

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@csu.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).

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.

2. Background

2.1. Textual information and text analytics

One of the goals of the narrative disclosure mandated by the SEC is to enhance the overall quality of financial disclosures by adding more meaningful and descriptive information to financial reports. Normatively, these disclosures should be analyzed and used to provide qualitative information about the company's performance (SEC, 2003, 2010). One such commonly referred-to disclosure is the MD&A, which allows investors to understand the company through the eyes of management. Such disclosures inform investors complementarily about the firm's future performance (Davis and Tama-Sweet, 2012; Sun, 2010; Bryan, 1997; Vincent, 1999), economic conditions (Li et al., 2013), and exhibit higher timeliness and predictability of the firm's value, which makes it a leading indicator of the firm's future performance (Liedtka, 1999; Maines et al., 2002).

Although the use of textual disclosures such as the MD&A is widely documented (Sutton et al., 2012), there are concerns about using such information for decision-making. Professional standards, such as SAS No. 99, suggest that auditors evaluate MD&A disclosures to determine if there are "overly optimistic annual report messages." Such "overly optimistic" disclosures should be considered a risk factor for potential fraud. Another issue surrounding textual disclosures is that both the MD&A content and structure are not optimized, meaning the texts used by firms can be complicated and excessive, which contributes to inefficient cognitive processing of the information. Engelberg (2008) concludes that text-based, qualitative information (i.e., "soft information") is valuable and relevant, but it could result in market anomalies, such as post-earnings announcement drift (Hirst et al., 2004; Maines and McDaniel, 2000; Plumlee, 2003; Engelberg, 2008), because of the high information processing costs. In contrast, quantitative information is "hard information," which does not require much cognitive efforts to process, so it is more easily incorporated into users' decision-making process.

The nature of underutilized textual disclosures has triggered research on how to retrieve textual information from documents automatically and with reasonable accuracy (Sutton et al., 2012; Brown and Tucker, 2011; Li, 2010; Fan et al., 2006). To achieve satisfactory results, prior studies focus on automatic information retrieval (Garnsey, 2006; Shirata et al., 2011; Visa et al., 2000), and subsequent information classification and knowledge management (Sutton et al., 2012; Boritz et al., 2013).

Brown and Tucker (2011) suggest that text analytics technology is one approach⁵ to analyze textual information. This technology, such as automatic content analysis, allows system developers to achieve the goal of retrieving and quantifying textual information for further processing. For example, Shirata et al. (2011) argue that simply relying on word frequency to extract information from textual data sources is not enough to associate textual disclosures with bankruptcy predictions. They use natural language processing to discriminate textual disclosure content made by bankruptcy firms with that of non-bankruptcy firms through morphological analysis and the conditional probability algorithm. Using the automatic content analysis technique, Boritz et al. (2013) analyze information technology weakness (ITWs) disclosures found in SOX Section 404 reports. After extracting textual data on ITWs from original disclosures, Boritz et al. (2013) apply the bottom-up approach to categorize terms and keywords used in ITW disclosures. The results are used to create a term dictionary to update automatic searches.

Humpherys et al. (2011) develop a classification algorithm to identify fraudulent disclosures by using a text analytics tool and machine-learning techniques to analyze MD&A disclosures. The goal of text analytics is to discover and extract useful information from unstructured textual sources to support human decision-making. Although automatic text analytics does not require data organized in a quantitative format, these text data must be manipulated before they can be analyzed (Shirata et al., 2011; Boritz et al., 2013). The textual data first have to be prepared so that the text analytics software tool can read the data. Followed by retrieving text from the document depository, pre-defined decision rules are applied to extract useful textual information for decision-making. Essentially, text analytics is a process of editing, organizing, and analyzing a huge amount of textual documents in different formats (Sullivan, 2001). This process can be used to discover the hierarchical relations among key concepts, such as who, what, when, and where information is typically found in textual documents. Such techniques are widely applied in areas such as text extraction/retrieval, natural language processing, and computational linguistics (Cimiano, 2006).

