Decision Support Systems 54 (2012) 76–86

Contents lists available at SciVerse ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

Improving financial data quality using ontologies

Jie Du, Lina Zhou *

Department of Information Systems, University of Maryland Baltimore County, Baltimore, MD 21250, United States

ARTICLE INFO

Article history:
Received 15 March 2011
Received in revised form 27 January 2012
Accepted 15 April 2012
Available online 12 June 2012

Keywords:
Data quality
Financial decision-making
Ontology
Ontology mapping
Portfolio management

ABSTRACT

The performance of financial decision-making directly concerns both businesses and individuals. Data quality is a key factor for decision performance. As the availability of online financial data increases, it also heightens the problem of data quality. In this paper, a taxonomy is created for data quality problems. More importantly, an ontology-based framework is proposed to improve the quality of online financial data. An empirical evaluation of the framework with the financial data of real-world firms provides preliminary evidence for the effectiveness of the framework. The framework is expected to support decision-making in finance and in other domains where data is spread across multiple sources with overlap but complementary in content.

1. Introduction

Today's widespread financial problems and the economic downturn highlight the importance of financial decision-making to individuals, businesses, and organizations. Intelligence gathering is the first stage of decision making [57], and data quality is a key factor for decision performance [29]. It is reported [63] that 20% of asset managers, investment bankers and hedge fund professionals spend between 25% and 50% of their time in validating data, which prevents them from focusing on tasks that contribute to the bottom line. According to a recent study of the costs and other consequences of dirty or inconsistent data in the secondary mortgage market in the U.S. [22], inaccurate data results in slow and expensive loan processing, weak underwriting, incorrect portfolio management, and other costs to lenders and mortgage investors. Given that financial data including financial statements, market data, and business news are being used increasingly by investors in stock market predictions [16,53], data quality has become an important and widespread issue in financial decision-making.

The problems with financial data come in a variety of forms. The main problems of financial data include ambiguity, inconsistency, missing values, inaccuracy, misrepresentation, incompletion, and so on [40]. For instance, missing values are not uncommon in Standard & Poor's Compustat North America dataset. Such problems can directly impact the performance of financial decision-making. Under this backdrop, this study aims to answer the following research question:

How should one address the quality problems of financial data so as to improve the performance of financial decision-making?

Both qualitative and quantitative approaches have been proposed to address various types of data quality problems [42]. For example, missing values can be replaced with global means or the most probable values [15]. Nevertheless, validating data quality is a challenging and time-consuming task [59]. This is especially true for financial data as an increasing amount of it becomes available on the Internet. The characteristics of the high frequency [25], high diversity and dependency of financial data render the conventional static approaches ineffective. Therefore, financial data calls for a synergic semantic alignment of various resources to improve financial data quality.

This study proposes a framework for addressing and identifying data quality problems following the design science research framework [28]. There are three types of artifacts created in our study. First, this study proposes an ontology-based framework to address the quality problems associated with online financial data. The framework is motivated by one unique feature of financial data, namely redundancy. Specifically, financial data about a firm is duplicated across multiple yet complementary online sources such as Yahoo!Finance, Google Finance, MSN Money Central, and Compustat. Yet the data are heterogeneous across different sources, even within the highly regulated financial domain. Our ontology is expected to address the above problem by enabling the mapping of data across different sources.

Second, this study creates a taxonomy and formalization of quality problems associated with financial data. The taxonomy, comprised of six types of quality problems such as missing values, is organized along two dimensions: the foundation and the abstraction level of ontology. Third, this study introduces a baseline method for evaluating the performance of financial decision-making that is based on fuzzy

* Corresponding author. Tel.: +1 410 4558628; fax: +1 410 4551073.
E-mail addresses: dujie1@umbc.edu (J. Du), zhoul@umbc.edu (L. Zhou).

0167-9236/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.dss.2012.04.016

theories. In view of the uncertainty involved in financial decision-making, the neuro-fuzzy approach is expected to be more robust when faced with data quality problems. The results of this study demonstrate that the proposed framework is effective for improving the quality of financial decision-making.

The remainder of this paper is organized as follows. Section 2 provides background information on financial decision-making, financial data quality, and ontology. Section 3 presents a taxonomy of problems associated with online financial data. Section 4 introduces the ontology-based framework for improving data quality in financial decision-making. The framework is evaluated in Section 5 and the results are presented and discussed in Section 6. Section 7 concludes the paper.

