Kivonat

Napjainkban egyre több az informatikától és a statisztikától korábban jórészt független szakma alkalmaz adatelemzést. Az adattárolás és feldolgozás fajlagos költségének csökkenésével egyre nagyobb mennyiségű és egyre nagyobb sokféleségű adaton végeznek elemzéseket, sokszor olyan szakemberek, akiknek fő szakterülete az adatminőség értékeléstől és javítástól igen messze esik. Mindeközben a heterogén forrásból származó, jellemzően tisztítatlan adatok által jelentett kockázatok jelentősége nem csökkent. A feldolgozási és elemzési folyamat érzékeny lehet a bemeneti adatok hibáira, valamint a szakértői feltételezések ellenőrzését is igényli. Ezt a feladatot nehezíti, hogy az adatok szerkezete és adott esetben az adatok forrása is időben változhat.

Az adatok rendszerezése, tisztítása és elemzése ma már gyakran adatelemző munkafolyamatok segítségével történik, melyek kezelését grafikus eszközök (pl. Rapid Miner, Knime) támogatják az adatelemzésben kevésbé jártas felhasználóknak is.

Amennyiben a feldolgozás / adattisztítás lépései során nem sikerül kiküszöbölni az összes adathibát, akkor ezek torzíthatják az adatelemzési lépések (pl. interaktív vizuális analízis, statisztikai módszerek) kimenetét. Ugyanakkor ezek a hibák sokszor kivédhetőek további adattisztító és konzisztencia-ellenőrző lépések beiktatásával, amelyek megakadályozhatják a hibás értékek továbbterjedését az adathibákra érzékeny lépésekig.

A dolgozat keretein belül megterveztünk egy ontológia alapú metamodellt, mely általános adatfeldolgozó folyamatokat ír le. Létrehoztunk reprezentatív példa adattisztító és adatelemző folyamatokat egy erre alkalmas grafikus eszközben (Rapid Miner) és biztosítottuk a folyamatokból (ontológia alapú) példánymodellek generálását. A téma szakirodalmának tanulmányozása alapján megalkottunk egy adathibákat leíró taxonómiát. Az adathibák terjedését leíró szabályokat definiáltunk az adatfeldolgozási, -tisztítási és -elemzési folyamatok különböző típusú lépéseire, és megvizsgáltuk, hogy mely lépés mely adathibákra érzékeny és hogyan tehető robusztussá. Példát adtunk arra, hogy a vizsgált környezet modelljének ismerete hogyan segítheti az adatok konzisztencia- és teljességellenőrzését.

Hibaterjedés alapú eszközt és módszert dolgoztunk ki, melyhez a fenti folyamatokból automatikusan komponens alapú hibaterjedési modelleket állítottunk elő. A vizsgálathoz megalkottunk egy általános komponens leíró modellt, amellyel más típusú rendszerek is leírhatóak. Megvalósítottunk egy eszközt, amely a generált modellen korlátkielégítési programozás alapú hibaterjedés vizsgálatot hajt végre, és képes felderíteni a lehetséges hibaokokat és jelenségeket a folyamatban, visszavezetve ezeket az eredeti modell szintjére. Az elkészült rendszerünk ezáltal képes rámutatni a folyamat azon lépéseire, ahol további ellenőrzésekre vagy adattisztításra van szükség.

A dolgozatban egy összetett felhő alapú alkalmazás teljesítmény és szolgáltatásbiztonsági mérési adatainak feldolgozásán és kezdeti elemzésén keresztül mutatjuk be módszerünk gyakorlati alkalmazhatóságát.

Eredményeink közvetlenül segíthetik adatelemzési projektek hatékony tervezését azáltal, hogy szisztematikus módon javaslatot teszünk bemeneti adatok és a köztes számítások hibáinak kiszűrésére a mért rendszer modelljének figyelembevételével. Ezzel időigényes és szakértői

tudást igénylő munkát váltunk ki és segítjük, hogy az elemző a lényegi problémák felderítésére koncentráljon. A megközelítésünk független az analízis során használt eszközöktől.

Abstract

Data analysis is an important activity which is performed by a growing number of non-technical users as well.

Data cleaning is necessary in most in data analysis processes. A data analysis process can be very sensitive to faults in the input data and often needs human check by experts. The structure and the sources of the data may also change over time, which makinges data cleaning more difficult. In data analysis projects, data processing, data cleaning and data mining areis often implemented by workflows. Graphical data analysis workflow tools (eg. Rapid Miner, Knime) support these activities for non-technical users as well who may not have deep experience in data cleaning and analysis.

If the data errors were are not eliminated completely during data processing / data cleaning, these errors can seriously distort the output result of the actual data analysis steps (eg. interactive visual analysis, statistical methods). However, these errors can be "caught" during the processavoided by inserting additional data cleaning and consistency checks. This way the further propagation of data errors can be stopped.steps to avoid propagating the wrong value across the process to the sensitive steps.

In our paper we describe an ontology-based metamodel to represent general data analysis processes. We created data cleaning and analysis processes in a popular graphical tool (Rapid Miner) and implemented the automatic generation of (ontology) instance models. We created a data fault-error-failure taxonomy based on literature research. For the various types of data cleaning, analysis and processing operations, we defined error propagation rules and analysed the data quality sensitivity and robustness of the steps. We show how system models can help the analysis of data consistency and completeness.

We examined how methods of dependable computing can be used within this field. We developed a method and a tool for error propagation analysis of the above mentioned processes which are transformed to component based system model. We created a method and a tool for the error propagation analysis based on constraint programming. Potential faults and failures of data. management workflows are traced back to the originall modemodel so our system can reveal weak points in the process where data cleaning steps should be inserted to prevent failures caused by poor data quality. In this paper we also introduce the practical applicability of our method in the case study of performance and dependability analysis of metrics measured in a complex cloud based application.

Our results can help to perform data analysis project more effectively by giving recommendations on how to eliminate input data errors and the effects of wrong calculations, also considering the model of the system under analysis. These recommendations can save a lot of expert effort and help data analysts to concentrate on the investigation of the business problem. Our methodology is independent from the tools being used for analysis.