάσεις της διαδικασίας KDD. Σε τι διαφέρει η διαδικασία KDD από την εξόρυξη γνώσης από δεδομένα;
2. Συγκεντρώστε δεδομένα θερμοκρασίας μιας περιοχής ανά ώρα, αρχίζοντας από τις 8:00 π.μ., για 12 συνεχόμενες ώρες και για 3 διαφορετικές ημέρες. Παρουσιάστε με γραφικές παραστάσεις τα τρία σύνολα με τα δεδομένα χρονοσειρών, στο ίδιο γράφημα. Αναλύστε τις τρεις καμπύλες. Συμπεριφέρονται με τον ίδιο τρόπο; Φαίνεται να υπάρχει μία γενική τάση όσον αφορά στη θερμοκρασία κατά τη διάρκεια της ημέρας; Είναι τα τρία διαγράμματα παρόμοια; Προβλέψτε ποια θα ήταν η τιμή της θερμοκρασίας την επόμενη ώρα σε κάθε μια από τις τρεις ημέρες. Συγκρίνετε τις προβλέψεις σας με τις πραγματικές τιμές που έχει τελικά η θερμοκρασία.
3. Προσδιορίστε την εργασία που πραγματοποιείται σε κάθε βήμα της διαδικασίας KDD στην Άσκηση 2. Ποιες ήταν οι δραστηριότητες εξόρυξης γνώσης που ολοκληρώσατε;
4. Περιγράψτε ζητήματα της εξόρυξης γνώσης από δεδομένα που αντιμετωπίσατε κατά την πραγματοποίηση της Άσκησης 2.
5. Περιγράψτε πώς τα ζητήματα της εξόρυξης γνώσης από δεδομένα που αναφέρο-

νται στην Ενότητα 1.3 γίνονται φανερά με τη χρήση πραγματικών βάσεων δεδομένων.

6. **(Έρευνα)** Βρείτε δύο εναλλακτικούς ορισμούς για την εξόρυξη γνώσης από δεδομένα. Συγκρίνετε αυτούς τους ορισμούς με τον ορισμό που αναφέρεται σε αυτό το κεφάλαιο.
7. **(Έρευνα)** Βρείτε σε εφημερίδες ή σε άλλα ειδησεογραφικά μέσα, τρία τουλάχιστον παραδείγματα εφαρμογών εξόρυξης γνώσης στον χώρο των επιχειρήσεων. Περιγράψτε τις εφαρμογές αυτές.

1.9 ΒΙΒΛΙΟΓΡΑΦΙΚΕΣ ΣΗΜΕΙΩΣΕΙΣ

Παρόλο που έχουν εκδοθεί πολλά εξαιρετικά βιβλία τα οποία εξετάζουν την εξόρυξη γνώσης και την ανακάλυψη γνώσης σε βάσεις δεδομένων, τα περισσότερα είναι εξειδικευμένα και απευθύνονται σε χρήστες των τεχνικών της εξόρυξης γνώσης και σε επαγγελματίες των επιχειρήσεων. Υπάρχουν όμως μερικά τεχνικά βιβλία που εξετάζουν προσεγγίσεις και αλγόριθμους εξόρυξης γνώσης. Ένα εξαιρετικό βιβλίο, το οποίο έχει γραφτεί από έναν από τους πρώτους ειδικούς στο χώρο, είναι το *Data Mining Concepts and Techniques* των Jiawei Han και Micheline Kamber [HK01]. Αυτό το βιβλίο δεν εξετάζει μόνο τους αλγορίθμους εξόρυξης γνώσης αλλά παρέχει επίσης και λεπτομερή κάλυψη των θεμάτων των αποθηκών δεδομένων, της επεξεργασίας OLAP, της προεπεξεργασίας των δεδομένων και της ανάπτυξης γλωσσών εξόρυξης γνώσης από δεδομένα. Άλλα βιβλία που παρέχουν τεχνική κάλυψη μερικών αλγορίθμων εξόρυξης γνώσης από δεδομένα είναι τα [Ada00] και [HMS01].

Πρόσφατα έχουν γίνει αρκετές έρευνες και επισκοπήσεις για την εξόρυξη γνώσης από δεδομένα, όπως είναι αυτές που δημοσιεύτηκαν στα ειδικά τεύχη των περιοδικών *Communications of the ACM* το Νοέμβριο του 1996 και το Νοέμβριο του 1999, *IEEE Transactions on Knowledge and Data Engineering* το Δεκέμβριο του 1996, και *Computer* τον Αύγουστο του 1999. Άλλα άρθρα επισκόπησης μπορούν να βρεθούν στα: [FPSS96c], [FPSS96b], [GGR99a], [Man96], [Man97] και [Cor97]. Ένα δημοφιλές φροντιστηριακό φυλλάδιο δημιουργήθηκε από την Two Crowns Corporation [Cor99]. Μία ολοκληρωμένη συζήτηση πάνω στη διαδικασία KDD βρίσκεται στο [BA96]. Στα άρθρα που εξετάζουν την τομή ανάμεσα στις βάσεις δεδομένων και την εξόρυξη γνώσης από δεδομένα ανήκουν τα [Cha97], [Cha98], [CHY96], [Fay98] και [HKMT95]. Υπάρχουν επίσης αρκετά tutorials με επισκοπήσεις των εννοιών της εξόρυξης γνώσης από δεδομένα: [Agr94], [Agr95], [Han96] και [RS99]. Ένα πρόσφατο tutorial, το [Kei97], περιέχει μία λεπτομερή επισκόπηση των τεχνικών οπτικοποίησης καθώς επίσης και μία ολοκληρωμένη βιβλιογραφία.

