7.4.1 Τα Στάδια της Εύρεσης Γνώσης

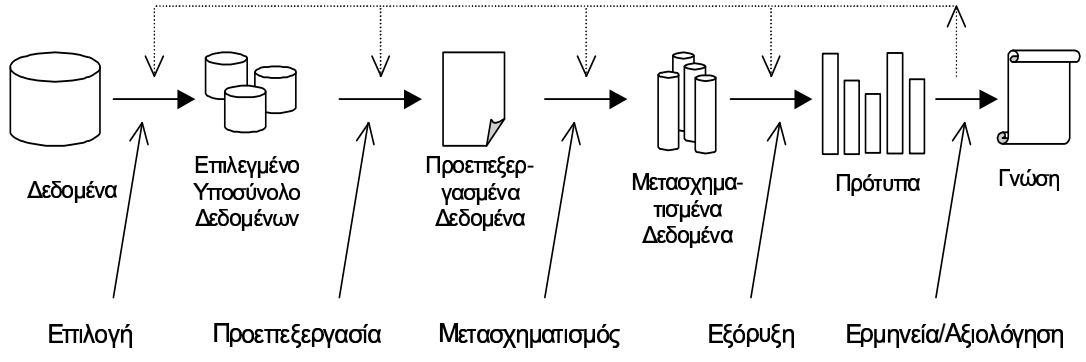
Η εύρεση γνώσης αρχίζει με την κατανόηση του τομέα στον οποίο θα εφαρμοστεί και τον προσδιορισμό του στόχου της από τη σκοπιά του χρήστη των αποτελεσμάτων. Ο ειδικός επί θεμάτων εύρεσης γνώσης πρέπει να συνεργαστεί με τον ειδικό του τομέα ώστε το πρόβλημα να καθοριστεί με αρκετή ακρίβεια και να είναι επιλύσιμο, τα αποτελέσματα να είναι μετρήσιμα και να είναι δυνατή η εφαρμογή τους (π.χ. σε αποδεκτά χρονικά όρια). Τα επιμέρους στάδια απεικονίζονται στο Σχήμα 7.6 και περιγράφονται στη συνέχεια. Πρέπει να σημειωθεί ότι τις περισσότερες φορές κάποια από τα επί μέρους βήματα είναι αναγκαίο να επαναληφθούν, καθώς στην πορεία ενδέχεται να προκύψουν προβλήματα που σχετίζονται με τις αρχικές επιλογές και τα οποία δεν ήταν δυνατό να εντοπιστούν αρχικά.

Επιλογή

Στο στάδιο της επιλογής, δημιουργείται το σύνολο δεδομένων στο οποίο θα εφαρμοστεί η αναζήτηση (*training data set selection*) με επιλογή στοιχείων (πινάκων, πεδίων) από σχεσιακές βάσεις δεδομένων εταιρειών. Επειδή τα δεδομένα είναι σχεδόν πάντα οργανωμένα για άλλη χρήση και οι αλγόριθμοι που εκτελούν την εύρεση γνώσης δεν μπορούν συνήθως να εργαστούν με πολλαπλούς πίνακες δεδομένων, απαιτείται η εξαγωγή των δεδομένων από αυτές και η οργάνωσή τους σε απλούστερες δομές. Στις μέρες μας αυτή η απαίτηση καλύπτεται από *συστήματα αποθήκευσης δεδομένων* (*data warehouse*) τα οποία παρέχουν στους αλγόριθμους εύρεσης γνώσης μία ευκολότερα προσβάσιμη όψη (view) των δεδομένων.

Προεπεξεργασία

Στο στάδιο της *προεπεξεργασίας* (*preprocessing*) των δεδομένων αντιμετωπίζονται περιπτώσεις ελλιπών δεδομένων (π.χ. άδεια πεδία), πεδίων με τιμές που ουσιαστικά τα καθιστούν κενά, (π.χ. Οδός = Άγνωστο), πεδίων με τιμές που υπονοούν (κατά σύμβαση) κάτι άλλο (π.χ. καταχώριση της ημερομηνίας "1/1/1900" σε πεδίο ημερομηνίας που απαιτούσε τιμή αλλά αυτή δεν ήταν διαθέσιμη), κλπ. Λόγω της φύσεως των εργασιών που πραγματοποιούνται, το στάδιο αυτό ονομάζεται και *στάδιο καθαρισμού* των δεδομένων (*data cleaning*).



Σχήμα 7.6: Τα βασικά στάδια της διαδικασίας εύρεσης γνώσης.