ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ

ΤΜΗΜΑ ΕΠΙΣΤΗΜΗΣ ΥΠΟΛΟΓΙΣΤΩΝ

ΠΑΡΟΥΣΙΑΣΗ / ΕΞΕΤΑΣΗ ΜΕΤΑΠΤΥΧΙΑΚΗΣ ΕΡΓΑΣΙΑΣ

Αδαμάκης Εμμανουήλ

Μεταπτυχιακός Φοιτητής

Τμήμα Επιστήμης Υπολογιστών, Πανεπιστήμιο Κρήτης

Επόπτης Μεταπτυχιακής Εργασίας: Καθηγητής, Κ. Στεφανίδης

Δρ. Γ. Μαργέτης (επιβλέπων)

Τρίτη, 21 Δεκεμβρίου 2021, ώρα 10:00 π.μ.

Join Zoom Meeting
https://zoom.us/j/99857484556

"Σχεδίαση και ανάπτυξη ενιαίου πλαισίου ανωνυμοποίησης, ανάλυσης, οπτικοποίησης και εξερεύνησης μεγάλων δεδομένων τα οποία αντλούνται μέσω ψηφιακών αγορών δεδομένων"

Περίληψη

Στη σημερινή εποχή που βασίζεται στα δεδομένα, η ανταλλαγή αυτών παίζει καθοριστικό ρόλο στην καθημερινότητα μας. Κάθε ψηφιακή συναλλαγή, από την πιο απλή έως την πιο περίπλοκη, απαιτεί ανταλλαγή δεδομένων μεταξύ των εμπλεκομένων μερών. Από ιδιώτες και μικρές επιχειρήσεις έως μεγάλες εταιρείες, οργανισμούς και κυβερνήσεις όλοι αποθηκεύουν, επεξεργάζονται και ανταλλάσσουν δεδομένα. Αυτή η πραγματικότητα, με την πάροδο του χρόνου, οδήγησε στη συσσώρευση τεράστιων όγκων δεδομένων, γνωστά και ως Μεγάλα Δεδομένα. Με την έλευση των Μεγάλων Δεδομένων, έγινε φανερό ότι υπήρχαν πολλές ευκαιρίες αναφορικά με την ανάλυσή τους και ότι τα αποτελέσματα τέτοιων αναλύσεων θα μπορούσαν να παρέχουν στους επεξεργαστές των δεδομένων αυτών εξαιρετικά ωφέλιμες πληροφορίες (insights). Επίσης, μεγάλη βοήθεια στη βελτίωση των αποτελεσμάτων τέτοιων αναλύσεων μπορεί να προσφέρει ο εμπλουτισμός και η συσχέτιση των υπαρχόντων, ιδιόκτητων, συνόλων δεδομένων με σύνολα που προήλθαν από εξωτερικές πηγές. Η απόκτηση συνόλων δεδομένων μέσω τρίτων, ήταν κατά κανόνα μια διαδικασία που απαιτούσε την άμεση προσέγγιση συγκεκριμένων κατόχων δεδομένων. Ωστόσο, τα τελευταία χρόνια, με την εμφάνιση των ψηφιακών αγορών δεδομένων, αυτή η κατάσταση έχει αρχίσει να αλλάζει.

Στο πρόσφατο παρελθόν, οι ανταλλαγές δεδομένων πραγματοποιούνταν με ελάχιστη έως καθόλου μέριμνα για το απόρρητο ή την προστασία των προσωπικών δεδομένων. Πρόσφατες νομοθετικές εξελίξεις, όπως η νομοθεσία της Ευρωπαϊκής Ένωσης για την προστασία των προσωπικών δεδομένων ΓΚΠΔ, ώθησαν πολλούς παρόχους και καταναλωτές Μεγάλων Δεδομένων να αναζητήσουν λύσεις τόσο για την προστασία του απορρήτου όσο και για την αξιολόγηση των κινδύνων απορρήτου των συνόλων δεδομένων που αυτοί διαχειρίζονται. Μετά από αυτές τις εξελίξεις, οποιαδήποτε δημοσιοποίηση δεδομένων πρέπει να εφαρμόζει κάποια μορφή προσαρμογής των δεδομένων αυτών πριν από τη δημοσιοποίηση, ώστε να προστατεύει το απόρρητο των ευαίσθητων πληροφοριών των ατόμων. Η ανωνυμοποίηση δεδομένων είναι ένα παράδειγμα μιας τέτοιας διαδικασίας προσαρμογής και περιλαμβάνει την αφαίρεση ή τη μετατροπή δεδομένων, με τρόπο που διατηρεί το απόρρητο, εξασφαλίζοντας έτσι ένα ορισμένο επίπεδο ανωνυμίας αυτών. Μία από τις πιο δύσκολες πτυχές οποιασδήποτε διαδικασίας ανωνυμοποίησης δεδομένων είναι η επίτευξη ισορροπίας μεταξύ της χρησιμότητας των δεδομένων και του απορρήτου. Υπό το πρίσμα αυτό, εργαλεία ανάλυσης κινδύνου και ανωνυμοποίησης είναι απαραίτητα προκειμένου να αυξηθεί η ευαισθητοποίηση σχετικά με τους κινδύνους απορρήτου, καθώς και να βοηθηθούν οι επεξεργαστές των δεδομένων αυτών τόσο στη συμμόρφωσή τους με τους κανονισμούς, όσο και στην ίδια την διαδικασία της ανωνυμοποίησης των δεδομένων αυτών. Αν και υπάρχουν κάποια εργαλεία στη βιβλιογραφία, αυτά δεν προσφέρουν ένα αρκετά ευρύ φάσμα επιλογών όσον αφορά τους τύπους δεδομένων που μπορούν να αναλύσουν, την υποστήριξη πολλαπλών διαστάσεων δεδομένων, καθώς και την οπτική εξερεύνηση των αποτελεσμάτων των αναλύσεων που διενεργούνται.