Το θέμα της παράλληλης και της κατανεμημένης εξόρυξης γνώσης από δεδομένα έχει αποτελέσει σημαντικό αντικείμενο έρευνας. Ένα συνέδριο, με θέμα τα μεγάλης κλίμακας παράλληλα KDD συστήματα, έγινε το 1999 [ZH00].

Η ιδέα να δημιουργηθεί μία προσέγγιση η οποία θα ενσωμάτωνε όλες τις δραστηριότητες της εξόρυξης γνώσης από δεδομένα, εκφράστηκε από στα [FPSS96b], [Man96] και [Man97]. Ο όρος *KDDMS* για πρώτη φορά προτάθηκε στο [IM96]. Ένα πρόσφατο ενοποιημένο μοντέλο και η άλγεβρα που υποστηρίζει όλες τις σημαντικές εργασίες της εξόρυξης γνώσης από δεδομένα, προτάθηκε από το [JLN00]. Το *3W μοντέλο* βλέπει τα δεδομένα σαν να έχουν χωριστεί σε τρεις διαστάσεις. Μία άλγεβρα, που ονομάζεται *άλγεβρα διαστάσεων*, προτάθηκε για την προσπέλαση σε αυτόν τον 3-διάστατο κόσμο.

Η DMQL αναπτύχθηκε στο Simon Fraser University [HFW+96].

Υπάρχουν αρκετές πηγές για τη διαδικασία KDD και την εξόρυξη γνώσης από δεδομένα. Στην ACM (Association for Computing Machinery) υπάρχει μία ομάδα ειδικού ενδιαφέροντος, η επονομαζόμενη SIGKDD, που ασχολείται αποκλειστικά με την προώθηση και τη διάδοση των πληροφοριών σχετικά με την KDD. Το *SIGKDD Explorations* είναι ένα δελτίο που διατίθεται δωρεάν και εκδίδεται από την ACM SIGKDD. Η ιστοσελίδα της ACM SIGKDD περιέχει μία πληθώρα από πηγές σχετικά με την KDD και την εξόρυξη γνώσης από δεδομένα (www.acm.org/sigkdd)

Μία ομάδα, της οποίας ηγούνται κατασκευαστές προϊόντων βάσεων δεδομένων και εξόρυξης γνώσης, η *Data Mining Group (DMG),* ασχολείται ενεργά με την ανάπτυξη βιομηχανικών προτύπων (standards) στην εξόρυξη γνώσης από δεδομένα. Πληροφορίες για την DMG υπάρχουν στην ιστοσελίδα www.dmg.org. Η ομάδα των ISO/IEC standards δημιούργησε ένα τελικό προσχέδιο από μία επιτροπή για ένα πρότυπο της SQL στο οποίο περιλαμβάνονται προεκτάσεις της εξόρυξης γνώσης από δεδομένα [Com01]. Επιπλέον, ένα πρόγραμμα από έναν όμιλο από κατασκευαστές και χρήστες προϊόντων εξόρυξης γνώσης από δεδομένα είχε σαν αποτέλεσμα τη δημιουργία του μοντέλου *CRISP-DM* (www.crisp-dm.org).

Σήμερα υπάρχουν αρκετά ερευνητικά περιοδικά σχετικά με την εξόρυξη γνώσης από δεδομένα. Σε αυτά συγκαταλέγεται το *IEEE Transactions on Knowledge and Data Engineering*, που εκδίδεται από την IEEE Computer Society, και το *Data Mining and Knowledge Discovery*, που εκδίδεται από την Kluwer Academic Publishers. Στα παγκόσμια συνέδρια που αναφέρονται στη διαδικασία KDD συγκαταλέγονται τα ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), Conference on Information and Knowledge Management (CIKM), IEEE International Conference on Data Mining (ICDM), European Conference on Principles of Data Mining and Knowledge Discovery (PKDD), Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD). Το *KDnuggets News* είναι ένα ενημερωτικό δελτίο που διανέμεται μέσω ηλεκτρονικού ταχυδρομείου και το οποίο κυκλοφορεί κάθε δύο εβδομάδες. Περιέχει άφθονη πληροφορία, σχετικά με την εξόρυξη γνώσης από δεδομένα, για επαγγελματίες των επιχειρήσεων, χρήστες και ερευνητές. Η εγγραφή είναι δωρεάν στο www.kdnuggets.com. Επιπρόσθετες πηγές για KDD μπορούν να βρεθούν στο Knowledge Discovery Central (www.kdcentral.com).