ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ
Π.Μ.Σ. « ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ ΚΑΙ ΥΠΗΡΕΣΙΕΣ»
ΚΑΤΕΥΘΥΝΣΗ: ΠΡΟΗΓΜΕΝΑ ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ

Μάθημα : ΑΠΟΘΗΚΕΣ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΕΠΙΧΕΙΡΗΜΑΤΙΚΗ ΕΥΦΥΪΑ

Ακ. έτος 2021–2022

1η Εργασία
Παράδοση: 10/06/2022

ΠΑΡΑΚΑΛΩ ΕΠΙΛΕΞΤΕ ΕΝΑ ΑΠΟ ΤΑ ΠΑΡΑΚΑΤΩ ΘΕΜΑΤΑ

Θέμα 1. Σχεδίαση και υλοποίηση αποθήκης δεδομένων

Ο στόχος 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 μπορείτε να χρησιμοποιήσετε το Σύστημα Διαχείρισης Βάσεων Δεδομένων (ΣΔΒΔ) Microsoft SQL Server και πιο συγκεκριμένα το λογισμικό Analysis Services.

Α Μέρος: Αποθήκη Δεδομένων – OLAP Λειτουργίες

Σας ζητείται να φτιάξετε μια αποθήκη δεδομένων που να βασίζεται σε μία βάση δεδομένων. Πιο συγκεκριμένα, θα πρέπει να ακολουθήσετε τα εξής βήματα:

Βήμα 1 (υλοποίηση ΒΔ): Θα ασχοληθείτε με ένα μεγάλο σύνολο δεδομένων (dataset) με πραγματικά δεδομένα. Μπορείτε να προτείνεται εσείς ένα σύνολο δεδομένων ή να επιλέξτε ένα από τα σύνολα δεδομένα που είναι διαθέσιμα στο

https://www.kaggle.com/datasets?sortBy=relevance&group=public&search=Retail+&page=1&pageSize=20&size=all&filetype=all&license=all

https://www.kaggle.com/c/instacart-market-basket-analysis/data

https://www.ecdc.europa.eu/en/covid-19/data

https://data.world/datasets/warehouse

https://data.gov/

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.

Παραδοτέο: *Περιγραφή του συνόλου δεδομένων. Τεκμηρίωση της ΒΔ που δημιουργήθηκε από το data set που σας έχετε επιλέξει.*

Βήμα 2 (προεπεξεργασία δεδομένων): Από την παραπάνω ΒΔ θα επιλέξετε τα δεδομένα που θα χρησιμοποιήσετε για αναλυτική επεξεργασία, και θα προχωρήσετε στην όποια προεπεξεργασία (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 προετοιμασία των δεδομένων και περιγραφή της διαδικασίας που ακολουθήθηκε.*

Βήμα 3 (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.*

Βήμα 4 (αναλυτική επεξεργασία - OLAP): Θα παρουσιάσετε πώς μπορεί να χρησιμοποιηθεί ο κύβος που σχεδιάσατε στο προηγούμενο βήμα για να γίνει αναλυτική επεξεργασία των δεδομένων. Πιο συγκεκριμένα, θα παρουσιάσετε παραδείγματα λειτουργιών 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, παρουσίαση σχετικών αναφορών (reports) και καταγραφή σχετικών συμπερασμάτων.*

Βήμα 5. Εξαγωγή *κανόνων συσχέτισης (Association rules)*
Θα πρέπει να εφαρμόσετε τεχνικές εξαγωγής κανόνων συσχέτισης από το σύνολο δεδομένων που επιλέξατε.

Παραδοτέο: Περιγράψτε την πιθανή διαδικασία προ-επεξεργασίας που χρειάστηκε να εφαρμόσετε. Επίσης, περιγράψτε τα αποτελέσματα και τα μέτρα ενδιαφέροντος που χρησιμοποιήσατε (support, confidence, lift κλπ) για την επιλογή των κανόνων.

Θέμα 2. Υλοποίηση collaborative filtering προσεγγίσεων

Θεωρούμε ένα διμερές γράφο user-item, όπου κάθε ακμή στο γράφο μεταξύ χρήστη u με το στοιχείο i, δηλώνει ότι στο χρήστη u αρέσει το στοιχείο i. Ορίζουμε επίσης τη μήτρα αξιολογήσεων για αυτό το σύνολο των χρηστών και των στοιχείων ως R, όπου κάθε γραμμή στο R αντιστοιχεί σε έναν χρήστη και κάθε στήλη αντιστοιχεί σε ένα στοιχείο. Εάν στο χρήστη i αρέσει το στοιχείο j, τότε R_{ij} = 1, αλλιώς R_{ij} = 0. Επίσης υποθέσουμε ότι έχουμε χρήστες **m** και **n** αντικείμενα, έτσι η μήτρα R είναι m x n.
Ας ορίσουμε έναν πίνακα P, m x m, ως διαγώνιος πίνακας του οποίου το i-οστό διαγώνιο στοιχείο είναι ο βαθμός του κόμβου χρήστη i, δηλαδή ο αριθμός των στοιχείων που αρέσουν στον χρήστη i. Ομοίως, Q ,n x n, είναι ένας διαγώνιος πίνακας του οποίου το i-οστό διαγώνιο στοιχείο είναι ο βαθμός του κόμβου αντικείμενο i ή ο αριθμός των χρηστών στους οποίους αρέσει το αντικείμενο i.

