Εξόρυξη δεδομένων

Εξόρυξη δεδομένων είναι η εξαγωγή μιας (ενδιαφέρουσας, αυτονόητης, μη προφανούς και πιθανόν χρήσιμης) πληροφορίας ή προτύπων από μεγάλες βάσεις δεδομένων, αντλώντας μεθοδολογίες από τη στατιστική , τη τεχνητή νοημοσύνη (artificial intelligence), τη μηχανική μάθηση (machine learning), τις βάσεις δεδομένων και τη χρήση αλγορίθμων ομαδοποιήσης. Γενικά, η Εξόρυξη Δεδομένων αναφέρεται στις διεργασίες για την ανάλυση δεδομένων από διαφορετικές πλευρές και την αποκόμιση χρήσιμων πληροφοριών από αυτά, δηλαδή πληροφορίες που μπορούν να χρησιμεύσουν στην πρόβλεψη μελλοντικών καταστάσεων και θα μπορέσουν να βοηθήσουν στη λήψη σωστών αποφάσεων.

Η «εξόρυξη δεδομένων» ανήκει στη γενικότερη μεθοδολογία της «ανακάλυψης της γνώσης απο τις βάσεις δεδομένων (Knowledge Discovery in Databases - KDD)», με την οποία θα ασχοληθούμε εκτενέστερα στο επόμενο κεφάλαιο. Οι διαχειριστές πληροφοριακών συστημάτων, οι αναλυτές δεδομένων και οι στατιστικολόγοι, χρησιμοποιούσαν τον όρο εξόρυξη δεδομένων (data mining (DM)) κυρίως όταν αναφερόταν στο πεδίο των βάσεων δεδομένων. Ο όρος (KDD) αναφέρθηκε από τους Piatetsky-Shapiro (1991) για να τονίστεί ότι η γνώση είναι το αποτέλεσμα της διαδικασίας εξόρυξης δεδομένων. Οι δύο αυτοί όροι ταυτίζονται μεταξύ τους, ενώ στην πραγματικότητα ο όρος KDD αναφέρεται στη συνολική διαδικασία ανακάλυψης προτύπων μέσα από μεγάλα και περίπλοκα σύνολα δεδομένων, ενώ ο όρος DM αναφέρεται στις τεχνικές που χρησιμοποιούνται για την ανακάλυψη της γνώσης.

Εισαγωγή στην Ανακάλυψη Γνώσης από Βάσεις Δεδομένων

Ο συνεχής και αυξανόμενος όγκος δεδομένων και πληροφοριών που καταγράφεται καθημερινά και αφορά πολλούς και διαφορετικούς τομείς, δημιούργησε την ανάγκη για ανάπτυξη νέων θεωριών και εργαλείων που βοηθούν την ανάλυση των δεδομένων αυτών. Το κύριο αντικείμενο της Ανακάλυψης Γνώσης σε Βάσεις Δεδομένων (Knowledge Discovery in Databases - KDD) είναι η εξεύρεση γνώσης και συμπερασμάτων για τη λήψη αποφάσεων μέσα από θεωρίες που αναπτύσσονται, καθώς και η δημιουργία νέων εργαλείων για το σκοπό αυτό. Ένα από τα συνήθη προβλήματα στην KDD διαδικασία είναι η μελέτη δεδομένων χαμηλής ποιότητας, όπου η εύρεση ενός μοντέλου που θα τα προσαρμόζει καθώς και η εξαγωγή συμπερασμάτων για τη δομή των δεδομένων αυτών, αποτελούν δύσκολη υπόθεση.

Με τον όρο KDD αναφερόμαστε στη συνολική διαδικασία και στα βήματα που ακολουθούνται ώστε τελικά να εξαχθούν χρήσιμες πληροφορίες μέσα από το μεγάλο όγκο των βάσεων δεδομένων. ΄Οπως αναφέρεται από τους Fayyad et al. (1996) ¨KDD είναι μία μη–τετριμμένη διαδικασία εύρεσης έγκυρων, νέων, χρήσιμων και πλήρως κατανοητών προτύπων από τα δεδομένα¨ ("Knowledge discovery in databases is the non–trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.").