Ultimately, text analytics is a technology used to transform unstructured, textual information into a more structured format so that it can be processed by either human or software tools. Achieving a satisfactory accuracy rate is essential for subsequent analysis. In the data extraction and categorization step, this study follows the two-step approach developed in prior research (Fisher et al., 2010; Boritz et al., 2013): (1) use software tools to extract textual data from company disclosures, and (2) categorize extracted data and prepare the data for further analysis. The complete methodology, including our approach to integrate textual data with XBRL quantitative data, is described in section three.

2.2. Issues in information integration

As a response to the keen interest of market participants for more transparent and reliable information, many new rules have been adopted worldwide to promote more accessible financial data. An example of such an initiative is the XBRL mandate in the U.S. The SEC (2009) claims that the interactive data initiative promotes efficient data integration and automates regulatory filings and business information processing. The role of the XBRL mandate on financial reporting and its effect on the SEC's filing program are analyzed in Debreceeny et al. (2005). Some empirical findings in the literature suggest that XBRL facilitates data

⁵ Brown and Tucker (2011) classify different ways to study the content of textual disclosures into three different approaches: (1) hand-coded content analysis, (2) survey rankings, and (3) automated text analysis. Each approach allows researchers to associate the content of textual disclosures with other financial information.

integration (Hodge et al., 2004). However, results of other empirical investigations on whether XBRL is able to reduce information asymmetry are mixed. Kim et al. (2012) conclude that XBRL decreases information risk and information asymmetry in both general and uncertain information environments. Yoon et al. (2011) find evidence that XBRL reduces information asymmetry in the Korean stock market. Blankespoor et al. (2014) argue that XBRL actually creates an adverse selection problem in the capital market. Specifically, their analysis indicates that more sophisticated investors in the market are more capable of using XBRL to integrate and analyze publicly available financial information than their counterparts. Overall, XBRL in its current form emphasizes standardizing quantitative information to minimize information processing costs; and empirical studies suggest that XBRL has permanently changed how financial information is prepared, filed, disseminated, and used by stakeholders.

To fully utilize the power of integrated information, decision aids are usually designed and used to help with a variety of decision-making tasks. Hodge et al. (2004) find that individuals who use XBRL-enhanced search engines are more likely to acquire and to incorporate the textual information in making investment decisions. Ghani et al. (2009) report that, when performing investment decision tasks, XBRL would be more useful tool to rely on compared to PDF and HTML. Nelson and Taylor (2007) suggest that information may have a greater effect on users' judgments if users are able to use technologies to perform the analysis necessary to transform textual information in the footnotes as if it had been recognized on the statement.

Overall, prior research findings suggest that: (1) relevant information (in quantitative XBRL format or in qualitative textual format) can be extracted from different sources; (2) the accessibility or usability of information can be limited due to the unstructured formats used to present and disclose information; and (3) the decision aids designed to assist users in acquiring and incorporating information can improve relevant decision-making tasks. This study builds on prior research surrounding XBRL to propose a methodology that integrates both quantitative information (in XBRL format) and qualitative disclosures (in textual format) for decision-making.

2.3. Semantic web

Semantic networking has been used by cognitive psychologists to describe the human memory structure. It uses a graphical notation to express the structure of human knowledge and serves as an expression of natural language. Semantic networking can also be applied on the Internet. As a collaborative project led by the Worldwide Web Consortium (W3C), the semantic web promotes the inclusion of more semantic content and creates a cross-platform framework that allows sharing and reusing data from different applications. The goal of the W3C is to develop a web environment that is different from current text-based unstructured web pages. During the web's early development stage, researchers placed more emphasis on resource discovery by developing a solution to disseminate financial information accurately and quickly in a networked environment (Debreceeny and Gray, 2001). As XBRL has become a standard to facilitate the exchange of financial information, the challenge now is to place XBRL in a broader semantic framework that integrates other data (e.g., the MD&A, sustainability reports, Internet forums) in the business information supply chain (Alles and Debreceeny, 2012; Vasarhelyi et al., 2012).