Εκτός από τα ζητήματα απορρήτου των δεδομένων που υπάρχουν σχετικά με τις δημοσιοποιήσεις και συναλλαγές Μεγάλων Δεδομένων, υπάρχουν επίσης προκλήσεις σχετικά με την ουσιαστική ανάλυσή τους. Η Οπτική Ανάλυση (Visual Analytics) είναι ένας τομέας έρευνας που εστιάζει στην παροχή αποτελεσματικών και διαφανών μεθόδων επεξεργασίας, οπτικοποίησης και ανάλυσης μεγάλου όγκου δεδομένων, έτσι ώστε οι αναλυτές να μπορούν να τους κατανοήσουν καλύτερα και να εξάγουν πληροφορίες που θα μπορούσαν να υποστηρίξουν τη λήψη αποφάσεων. Στη βιβλιογραφία, υπάρχει αρκετή ποικιλία εφαρμογών Οπτικής Ανάλυσης. Ανάμεσα στα κοινά χαρακτηριστικά των εφαρμογών αυτών, υπάρχει η δυνατότητα δημιουργίας ταμπλό (dashboards) για την υποστήριξη της εξερεύνησης Μεγάλων Δεδομένων. Τα ταμπλό (dashboards) είναι μια συλλογή οπτικοποιήσεων δεδομένων και επιλογών φιλτραρίσματος, σχεδιασμένα για να βοηθούν τους αναλυτές με την ανάλυση των δεδομένων παρέχοντας τους έναν διαδραστικό τρόπο για τη διεξαγωγή της. Ωστόσο, οι περισσότερες από τις διαθέσιμες λύσεις, επί του παρόντος, υπολείπονται όσον αφορά τη εμβάθυνση (drill-down) και γενίκευση (roll-up) των δεδομένων που οπτικοποιούνται στο ταμπλό. Η εμβάθυνση στα δεδομένα αναφέρεται στη διαδικασία κατά την οποία ένας αναλυτής μπορεί να μεταβεί από μια ομαδοποίηση δεδομένων σε μια πιο λεπτομερή ομαδοποίηση, ενώ η γενίκευση αφορά στη διερεύνηση των δεδομένων σε σταδιακά λιγότερο λεπτομερές επίπεδο.

Όμως, οι εφαρμογές παρέχουν αυτήν τη λειτουργικότητα με περιορισμένο τρόπο και μόνο σε συγκεκριμένα γραφήματα ή γράφους, χωρίς να μπορούν να υποστηρίξουν τη διάδοση των ενεργειών της εμβάθυνσης ή γενίκευσης στις υπόλοιπες οπτικοποιήσεις του ταμπλό.

