

Εργασία εξαμήνου (συμμετοχή στον τελικό βαθμό: 50%)

Ομάδες των 2 ατόμων

1ο μέρος: Υλοποίηση αποθήκης δεδομένων – Ανάλυση τύπου OLAP

Στόχος του 1^{ου} μέρους της εργασίας είναι η εξοικείωση με τις αποθήκες δεδομένων (ΑΔ) και ειδικότερα την κατασκευή κύβων από μεγάλες βάσεις δεδομένων (ΒΔ) και την αναλυτική επεξεργασία των δεδομένων του κύβου (OLAP). Για τους σκοπούς της άσκησης θα χρησιμοποιηθεί το Σύστημα Διαχείρισης Βάσεων Δεδομένων (ΣΔΒΔ) Microsoft SQL Server 2005 και πιο συγκεκριμένα το λογισμικό Analysis Services.

Έστω ότι θέλουμε να φτιάξουμε μια ΑΔ που να βασίζεται σε μία μεγάλη ΒΔ. Σε πρώτη φάση αντιμετωπίζουμε το πρόβλημα της προπαρασκευής δεδομένων (*data preprocessing*) ώστε να κρατήσουμε όσα μας ενδιαφέρουν για τους σκοπούς της ανάλυσης. Στη συνέχεια, ορίζουμε πάνω στα δεδομένα αυτά μια αποθήκη δεδομένων - κύβο - ακολουθώντας το σχήμα αστέρα (*star schema*) ή χιονονιφάδας (*snowflake schema*). Τέλος, εκτελούμε στην ΑΔ διάφορες λειτουργίες OLAP (roll-up, drill-down, slice, dice και pivot).

Πιο συγκεκριμένα, θα πρέπει να ακολουθήσετε τα εξής βήματα:

1ο βήμα (υλοποίηση ΒΔ): Θα ασχοληθείτε με ένα μεγάλο σύνολο δεδομένων (dataset) με πραγματικά δεδομένα (περιγράφονται στη συνέχεια της εκφώνησης). Αφού κάνετε την απαραίτητη επεξεργασία μεταφοράς δεδομένων από αρχεία text ή από άλλες ΒΔ (ανάλογα με την περίπτωση), θα εισάγετε τα δεδομένα σε μια κατάλληλα σχεδιασμένη ΒΔ στον SQL Server 2005.

Παραδοτέο: Τεκμηρίωση της ΒΔ που δημιουργήθηκε από το dataset που σας έχει δοθεί (πίνακες, χαρακτηριστικά, περιορισμοί, εξαρτήσεις κλπ.).

2ο βήμα (προπαρασκευή δεδομένων): Από την παραπάνω ΒΔ θα επιλέξετε τα δεδομένα που θα χρησιμοποιήσετε για αναλυτική επεξεργασία, και θα προχωρήσετε στην όποια προπαρασκευή (καθαρισμό, μετασχηματισμό κλπ.) θεωρείτε απαραίτητη ώστε να αντιμετωπίστε π.χ. το ζήτημα των προβληματικών δεδομένων (ελλιπείς ή εσφαλμένες - μη λογικές - τιμές), των συνεχών πεδίων τιμών κλπ.

Παραδοτέο: Παρουσίαση των ζητημάτων που κλήθηκε να αντιμετωπίσει η προπαρασκευή των δεδομένων και περιγραφή της διαδικασίας που ακολουθήθηκε.

3ο βήμα (υλοποίηση της ΑΔ): Πάνω στα δεδομένα που επιλέξατε στο προηγούμενο βήμα, θα ορίσετε μια αποθήκη δεδομένων – κύβο – με τις κατάλληλες διαστάσεις, ιεραρχίες και μέτρα. Η ΑΔ θα πρέπει να ακολουθεί ως μοντέλο το σχήμα αστέρα (*star schema*) ή χιονονιφάδας (*snowflake schema*).

