Εξόρυξη δεδομένων

Εξόρυξη δεδομένων (ή ανακάλυψη γνώσης από βάσεις δεδομένων) είναι η εξεύρεση μιας (ενδιαφέρουσας, αυτονόητης, μη προφανούς και πιθανόν χρήσιμης) πληροφορίας ή προτύπων από μεγάλες βάσεις δεδομένων με χρήση αλγορίθμων ομαδοποίησης ή κατηγοριοποίησης και των αρχών της στατιστικής, της τεχνητής νοημοσύνης, της μηχανικής μάθησης και των συστημάτων βάσεων δεδομένων. Στόχος της εξόρυξης δεδομένων είναι η πληροφορία που θα εξαχθεί και τα πρότυπα που θα προκύψουν να έχουν δομή κατανοητή προς τον άνθρωπο έτσι ώστε να τον βοηθήσουν να πάρει τις κατάλληλες αποφάσεις.

Ο στόχος

Ο πραγματικός στόχος της εξόρυξης δεδομένων είναι η αυτόματη ή ημιαυτόματη ανάλυση μεγάλων ποσοτήτων δεδομένα για την εξαγωγή κάποιου ενδιαφέροντος προτύπου που ήταν άγνωστο μέχρι εκείνη τη στιγμή, όπως ομάδες από εγγραφές δεδομένων (συσταδοποίηση), ασυνήθιστες εγγραφές (*anomaly detection*) και εξαρτήσεις (κανόνες συσχετίσεων). Αυτό συνήθως συμπεριλαμβάνει τη χρήση βάσης δεδομένων όπως χωρικά ευρετήρια. Αυτά τα πρότυπα ύστερα μπορούν να θεωρηθούν ως μία περιγραφή των δεδομένων εισαγωγής και να χρησιμοποιηθούν για περαιτέρω ανάλυση ή για παράδειγμα στην εκμάθηση μηχανής και στην προγνωστική ανάλυση. Για παράδειγμα, η εξόρυξη δεδομένων θα μπορούσε να προσδιορίσει πολλαπλά σύνολα στα δεδομένα, τα οποία μπορούν να χρησιμοποιηθούν μετά για να εξασφαλίσουν περισσότερο ακριβή αποτελέσματα από ένα σύστημα υποστήριξης αποφάσεων. Παρότι η συλλογή δεδομένων και η προετοιμασία δεδομένων, αλλά και η ερμηνεία των αποτελεσμάτων και εκθέσεων δεν αποτελούν μέρος της εξόρυξης δεδομένων, παρ' όλα αυτά ανήκουν στην ανακάλυψη γνώσης από βάσεις δεδομένων σαν κάποια επιπρόσθετα βήματα.

Ιστορία και Εξέλιξη

Η χειροκίνητη εξαγωγή προτύπων από δεδομένα συμβαίνει εδώ και αιώνες. Οι πρώτες μέθοδοι για τον προσδιορισμό προτύπων ήταν αυτές της θεωρίας Bayes και της ανάλυσης της παλινδρόμησης. Ο πολλαπλασιασμός, η ευρεία διαθεσιμότητα και η εξέλιξη της τεχνολογίας υπολογιστών έχουν αυξήσει τον όγκο των συγκεντρωμένων δεδομένων και την ζήτηση για αποδοτικούς και αποτελεσματικούς χειρισμούς. Καθώς οι συλλογές δεδομένων αυξήθηκαν τόσο σε όγκο όσο και σε πολυπλοκότητα, η χειρωνακτική ανάλυση των δεδομένων έχει αντικατασταθεί από την αυτόματη επεξεργασία δεδομένων. Σε αυτό συνέβαλαν άλλες ανακαλύψεις της επιστήμης των υπολογιστών, όπως τα νευρωνικά δίκτυα, η συσταδοποίηση, οι γενετικοί αλγόριθμοι (1950), τα δέντρα απόφασης (1960) και η μηχανή υποστήριξης διανυσμάτων(1990). Η εξόρυξη δεδομένων είναι η διαδικασία εφαρμογής αυτών των μεθόδων στα δεδομένα με σκοπό την αποκάλυψη άγνωστων προτύπων σε μεγάλα σύνολα δεδομένων. Αυτό γεφυρώνει το χάσμα της εφαρμοσμένης στατιστικής και της τεχνητής νοημοσύνης (τα οποία συνήθως παρέχουν το μαθηματικό υπόβαθρο) με την διαχείριση βάσης δεδομένων κάνοντας χρήση του τρόπο με τον οποίο αποθηκεύονται και κατατάσσονται στη βάση δεδομένων για να εκτελέσουν την θεωρία και τους διαθέσιμους αλγορίθμους περισσότερο αποτελεσματικά, επιτρέποντας σε τέτοιες μεθόδους να εφαρμόζονται σε μεγάλα σύνολα δεδομένων.