Η προτεινόμενη μεθοδολογία μας για την αντιμετώπιση των προαναφερθέντων ζητημάτων περιλαμβάνει τον σχεδιασμό και την ανάπτυξη ενός ενιαίου πλαισίου εφαρμογών που στοχεύουν στην ανάλυση, οπτικοποίηση και εξερεύνηση μεγάλων δεδομένων, διασφαλίζοντας παράλληλα την ασφάλεια και την ιδιωτικότητα. Αυτές οι εφαρμογές παρέχουν τη δυνατότητα ανάλυσης του κινδύνου διαρροής προσωπικών δεδομένων που μπορεί να διαρρεύσουν μέσα από ένα σύνολο δεδομένων, καθώς και τη δυνατότητα ανωνυμοποίησής τους. Επιπλέον, διευκολύνουν την απεικόνιση και την εξερεύνηση μεγάλων συνόλων δεδομένων συνδυάζοντας ιδιόκτητα σύνολα δεδομένων με σύνολα που αποκτήθηκαν από αγορές ψηφιακών δεδομένων και οπτικοποιώντας τα μέσω διαδραστικών ταμπλό. Τα ταμπλό αυτά μπορούν να προσαρμόζονται στις απαιτήσεις του πλαισίου ανάλυσης του χρήστη και να παρέχουν λειτουργίες εμβάθυνσης ή γενίκευσης των δεδομένων με βάση τον τύπο των δεδομένων υπό ανάλυση, επιτρέποντας έτσι στους χρήστες να εντοπίσουν νέα γνώση την οποία δεν κατείχαν πριν αναλύσουν τα συγκεκριμένα σύνολα δεδομένων.

University of Crete

Computer Science Department

M.Sc. Thesis

Emmanouil Adamakis

Master's Thesis Supervisor: Professor, C. Stephanides

Dr. G. Margetis (Co-Advisor)

Tuesday, 21 December 2021, 10:00 a.m.

Join Zoom Meeting
https://zoom.us/j/99857484556

“Design and development of a unified framework for the anonymization, analysis, visualization and exploration of big data acquired from digital market places”

Abstract

In today's data-driven world, data interchange plays a pivotal role in our daily lives. Every digital transaction, from the simplest to the most complex, requires data exchanging between the parties involved. From individuals and small businesses to large corporations, organizations, and governments all store, process and exchange data. This situation, over time, has led to the accumulation of large volumes of data, called Big Data. With the emergence of Big Data, it became apparent that there were numerous opportunities in terms of their analysis and the information results (insights) of such analyses, which could be highly beneficial to the data processors' goals. Of great assistance at improving the outcomes of such analyses was also identified to be the enrichment and correlation of existing internal datasets with datasets acquired from external sources. Obtaining third-party datasets used to entail approaching specific data owners directly; however, with the emergence of digital data market places in recent years, this situation has begun to change.

Until recently, data exchanges were carried out with little to no regard for privacy or the protection of personal data. Recent legislative developments, such as the European Union's GDPR data protection laws, have prompted many data providers and consumers to seek solutions for both protecting individuals' privacy and assessing the privacy risks of the datasets under their management. Following these developments, any data disclosure has to employ some form of data sanitization prior to release, in order to protect the privacy of individuals' sensitive information. Anonymization of data is an example of such a sanitization process, and it involves the deduction or transformation of data in a privacy-preserving manner in order to achieve a certain level of anonymity. One of the most difficult aspects of any anonymization process is striking a balance between data utility and privacy. Under that scope, risk analysis and anonymization tools are required in order to increase awareness of the privacy risks, aid in regulatory compliance, and assist data processors with the anonymization process. Although there are a few tools reported in literature, they do not offer a wide range of options in terms of the types of data that can be analyzed, the support of data multidimensionality, and visual exploration of the risk analysis results.

Aside from data privacy issues regarding the disclosures and exchanges of Big Data, there are also challenges over their meaningful analysis. Visual analytics is a research area that focuses on offering efficient and transparent methods of processing, visualizing, and analyzing large volumes of data so that analysts may better understand them and extract insights that could support data-driven decision making. In the literature, a variety of visual analytics applications are available. Among the most common features of such applications is the ability to create dashboards in order to support Big Data exploration. Dashboards are a collection of data visualizations and filtering options designed to assist analysts and provide an interactive way for them to conduct their analysis. However, most of the currently available solutions fall short when it comes to dashboard-wide data exploration through drill-down or roll-up analysis. Data drill-down refers to the process by which an analyst can shift from a grouping of data to a more detailed and granular group of data, whereas roll-up refers to investigating data in progressively less detailed levels. The applications offering this functionality only provide it in a limited fashion and for specific charts or graphs,

without being able to support propagation of the drilling or rolling actions to the rest of the dashboard's visualizations.

Our proposed methodology for dealing with the aforementioned issues involves the design and development of a unified framework of applications aimed at the analysis, visualization, and exploration of big data, while ensuring security and privacy. These applications provide the ability to analyze the risk of leaking personal data that may pass through a set of data, and also the ability to anonymize them. Furthermore, they facilitate the visualization and exploration of large datasets by combining previously owned datasets with those obtained from digital data marketplaces and displaying them through interactive dashboards. These dashboards can be adapted to the user requirements and provide data-drilling functionalities based on the type of data under analysis, thus allowing users to gain new insights that they could not have gained otherwise.