Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Τεχνολογιας Πληροφορικης και Υπολογιστων
Εργαστηριο Υπολογιστικων Συστηματων

Διπλωματική Εργασία 2011-2012

Ανωνυμοποίηση Σχεσιακών Δεδομένων με Χρήση Κατανεμημένων Τεχνικών Μηχανικής Εκμάθησης

Εισαγωγή

Την τελευταία δεκαετία έχουμε γίνει μάρτυρες μιας έκρηξης δεδομένων, που ακόμα βρίσκεται σε εξέλιξη. Πληθώρα εταιριών, οργανισμών, ερευνητικών και ακαδημαϊκών κέντρων παράγουν τεράστια ποσά δεδομένων και στηρίζονται στη συνεχή ανάλυσή τους για να ανιχνεύουν μοτίβα συμπεριφοράς, να ανακαλύπτουν ενδιαφέρουσες τάσεις και συσχετίσεις ή να προσφέρουν προσωποποιημένες υπηρεσίες σε χρήστες. Για την εξυπηρέτηση αυτών των αναγκών, έχουν προταθεί και χρησιμοποιούνται ευρέως κατανεμημένα συστήματα που στηρίζονται σε αρχιτεκτονικές χωρίς διαμοιραζόμενους πόρους (shared-nothing), τα οποία προσφέρουν το πλεονέκτημα της κλιμακωσιμότητας και της ανοχής σε σφάλματα με χαμηλό κόστος (π.χ. P2P, Cloud κτλ).

Ωστόσο, η πρόσβαση σε μεγάλο όγκο δεδομένων που συχνά παράγονται από πολλές διαφορετικές πηγές εγείρει σημαντικά ζητήματα προστασίας της ανωνυμίας των ατόμων και της ιδιωτικότητας των πληροφοριών που τα αφορούν. Ακόμα κι αν κάποια χαρακτηριστικά όπως το όνομα ή το ΑΦΜ απομακρυνθούν από το σύνολο των δεδομένων, ο συνδυασμός κάποιων από τα εναπομείναντα γνωρίσματα (που ονομάζονται quasi-identifier attributes ή απλά QIDs) με εξωτερικά, δημοσίως διαθέσιμα δεδομένα (π.χ. εκλογικούς καταλόγους) ενδέχεται να οδηγήσουν στην ταυτοποίηση ατόμων και να αποκαλύψει ευαίσθητες προσωπικές τους πληροφορίες.

Η αρχή του k-anonymity έχει προταθεί για την προστασία προσωπικών δεδομένων από τέτοιους κινδύνους [5]. Ο σκοπός της είναι κάθε συνδυασμός τιμών των QIDs να εμφανίζεται στα δημοσιευμένα δεδομένα σε τουλάχιστον k εγγραφές. Συνήθως αυτό επιτυγχάνεται με μεθόδους γενίκευσης (generalization) της τιμής κάποιων attributes, για παράδειγμα αντικαθιστώντας την ηλικία ενός ασθενούς 25 ετών με το πιο γενικό διάστημα 20-30. Κάθε γενίκευση τιμών βέβαια συνεπάγεται απώλεια πληροφορίας στα τελικά δεδομένα. Επομένως ο αλγόριθμος ανωνυμοποίησης θα πρέπει να επιτυγχάνει την ελάχιστη απώλεια πληροφορίας, διατηρώντας τα δεδομένα όσο το δυνατόν πιο χρήσιμα.

Υπάρχει πληθώρα αλγορίθμων που επιτυγχάνουν k-anonymity στη σχετική βιβλιογραφία (π.χ. [3], [4]). Ωστόσο οι αλγόριθμοι αυτοί λειτουργούν κεντρικά και παρουσιάζουν περιορισμούς στη χρήση τους για την ανωνυμοποίηση μεγάλου όγκου δεδομένων. Στην παρούσα διπλωματική θα ασχοληθούμε με την αποδοτική ανωνυμοποίηση μεγά-

λου όγκου σχεσιακών δεδομένων κάνοντας χρήση κατανεμημένων τεχνικών μηχανικής εκμάθησης.

Σκοπός
Σκοπός της διπλωματικής είναι η ανάπτυξη μιας κατανεμημένης μεθόδου ανωνυμοποίησης, η οποία θα εξασφαλίζει την ιδιωτικότητα κατανεμημένων δεδομένων με τη μικρότερη δυνατή απώλεια πληροφορίας. Αυτό θα γίνει με τη χρήση κατανεμημένων εργαλείων όπως το Hadoop [1], το οποίο αποτελεί το πλέον διαδεδομένο framework για κατανεμημένη επεξεργασία και το Apache Mahout [2], το οποίο προσφέρει μια βιβλιοθήκη για αλγορίθμους μηχανικής εκμάθησης πάνω από Hadoop.
Επικοινωνία:
Νεκτάριος Κοζύρης, Καθηγητής nkoziris@cslab.ece.ntua.gr
Κατερίνα Δόκα, Μεταδ. Ερευνήτρια katerina@cslab.ece.ntua.gr

Βιβλιογραφία

[1] Apache Hadoop. http://hadoop.apache.org/. 2

[2] Apache Mahout: Scalable machine learning and data mining. http://mahout.apache.org/. 2

[3] K. LeFevre, D.J. DeWitt, and R. Ramakrishnan. Incognito: Efficient Full-Domain k-Anonymity. In *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data (SIGMOD'05)*, page 60. ACM, 2005. 1

[4] K. LeFevre, DJ DeWitt, and R. Ramakrishnan. Mondrian Multidimensional k-Anonymity. In *Proceedings of the 22nd International Conference on Data Engineering (ICDE'06)*, pages 25–25. IEEE Computer Society Press, 2006. 1

[5] L. Sweeney et al. k-anonymity: A Model for Protecting Privacy. *International Journal of Uncertainty Fuzziness and Knowledge Based Systems*, 10(5):557–570, 2002. 1