Διαδικασία

Η διαδικασία ανακάλυψης γνώσης από βάσεις δεδομένων(KDD) συνήθως ορίζεται από τα εξής στάδια:

1. Συλλογή
2. Προεπεξεργασία
3. Μετασχηματισμός
4. Εξόρυξη δεδομένων
5. Ερμηνεία/Αξιολόγηση.

Υπάρχουν όμως κι άλλες παραλλαγές για τον ορισμό των σταδίων αυτών σύμφωνα και με το CRoss Industry Standard Process for Data Mining (CRISP-DM) όπου τα στάδια έχουν ως εξής:

1. Κατανόηση Θέματος
2. Κατανόηση δεδομένων
3. Προετοιμασία δεδομένων
4. Μοντελοποίηση
5. Αξιολόγηση
6. Ανάπτυξη ή απλοποιημένη διαδικασία όπως
    1. Προ-επεξεργασία
    2. Εξόρυξη δεδομένων
    3. Επικύρωση αποτελέσματος.

Προ-επεξεργασία

Πριν την εφαρμογή των αλγορίθμων εξόρυξης δεδομένων, το ερευνώμενο σύνολο δεδομένων πρέπει να συναρμολογείται. Καθώς η εξόρυξη δεδομένων μπορεί να αποκαλύψει μόνο τα πρότυπα που πράγματι εμφανίζονται στα δεδομένα, το σύνολο δεδομένων που ερευνούμε, πρέπει να είναι αρκετά μεγάλο για να περιέχει αυτά τα πρότυπα παραμένοντας να εξορυχθεί σε ένα αποδεκτό χρονικό διάστημα. Μία συνηθισμένη πηγή για δεδομένα είναι η data mart ή η data warehouse. Η προεπεξεργασία είναι απαραίτητη για την ανάλυση πολυπαραγοντικών συνόλων δεδομένων πριν την εξόρυξη δεδομένων.

Έτσι το ερευνώμενο σύνολο καθαρίζεται.Το καθάρισμα δεδομένων διαγράφει τις παρατηρήσεις που περιέχουν θόρυβο και αυτές με ελλειπή ή ελλείποντα δεδομένα.

Τεχνικές

Η εξόρυξη δεδομένων περιλαμβάνει κάποιες από τις ακόλουθες τάξεις διαδικασιών:

- Ανίχνευση ανωμαλιών (Anomaly detection) - Ο προσδιορισμός ασυνήθιστων εγγραφών δεδομένων, που μπορεί να παρουσιάζουν κάποιο ενδιαφέρον ή λάθη στα δεδομένα που απαιτούν περαιτέρω έρευνα.
- Κανόνες συσχέτισης (Μοντέλο αλληλεξάρτησης) - Αναζητήσεις για σχέσεις μεταξύ των μεταβλητών. Για παράδειγμα, ένα σούπερ μάρκετ μπορεί να συλλέξει δεδομένα που αφορούν τις αγοραστικές τους συνήθειες. Χρησιμοποιώντας τους κανόνες συσχέτισης, το σούπερ μάρκετ μπορεί να υπολογίσει ποια προϊόντα αγοράζονται