The Simple Knowledge Organization System (SKOS) data model provides a standard, low-cost migration path for porting existing knowledge organization systems (i.e., unstructured web pages) to the content-rich semantic web. It is a common data model, based on a Resource Description Framework (RDF), for sharing and linking knowledge organization systems through the web (Debreceeny and Gray, 2001). Using the SKOS model, conceptual resources (concepts) are identified with Uniform Resource Identifiers (URIs). These concepts are labeled with strings in one or more natural languages and are semantically related to each other in informal hierarchies and association networks.⁶

Many knowledge-organization systems, taxonomies, classification schemes, and subject-heading systems share a similar structure and are used in similar applications. SKOS captures much of this similarity and makes it explicit, which enables data and technology sharing across diverse applications. Therefore, in this study, we use SKOS to capture the necessary information integration between conceptualized textual narratives and formatted XBRL financial elements.

3. Method: design and development

We use an approach similar to Boritz et al. (2013) and the design science research (DSR) schema suggested by Gregor and Hevner (2013) to illustrate our research design. Gregor and Hevner (2013) provide a taxonomy to classify potential DSR contributions to the making of prescriptive knowledge. Their DSR framework posits that DSR contributes to new knowledge when a new solution is proposed for a known problem. It is critical, however, for this type of research to clearly present and communicate the new design. The contribution of our research design is the method that we used to associate tagged XBRL instances with relevant knowledge found in textual disclosures. Traditional textual data exists in scattered paragraphs in a number of different disclosures that are very difficult to integrate systematically. O'Riain et al. (2012) suggest that the key to successfully achieving data integration is semantic representation and semantic information access. That is, the two main tasks to integrate textual and numeric data are (1) locating and categorizing textual data from different sources, and (2) creating linkage between the extracted textual data concepts and the structured quantitative data. The former is achieved through text analytics, while the latter refers to developing a mechanism that links textual and quantitative datasets in a Resource Description Framework (RDF). Accordingly, in this study, we present the process of developing our prototype system in the following steps: (1) define a knowledge

⁶ <http://www.w3.org/TR/2009/NOTE-skos-primer-20090818/> (accessed on January 19, 2016).

framework used to categorize textual data; (2) develop a systematic approach to automate text extraction; (3) convert XBRL taxonomy into a semantic web-based vocabulary dataset; (4) create the textual information vocabulary dataset; and (5) establish a semantic link to connect financial and textual vocabulary datasets.

The advantages of applying this approach to integrate quantitative data with textual information are as follows: (1) Automation: users only need to exert minimum effort to locate financial and relevant qualitative data because the process is automatic; (2) Task independence: the same approach can be applied to analyzing textual information other than MD&A, such as footnote disclosures and audit opinions; (3) Cross-language applicability: although the demonstrated system is in traditional Chinese, this approach can be easily modified to apply to other languages; and (4) Verifiability: we develop an objective evaluation method to verify the accuracy of the information integration, available both in the text mapping process and the presentation of results.

