DATA MINING CONSULTING IMPROVE DATA QUALITY

Xingsen Li[1,2*], Yong Shi[2,3], Jun Li[1,2] , Peng Zhang[2,4]

*[*1]School of Management, Graduate University of the Chinese Academy of Sciences, Beijing, 100080*

Email: lixingsen@sohu.com

[2] Research Center on Fictitious Economy & Data Science, CAS, Beijing, 100080, China

[3]Graduate University of the Chinese Academy of Sciences, Beijing, 100039

[4]School of Information science and engineering, Graduate University of the Chinese Academy of Sciences, Beijing, 100080

ABSTRACT

Data are important for making decisions. However, the quality of the data affects the quality of decisions. Data mining as one of the most important sources of knowledge needs high quality data to mine, but there are not enough good quality data in many enterprises. By analyzing the reasons for low data quality systematically, a new method called data mining consulting for improving data quality has been established. It defines data quality in a wider sense from the view of data mining, finds data quality problems, and solves data quality problems by a series of methods. Its application shows that it has good practicality and can increase data quality considerable.

Keywords: Data quality, Data mining consulting, Data mining, Software engineering, Information system

1 INTRODUCTION

In recent years, data has become more and more important in the information age. Individuals, companies, and organizations usually make decisions based on data. Data accumulate so quickly that data mining technology has to be used for analyzing data. Today data mining as a knowledge resource has been widely accepted especially in large sized-enterprises, government, and financial departments (Han & Micheline, 2006; Shi, 2002). More and more leading-edge organizations are realizing that data mining provides them with the ability to reach their goals in customer relationship management, risk management, fraud and abuse detection, etc. Also, data mining is becoming a key technology to e-business (Noonan, 2000). Data mining could help enterprises establish a knowledge base during the development phase and aid in making the right decisions instead of making mistakes, gaining an early benefit from the informationalization process. Additionally, it could bring added value from data services and new revenue. However, following the rule of "garbage in, garbage out," data mining needs high quality data while there are often not enough good data in many enterprises for data mining to yield credible conclusions. From PriceWaterhouseCoopers' survey in New York in 2001, 75% of 599 companies had economic losses because of data quality problems (Pierce, 2003).

Previous research on improving data quality was from the view of information systems (IS) (Wang, 1993; Aebi & Perrochon, 1993; Wang, 1995; Missier, 2003; Dasu, 2003; Scannapieco, 2004) or from the view of data warehousing (Rahm, 2000). Their ranges were not large enough for the needs of data mining. Data cleaning and Extraction-Transformation-Loading tools (Hernandez, 1998; Lee, 1999; Galhardas, 2000; Galhardas, 2001; Raman, 2001; Guo & Zhou, 2002; Dasu , 2003) or tolerance algorithms (Zhu & Wu, 2004) have been used to mine low quality data. However, those methods only can improve the current data quality for mining. Low quality data are created every day and will make the data dirty again. Moreover, data cleaning conceals the source of dirty data, so not enough actions are taken to improve the system, therefore forming a vicious circle as shown in Figure 1.

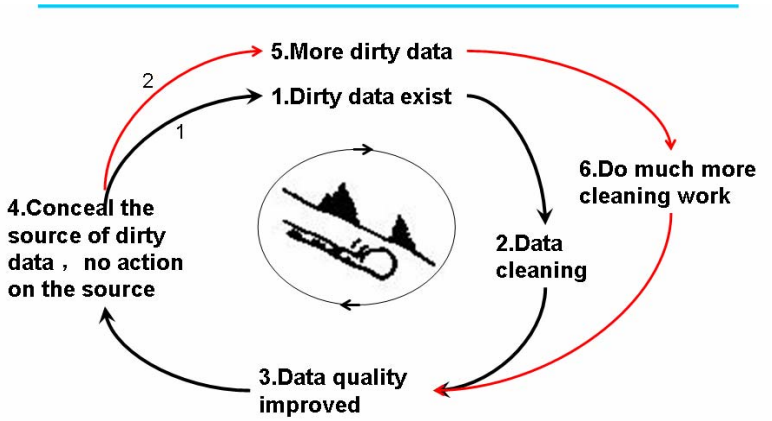

Figure 1. Vicious circle of data cleaning

In order to get credible conclusions, data cleaning and processing accounted for 80-90% of the workload of a data mining project (Johnson, 2003). That makes data mining so difficult that most small and medium businesses cannot afford to do it. Data quality problems have become an important factor in data mining applications (Dasu, etc., 2003). Dasu (2003) gave a good start to data cleaning from the earlier phase by expert systems.