Παραδοτέο: Τεκμηρίωση της ΑΔ που δημιουργήθηκε: το πολυδιάστατο σχήμα του κύβου, οι διαστάσεις (*dimensions*) και τα μέτρα (*measures*) του, ιεραρχίες διαστάσεων, κλπ.

4ο βήμα (αναλυτική επεξεργασία - OLAP): Θα επιδείξετε πώς μπορεί να χρησιμοποιηθεί ο κύβος που σχεδιάσατε στο προηγούμενο βήμα για να γίνει αναλυτική επεξεργασία των δεδομένων. Πιο συγκεκριμένα, θα παρουσιάσετε παραδείγματα λειτουργιών OLAP (roll-up, drill-down, slice, dice, pivot) μαζί με επεξήγηση των αποτελεσμάτων και πιθανά συμπεράσματα στα οποία θα μπορούσε να οδηγηθεί ο αναλυτής των δεδομένων.

Παραδοτέο: Παραδείγματα εκτέλεσης λειτουργιών OLAP πάνω στον κύβο και καταγραφή σχετικών συμπερασμάτων.

2ο μέρος: Εξόρυξη γνώσης από δεδομένα (data mining)

Στόχος του 2^{ου} μέρους της εργασίας είναι η εξοικείωση με τις τεχνικές εξόρυξης γνώσης από βάσεις δεδομένων (*data mining*). Για τους σκοπούς της άσκησης θα χρησιμοποιηθεί κάποιο από τα εργαλεία: Analysis Services του Microsoft SQL Server 2005, WEKA, Oracle BI.

Πιο συγκεκριμένα, θα πρέπει να ακολουθήσετε τα εξής βήματα (πάνω στην ΑΔ που υλοποιήσατε στο 1^ο μέρος):

5ο βήμα: Θα σκεφτείτε κάποια σενάρια κατηγοριοποίησης (*classification*) που θα μπορούσαν να ενδιαφέρουν τους αναλυτές της συγκεκριμένης εφαρμογής και θα χρησιμοποιήσετε κάποια τεχνική δέντρων απόφασης (*decision trees*). Περιγράψτε τη διαδικασία και εξηγήστε τα αποτελέσματα που προκύπτουν. Πως θα μπορούσε να εκμεταλλευτεί η εταιρία την γνώση αυτή;

Παραδοτέο: Περιγραφή διαδικασίας κατηγοριοποίησης (*classification*) και καταγραφή σχετικών συμπερασμάτων.

6ο βήμα: Θα επαναλάβετε τη διαδικασία, αυτή τη φορά για συσταδοποίηση (*clustering*).

Παραδοτέο: Περιγραφή διαδικασίας συσταδοποίησης (*clustering*) και καταγραφή σχετικών συμπερασμάτων.

7ο βήμα: Θα επαναλάβετε τη διαδικασία, αυτή τη φορά για ανάλυση συσχετίσεων (*association mining*).

Παραδοτέο: Περιγραφή διαδικασίας ανάλυσης συσχετίσεων (*association mining*) και καταγραφή σχετικών συμπερασμάτων.

Διαθέσιμα datasets

Κάθε ομάδα θα επιλέξει ένα dataset πάνω στο οποίο θα εργαστεί. Τα datasets που διατίθενται για αυτή την εργασία είναι τα εξής¹:

1. **Movies data:** δεδομένα ταινιών κινηματογράφου. Υπάρχει κρυμμένη γνώση πίσω από τα πρωτογενή δεδομένα; Παρατηρείτε συσχετίσεις μεταξύ των μεταβλητών;
2. **Income data:** δεδομένα σχετικά με εισοδήματα και επαγγέλματα από δυο πόλεις των ΗΠΑ. Παρατηρείτε συσχετίσεις μεταξύ των μεταβλητών; Μπορείτε να προβλέψετε τάσεις;
3. **Census data:** δημογραφικά δεδομένα σε παγκόσμιο επίπεδο. Υπάρχει κρυμμένη γνώση πίσω από τα πρωτογενή δεδομένα; Μπορείτε να προβλέψετε τάσεις;
4. **Quakes data:** σεισμολογικά δεδομένα σε παγκόσμιο επίπεδο. Παρατηρείτε συσχετίσεις; Μπορείτε να προβλέψετε τάσεις;

