International Journal of Accounting Information Systems 21 (2016) 32–46

Contents lists available at ScienceDirect

International Journal of Accounting Information Systems

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

Integrating XBRL data with textual information in Chinese: A semantic web approach

Chi-Chun Chou [a], C. Janie Chang [b,*], Jacob Peng [c]

[a] College of Business, California State University, Monterey Bay, 100 Campus Center, Seaside, CA 93955, United States
[b] The Charles W. Lamden School of Accountancy, San Diego State University, 5500 Campanile Drive, San Diego, CA 92182-8221, United States
[c] Department of Accounting and Taxation, School of Business, Robert Morris University, 6001 University Boulevard, Moon Township, PA 15108, United States

ARTICLE INFO

Article history:
Received 25 November 2014
Received in revised form 17 April 2016
Accepted 18 April 2016
Available online 11 May 2016

Keywords:
Text analytics
Business information retrieval
Integration of financial and non-financial information
XBRL

Due to formatting differences, the difficulties of processing the textual disclosures and integrating them with quantitative financial data are well documented in the literature. Using a design science methodology, this paper describes a method that automatically extracts relevant textual data from annual reports published in Chinese. These extracted words are then mapped to a knowledge framework we proposed. This paper shows that it is technologically feasible to reorganize the MD&A contents into any given knowledge structure to improve the search capability, readability, and cohesiveness of the MD&A contents. Finally, we demonstrate a prototype system that uses semantic web technology to achieve information integration that presents XBRL formatted accounting data with relevant textual disclosures together to assist user decision making.

Published by Elsevier Inc.

1. Introduction

In this study, we propose and analyze a methodology to integrate textual disclosures with quantitative financial information using automatic text analysis and a semantic web. Nonfinancial text-based information has been documented to be value relevant (Antweiler and Frank, 2004; Tetlock, 2007; Previts et al., 1994; Abrahamson and Amir, 1996; Li, 2010; Brown and Tucker, 2011), which has led to the Securities and Exchange Commission's (SEC) elicitation of publicly traded companies to prepare more meaningful management discussion and analysis (MD&A) disclosures. The recent requirement of tagging footnotes in Extensible Business Reporting Language (XBRL) documents provides further evidence that readers of financial reports demand more meaningful and easily accessible textual information. As suggested by SEC Commissioner Gallagher (2013), "disclosure reform" is a prerequisite for capital markets to function effectively and efficiently.[1] The emphasis of such reform is not to increase the amount of required disclosures. Rather, these disclosures should provide investors with a means to discern the most critical information. The process of capturing, storing, and reusing various forms of unstructured knowledge, such as textual disclosures, often involves transformation of such knowledge into semi-structured or structured documents (Huang and Kuo, 2003; Vasarhelyi et al., 2012).

Despite the call for companies' preparation of more meaningful disclosures, the increased amount of both quantitative and textual information creates an information overload (Plumlee, 2003; Sun, 2010) and causes users to ignore some textual and qualitative information because the information is not easy to process cognitively (Engelberg, 2008). This problem of ignoring or underutilizing textual information in decision-making is also seen in XBRL-enabled financial reports. The SEC has required U.S.

* Corresponding author.
E-mail addresses: cchou@csumb.edu (C.-C. Chou), jchang@mail.sdsu.edu (C.J. Chang), peng@rmu.edu (J. Peng).

[1] Please refer to *Remarks at the 2nd Annual Institute for Corporate Counsel.* Washington, DC: Security and Exchange Commission at https://www.highbeam.com/doc/1G1-352019289.html (accessed on January 19, 2016).

http://dx.doi.org/10.1016/j.accinf.2016.04.002
1467-0895/Published by Elsevier Inc.

public companies to file their 10-K and 10-Q reports in XBRL format since 2009. This recent development in the U.S. GAAP Taxonomy (UGT) incorporates tags to be used for both financial data and nonfinancial information when submitting XBRL-formatted financial reports. Currently, there are at least 10,000 tags (out of 18,500) in the UGT that are designed to be used for narrative information.[2] Prior research indicates most public companies use more than 1100 tags for footnote disclosure when submitting their XBRL-based financial reports to the SEC under the "Level 4" detailed-tagging requirement (Blankespoor, 2012). Despite the in-depth coverage of the attempts of XBRL tags to incorporate textual information, the current UGT does not cover any information from MD&A.[3] It also does not provide a solution to combine information from different sources or to present this information in different formats. Theoretically, the textual information, such as the MD&A disclosures, has attributes in common with what is presented in the quantitative sections of financial statements. These shared attributes can be extremely valuable because they complement each other in aiding users to cross-reference relevant data from different sources. Unfortunately, these shared attributes cannot be linked because the textual information (mainly MD&A disclosures) is found in either PDF format or within the block-tagged XBRL instance, while the quantitative financial information is available in XBRL format. The main purpose of this paper is to describe a method to integrate the XBRL financial data and textual information from different sources. In addition, we demonstrate that this integrated information can be presented using a user-friendly interface to aid users' decision making.