We have done more research on data quality from the view of data mining. The purpose of this paper is to propose a systemic solution for improving data quality. The rest of our paper is organized as follows. In Section 2, we introduce the definition of data mining consulting methodology. Section 3 presents the details of data mining consulting. We give an example of applying our model to a real application leading to a satisfactory answer in Section 4. The paper is summarized in Section 5.

2 WHAT IS DATA MINING CONSULTING?

Data mining consulting was put forward by Extension Theory, which was established by Prof. Wen Cai in China in 1976. Extension Theory is a discipline that studies the extensibility of things, the laws and methods of exploitation, and the innovation needed to solve all kinds of contradiction problems in the real world with formalized models (Cai, 2005). Extension theory establishes matter-element, affair-element, and relation-element to describe matter, affairs, and relations. From the view of matter-element analysis in theory, matter can be divided into two parts: an imaginary part and a real part from the view point of material nature of matter. Or the division can be into soft parts and hard parts from the view point of systems, which is called the conjugate nature of matter-elements. The theory says that the real part is the base and the imaginary part is what we use (Cai, 1999).

According to Extension Theory, data mining is conjugated. The real part is the data mining techniques and software tools, and the imaginary part is the idea of data mining and the methodology. As usual, the imaginary part plays a very important role. Based on the integration of the real and imaginary parts, we put forward the following methodology for improving data quality called *data mining consulting method*:

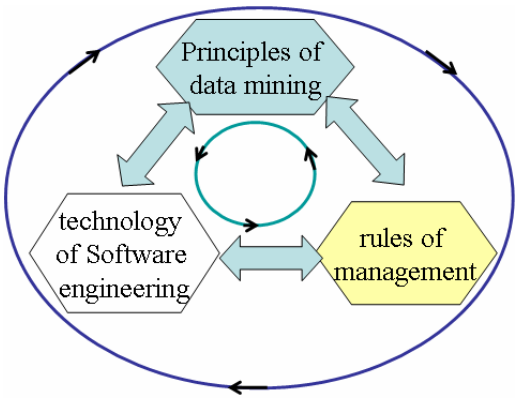

Figure 2. Structure of data mining consulting

Data mining consulting consists of three parts: the principles, the technology of software engineering, and the rules of management, as shown in Figure 2. These three parts interact with each other and form two circles: principles of data mining give a new view for data quality management and other related rules; then the new management rules make software engineering more effective and therefore the software becomes more suitable for data mining. This is a big circle outside. At the same time, a series of new management rules from the view of data mining can enhance the principles. Strong principles of data mining can then guide the software design and implementation. Good software design and implementation can decrease the workload of management. This forms the small co-adjustment circle seen in Figure 2. From principles of data mining, the conditions that data mining needs and its standard rules are listed, and then traced back to the use of the software. Actions are taken to prevent the creation of dirty data. Through the whole cycle, a series of management rules are used to reduce human mistakes. This series of principles, rules, and actions from requirements analysis process, data base design, software development to data integration, cleaning, and mining is called data mining consulting. Its aim is to improve data quality and to make the implementation of data mining projects efficient and easy to run.

3 HOW DOES DATA MINING CONSULTING IMPROVE DATA QUALITY?

3.1 Framework of data mining consulting

In order to improve the accuracy and integrity of the data, a data mining consulting solution framework is presented in Figure3.