Παραδοτέο εργασίας, προθεσμία και τόπος παράδοσης

Το τελικό παραδοτέο θα αποτελείται από ένα φάκελο που θα περιέχει τα εξής:

1. CD που θα περιέχει τα εξής: το αρχικό dataset, τα περιεχόμενα της αρχικής ΒΔ (βλ. βήμα 1), τα script files της προπαρασκευής των δεδομένων (βλ. βήμα 2), τα περιεχόμενα της ΑΔ (βλ. βήμα 3), τα μοντέλα που προέκυψαν από την κατηγοριοποίηση (βλ. βήμα 5), τα μοντέλα που προέκυψαν από την συσταδοποίηση (βλ. βήμα 6) και τα μοντέλα που προέκυψαν από την ανάλυση συσχετίσεων (βλ. βήμα 7).
2. Report με τα επιμέρους παραδοτέα των βημάτων 1 έως 7.

Στο εξώφυλλο του φακέλου θα υπάρχουν τα στοιχεία:

¹ Κάθε ομάδα θα δηλώσει στον κ. Γιατράκο (ngiatrak@unipi.gr) το dataset που προτιμά (1^η, 2^η, ... προτίμηση) και αφού γίνει η ανάθεση θα της υποδειχθεί σύνδεσμος με το σχετικό σύνολο δεδομένων.

«Αποθήκες Δεδομένων και Εξόρυξη Γνώσης (εργασία εξαμήνου)»

Συλλογή δεδομένων: XXX

Ομάδα εργασίας: (ΑΜ, ονοματεπώνυμο)

Ημερομηνία παράδοσης: ΧΧ/ΧΧ/2010

Η εργασία θα πρέπει να έχει παραδοθεί στη θυρίδα του κ. Θεοδωρίδη (γρ. 501) μέχρι την **Τρίτη 12/1/2010**.

Απορίες σχετικά με την άσκηση

Για οποιαδήποτε απορία σχετικά με την άσκηση μπορείτε να απευθύνεστε στον κ. Νίκο Πελέκη (npelekis@unipi.gr) για θέματα που αφορούν το Oracle BI, την κ. Δέσποινα Κοπανάκη (dkopanak@unipi.gr) για θέματα που αφορούν το WEKA και τον κ. Νίκο Γιατράκο (ngiatrak@unipi.gr) για θέματα που αφορούν το SQL Server – Analysis Services είτε μέσω email, είτε μέσω του forum του μαθήματος (<http://195.251.230.57/forum/>).

Ζητήματα δεοντολογίας

Είναι προφανές ότι η βαθμολογία πρέπει να αντικατοπτρίζει το επίπεδο της γνώσης που αποκόμισε ο φοιτητής μέσα από το μάθημα και κατάφερε να μεταφέρει αυτή τη γνώση στην άσκηση. Για να εξασφαλιστεί όσο είναι δυνατό η παραπάνω αρχή, (α) σε περίπτωση αντιγραφής οι εμπλεκόμενες εργασίες μηδενίζονται, (β) σε περίπτωση αμφιβολίας για το κατά πόσο η ομάδα που αναγράφεται ήταν εκείνη που ανέπτυξε την εργασία, ενδέχεται να της ζητηθεί να την παρουσιάσει για τυχόν διευκρινίσεις, (γ) σε περίπτωση μεγάλης απόκλισης του βαθμού της εργασίας από το βαθμό της γραπτής εξέτασης, ο πρώτος δεν θα λαμβάνεται υπόψη για τον τελικό βαθμό του φοιτητή.