Κεφάλαιο 20

Ανακάλυψη Γνώσης σε Βάσεις δεδομένων

Τεχνητή Νοημοσύνη - Β' Έκδοση

Ι. Βλαχάβας, Π. Κεφαλάς, Ν. Βασιλειάδης, Φ. Κόκκορας, Η. Σακελλαρίου

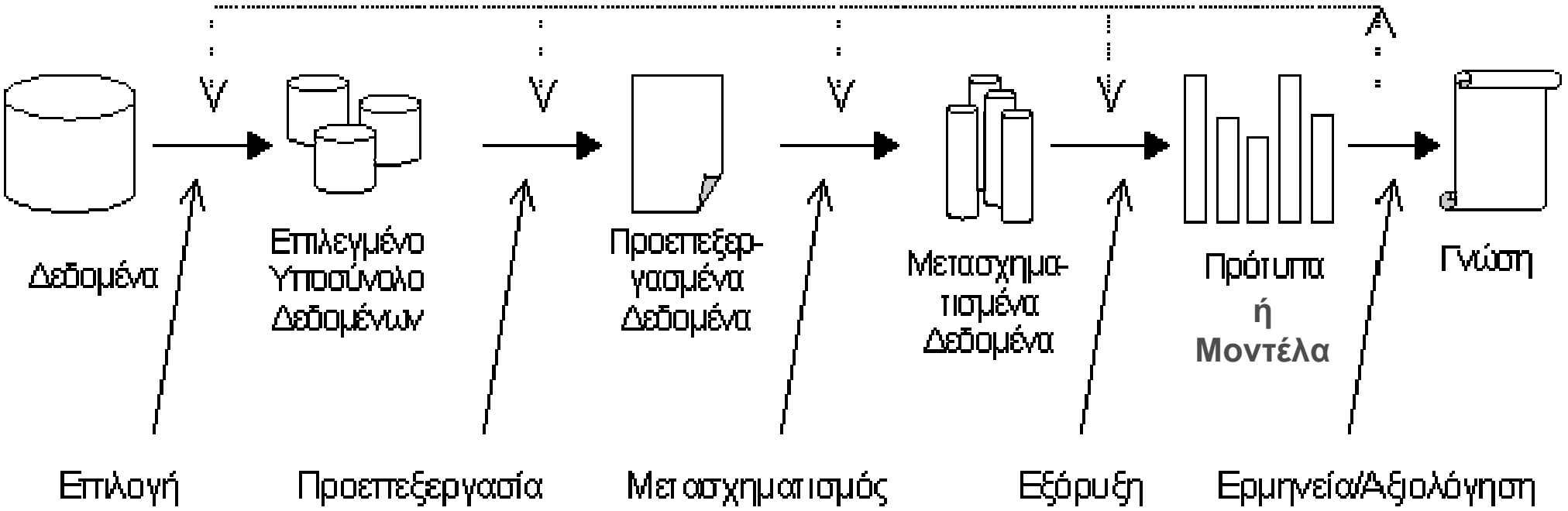
Ανακάλυψη Γνώσης σε Βάσεις Δεδομένων

- Σύνθετη διαδικασία για τον προσδιορισμό έγκυρων, νέων, χρήσιμων και κατανοητών σχέσεων-προτύπων σε δεδομένα (*Knowledge Discovery in Databases - KDD*).
- Αποτελεί μια σημαντική εφαρμογή σε πραγματικές συνθήκες και σε μεγάλη κλίμακα των ερευνητικών αποτελεσμάτων της Στατιστικής, των Βάσεων Δεδομένων (ΒΔ) και της Μηχανικής Μάθησης.
- Είναι μια ολοκληρωμένη διαδικασία που περιλαμβάνει:
  - την επεξεργασία των δεδομένων
  - την εφαρμογή των αλγορίθμων ανακάλυψης γνώσης και τέλος
  - την ερμηνεία των αποτελεσμάτων.

Παραδείγματα Εφαρμογής

- Εταιρία κινητής τηλεφωνίας θέλει να προβλέψει ποιοι από τους συνδρομητές της δε θα ανανεώσουν τη συνδρομή τους, ώστε ενδεχομένως να τους κάνει κάποια περισσότερο ελκυστική προσφορά.
- Ανακάλυψη συσχετίσεων μεταξύ ασθενειών και άλλων χαρακτηριστικών (π.χ. τόπο διαμονής, διατροφικές συνήθειες, παλαιότερες ασθένειες) που μπορούν να οδηγήσουν σε ιατρική πρόοδο.
- Ασφαλιστική εταιρία θέλει να μελετήσει το ιστορικό των πελατών της ώστε να σχεδιάσει περισσότερο ελκυστικά προϊόντα (ασφαλιστικά πακέτα).
- Τράπεζα θέλει να μπορεί να εντοπίσει κακόβουλη χρήση πιστωτικών καρτών.

Τα Στάδια της Ανακάλυψης Γνώσης (1/3)



Τα Στάδια της Ανακάλυψης Γνώσης (2/3)

1. Επιλογή Δεδομένων

- Δημιουργείται το σύνολο δεδομένων στο οποίο θα εφαρμοστεί η αναζήτηση (*training data set selection*) με επιλογή στοιχείων (πινάκων, πεδίων) από σχεσιακές βάσεις δεδομένων εταιρειών.

2. Προεπεξεργασία (*preprocessing*) 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.
- Ονομάζεται και *στάδιο καθαρισμού* των δεδομένων (*data cleaning*).