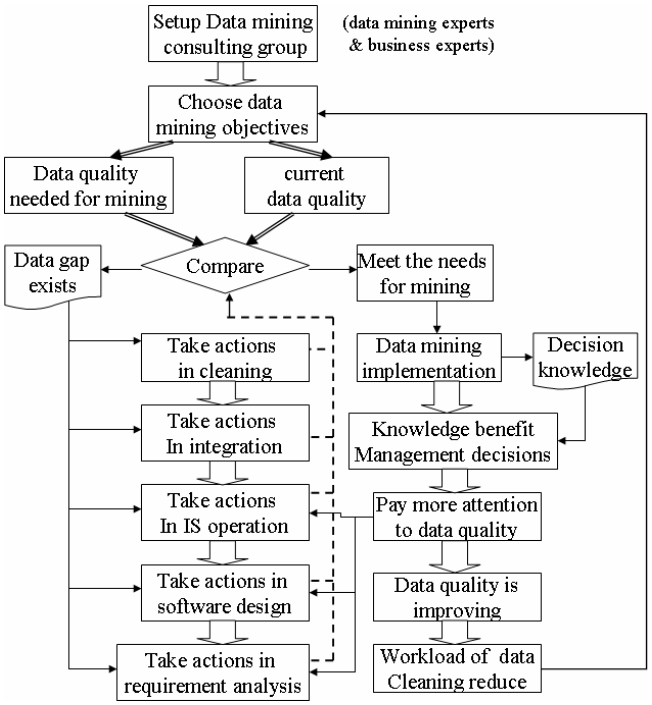

Figure 3. The framework of data mining consulting

In this framework, data quality needed in data mining is first listed (given in detail in 3.2). We collect the data set and identify the gap between the present data and objective data in the view of data mining, using data mining consulting actions including data mining testing, data quality analysis, and data structure adjustment, storage and integration, time remaining, etc. until the data meet standard quality requirements.

By recycling data mining experiments and taking improvement measures, the data gap will be decreased and high quality data will take the place of poor data. Once the conclusions of data mining begin to benefit a business's decision-making, senior management will pay more attention to data's accuracy and take effective measures that will boost information system development, such as increasing input, improving management, emphasizing data analysis, etc. With the above measures, we can augment the demands of data, integrate more data, deal with the relevant quality issues, and come to the next phase of data mining consulting and

implementation. This kind of spiral-recycling implementation will accelerate the transformation from unready mining data to ready mining data and also enhance the quality of corporation information systems (Li, 2006).

3.2 Data quality from the view of data mining principle

Aebi and Perrochon (1993) give a definition of data quality from the view of an information system. Data quality measures the amount of consistency, instance correctness, completeness, and minimality are achieved in a certain system. That is true in information systems (IS) but not really from the view of data mining.

Table 1. Data quality added from the view of data mining

Requirement	Explanation	Dirty data examples
Correctness	Data reflect its true reality	Age=120 or input birthday= "11/11/1911" when birthday is unknown
Completeness	Data sets contain all data mining needs	Lack lost customers' information while mining customer retaining
Consistency	Codes in different systems are consistent; no conflict while integrating	A customer's ID in CRM system is "1100", while in POS system is "021233"
Minimality	No repeat records after integration	A sales record became 3 records after integration
Reliability	Results of integrating stable regardless of who or when did	Attributes in product table changed between two integrating process resulting in confused sales information

If the information system quality does not meet the needs of data mining, improvements should be done while doing the data mining trial or action should be taken to first improve the data quality.

3.3 Reasons for dirty data

Rahm, et al. (2000) classified the major data quality problems that could be solved by data cleaning and data transformation. As can be seen, these problems are closely related and should thus be treated in a uniform way. Data transformations are needed to support changes in the structure, representation, or data content. These transformations become necessary in many situations, e.g., to deal with schema evolution, migration of a legacy system to a new information system, or when multiple data sources are to be integrated. Classification of data quality problems (Rahm, 2000) is shown in Figure 4:

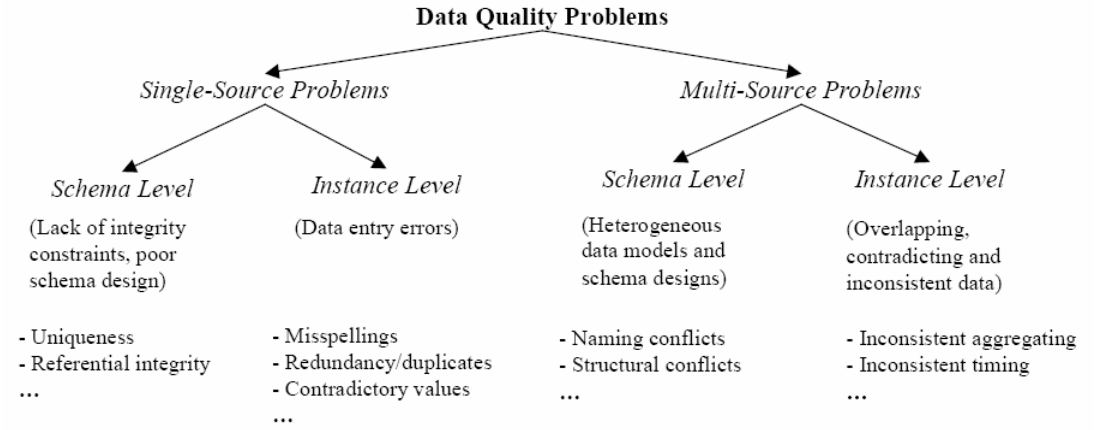

Figure 4. Classification of data quality problems (Rahm, 2000)

In this paper, data quality problems are distinguished between single-source and multi-source problems and between schema- and instance-related problems. Schema-level problems, of course, are also reflected in the

instances; they can be addressed at the schema level by an improved schema design (schema evolution), schema translation, and schema integration. Instance-level problems, on the other hand, refer to errors and inconsistencies in the actual data contents, which are not visible at the schema level. They are the primary focus of data cleaning. These data quality problems cover the ETL process (extraction, transformation, loading), basically used in data warehousing. To meet the needs for data mining, two more things must be considered: first, once the data mining objective is selected, will the IS providing all the data for it; and even if single-source and multi-source problems and between schema- and instance-related problems have been solved, will the data meet the needs then? To trace back all possible problems, Li (2007) gave a flowchart as shown in Figure 5.

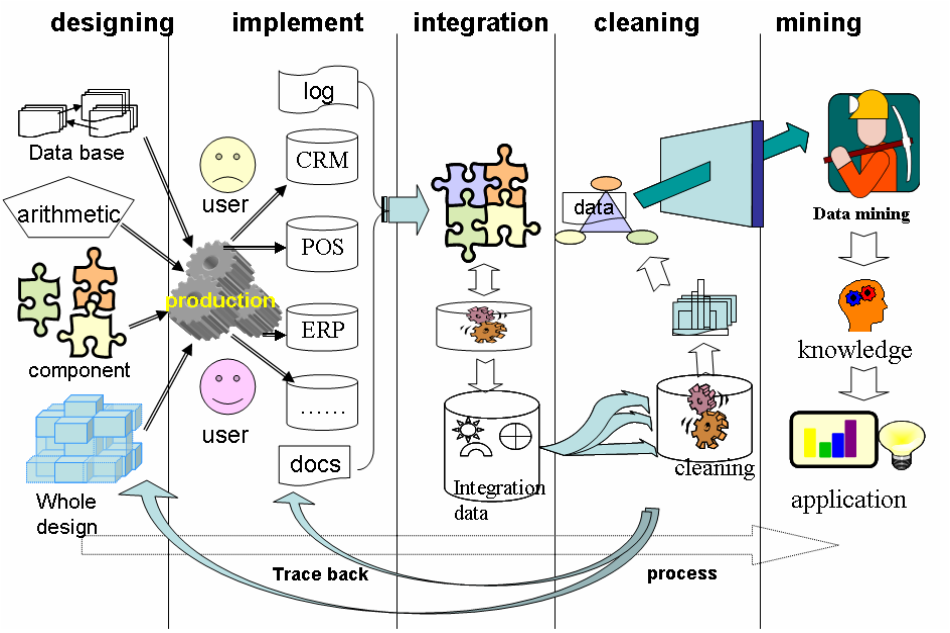

Figure 5. Flowchart for tracing all possible data quality problems

Starting from the mining process, we can trace back for more processes: data cleaning, data integration, software implement, and software designing. Such analysis finds additional problems as listed in Table 2.

