A Critical Review of SQL-Based Mining Relational Database

Yuexiao Teng*
East China University of Science and Technology, Shanghai, China.

* Corresponding author. Email: roger.yxt@aliyun.com
Manuscript submitted March 21, 2021; accepted May 5, 2021.
doi: 10.17706/ijcce.2021.10.3.68-74

Abstract: Mining database aims to discover hidden, yet potentially useful knowledge in database and it has many applications. SQL-based mining relational database is one of its branches, which takes advantage of Standard Query Language to mine datasets. However, due to the little survey research emphasis on this very specific area, currently few literature review has been conducted to investigate, summarize or critique this field. In order to fill this existing gap, in this paper, an original critical review has been carried out, which is based on academic research papers since 1990. To author's best knowledge, it is the first time to conduct such a review in a critical way regarding SQL-based mining relational database. This review highlights a strong point and a point of improvement and organized in a reversely chronologic order. The review result shows that since 1990, there have been some research work done to mine relational database by SQL. Almost all of it was empirical and some of it was based on the extension of the standard SQL, however all of research or proposed systems did not support distributed relational databases.

Key Words: Data mining, mining relational database, SQL, review.

1. Introduction

Relational database plays a critical role in management information system a.k.a. MIS (A Management Information System (MIS) is an information system that is intended to be used by the higher management of an organization [1].) such as Health Information Systems, Sales Information System etc. Mining relational database, seen as relational database knowledge discovery is the automated or manual extraction of patterns which represent database knowledge stored in large relational databases. Mining Relational Database is more than just kind of data mining whose sources come from relational databases. It is due to the fact that it requires domain-specific knowledge including relational database theory for instance schemas, domains, attributes, tuples and relations [2] in relational databases. Additionally, it also requires business knowledge in particular application scenarios. At the moment, a number of Mining Relational Database techniques can be made use of extracting and analyzing relational database data.

As the name suggests, SQL-based mining relational database takes advantage of Structured Query Language (SQL) to mine relational databases. SQL is a standard query language interacting with relational database management system including create, delete, update as well as query operations. A majority of mainstream relational databases have its own SQL dialects e.g. PL/SQL implemented by Oracle Corporation, MSSQL customized by Microsoft Corporation. International Organization for Standardization (ISO) has defined SQL standards for industry over the past decades for example, SQL 2016 [3]. Thus, based on SQL, data

mining related techniques naturally can be applied in relational database management systems.

Despite some reviews on data mining techniques and applications, the authors rarely make investigation-oriented research efforts in SQL-based mining relational database. It may be because SQL-based mining relational database is a narrow and unusual research area. Therefore, researchers and scientists have not paid much attention to this field by far. In a word, in terms of SQL-based mining relational database, the current research work lacks reviews from a balanced perspective, or even simple reviews in the past twenty or thirty years approximately. Thus, this paper would review SQL-based mining relational database in a critical way since 1990.

The contribution of the paper is that to author's best knowledge, it is the first time to give a critical review on SQL-based mining relational database.

Paper Organization: the rest of this paper is structured as follows. Section 2 presents the related research work. Section 3 provides an overview of mining relational database. Section 4 reviews SQL-based mining relational database in a critical way. Section 5 a conclusion is drew.

2. Related Work

Date mining also known as knowledge discovery in database is defined as the process of finding valuable patterns from datasets by extraction or the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner [4]. It is likewise defined as non-trivial extraction of implicit, previously unknown and potentially useful information from data [4]. Mining database is one of the important branches in data mining and it involves many techniques. SQL-based mining database is a popular one, however there is no dedicated review regarding its research progress and status.

Shu-Hsien Liao, Pei-Hui Chu and Pei-Yuan Hsiaod [5] conducted a literature survey regarding Data Mining Techniques and its application from 2000 to 2011. The paper reviewed neural networks, dynamic prediction, knowledge-based systems, intelligent agent systems etc. and their applications. It concluded that the development of Data Mining Techniques is tending to be more specialized and the development of Data Mining Application is more problem-focused.

Sreekumar Pulakkazhy and R. V. S. Balan [6] reviewed data mining techniques and procedures specially applied in banking areas. It summarized that Data Mining Techniques are used by banks in a variety of application areas such as marketing, fraud detection as well as risk management, which can help banks to make better decisions. Thus, there are increasingly banks that are investing in data mining techniques.