3. Μετασχηματισμός Δεδομένων (*transformation*)

- Τα δεδομένα μετασχηματίζονται ώστε να διευκολύνουν την ανακάλυψη γνώσης.
- Τέτοιοι μετασχηματισμοί μπορεί να περιλαμβάνουν για παράδειγμα:
  - τη μείωση του αριθμού των υπό εξέταση χαρακτηριστικών (dimensionality reduction) με επιλογή ορισμένων εξ' αυτών (feature selection ή attribute selection),
  - την ομοιόμορφη κωδικοποίηση της ποιοτικά ίδιας πληροφορίας (π.χ. ενοποίηση ενός πεδίου με τίτλο salary σε έναν πίνακα με το πεδίο payment σε κάποιον άλλο πίνακα),
  - τη μετατροπή συνεχόμενων αριθμητικών τιμών σε διακριτές τιμές, (διακριτοποίηση),
  - κλπ.

Τα Στάδια της Ανακάλυψης Γνώσης (3/3)

4. Επιλογή Αλγορίθμου και Εφαρμογή του

- Καθορίζεται τι είδους γνώση θα αναζητηθεί, κάτι που έμμεσα προσδιορίζει και την κατηγορία αλγορίθμου που θα χρησιμοποιηθεί.
- Τα παράγωγα της διαδικασίας ανακάλυψης γνώσης μπορεί να είναι:
  - πρότυπα πληροφόρησης - informative patterns *(μάθηση χωρίς επίβλεψη)*
  - μοντέλα πρόβλεψης - predictive models *(μάθηση με επίβλεψη)*.
  - ✓ **ΣΗΜΕΙΩΣΗ**: Πολλές φορές γίνεται ισοδύναμη χρήση των όρων σχέσεις-προτύπα-μοντέλα.
- Είναι ένα καθαρά υπολογιστικό στάδιο, στο οποίο γίνεται η ουσιαστική αναζήτηση της γνώσης στα δεδομένα. Περιγράφεται και με τον όρο **εξόρυξη σε δεδομένα (data mining)**
- **Καταχρηστικά** ο όρος έχει επικρατήσει να χρησιμοποιείται για να περιγράψει ολόκληρη τη διαδικασία ανακάλυψης γνώσης.

5. Ερμηνεία (*interpretation*) – Αξιολόγηση (*evaluation*)

- Γίνεται *ερμηνεία* και *αξιολόγηση* των ευρεθέντων προτύπων, πιθανώς με υποβοήθηση γραφικών απεικονίσεων των προτύπων ή/και των δεδομένων που περιγράφονται από το πρότυπο (*pattern/data visualization*).

Προβλήματα στην Ανακάλυψη Γνώσης (1/2)

- Τα συστήματα ανακάλυψης γνώσης, βασίζονται στην παροχή δεδομένων εισόδου από βάσεις δεδομένων οι οποίες τείνουν να είναι δυναμικές, μεγάλου μεγέθους, ελλιπείς και να περιέχουν εσφαλμένα δεδομένα.
- Επιπλέον προβλήματα προκύπτουν από το πόσο σχετική και επαρκής είναι η αποθηκευμένη πληροφορία.
- Τα σημαντικότερα προβλήματα που υπεισέρχονται στην ανακάλυψη γνώσης σε βάσεις δεδομένων είναι:

- 1. Ακατάλληλα δεδομένα
  - Οι βάσεις δεδομένων δεν είναι πάντοτε σχεδιασμένες για ανακάλυψη γνώσης και συχνά οι ιδιότητες και τα πεδία που θα απλοποιούσαν τη διαδικασία αναζήτησης όχι μόνο λείπουν αλλά και δεν είναι δυνατόν να συλλεχθούν από το χρήστη.

- 2. Ελλιπή δεδομένα
  - Πολλές φορές η τιμή κάποιων πεδίων απουσιάζει (ελλιπή δεδομένα – missing data).
  - Για παράδειγμα, κάποιο μέγεθος μπορεί να μη μετρήθηκε, κάποια ερώτηση να μην απαντήθηκε, κάποια καταχωρημένη τιμή να διαγράφηκε, κτλ.

Προβλήματα στην Ανακάλυψη Γνώσης (2/2)

- 3. Θόρυβος
  - Τα λάθη στις τιμές των πεδίων είναι γνωστά ως θόρυβος (noise - noisy data) και είναι σημαντικό να αποβάλλεται, γιατί επηρεάζει τη συνολική ακρίβεια της παραγόμενης γνώσης.
- 4. Αραιά Δεδομένα
  - Στην αναζήτηση γνώσης σε βάσεις δεδομένων, ο χώρος αναζήτησης ορίζεται από το δυναμοσύνολο των συνόλων στα οποία ορίζονται τα πεδία.
  - Υπάρχουν περιπτώσεις που για διάφορους λόγους, τα διαθέσιμα δεδομένα καλύπτουν μικρό ποσοστό του χώρου αναζήτησης (αραιά δεδομένα – sparse data), με αποτέλεσμα να δημιουργούνται προβλήματα στην αναζήτηση γνώσης.
- 5. Μέγεθος Βάσης Δεδομένων
  - Ο μεγάλος αριθμός εγγραφών στη ΒΔ κάνει χρονοβόρα την εκτέλεση του αλγορίθμου για την ανακάλυψη της γνώσης και τον έλεγχο της ποιότητας της γνώσης που προκύπτει.
- 6. Δείγματα
  - Η χρήση δείγματος είναι σχεδόν πάντα επιβεβλημένη. Η λήψη ενός δείγματος απαιτεί μεγάλη προσοχή και εφαρμογή στατιστικών τεχνικών, ώστε να αντιπροσωπεύει ικανοποιητικά την αρχική βάση.
- 7. Πρόσφατα Δεδομένα
  - Κατά πόσο μπορεί να θεωρηθεί ότι οι κανόνες που κάποτε παρήχθησαν ανταποκρίνονται στην πλέον ενημερωμένη και πρόσφατη έκδοση της βάσης δεδομένων;