3.1. Step 1: define a knowledge framework used to categorize textual data

Step 1 in our model development is to construct a knowledge structure to organize textual disclosures. The knowledge structure can be constructed either through the integration of existing hierarchy (top-down approach) or clustering terms present in a current set of documents to form a domain-specific knowledge base (bottom-up approach). For example, Sutton et al. (2012) apply an existing Enhanced Business Reporting Consortium (EBRC) framework in an effort to structure the knowledge contained in the MD&A disclosures. Garnsey (2006) uses the vocabulary present in official accounting pronouncements as the starting point to group terms extracted from textual disclosures. Alternatively, Boritz et al. (2013) extract terms from SOX Section 404 disclosures and use the bottom-up approach to construct the search dictionary for use in automatic text retrieval in later analyses. Both the top-down and bottom-up approach are designed to allow researchers to efficiently and effectively extract textual data and apply pre-defined rules to categorize previously unstructured, textual data for further analyses. Not only can knowledge construction be used to map information extracted from textual sources, it can also be used to provide feedback to the extraction process to enhance the relevance of the information that is being extracted from sources. Since a gold standard for building a taxonomy of annual report information does not currently exist, we manually create the knowledge structure in this study. Gómez-Pérez et al. (2004) suggest that domain knowledge is needed to conceptualize the collection of terms without a hierarchical framework. Based on SWOT analysis, balanced scorecard (BSC), and PwC's Business Value Reporting Framework (Eccles et al., 2001), we create a five-level knowledge hierarchy that users can drill down to detailed concepts based on term associations. This hierarchy is used later to categorize terms and concepts extracted from textual disclosures made by our sample firms.

3.2. Step 2: develop a systematic approach to automate text extraction

The main objective in Step 2 of our method is to form the keyword base for terms extracted from textual disclosures. We use text analytics technology to segment words and to tag parts of speech (POS) in sentences and terms. In text analytics literature, a "word" is often defined as the smallest element that can be used to create semantic or pragmatic content in natural language processing. Similar to Li (2010) and Brown and Tucker (2011), we apply automatic text analysis in this research. Our automatic text analysis process faces additional challenges due to the fundamental differences between the Chinese and English languages.

A Chinese word is composed of at least two Chinese characters, or a bigram. This creates a problem in automatic word segmentation and tagging (Sproat and Emerson, 2003). One way to segment a Chinese word is to do so according to the number of characters from a given sequence of text. An "n" gram sequence can thus be expressed in an "n-gram" Chinese word. For example, a bigram can be "dong shi" (a two-character word, which means "a member of a board of directors"), and "tou zi zhe" (a three-character word, which means "investors") is an example of a trigram (3-gram sequence). To establish the keyword base, we produce multiple n-gram words using the textual data from annual reports. However, not all n-gram words automatically segmented from textual data are accurate. For example, a sentence such as "Risk assessment is one of the elements in evaluating internal controls" in Chinese can produce multiple bigram words with different meanings. This is due to the fact that Chinese lacks morphological inflections that provide cues for word boundaries (Forst and Fang, 2009), and only some segmented bigram words are meaningful and accurate, such as "risk," "evaluation," and "is one of the." To obtain accurate n-gram words in this process, and to improve the results, we use mutual information and term frequency to measure morphological intensity (Yang et al., 2000):

- First, we calculate the total number of characters from the back-end lexicon and use N_c to represent this number.
- Assuming that a special term that we are looking for is w (for example, "corporate governance", i.e., "gong si zhi li"), and it is composed of $c_1c_2...c_n$ Chinese characters (for example, $w = c_1c_2c_3c_4$; $c_1 = \text{gong}$, $c_2 = \text{si}$, $c_3 = \text{zhi}$, and $c_4 = \text{li}$) with $f(c_1), f(c_2) \dots f(c_n)$ as its respective frequency, then the probability of each character in the word w can be expressed as:

$$\frac{f(c_1)}{N_c}, \frac{f(c_2)}{N_c} \dots \frac{f(c_n)}{N_c}.$$

- Assuming each character in the lexicon is independent from each other, the probability of characters needed for the word w can be expressed as:

$$\frac{f(c_1)}{N_c} \times \frac{f(c_2)}{N_c} \times \dots \times \frac{f(c_n)}{N_c}.$$

```

01 <skos:Concept rdf:about=" http://xbrl.fasb.org/us-gaap/2015/elts/us-gaap-std-2015-01-
02 31.xsd#TechnologyServicesRevenue">
03
04 <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Thing"/>
05
06 <skos:notation>
07 us-gaap: TechnologyServicesRevenue
08 </skos:notation>
09
10 <skos:definition>
11 Revenue from providing technology services. The services may include training, installation,
12 engineering or consulting. Consulting services often include implementation support,
13 software design or development, or the customization or modification of the licensed
14 software.
15 </skos:definition>
16
17 <skos:prefLabel xml:lang="en-US">
18 Technology Services Revenue
19 </skos:prefLabel>
20
21 <skos:inScheme rdf:resource=" http://xbrl.fasb.org/us-gaap/2015"/>
22 </skos:Concept>