Hussain Ahmad Madni, Zahid Anwar and Munam Ali Shah [7] focused on academic publication over the period of from 2007 to 2017 – a decade survey, aiming to review and categorize Data Mining Techniques in addition to its applications. The authors presented a simple and clear view of various models such as dynamic predication-based model, decision-tree based model etc. used in data mining.

Mrs. S. Revathi and Dr. K. Nandhini [8] carried out a review of Data Mining techniques applied in the log processing in the context of cloud computing i.e. Hadoop environment. An analysis has been made on the log data processing by different researchers. By comparing these log data processing methodologies, the authors summarized the strengths and weaknesses.

Ming-Syan Chen, Jiawei Han and Philip S. Yu [9] provided an overview of data mining techniques including classification and contrast. As claimed in their paper, from a database scientist's perspective, the authors investigated a number of data mining systems such as health care data mining system, image data mining system etc.

3. Mining Database

Mining database is one of the variances of data mining as the data source is database. The target of data mining must contain a volume of data is self-evident. Since a database is a collection of related data [2], data mining techniques are naturally applied in database, serving as a role in knowledge discovery. According to the R. Agrawal M. Carey *et al.*'s definition [10], database mining refers to the efficient construction and verification of models of patterns embedded in large databases, and is emerging as a major application area for databases.

There are a number of commercial or open-source database products such as IBM DB2, Oracle, Microsoft SQL Server, MySQL and PostgreSQL. All of them could be mined for research purposes. With the rise of information technology, database is more and more playing an indispensable role in modern information management. Additionally, database is an essential part of our life in modern society such as deposit, booking airline reservation, online shopping and so forth. Thus, mining database could have practical value and at same time, these applications offer abundant datasets for data mining studies [4], [11].

Mining database is a process of discovering knowledge in database.

Fig. 1. The mining database process.

Fig. 1 shows a general version of the mining database process.

4. SQL-Based Mining Relational Database

In this section, SQL-based mining relational database research work starting from 1990 is summarized critically i.e. highlighting one strong point and one weak point per paper every research paper in a reversely chronological way.

In 2012, Carlos Ordonez and Zhibo Chen, in the paper "Horizontal Aggregations in SQL to Prepare Data Sets for Data Mining Analysis [12]", have put forward powerful methods to automatically generate SQL code and return horizontal aggregations rather than traditionally one number every row. The strong point of this paper is the authors have proposed three fundamental methods rigorously proving the effectiveness of the horizontal aggregations. Due to the nature of horizontal aggregation techniques, there is one weakness that it maybe generate more columns, especially in terms of large relational database usage, which might impair relational database throughput, one significant indicator to measure relational database management system. Another improvement, in my view, might be that the authors could explore non-relational databases i.e. NoSQL (Not Only SQL) because this key-value storing method due to the high performance has increasingly usage scenarios in today's Internet-based world.

Carlos Ordonez and Carlos Garcia-Alvarado published a paper titled "A Data Mining System Based on SQL Queries and UDFs for Relational Databases [13]" in 2011. In order to overcome a performance bottleneck such as query processing, the authors have put forward an original system based on SQL. The strength of the paper is that even with large and high dimensional data sets, the system can make an analysis of relational database tables for building and storing statistical models. One improvement could be enabling more database connectivity (e.g. JDBC) except ODBC with innate weaknesses including poor extensibility, which can support more user application scenarios.

In the paper "A Framework for SQL-Based Mining of Large Graphs on Relational Databases [14]", Sriganesh Srihari *et al*. have proposed a nimble framework in terms of time and space complexity to mine graphs by

SQL-based data mining techniques in relational database. One strong point is that the efficiency of this approach has been proved by the authors' empirical and experimental evaluation. The framework, however, could be enhanced to support grid relational database technologies e.g. Oracle 10g, MySQL Cluster in support of large graph mining. Since the Internet and mobile applications require a great deal of parallel and distributed computing, these data mining technologies might be in demand in the future.

Alfredo Ferro *et al*. integrated data mining primitives with the popular relational database MySQL 5.1 in their paper "MySQL Data Mining: Extending MySQL to support data mining primitives (demo) [15]", extending MySQL to support novel Apriori-like SQL statements based on the decision tree learning algorithm C4.5. The strength of the paper is a web-based graphic interface has been developed and presented for the convenience of data mining researchers and practitioners. In addition, from my point of view, it was challenging for the authors to define some new SQL statements described by BNF (Backus Naur Form) according to the SQL standard i.e. SQL1992 because the novel database reserved words e.g. apriori, classify were used. On the other hand, the limitation of this research is obvious. There are a wide variety of relational database management system (RDBMS) either commercial or non-commercial, whereas the paper only supports MySQL, one of the open source relational database management system.