Table 2. Data quality problems from the view of data mining

Process	**Problems**
Data mining	Lack of attributes for a specific mining objective Data set cannot support objective-oriented mining
Data cleaning	Filtering, Filling in real information that may be valuable for mining
Data integration	Null value created when connect tables Conversion of key attributes caused conflicts
Software implement	Mistakes during data inputting process System updating caused data inconsistency Default format misunderstand
Software designing	Necessary attributes missed or did not constrain input must Coding disaccord between ISs Constraint rules missed or wrong

Two data quality problems need to be solved systematically: when the data set cannot support objective-oriented mining and human errors during data inputting process. The earlier we discover these problems, the more easily we can solve them.

3.4 Actions for improving data quality by using an information system

If a data set cannot support objective-oriented mining, we can improve the Information Systems by analyzing

decision objectives with the current data set, identifying the data gap by a data map, and then taking action (Figure 6):

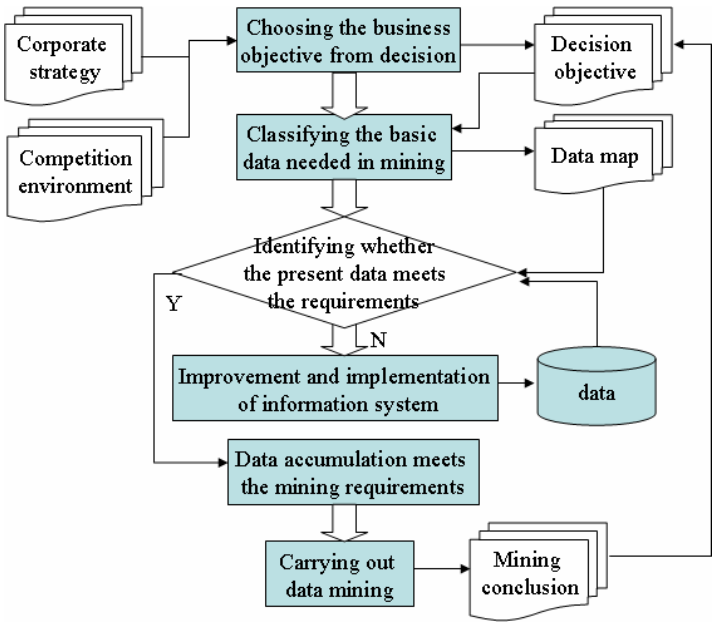

Figure 6. Improving the information system based on data mining consulting

The process includes five steps as follows:

Step 1. Defining the business objects that are supported by decision-making according to corporate strategy and competition environment.
Step 2. Based on the data that is required by business objectives, making the data map.
Step 3. Identifying whether the present information system can satisfy the requirements of data mining based on the business objective and the data map. If not, choosing and applying a complementary software system.
Step 4. Using out the complementary information system and accumulating data.
Step 5. Starting the data mining project to gain data mining results for decision-making when the mined data has accumulated to the required amount.

This process can help organizations to improve efficiently their information systems according to their business requirements. In other words, it could have an effect in a very short time.

3.5 Actions for improving data quality by management

Concerning human error during the data inputting process, it is very important to let the persons doing inputting know that the data will be used for mining and can produce valuable knowledge for the business and that it is not just being stored in the database with the possibility of deletion. The process for improving data quality divides the daily working process into four main phases:

(1) Planning: set data quality objectives and propose key action measures; make IS users understand the goals. The goal and action measures are communicated to workers and are related to rewards.
(2) Responsibility: define the responsibility of each position concerned with data quality and its Key Performance Indicators (KPIs).
(3) Results tracking: set up an inquiry system based on facts and data analysis, periodically inquire about data quality achievement, and find dirty data in earlier phases. Finally propose improved methods to ensure the implementation of data quality objectives through meetings or in other ways.
(4) Performance evaluation: evaluate each employee's achievement or contribution to data quality; then grade the evaluation results and give rewards.