```

Fig. 1. An Excerpt of Hierarchical Association in SKOS (US GAAP Example). This is an excerpt of a partial fragment of a U.S. GAAP Taxonomy SKOS document based on the principles of FASB's working draft of converting the U.S. GAAP Taxonomy into a SKOS document (O'Riain et al., 2012). This SKOS link is used to create a vocabulary set for a financial reporting concept named "TechnologyServicesRevenue". The concept name is defined in lines 6–8, the concept definition is defined in lines 10–15, and the presentation label is declared in lines 17–19.

- Similarly, if we can express the total number of words in the lexicon as N_w , and its frequency as $f(w)$, then the probability of w detected in the data text can be expressed as:

$$\frac{f(w)}{N_w}.$$

- The mutual information formula then can be expressed as the logarithm of:

$$\frac{f(w)}{N_w} / \frac{f(c_1)}{N_c} \times \frac{f(c_2)}{N_c} \times \dots \times \frac{f(c_n)}{N_c}.$$

$$\text{The result is: } MI(w) = \log_2 \frac{(N_c)^n \times f(w)}{N_w \times f(c_1) \times f(c_2) \times \dots \times f(c_n)}.$$

A larger $MI(w)$ means the probability that the word w detected is greater than the probability of each independent character that composes the word being detected individually. This implies that characters are not independent from each other; rather, they are associated and are composed as an n -gram word. In conjunction with the term frequency dictionary, this value can be applied to obtain accurate terms from textual data.

3.3. Step 3: convert XBRL taxonomy into a semantic, web-based vocabulary dataset

Step 3 in our model development is to form a relationship between XBRL and instantiations of the annual report knowledge structure. Spies (2010) suggests that the Ontology Web Language (OWL) can be used to represent the generally accepted accounting principles taxonomies in XBRL through the use of a set of logical principles of the business reporting metadata and classification systems. Vasarhelyi et al. (2012) also suggest that knowledge from various sources can be embedded into OWL so the concepts and relationships among information are organized and defined. In our model, the instantiated vocabulary dataset is based on SKOS, the semantic web technology developed by W3C. SKOS is similar to OWL; it also uses the RDF schema to develop structured, controlled, and domain-specific vocabularies. SKOS can be used to form a hierarchical knowledge structure and makes the structure (i.e., the decision topic tree established in the previous step) machine-understandable. Compared to OWL, SKOS is more suitable for semi-formal conceptualization⁷ such as a thesaurus-like structure, e.g., the XBRL taxonomy. For example, SKOS is used by FASB to create the definitions and preferred labels for XBRL elements (FASB, 2012). Fig. 1 provides an excerpt of a partial fragment of a U.S. GAAP Taxonomy SKOS document based on the principles of FASB's working draft of converting the U.S. GAAP Taxonomy into a SKOS document (O'Riain et al., 2012).

In this example, a U.S. GAAP XBRL element "TechnologyServicesRevenue" is converted to a SKOS semantic link, which utilizes the RDF schema to locate a financial reporting concept defined in the U.S. GAAP Taxonomy. This link includes semantic

⁷ WWW Consortium (W3C). 2008. Using OWL and SKOS. (<https://www.w3.org/2006/07/SWD/SKOS/skos-and-owl/master.html>; accessed on January 23, 2016)