Η χρησιμοποίηση των τεχνικών KDD εφαρμόζεται σε πολλούς κλάδους, όπως το:

- Marketing (π.χ. για εξαγωγή κανόνων συσχέτισης μεταξύ των προϊόντων ενός super market και την καλύτερη τοποθέτησή τους στα ράφια)
- στο χρηματιστήριο (π.χ. για εύρεση κοινού προφίλ μετοχών)
- στην ανίχνευση απάτης (π.χ. στις οικονομικές συναλλαγές όπου μπορεί να εντοπιστούν κλοπές ή παρατυπίες)
- στις επικοινωνίες, στον καθαρισμό των δεδομένων

και σε πολλούς άλλους τομείς που οι τεχνικές εξόρυξης δεδομένων δίνουν πολύτιμες πληροφορίες μέσα από τις τεράστιες αυτές ποσότητες δεδομένων.

Στην εξόρυξη δεδομένων εφαρμόζονται συγκεκριμένοι αλγόριθμοι για την ανάλυση των δεδομένων. Στη διαδικασία όμως ανακάλυψης ή αλλιώς εξόρυξης γνώσης συμπεριλαμβάνονται και επιπλέον βήματα όπως είναι η προετοιμασία των δεδομένων, ο καθαρισμός δεδομένων, η επιλογή των καταλληλότερων χαρακτηριστικών, η σωστή αποκρυπτογράφηση και μελέτη των αποτελεσμάτων.

Διάφορα επιστημονικά πεδία όπως η μηχανική μάθηση, η αναγνώριση προτύπων, η στατιστική, η τεχνητή νοημοσύνη καθώς και η παρουσίαση των δεδομένων, ενσωματώνονται διαρκώς στις διαδικασίες ανακάλυψης γνώσης με αποτέλεσμα να τις βελτιώνουν και να τις εξελίσσουν ώστε το παραγόμενο προϊόν που είναι η γνώση, να είναι υψηλού επιπέδου και να προκύπτει από χαμηλού επιπέδου δεδομένα των βάσεων δεδομένων. Το αντικείμενο της εξόρυξης δεδομένων αφορά κυρίως γνωστές και νέες τεχνικές του πεδίου της μηχανικής μάθησης, της αναγνώρισης προτύπων, της στατιστικής,

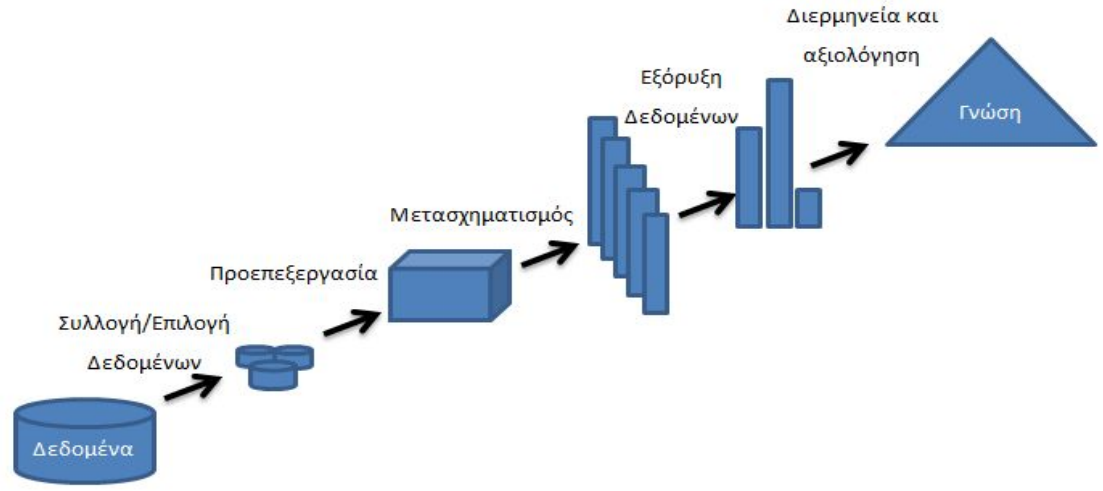
της τεχνητής νοημοσύνης που θα χρησιμοποιηθούν για να παράγουν μοντέλα που θα περιγράψουν σωστά τα δεδομένα.