In the research paper titled "Data mining using Relational Database Management Systems [16]", Beibei Zou *et al*. carried out empirical studies on the integration of machine learning techniques and SQL-based mining relational database by Weka, a well-loved open source machine learning based data mining software package. One advantage of this paper is, in order to break through the limitation of the amount of main-memory data structures, the authors have proposed a unified interface interacting with relational databases supporting various data mining algorithms rather than optimizing just a single algorithm. The improvement of this paper could be offering an option to enable PreparedStatement or Statement rather than using PreparedStatement in most cases. It is because PreparedStatement may be time-consuming in compile time, which means it is not efficient all the time.

Xuequn Shang *et al*. in the paper "SQL Based Frequent Pattern Mining without Candidate Generation [17]", have proposed a novel frequent pattern growth method for mining long and short SQL-based frequent pattern mining. One strength of this paper is, compared with the traditional research, this research has achieved good performance at the same time examined and evaluated by the authors. One further investigation could be the original SQL-based data mining techniques could also be implemented in distributed relational database management system i.e. parallel relational database management system.

In 1997, two researchers Maciej Zakrzewicz and Tadeusz Morzy from Poznan University of Technology published a paper named "SQL-like language for database mining [18]". The two authors participated in a data mining research project and put forward an unexampled SQL-like language called MineSQL that is the extension of standard SQL. One shining point of this paper is the paper not only gave the syntax of the new SQL language MineSQL but also the usage examples. For instance, the rule query is coded as below:

MINE rule, support (rule) s., confidence (rule) c.

FOR product

TO day

Thus, it might be easier and faster for practitioners and researchers to learn how to use the new SQL language. Additionally, since the paper presents only one new statement, namely MINE combined with the usage of traditional SQL e.g. insert statement, where clause, group by clause to achieve its database mining goal, my view is it has a smoother learning curve for beginners. One weak point of the paper might be only laying emphasis on rule query language not all Data Definition Language (DDL) or Data Manipulation Language (DDL).

Back to 1996, Tomasz Imielin *et al*. conducted empirical studies on SQL-based database mining in the paper

"MSQL: A Query Language for Database Mining [19]". The authors have proposed an original version of rule-based query language which composes the Discovery Board system. One strength of this paper is that unlike other previously reviewed research papers, this paper defined basic mathematical notions in terms of MSQL such as descriptor, conjunctset, propositional rule and confidence, making the work more logical and meticulous. However, the authors have implemented only a subset of their MSQL in their system not the full set of MSQL. Thus, from a technical view, the system has not fully conformed to the International SQL standard, it may not support dynamic or embedded SQL. For that matter, the design of this new data mining language had not been completely justified and thus the contribution of the paper is qualified.

In the paper "Quest: A Project on Database Mining [10]", a group of International Business Machines Corporation (IBM) researchers developed a data mining system for large relational databases based on a business-related prototype in 1994. The strong point of this paper is that the proposed system supported a variety of knowledge discovery features such as classification rules, pattern matching, association rules, sequential patterns as well as pattern analysis. One improvement could be that all the data for building the prototype came from business data i.e. sales data, which possibly would produce a biased output. If the researchers had gathered the data from a wide range of systems including non-commercial, the system based on the prototype might have been more adapted.

Rakesh Agrawal, Tomasz Imielinski and Arun Swami from IBM Almaden Research Center have designed and implemented a novel algorithm generating all association rules in the relational database presented in the paper "Mining Association Rules between Sets of Items in Large Databases [20]". The strength of their research could be that in order to show the effectiveness of the proposed algorithm, the authors applied it in a large retail company, which had practical value. One weakness of this paper is that the work could be more novel because the research merely focused on the enhancement of database functionality. In addition, the work was based on the existing research project called Quest [10], which was previously developed and implemented well.

5. Concluding Remarks

This paper reviews SQL-based mining relational database research from an unbiased view since 1990 due to that few research review efforts made in this narrow field, compared with some previous reviews in data mining techniques and its applications. Yet all of these previous reviews have not focused on the SQL-based mining relational database. This paper is the first paper to conduct a chronological literature review in terms of SQL-based mining relational database, which makes an analysis of its strength and weakness. The implication and finding are that since 1990, there have been some research efforts made in SQL-based mining relational database, almost all of which were empirical and some of which extended the standard SQL as their data mining techniques. But all of research work or systems did not support distributed or parallel environments. Therefore, this critical review paper will facilitate further empirical or possibly theoretic research on SQL-based mining relational database.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Yuexiao Teng is the only author. He conducted the literature survey research and wrote this paper.