The processes above can also make information system users (often in charge of data input), managers, software

designers, and database administrators share their experiences in daily work and form a responsible and cooperative culture among the team. Moreover, a series of well-written documents on how to improve data quality will greatly help.

4 CASE STUDIES

In web companies, the number of registered customers and ordinary visitors has increased rapidly since 1998. They are providing richer information and more products, and the accumulated data of each business unit has become more abundant as well. These data become useful if they are analyzed and mined in the future. However, OLAP statistically analyses are quite descriptive, and they lack illustrations of the rules and the business value behind the data. Therefore, it is very difficult to know the intrinsic relationships among the data, understand the real demand of customers, and even forecast future requirements. The reason for this could be that the information customers' supply is not sufficient or has low validity and reliability. In this case, many data mining enterprises are not likely to do data mining projects with this company.

Table 3. Data mining precision contrast after data mining consulting

	Precision			
	Before data mining consulting		After data mining consulting	
No.	Training set	Testing set	Training set	Testing set
1	64.13%	56.25%	89.82%	78.55%
2	59.84%	68.15%	90.51%	76.30%
3	74.01%	43.55%	91.97%	76.99%
4	56.33%	66.04%	88.91%	75.23%
5	65.71%	63.04%	87.73%	76.12%
6	66.31%	62.64%	87.64%	79.01%
7	67.65%	50.96%	90.40%	77.03%
8	61.34%	65.15%	88.44%	77.88%
9	65.23%	58.30%	90.08%	75.99%
10	59.74%	68.57%	90.08%	75.99%
Avg.	64.03%	60.27%	89.56%	76.91%

To discover the characteristics and the real requirements of clients and to develop the corresponding product as soon as possible along with more and more intense competition, one company cooperated with us to solve their problem. Our team analyzed the operation of customer data in detail with the help of Extension theory and rich data mining experience and proposed to use data mining consulting to improve data quality and carry out the project by phases. At present, a multi-objective linear programming method based software has been used in experimental data mining. Correspondingly, some primarily conclusions are deduced and have a good effect on the application.

5 CONCLUSIONS

This paper provides an overview of data quality problems and enlarges the concept of data quality from the view of data mining. By systemic analysis, we find that it is necessary to redefine data quality for the needs of data mining. Based on Extension Theory, data mining consulting provides a novel solution for improving data quality from earlier phases, such as requirement analysis, software design, software implementation, and data integration by data mining consulting, Some small and medium businesses that have poor data quality, or even no information systems, can be supplied with specific solutions to attempt data mining projects. Nonetheless, there are still some limitations on the implementation of data mining consulting. There is still much work to do. For instance, teams of data mining consultants need experts skilled both in business and data mining. The

process of cleaning data from the beginning is challenging and needs time. Cleaning data is effective for those companies that accumulate data in a short time, but it can do less for old data. However, the process has good practicality for data mining in low quality data enterprises and can help find data quality problems earlier, making all workers realize the value of data. We are sure that our process provides a good base for collecting high-quality data and can solve data quality problems from A to Z in the future.

6 ACKNOWLEDGMENTS

This research has been partially supported by a grant from National Natural Science Foundation of China (#70621001, #70531040, #70501030, #10601064, #70472074), National Natural Science Foundation of Beijing #9073020, 973 Project #2004CB720103, Ministry of Science and Technology, China, National Technology Support Program #2006BAF01A02, Ministry of Science and Technology, China, and BHP Billiton Co., Australia.

7 REFERENCES

Aebi, D. & Perrochon, L. (1993) Towards improving data quality. In: Sarda, N.L., ed. *Proceedings of the International Conference on Information Systems and Management of Data,* 273-281. Delhi.

Cai, W. (1999) Extension theory and its application .*Chinese Science Bulletin* 44(17), 1538-1548.