A) RDDF Tags

```

01      <owl:Ontology>
02      <owl:imports>
03          <owl:Ontology rdf:about="MDA.rdf"/>
04          <owl:Ontology rdf:about="TW-GAAP.rdf"/>
05      </owl:imports>
06      </owl:Ontology>
07
08      <owl:Thing rdf:about="#FinancialResource">
09          <rdf:type rdf:resource="&mda;FinancialResource"/>
10          <skos:prefLabel xml:lang="en">Financial Resource</skos:prefLabel>
11          <skos:altLabel xml:lang="zh">財務資源</skos:altLabel>
12          <skos:inScheme rdf:resource="#ISHBIS"/>
13          <skos:relatedMatch rdf:resource="&tw-gaap;AvailableSaleFinancialAssets"/>
14          <skos:relatedMatch rdf:resource="&tw-gaap;CashHand"/>
15          <skos:relatedMatch rdf:resource="&tw-gaap;FinancialAssetsCarriedCost"/>
16          <skos:relatedMatch rdf:resource="&tw-gaap;FinancialLiabilitiesCarriedCost"/>
17          <skos:relatedMatch rdf:resource="&tw-gaap;HeldMaturityFinancialAssets"/>
18      </owl:Thing>

```

B) RDF Graph

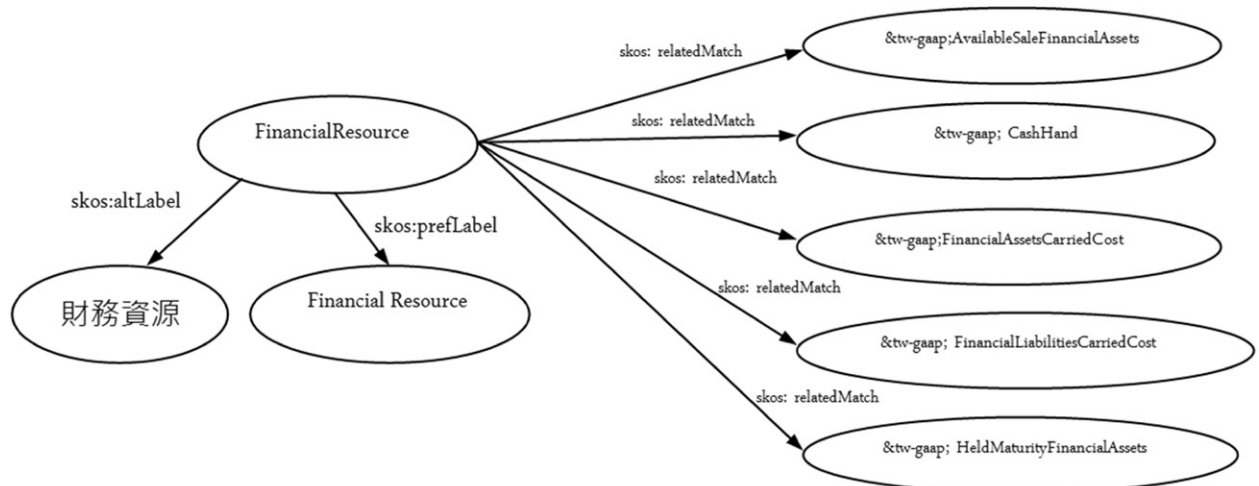


Fig. 2. An illustration of using RT to integrate a knowledge concept to XBRL elements. Panel A. RDF Tags: This example shows how knowledge concepts integrate with XBRL elements. Respective knowledge concepts and XBRL elements defined in existing SKOS files are imported (lines 2–5) to create a target SKOS file, where knowledge concepts and XBRL elements are linked. For example, in line 9–11 a knowledge concept “FinancialResource” is connected with XBRL elements: AvailableSaleFinancialAssets, CshHand, FinancialAssetsCarriedCost, FinancialLiabilitiesCarriedCost, and HeldMaturityFinancialAssets (lines 13–17). Panel B. RDF Graph: The graphical representation of the RT link that links the knowledge concept FinancialResource to XBRL elements is demonstrated in this figure.

information such as the concept name (line 6–8), concept definition (line 10–15), and presentation label (line 17–19). This SKOS link is used to create the vocabulary set for all financial reporting concepts. Each link can later be used to map with text extracted from financial statements to provide the linkage between XBRL elements and keywords extracted from textual disclosures.