συνήθως μαζί και να χρησιμοποιήσει αυτή την πληροφορία για αγοραστικούς σκοπούς.
- Συσταδοποίηση - είναι η διαδικασία ανακάλυψης ομάδων και δομών στα δεδομένα που είναι "παρόμοια" κατά κάποιο τρόπο,χωρίς να χρησιμοποιούνται γνωστές δομές στα δεδομένα.
- Κατηγοριοποίηση - είναι η διαδικασία γενίκευσης γνωστών δομών για την εφαρμογή τους πάνω σε νέα δεδομένα. Παραδείγματος χάριν, ένα πρόγραμμα ηλεκτρονικού ταχυδρομείου ενδέχεται να προσπαθήσει να χαρακτηρίσει ένα μήνυμα ηλεκτρονικού ταχυδρομείου ως νόμιμο ή spam.
- Παλινδρόμηση (στατιστική) - Προσπαθεί να βρει μία συνάρτηση που μοντελοποιεί τα δεδομένα με το λιγότερο λάθος.

Επικύρωση αποτελέσματος

Το τελικό βήμα της ανακάλυψης γνώσης από δεδομένα είναι η επικύρωση των προτύπων που εξήχθησαν από τους αλγορίθμους της εξόρυξης δεδομένων που απευθύνονται σε ευρύτερο σύνολο δεδομένων. Δεν είναι όλα τα πρότυπα που βρέθηκαν απαραίτητα έγκυρα. Είναι συνηθισμένο για τους αλγορίθμους της εξόρυξης δεδομένων να βρίσκουν πρότυπα στο σύνολο εκπαίδευσης, τα οποία δεν υπάρχουν στο γενικό σύνολο δεδομένων. Αυτό καλείται υπερφόρτωση(overfitting).Για να ξεπεραστεί αυτό, στην εκτίμηση χρησιμοποιείται ένα δοκιμαστικό σύνολο δεδομένων στο οποίο δεν έχουν εφαρμοστεί οι αλγόριθμοι της εξόρυξης δεδομένων. Τα πρότυπα, που έχουν προκύψει, εφαρμόζονται σε αυτό το δοκιμαστικό σύνολο και το προκύπτον αποτέλεσμα συγκρίνεται με το επιθυμητό. Για παράδειγμα, ένας αλγόριθμος της εξόρυξης δεδομένων που ξεχωρίζει τα ανεπιθύμητα μηνύματα με τα "επιθυμητά" θα εφαρμοζόταν σε ένα σύνολο εκπαίδευσης από δείγματα ηλεκτρονικών μηνυμάτων.Μόλις εφαρμοζόταν, τα εξαχθείσα πρότυπα θα εφαρμόζονταν στο δοκιμαστικό σύνολο μηνυμάτων στο οποίο δεν είχε εφαρμοστεί πριν. Η ευστοχία αυτών των προτύπων μπορεί τώρα να μετρηθεί από τα πόσα μηνύματα έχουν καταταχθεί-ταξινομηθεί σωστά. Ένας αριθμός από στατιστικές μεθόδους μπορεί να χρησιμοποιηθεί για την αξιολόγηση του αλγορίθμου, όπως το ROC curves.

Αν τα πρότυπα δεν ανταποκρίνονται με τα επιθυμητά κριτήρια, τότε είναι απαραίτητο να εκτιμηθεί ξανά και να αλλαχθεί η προ-επεξεργασία και η εξόρυξη δεδομένων. Στην αντίθετη περίπτωση που ανταποκρίνονται με τα επιθυμητά κριτήρια, το τελικό στάδιο είναι να ερμηνευτούν τα πρότυπα και να τα μετατρέψουμε σε γνώση.

Εφαρμογές

- **Ιατρική**
- **Οικονομία**
- **Τηλεπικοινωνία**

Πηγή :
https://el.wikipedia.org/wiki/%CE%95%CE%BE%CF%8C%CF%81%CF%85%CE%BE%CE%B7_%CE%B4%CE%B5%CE%B4%CE%BF%CE%BC%CE%AD%CE%BD%CF%89%CE%BD