Our method demonstrates that a knowledge framework can be constructed and used to organize textual disclosures. The extracted words from footnotes and MD&A disclosures are mapped with quantitative, XBRL-formatted financial data automatically using the algorithm we developed. Additionally, we use semantic tagging based on a Simple Knowledge Organization System (SKOS) to integrate unstructured MD&A information with XBRL-instance documents to provide users with more complete information when using financial reports in making decisions. Using annual reports of publicly traded companies in Taiwan,[4] we demonstrate an integrated system that semantically links textual information with financial data in XBRL format to provide users with integrated information.

This paper contributes to the information integration literature in methodology. This study designs and demonstrates a proof-of-concept prototype system that integrates previously scattered financial reporting line items, footnotes, and textual disclosures in an annual report. In addition, this study contributes to the enhancement of emerging text analytics literature by applying automatic text analysis in the Chinese language. Unlike English and other Western languages, the Chinese language does not delimit words by space (Peng et al., 2004), which makes the word segmentation and part-of-speech tagging challenging. Therefore, text analytics should differ greatly between regions that primarily use Western languages (United States, England, France, etc.) and regions that primarily use Chinese (China, Taiwan, Hong Kong, etc.). Because the Greater China Area plays an increasingly significant economic role in the global market today, it is essential and time-relevant to study how we can use current technologies, such as text analytics, to retrieve and integrate business information in Chinese.

This paper has important practical implications. In addition to the algorithm that automatically extracts Chinese words and maps with knowledge concepts in the extendable knowledge framework, the methodology we developed through this research proves it is technologically feasible to reorganize the MD&A contents into any given knowledge structure to improve the search capability, readability, and cohesiveness of the MD&A contents.

Our research follows the design science research (DSR) methodology (Gregor and Hevner, 2013; Hevner et al., 2004; Sedbrook and Newmark, 2008) to first identify a challenge in integrating accounting information. However, the approach described in this paper is not without limitations and as such provides future research opportunities. First, this paper is descriptive in nature that its goal is to demonstrate "how things ought to be" (Geerts, 2011). We acknowledge that only a limited number of sample annual reports are used to test our prototype system. The second limitation relates to the technical barrier that as the number of extracted words from annual reports increases the system performance decreases dramatically. Finally, although the knowledge framework proposed in this paper can be used to integrate financial and textual information, it may not be generalizable when a different set of annual reports is used.

This paper is structured into the following sections. After the introduction is the second section that provides a summary of prior research on the role of textual information and its impact on information integration. This is followed by a review of the literature in text analytics and semantic web and a description of our methodology to solve information retrieval and integration issues. Next, we present the two main activities of DSR: building and evaluating the proposed system design. Specifically, we illustrate a system that combines Chinese text analytics and semantic web technologies to extract and integrate quantitative and textual information. Finally, in the last section, we conclude this study and discuss future research directions.

[2] We use 2015 US GAAP Taxonomy to count the numbers of tags. Please refer to http://www.fasb.org/cs/ContentServer?c=Page&pagename=FASB%2FPage%2FSectionPage&cid=1176164649716 (accessed on April 14, 2016).

[3] Please refer to https://www.sec.gov/rules/final/2009/33-9002.pdf (accessed on April 14, 2016).

[4] We selected companies from Taiwan for this study because of the comparability of MD&A given the rules required by Taiwan's Financial Supervisory Commission (the SEC equivalence in Taiwan). In Taiwan, filers need to follow a mandatory reporting rule as the general guidance in preparing annual reports, including the MD&A section. The rule also recommends that preparers use a boilerplate with predetermined subdivisions. Based on the boilerplate, most filers are directed to present a standardized knowledge structure in their MD&A. This particular institutional setting provides an opportunity for this research to demonstrate the possibility of reorganizing the MD&A section using text analytics.