2. Background

2.1. Data quality to financial decision-making

Financial decision-making applications can be classified into five major categories, including stock forecasting, portfolio management, bankruptcy prediction, foreign exchange market, and fraud detection [69]. All of these applications require the collection of data whose quality is of great importance to their success. According to IBM Business Consulting's 2008 CFO Survey, data management remained a high priority for integrated finance organization. Issues around data consistency, data accuracy, and data integrity are primary concerns. Poor data quality can have substantially negative, social and economic impacts [55,65]. Since the financial data is highly time-variant, nonlinear, and noisy, data quality especially impacts financial decision-making [4].

2.2. Data quality dimensions

Traditionally, data quality is measured from multiple dimensions, including accuracy, consistency, completeness, and so on [51,54]. Based on the framework of Madnick et al. [42], data quality research can be characterized by two dimensions: topics and methods. Data quality covers a wide range of topics, which include data quality impact [21,39], database-related technical solutions for data quality [13,19], data quality in the context of computer science and information technology [38,50], and data quality in curation [9]. The quality of internal financial data can be managed by enhancing internal controls and reporting processes, which may be supported by application software such as Oracle Hyperion Financial Management (FDQM) [49]. Much research on financial decision-making has focused on improving the performance/outcome by developing and enhancing algorithms and decision models [5,62]. However, little research has focused on addressing the quality of financial data.

There are two types of methods that have been used to address data quality problems: the quantitative method and the qualitative method. The quantitative method is dominant. For example, Madnick and Zhu [41] improve data quality with context interchange technology. Thatcher and Pingry [60] present an econometric model to formalize the complex relationships among IT investments, product quality, and economic performance. Ballou et al. [3] use a mathematical model to analyze how data quality dimensions change within an information manufacturing system. Other studies deal with data quality problems using qualitative methods. For example, Davidson et al. [14] explore how the information maps can be used to improve data quality using a longitudinal case study. Kerr [33] adopts an ethnography method to study the data quality problems in the health sector. This study aims to improve the financial data quality by combining both quantitative and qualitative methods.

2.3. Ontology application in finance domain

Ontology is defined as an explicit specification of conceptualization [26]. Conceptualization refers to an abstract model of a particular domain of knowledge. Explicit specification means the concepts, their attributes and the relationship between concepts. Classes and instances are common components of ontologies [26]. Classes, also known as concepts, are used to model the domain structure. Instances belong to classes and are used to model the "ground level" objects. Ontology research centers on two issues: ontology building and ontology mapping [17,18]. The building process could be manual, semi-automated, or fully automated. As more and more ontologies are being generated, how to reuse these existing ontologies becomes essential. Ontology mapping expands and combines existing ontologies in support of communication between existing and new domains. An evaluation of existing ontology mapping techniques is given by Kaza and Chen [32].

Based on a shared and common understanding of a specific domain, ontology plays a key role in improving information consistency, reusability, systems interoperability, and knowledge sharing. Additionally, ontology has a huge potential to improve information organization, management, and understanding [20]. Some financial applications have benefited from ontology. For instance, an ontology is proposed to facilitate the communication among agents in a multi-agent financial investment system [58,70]; and ontologies are used to facilitate the predictions of firms which will have fraudulent financial statements [31], and to support investigators in the detection of fraudulent financial sites [36,71]. Nevertheless, applying ontology to address the problems of financial data has not been explored.

2.4. Schematic and data heterogeneity

Schematic and data heterogeneity makes communication among heterogeneous sources difficult and may cause many issues about data quality [34]. Database schema matching has received a long standing research attention. According to Kim and Seo [34], schema conflicts result from the use of different schema definitions from different sources. In other words, it is caused by the use of different concepts for semantically equivalent information. Data conflicts, on the other hand, are due to inconsistent data in the absence of schema conflicts [34]. There are many types of inconsistent data. Different representations might be used for the same data while the same representation might be used for different data. Computing errors or incorrect entry is another cause of data quality problems. Both types of heterogeneity are significant to online financial data. For instance, one data source represents a company's revenue as '*sales*', while another represents it as '*revenue*'. The unit for dollar amount is a thousand in Yahoo!Finance and a million in Compustat.

3. An ontology-anchored classification schema for data quality problems

The financial market is characterized by noisy, nonlinear data [66]. The noise of financial data includes dynamic noise, which disturbs the information obtained, and observation noise, which negatively impacts the accuracy of measurement. From the information system's perspective, data quality problems can be treated as representation deficiencies, which are defined in terms of the difference between the view of the real-world system as inferred from the information system, and the view that is obtained by directly observing the real-world system [64].

Data quality is a multidimensional concept that concerns both objective aspects that are intrinsic to the data (e.g., completeness) and contextual aspects that vary across tasks and users (e.g., ambiguity). It is important to address both aspects of the data quality for improved support of decision-making [56,67]. Therefore, this study