Βήματα μιας διαδικασίας KDD

Κατά την διάρκεια της διαδικασίας KDD, ακολουθούνται πολλά βήματα και λαμβάνεται πλήθος αποφάσεων. Τα βήματα αυτά μπορούν να χωριστούν ώς εξής :

1. **Αναγνώριση του στόχου και των απαιτήσεων του προβλήματος.**
   Το βήμα αυτό περιλαμβάνει την κατανόηση, την αξιολόγηση της τρέχουσας κατάστασης, και την αναγνώριση των στόχων του προβλήματος, ώστε να γίνει σαφές ποιες αποφάσεις θα ληφθούν σχετικά με μετασχηματισμούς,αλγορίθμους κλπ.
2. **Επιλογή των πιο αντιπροσωπευτικών δεδομένων που θα χρησιμοποιηθούν στη διαδικασία ανακάλυψης γνώσης.** Σε αυτό το βήμα δημιουργείται το σύνολο δεδομένων στο οποίο θα εφαρμοστούν οι αλγόριθμοι ανακάλυψης γνώσης. Τα δεδομένα αυτά είναι συνήθως ανομοιογενή και αποθηκευμένα σε πολλές διαφορετικές πηγές, όπως ανεξάρτητες βάσεις δεδομένων, αρχεία, εξωτερικές πηγές κλπ. Οπότε θα πρέπει να ενσωματωθούν σε μια κοινή βάση δεδομένων ώστε να γίνει ευκολότερη η πρόσβαση των αλγορίθμων στα δεδομένα αυτά.
3. **Καθαρισμός των δεδομένων.** Το τρίτο και πιο σημαντικό βήμα είναι η προεπεξεργασία του συνόλου των δεδομένων. Τα δεδομένα που έχουν συλλεχθεί πολλές φορές έχουν σφάλματα, ή ελλειπές τιμές, τα οποία μπορεί να αποπροσανατολίσουν και να οδηγήσουν τους αλγόριθμους εξόρυξης στην εξαγωγή άκυρων και λανθασμένων προτύπων. Για τον λόγο αυτό είναι σημαντικό να γίνει καθαρισμός των δεδομένων, ώστε να έχουμε ένα τελικό αξιόπιστο σύνολο δεδομένων.

