Wong Lim Soon

From: Loo Line Fong [loolf@comp.nus.edu.sg]

Sent: Friday, July 27, 2007 8:53 AM

To: rs-l@comp.nus.edu.sg; acad-cs-l@comp.nus.edu.sg

Cc: ksung@comp.nus.edu.sg; Ling Tok Wang; Tan Kian Lee; leeml@comp.nus.edu.sg; judice.koh@gmail.com; Loo Line Fong

Subject: Postgraduate Seminar by Ms Koh Lie Yong

SCHOOL OF COMPUTING,NUS

POSTGRADUATE SEMINAR BY

Ms Koh Lie Yong

Correlation-based Methods for Biological Data Cleaning

SR3A (COM1 #02-12)

6 August1 2007, 10.00am

Abstract:

Data overload combine with widespread use of automated large-scale analysis and mining result in a rapid depreciation of the World's data quality. Data cleaning is an emerging domain that aims at improving data quality through the detection and elimination of data artifacts. These data artifacts comprise of errors, discrepancies, redundancies, ambiguities, and incompleteness that hamper the efficacy of analysis or data mining.

Despite the importance, data cleaning remains neglected in certain knowledge-driven domains. One such example is Bioinformatics; biological data are often used uncritically without considering the errors or noises contained within, and research on both the "causes" of data artifacts and the corresponding data cleaning remedies are lacking. In this thesis, we conduct the an in-depth study of what constitutes data artifacts in real-world biological databases. To the best of our knowledge, this is the first complete investigation of the data quality factors in biological data.The result of our study indicates that biological data quality problem is by nature multifactorial and requires a number of different data cleaning approaches. While some existing data cleaning methods are directly applicable to certain artifacts, others such as annotation errors and multiple duplicate relations have not been studied. This provides the inspirations for us to devise new data cleaning methods.

Current data cleaning approaches derive observations of data artifacts from the values of independent attributes and records. On the other hand, the correlation patterns between the attributes provide additional information of the relationships embedded within a data set among the entities. In this thesis, we exploit the correlations between data entities to identify data artifacts that existing data cleaning methods fall short of addressing. We propose 3 novel data cleaning methods for detecting outliers and duplicates, and further apply them to real-world biological data as proof-of-concepts.

Traditional outlier detection approaches rely on the rarity of the target attribute or records. While rarity may be a good measure for class outliers, for attribute outliers, rarity may not equate abnormality. The ODDS (Outlier Detection from Data Subspaces) method utilizes deviating correlation patterns for the identification of common yet abnormal attributes. Experimental validation shows that it can achieve an accuracy of up to 88%.

The ODDS method is further extended to XODDS, an outlier detection method for semi-structured data models such as XML which is rapidly emerging as a new standard for data representation and exchange on the World Wide Web (WWW). In XODDS, we leverage on the hierarchical structure of the XML to provide addition context information enabling knowledge-based data cleaning. Experimental validation shows that the contextual information in XODDS elevates both efficiency and the effectiveness of detecting outliers.

Traditional duplicate detection methods regard duplicate relation as a boolean property. Moreover, different types of duplicates exists, some of which cannot be trivially merged. Our third contribution, the correlation-based duplicate detection method induced rules from associations between attributes in order to identify different types of duplicates.

Correlation-based methods aimed at resolving data cleaning problems are conceptually new. This thesis demonstrates they are effective in addressing some data artifacts that cannot be tackled by existing data cleaning techniques, with evidence of practical applications to real-world biological databases.

--
regards,

Line Fong (Ms)
Graduate Division
School of Computing, NUS
email: loolf@comp.nus.edu.sg