3.4. Step 4: create the textual information vocabulary dataset

In Step 4, we create a SKOS dataset for textual disclosures that are organized into the knowledge structure elements. In order for the combined use of financial and nonfinancial information to improve the usefulness of reporting data, textual data needs to be tagged so the contexts can be used and understood by a decision aid (Vasarhelyi et al., 2012). The semantic web technology provides concept declaration and association mechanisms to define concepts, link related knowledge concepts, and build up the

full knowledge structure. For example, SKOS provides the following association concepts to capture different types of semantic relations:

- (1) Equivalence: when two concepts have a horizontal association, the association is linked by either used for (UF) or to use (USE).
- (2) Hierarchical: When two concepts have a hierarchical association, it is represented by broader term (BT) or narrower term (NT).
- (3) Associate: When a concept is associated with a financial reporting item, it is linked by related term (RT).
- (4) Lexical labels: Preferred label (prefLabel), alternative label (altLabel), and hidden label (hiddenLabel) are used to link concepts. Lexical labels also have multilingual support.

3.5. Step 5: establish a semantic link to connect financial and textual vocabulary datasets

In the final step of building our integrated model, we use the “associate” relation (RT) to link knowledge structure terms (textual data from MD&A) and XBRL elements (financial reporting concepts). The RT relation allows users to connect related concepts. However, RT does not automatically form a “bi-directional” relation. Concerning the possibility of “cross-reference,” it is a better design to create a new dataset that declares no new concepts and only defines the bi-directional associations between the two datasets. Therefore, we use “owl:imports” to import both the MD&A SKOS file and the XBRL SKOS file and use RT to associate knowledge concepts and the related XBRL elements. Panel A in Fig. 2 provides an example of this design (lines 2 to 5). The “owl:imports” element is used to import the two existing SKOS files (MDA.rdf for the MD&A file and TW-GAAP.rdf for the XBRL elements) into the target SKOS so that the elements defined in the two existing files can be used in the target SKOS. In the target SKOS, we create a “relation” (RT) between one defined element from the MD&A SKOS and the other defined element from the XBRL SKOS using the “skos:relatedMatch” element. Lines 8 to 18 in Fig. 2 presents an example that shows when a sentence from MD&A disclosure has representative terms that are considered a “financial resource” in our knowledge structure, this sentence will be linked to the following XBRL elements: “AvailableSaleFinancialAssets” (line 13), “CashHand” (line 14), “FinancialAssetsCarriedCost” (line 15), “FinancialLiabilitiesCarriedCost” (line 16), and “HeldMaturityFinancialAssets” (line 17). The reported values of these XBRL elements are linked to “financial resource” because the element label or definition contains one of representative terms. The graphical representation of this semantic link is also shown in Panel B of Fig. 2. The RT association that links related textual terms to financial reporting concepts allows users to connect from reading a financial reporting line item in the XBRL instance document to related keywords in textual disclosures.

4. Results

Initially, we collected a random sample of 40 annual reports published in PDF format and their corresponding financial data from Taiwan's Taiwan 50 Index (TW50) Stocks for the period between 2003 and 2005. Financial services firms are excluded from this study. We extracted textual business information related to operations and strategies from the “letters to stockholders,” “results of operations,” “analysis of sales by segments,” and “risk factors related to our business” sections of the PDF annual report. The total number of raw data in extracted sentences from these 40 annual reports is 3507 sentences. To ensure the generalization of our method, we further collected 20 PDF annual reports and XBRL instances from the TW50 between 2009 and 2012.⁸ The total number of sentences from the new sample is 6167 sentences.