Cai, W., Yang, C.-Y. et al. (2005) A New Cross Discipline —Extenics. *Science Foundation In China*, 13(1), 55-61.

Dasu, T. & Johnson, T. (2003) *Exploratory Data Mining and Data Cleaning*. New York, NY: John Wiley & Sons, Inc.

Dasu, T., Vesonder, G., & Wright, J. (2003) Data quality through knowledge engineering. *Conference on Knowledge Discovery in Data archive, Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, 705-710. Washington, DC.

Galhardas, H., Florescu, D., Shasha, D., et al. (2000) AJAX: an extensible data cleaning tool. In: Chen, W.D., Naughton, J.F., Bernstein, P.A., eds. *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, 590. Texas.

Galhardas, H., Florescu, D., Shasha, D., et al. (2001) Declarative data cleaning: language, model and algorithms. In: Apers, P., Atzeni, P.,Ceri, S., et al, eds. *Proceedings of the 27th International Conference on Very Large Data Bases*, 371-380. Rome.

Guo, Z., Zhou, A. (2002) Research on Data Quality and Data Cleaning: a Survey. *Journal of Software* 13(11), 2076-2082.

Hernandez, M.A. & Stolfo, S.J. (1998) Real-World data is dirty: data cleansing and the merge/purge problem. *Data Mining and Knowledge Discovery* 2(1), 9-37.

Johnson, T. & Dasu, T. (2003) Data Quality and Data Cleaning - An Overview. *International Conference on Management of Data Archives, Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data*, pp: 681-681. San Diego, CA.

Lee, M.L., Ling, T.W., Lu, H.J., et al. (1999) Cleansing data for mining and warehousing. In: Bench-Capon, T., Soda, G., Tjoa, A.M., eds. *Database and Expert Systems Applications*, 751~760. Florence

Li, X.-S., Shi, Y., & Li, A.-H. (2006) Application Study on Enterprise Data Mining Solution Based on Extension Set, *Journal of Harbin Institute of Technology* 38(7), 1124-1128 (in Chinese).

Li, X.-S., Shi, Y., Lu, M., et al. (2007) Knowledge Acquisition under the Low-quality Data, *Contemporary*

Economy & Management 29 (3), 78-83 (in Chinese).

Missier, P., Lalk, G. V. Verykios, V., etc. (2003) Improving Data Quality in Practice: A Case Study in the Italian Public Administration. *Distributed and Parallel Databases* 13 (2), 135-160.

Noonan, J. (2000) Data Mining Strategies. *DM Review*, July. Retrieved from the WWW, July 16, 2007: http://www.dmreview.com/

Pierce, E. (2003) *A Progress Report from the MIT Information Quality Conference*. Retrieved from the WWW, July 18, 2007: Http://www.Iqconference.org

Rahm, E. & Do, H. (2000) Data cleaning: problems and current approaches. *IEEE Data Engineering Bulletin* 23 (4), 3-13.

Raman, V. & Hellerstein, J. (2001) Potter's wheel: an interactive data cleaning system. In: Apers, P., Atzeni, P., Ceri, S., et al, eds. *Proceedings of the 27th International Conference on Very Large Data Bases*, 381-390. Rome.

Scannapieco, M. Virgillito, A., Carlo Marchetti, C., etc. (2004) The DaQuinCIS architecture: a platform for exchanging and improving data quality in cooperative information systems. *Information Systems* 29, 551–582.

Shi, Y. & Zeleny, M. (Ed.) (2002) *Data mining*, IEBM Handbook of Information Technology in Business. England: International Thomson Publishing.

Wang, R.Y., Kon, H.B., Madnick, S.E. (1993) Data quality requirements analysis and modeling. In: *Proceedings of the 9th International Conference on Data Engineering*, 670-677. Vienna.

Wang, R.Y., Storey, V.C., Firth, C.P. (1995) A framework for analysis of data quality research. *IEEE Transactions on Knowledge and Data Engineering* 7(4), 623-640

Zhu, X. & Wu, X. (2004) Class noise vs. attribute noise: a quantitative study of their impacts. *Artificial Intelligence Review* 22 (3-4), 177-210.