4. **Μετασχηματισμός δεδομένων.** Στο στάδιο αυτό τα δεδομένα προσαρμόζονται στις απαιτήσεις των μεθόδων ανάλυσης. Για παράδειγμα μπορεί να γίνει μετατροπή αριθμητικών τιμών σε ονομαστικές τιμές ή αριθμητικών τιμών σε άλλες αριθμητικές τιμές. Για το σκοπό αυτό εφαρμόζονται μέθοδοι μείωσης των διαστάσεων του χώρου αναπαράστασης των δεδομένων είτε μέσω μετασχηματισμού (βάσης αναπαράστασης) ή με επιλογή κατάλληλου μικρότερου αριθμού διαστάσεων.
5. **Επιλογή της κατάλληλης μεθόδου εξόρυξης δεδομένων.** Σε αυτό το βήμα καθορίζεται το είδος της γνώσης που θα αναζητηθεί το οποίο σημαίνει ότι επιλέγεται η μέθοδος εξόρυξης δεδομένων που θα χρησιμοποιηθεί όπως για παράδειγμα αν θα είναι κατηγοριοποίηση, συσταδοποίηση, παλινδρόμηση κ.ά.
6. **Επιλογή αλγορίθμου εξόρυξης δεδομένων.** Αφού επιλέχτηκε η μέθοδος εξόρυξης δεδομένων πρέπει να επιλεχθεί και ο κατάλληλος αλγόριθμος.
7. **Παραμετροποίηση και εκτέλεση του αλγορίθμου που επιλέχθηκε.** Σε αυτό το βήμα εκτελείται κάποιος αλγόριθμος για την δημιουργία ενός μοντέλου.
8. **Έλεγχος της διαδικασίας και πιθανή επανάληψη και βελτίωση προηγούμενων βημάτων.** Σε αυτό το στάδιο γίνεται η ερμηνεία του προτύπου που προέκυψε. Είναι πολύ πιθανό σε αυτό το στάδιο να χρειαστεί να επιστρέψουμε σε ένα από τα παραπάνω βήματα.
9. **Αναπαράσταση και χρήση της ανακαλύφθεισας γνώσης.** Στο τελικό αυτό στάδιο η γνώση που έχει ανακαλυφθεί παρουσιάζεται στον χρήστη με τη βοήθεια γραφικών παραστάσεων, βοηθώντας τον έτσι να κατανοήσει και να ερμηνεύσει τα αποτελέσματα της εξόρυξης δεδομένων. Επιπλέον μέσα από αυτό το βήμα γίνεται έλεγχος για τυχόν συγκρούσεις μεταξύ της υπάρχουσας γνώσης και της παραγόμενης. Εάν τα αποτελέσματα δεν είναι ικανοποιητικά μπορεί να επανέλθουμε σε προηγούμενα βήματα, να τα επαναλάβουμε και πιθανόν να τροποποιηθεί το σύνολο δεδομένων ή να χρησιμοποιηθεί διαφορετική μέθοδος εξόρυξης δεδομένων.



Βασικά στάδια Ανακάλυψης Γνώσης από Βάσεις Δεδομένων.

Για την πετυχημένη ολοκλήρωση μιας διαδικασίας ανακάλυψης γνώσης, τα περισσότερα από τα βήματα που ακολουθούνται, αφορούν τις μεθόδους και τους αλγορίθμους εξόρυξης δεδομένων, χωρίς όμως να παραβλέπουμε και την αξία των υπολοίπων βημάτων.

Εξόρυξη Δεδομένων από Βάσεις Δεδομένων (Data Mining)

Τα μοντέλα που παράγονται από το στάδιο της Εξόρυξης Δεδομένων διακρίνονται σε δυο βασικούς τύπους:

- Προβλεπτικά μοντέλα (predictive models)
- Περιγραφικά μοντέλα (descriptive models)

Ένα **προβλεπτικό μοντέλο** κάνει μια 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. Ένα προβλεπτικό μοντέλο περιλαμβάνει μεθόδους όπως κατηγοριοποίση, ανάλυση χρονολογικών σειρών, παλινδρόμηση και πρόβλεψη.

Σε ένα **περιγραφικό μοντέλο** το ζητούμενο είναι η ομαδοποίηση των δεδομένων σε κατηγορίες χωρίς την προγενέστερη γνώση των κατηγοριών αυτών. Η συσταδοποιήση, η παρουσιάση συνόψεων, οι κανόνες συσχετίσεων και η ανακάλυψη ακολουθιών ανήκουν σε αυτό το μοντέλο.

Μοντέλα και μέθοδοι στην εξόρυξη γνώσης από δεδομένα

Παρακάτω θα αναφέρουμε περιγραφικά τη κάθε τεχνική.