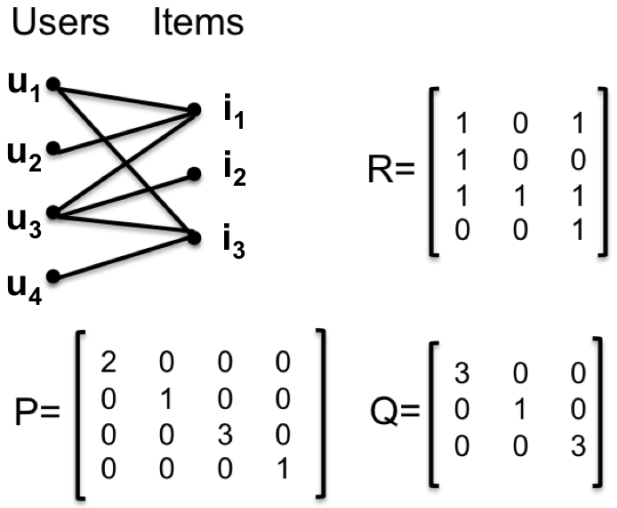

Figure1.

A) Ορίστε την μη-κανονικοποιημένη ομοιότητα ανάμεσα σε χρήστες $T = R*R^T$. Εξηγήστε την έννοια του T_{ii} και T_{ij}, με βάση έναν bipartite γράφο (See Figure 1) (π.χ.. node degrees, path between nodes, κλπ.).

Β) Ας ορίσουμε τη μήτρα ομοιότητας, S_I, nxn, έτσι ώστε το στοιχείο στη γραμμή i και στη στήλη j είναι η ομοιότητα (cosine similarity) του αντικειμένου i με το αντικείμενο j που αντιστοιχεί στη στήλη i και στη στήλη j του πίνακα R. Δείξτε ότι $S_I = Q^{-1/2} R^T R Q^{-1/2}$ όπου $Q^{-1/2}$ με $Q_{rc}^{-1/2} = 1/\sqrt{Q_{rc}}$ για όλες τις μη μηδενικές εισόδους του πίνακα και 0 για όλες τις άλλες θέσεις.

Γ) Εάν θεωρήσουμε τη μήτρα συστάσεων, Γ, m x n, έτσι ώστε Γ (i, j) = rij. Θέλουμε να βρούμε το Γ και για τις item-item, user-user collaborative filtering προσεγγίσεις με βάση του R, P και Q. Για την item-item προσέγγιση ισχύει $\Gamma = RQ^{-1/2} R^T R Q^{-1/2}$.

Υλοποιήστε μια εφαρμογή που θα βασίζεται στην προσέγγιση user-user και item-item collaborative filtering και αξιολογήστε τα αποτελέσματα επιλέγοντας κάποιους ενδεικτικά

χρήστες μέσα από το σύνολο δεδομένων. Επίσης υπολογίστε μέσο τετραγωνικό σφάλμα για ένα υποσύνολο χρηστών που θα επιλέξετε ως σύνολο ελέγχου (test set).

Σε αυτή την ερώτηση θα εφαρμόσετε αυτές τις μεθόδους σε ένα πραγματικό σύνολο δεδομένων. Τα δεδομένα περιέχουν πληροφορίες σχετικά με τηλεοπτικές εκπομπές. Πιο συγκεκριμένα, για 9985 χρήστες και 563 δημοφιλείς τηλεοπτικές εκπομπές, γνωρίζουμε αν ένας δεδομένος χρήστης παρακολούθησε μια συγκεκριμένη εκπομπή σε μια περίοδο 3 μηνών.

Κατεβάστε το σύνολο δεδομένων (DATA_S2.ZIP) από:
https://lefkippos.ds.unipi.gr/modules/document/index.php?course=DSERV107&openDir=%2F603d3de5VvHL%2F60815e80lu5l%2F60815e92PIGx

Παρατηρήσεις
Η μέθοδος συστάσεων που χρησιμοποιεί user-user collaborative filtering για το u, μπορεί να περιγραφεί ως εξής: Για όλα τα αντικείμενα(items) s, υπολογίζουμε

$$r_{u,s} = \Sigma_{x \in users} \text{cos-sim}(x, u) * R_{xs}$$

και προτείνουμε τα k items για τα οποία το $r_{u,s}$ είναι το μεγαλύτερο.
Παρόμοια , για την προσέγγιση item-item collaborative filtering για το χρήστη u θα έχουμε: για όλα τα items s,

$$r_{u,s} = \Sigma_{x \in items} R_{ux} * \text{cos-sim}(x, s)$$

Αναφορές
1. http://infolab.stanford.edu/~ullman/mmds/ch9.pdf
2. Διαφάνειες Μαθήματος

Παρατηρήσεις

1. **Η εργασία θα γίνει σε ομάδες 2-3 ατόμων.**

2. **Η εργασία θα υποβληθεί μέσω e-class (https://lefkippos.ds.unipi.gr/) στην περιοχή «Εργασίες»**
   Θα πρέπει να παραδώσετε ένα αρχείο AM1-AM2-AM3.zip (AM είναι ο αριθμός μητρώου σας) το οποίο θα περιλαμβάνει:
   Α. Τεκμηρίωση με βάση τις απαιτήσεις που αναφέρονται σε κάθε ερώτημα (σε μορφη pdf).
   Β. Πηγαίο κώδικα, σύνολα δεδομένων, αποτελέσματα ανάλυσης.