ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
"ΤΕΧΝΟΛΟΓΙΕΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΣΤΗΝ ΙΑΤΡΙΚΗ ΚΑΙ ΤΗ ΒΙΟΛΟΓΙΑ"

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Εξερεύνηση και Καθαρισμός Δεδομένων σε Βιοϊατρικές Βάσεις

Άννα Α. Γόγολου

Επιβλέπων: **Ιωάννης Ιωαννίδης,** Καθηγητής ΕΚΠΑ

ΑΘΗΝΑ

ΦΕΒΡΟΥΑΡΙΟΣ 2016

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Εξερεύνηση και Καθαρισμός Δεδομένων σε Βιοϊατρικές Βάσεις

Άννα Α. Γόγολου
Α.Μ.: ΠΙΒ0109

ΕΠΙΒΛΕΠΩΝ: **Ιωάννης Ιωαννίδης,** Καθηγητής ΕΚΠΑ

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ: **Άρτεμις Χατζηγεωργίου,** Καθηγήτρια
Ζωή Κούρνια, Ερευνήτρια ΙΙΒΕΑΑ

Φεβρουάριος 2016

ΠΕΡΙΛΗΨΗ

Ως ο ορισμός της διαδικασίας ανίχνευσης και διόρθωσης ασυνεπειών και λαθών στα δεδομένα, ο καθαρισμός δεδομένων αποτελεί ένα ουσιαστικό βήμα προεπεξεργασίας που σχετίζεται με πολλά θέματα βάσεων δεδομένων και βιοπληροφορικής. Καθαρά κι έγκυρα δεδομένα αποτελούν σημαντική προϋπόθεση για την εξέλιξη της έρευνας οποιουδήποτε ακαδημαϊκού ερευνητή και όχι μόνο. Πιο συγκεκριμένα, η ανάγκη για καθαρά δεδομένα κυριαρχεί σε κάθε έκφανση επιστημονικής δραστηριότητας, αλλά και στις δραστηριότητες της σημερινής οικονομίας. Για το σκοπό αυτό, υπάρχει πλήθος εργαλείων καθαρισμού δεδομένων, με διαφορετικό βαθμό επιτυχίας το καθένα στην αντιμετώπιση των υφιστάμενων προκλήσεων της εκάστοτε διαδικασίας.

Στην παρούσα διπλωματική εργασία, παρουσιάζω την υλοποίηση και ανάπτυξη ενός ολοκληρωμένου εργαλείου καθαρισμού δεδομένων. Το παρόν εργαλείο είναι μια φιλική προς το χρήστη διαδικτυακή εφαρμογή που προσφέρει προηγμένη (ημι)-αυτόματη διαδικασία καθαρισμού για μεγάλο όγκο ετερογενών δεδομένων. Το εργαλείο τρέχει πάνω από το σύστημα madIS, το οποίο παρέχει επεξεργασία και ανάλυση των δεδομένων με τη χρήση συναρτήσεων (τελεστών) γραμμένων σε Python που επεκτείνουν την SQL-λειτουργικότητα ενός σχεσιακού συστήματος βάσης δεδομένων.

Αυτόματη ανίχνευση σφαλμάτων τύπου και ακραίων αριθμητικών τιμών (αριθμητικών εκτόπων) επιτυγχάνεται κατά τη διαδικασία σύνθεσης του προφίλ των δεδομένων. Μέσω του παρόντος εργαλείου, μια εκτεταμένη σουίτα εξερεύνησης και ανάλυσης δεδομένων, ικανοποίησης περιορισμών, διαδραστικών οπτικοποιήσεων στατιστικών αποτελεσμάτων και αποτελεσμάτων τεχνικών εξόρυξης δεδομένων προσφέρεται στο χρήστη, προκειμένου να εντοπίζει με διαδραστικό τρόπο πιθανά σφάλματα, ακραίες τιμές, τυπογραφικά λάθη και παραβάσεις κανόνων καθαρισμού. Επιπλέον, προτεινόμενες διορθώσεις μπορούν εύκολα να γίνονται αποδεκτές ή να απορρίπτονται.

Ως μέρος της διαδικασίας καθαρισμού, το εργαλείο υποστηρίζει, επίσης, την επεκτασιμότητα των δεδομένων μέσω του υπολογισμού νέων παράγωγων μεταβλητών. Τέλος, διατηρείται ιστορικό των ενεργειών του χρήστη, επιτρέποντάς του να αναιρέσει ένα ή περισσότερα βήματα, να εξάγει μια ροή εργασίας (workflow) και να την επανεκτελέσει σε διαφορετικά ή επιπρόσθετα δεδομένα.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Καθαρισμός Δεδομένων

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: εντοπισμός ακραίων αριθμητικών τιμών, τυπογραφικά λάθη, κανόνες καθαρισμού δεδομένων, διαδραστικές οπτικοποιήσεις, υπολογισμός νέων παράγωγων μεταβλητών

ABSTRACT

Defined as the process of detecting and correcting inconsistencies and errors in data, data cleaning constitutes an essential pre-processing step in many database- and bioinformatics-related tasks. Curated and valid data is a prerequisite for the upcoming research activity of any academic researcher and not only. In particular, the need for cleaned data dominates in every scientific activity and in today's economy. There are several existing data cleaning tools, with a varying degree of success in dealing with the challenges of this process.

In this thesis, I present the development and functionality of a completed data cleaning tool. This tool is a user-friendly web application offering an advanced (semi)-automatic data cleaning process on large volumes of heterogeneous data. The tool runs on top of the madIS system, which provides data processing and analysis functionality via an extended relational database system.

Automatic detection of type errors and numeric outliers is achieved during the data profiling process. An extensive suite of data analysis, constraint satisfaction, interactive data mining and statistical visualization results is offered to the user in order to identify potential errors, outliers, misspellings and violations. In addition, the tool suggests corrections that are easily accepted or rejected.

As part of its data curation functionality, the tool also supports data extensibility with row and aggregate operations being available to compute new derived variables in the data. Finally, the tool keeps history of users' actions allowing them to undo/redo history, extract workflows and re-execute them on different or additional data.

SUBJECT AREA: Data Cleaning
KEYWORDS: numeric outliers, misspellings, data cleaning rules, interactive visualizations, new derived columns