References

[1] Mishra, U. (2013). Introduction to management information system. *SSRN Journal*. Retrieved from

https://ssrn.com/abstract=2307474

[2] Elmasri, R., & Navathe, S. B. (1994). *Fundamentals of Database Systems*. Benjamin Cummings Publishing Company.

[3] ISO/IEC 9075-1:2016. ISO. Retrieved from https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/06/35/63555.html

[4] Hand, D. J., Mannila, H., & Smyth, P. (2001). *Principles of Data Mining*. Cambridge, Mass: MIT Press.

[5] Liao, S.-H., Chu, P.-H., & Hsiao, P.-Y. (Sep. 2012). Review: Data mining techniques and applications — A decade review from 2000 to 2011. *Expert Syst. Appl., 39(12)*, 11303–11311.

[6] Pulakkazhy, S., & Balan, R. V. S. Data mining in banking and its applications-A review. *Journal of Computer Science, 9(10)*, 1252-1259.

[7] Madni, H. A., Anwar, Z., & Shah, M. A. (Sep. 2017). Data mining techniques and applications — A decade review. *Proceedings of the 2017 23rd International Conference on Automation and Computing (ICAC)* (pp. 1–7).

[8] Revathi, S., & Nandhini, D. K. (Feb. 2018). Review of mining techniques used in the log data processing based on Hadoop and cloud computing environment. *International Journal of Advanced Research in Computer Science, 9(1)*.

[9] Chen, M.-S., Han, J., & Yu, P. S. Data mining: An overview from database perspective. *IEEE Transactions on Knowledge and Data Engineering, 8*, 866-883.

[10] Agrawal, R., Carey, M., Faloutsos, C., Ghosh, S., & Swami, A. (1994). Quest: A project on database mining. *ACM SIGMOD Record, 23(2)*, 514.

[11] Pareek, A., & Gupta, D. M. (2012). Review of data mining techniques in cloud computing database. *International Journal of Advanced Computer Research, 2(2)*, 5.

[12] Ordonez, C., & Chen, Z. (Apr. 2012). Horizontal aggregations in SQL to prepare data sets for data mining analysis. *IEEE Transactions on Knowledge and Data Engineering, 24(4)*, 678–691.

[13] Ordonez, C., & Garcia-Alvarado, C. (2011). A data mining system based on SQL queries and UDFs for relational databases. *Proceedings of the 20th ACM International Conference on Information and Knowledge Management - CIKM '11* (p. 2521), Glasgow, Scotland, UK.

[14] Srihari, S., Chandrashekar, S., & Parthasarathy, S. (2010). A framework for SQL-based mining of large graphs on relational databases. *Advances in Knowledge Discovery and Data Mining*, 160–167.

[15] Hutchison, D., *et al.* (2010). MySQL data mining: Extending MySQL to support data mining primitives (demo). *Knowledge-Based and Intelligent Information and Engineering Systems*, 438–444.

[16] Zou, B., Ma, X., Kemme, B., Newton, G., & Precup, D. (2006). Data mining using relational database management systems. *Lecture Notes in Computer Science*.

[17] Shang, X., Sattler, K. U., & Geist, I. (2004). SQL based frequent pattern mining without candidate generation. *Proceedings of the 2004 ACM Symposium on Applied Computing-SAC '04* (p. 618).

[18] Morzy, T., & Zakrzewicz, M. (1997). SQL-like language for database mining. *Proceedings of the First East-European Symposium on Advances in Databases and Information Systems (ADBIS)*.

[19] Imieliński T., & Virmani, A. (1999). MSQL: A query language for database mining. *Data Mining and Knowledge Discovery, 3*, 373–408.

[20] Agrawal, R., Imieliński, T., & Swami, A. (Jun. 1993). Mining association rules between sets of items in large databases. *SIGMOD Rec., 22(2)*, 207–216.

Yuexiao Teng received his computer science B.S. degree and computer application technology M.S. degree both from East China University of Science and Technology, Shanghai, China in 2009 and in 2012 respectively. He has several years working experience in information technology industry. Currently, his research interests include software engineering, distributed computing and data mining.