- **Κατηγοριοποίηση (Classification).** Είναι η πιο γνωστή τεχνική της εξόρυξης δεδομένων. Σύμφωνα με την τεχνική αυτή τα δεδομένα κατατάσσονται σε προκαθορισμένες ομάδες ή κατηγορίες κλάσεις. Μέσα από μια διαδικασία μάθησης που εφαρμόζεται σε ένα σύνολο δεδομένων εκπαίδευσης (training set) χωρισμένο σε γνωστές κατηγορίες, ορίζεται ένα μοντέλο. Στη συνέχεια και σύμφωνα με το μοντέλο αυτό, τα σύνολα δεδομένων ελέγχου (training set) ταξινομούνται στις αντίστοιχες κατηγορίες. Ο τελικός στόχος της διαδικασίας ταξινόμησης είναι να τοποθετηθούν νέα και άγνωστα δεδομένα στις σωστές κατηγορίες.
- **Συσταδοποίηση (Clustering).** Η τεχνική αυτή διαμοιράζει ένα σύνολο δεδομένων σε ομάδες (ή αλλιώς κλάσεις ή συστάδες), όπου τα στοιχεία της κάθε ομάδας παρουσιάζουν ομοιότητες μεταξύ τους. Η ομαδοποίηση γίνεται με την εφαρμογή κάποιου δείκτη ομοιότητας. Στη διαδικασία αυτή οι κατηγορίες που προκύπτουν δεν είναι από πριν γνωστές.
- **Παλινδρόμηση (Regression).** Είναι παρόμοια με την τεχνική της κατηγοριοποιήσης και χρησιμοποιείται για να απεικονίσει ένα στοιχειώδες δεδομένο σε μια πραγματική μεταβλητή πρόβλεψης. Στόχος είναι με βάση κάποιες ανεξάρτητες μεταβλητές (independent variables) να προβλεφθούν οι τιμές μιας εξαρτημένης μεταβλητής (dependent variable).

- **Ανάλυση Χρονοσειρών (Time Series Analysis).** Με αυτή τη τεχνική, 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.
- **Πρόβλεψη (Prediction).** Η πρόβλεψη (Prediction) μπορεί να θεωρηθεί σαν ένα είδος κατηγοριοποίησης. Η διαφορά έγκειται στο γεγονός ότι ως πρόβλεψη θεωρείται περισσότερο το να δίνεται τιμή σε μία μελλοντική κατάσταση παρά σε μία τρέχουσα. Οι εφαρμογές πρόβλεψης περιλαμβάνουν πρόγνωση πλημμύρων, αναγνώριση ομιλίας, μηχανική μάθηση και αναγνώριση προτύπων.
- **Παρουσίαση Συνόψεων.** Η Παρουσίαση Συνόψεων (summarization) απεικονίζει τα δεδομένα σε υποσύνολα 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, στη 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.
- **Κανόνες συσχέτισης (Link Analysis).** Η ανάλυση συνδέσμων (link analysis), ή αλλιώς ανάλυση συγγένειας (affinity analysis) ή συσχέτιση (association), είναι διαδικασία εκείνη της εξόρυξης γνώσης που αποκαλύπτει συσχετίσεις μεταξύ των δεδομένων. Ο προσδιορισμός των κανόνων συσχετίσεων είναι ένα καλό παράδειγμα για αυτού του είδους της εφαρμογής. Ένας κανόνας συσχέτισης (association rules) είναι ένα μοντέλο που αναγνωρίζει ειδικούς τύπους συσχέτισης μεταξύ των δεδομένων. Αυτές οι συσχετίσεις συχνά χρησιμοποιούνται στις λιανικές πωλήσεις για να αναγνωριστούν προϊόντα που συχνά αγοράζονται μαζί.
- **Ανακάλυψη Ακολουθιών (Sequential Analysis).** 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 παρόμοια με τις συσχετίσεις στο ότι συσχετίζονται τα δεδομένα (ή τα γεγονότα) που εξάγονται, με την διαφορά ότι η συσχέτισή τους αυτή βασίζεται στο χρόνο.