4.1. Result of step 1: define a knowledge framework used to categorize textual data

Following the hierarchy we describe in Section 3.1, we created a 5-level hierarchy that is used to categorize extracted terms from textual disclosures:

Level 1: The first level in our financial reporting knowledge construct represents a firm's strengths and weaknesses (S/W) in its strategies and operations, as related to its ability to compete in the industry, as well as the opportunities and threats (O/T) from the external environment, such as labor market, economy, and changes in regulations. The S/W and O/T elements are at the highest level in the hierarchy.

Level 2: Under Level 2, we identify four additional elements based on the BSC approach to performance measurement: financial, customer, learning and growth, and internal business process. According to BSC, to advance firm performance, managers should focus on these four perspectives to create values for the organization, based on the guidance of the organization's vision and strategy. Hence, at Level 2, we have eight elements (i.e., two elements at Level 1 times 4 elements at Level 2).

Level 3: At Level 3, we classify the eight elements into more detailed categories based on PwC's Business Value Reporting framework. For example, the “Internal Process” element under S/W can be further classified as product innovation, operational process, and customer service process. The “Customer/Market” element under O/T can be further classified into customer demand

⁸ We selected annual reports from 2009 to 2012 for our robustness test. Starting in 2009, Taiwan's SEC requires all publicly traded companies to file their annual reports using XBRL. By selecting financial reports filed during this period, we could use readily available XBRL instance documents submitted by firms in our research. We did not use annual reports after 2013 to maintain comparability between our original sample and the new sample, because 2013 was the first year Taiwan adopted IFRS.

and industry supply, future trends in the industry development, firm strategies for future development, and competitors and market competition. The selection of the 26 elements of this level is not as objective as are the selections made in the prior two levels.

Level 4: Level 4 categorization is performed after examining the textual portions of the sampled annual reports. We further break down elements from the previous level. For example, the “Product Innovation” element in Level 3 is separated into R&D capability, R&D process, R&D future trend, and manufacturing technology. Another example is the “Economy” element. At Level 4, it is separated into five categories: overall economy performance, interest rate fluctuations, exchange rate fluctuations, money supply, and commodity price level. There are a total of 33 elements in our Level 4 category.

Level 5: In Level 5, elements are broken down further to provide more details. For instance, “Operations” is broken down to percentage of revenues from operations and market share. “R&D capability” is further separated into R&D goals, R&D investment, and R&D results. The approach to selecting elements after Level 2 is more ad hoc, and there could be other elements to be chosen in establishing the structure of the financial reporting knowledge construct.

4.2. Result of step 2: develop a systematic approach to automate text extraction

The 3507 sentences extracted from 40 annual reports between 2003 and 2005 formed our initial word database. We used the Chinese Word Segmentation System for POS tagging, the system developed by the Chinese Knowledge Information Processing (CKIP) project affiliated with Taiwan’s National Archives Program. This system has a dynamic adaptive ability in discipline-based lexicon building and an online real-time word segmentation feature. Using this system, the sentence, “Risk assessment is one of the elements in evaluating internal controls” can be segmented to individual terms “risk,” “evaluate,” “is one of,” “internal,” “control,” and “element.” The system also allows new lexicon building, so the new terms, “evaluate risks,” “internal control,” and “compose elements” can be added to the lexicon and replace the original segmentation results.

After basic language units (words and sentences) were successfully extracted and segmented from annual report textual disclosures, we manually mapped each sentence with an element in the knowledge structure using the Vector Space Model (VSM) described by Brown and Tucker (2011). Using a training sample of all textual disclosures from 10 randomly selected annual reports, we manually mapped sentences to elements in the knowledge structure and constructed the concept feature vector. The frequency of each term segmented from the textual data sources corresponds to a dimension in a space vector, which represents the significance of this term on its particular dimension. In this study, we modify a commonly used Term Frequency/Inverse Document Frequency (TFIDF) concept and use the Term Frequency/Inverse Concept Frequency (TFICF) approach to calculate the feature vector. This system indicates the importance of the term in its dimension and forms a feature vector that represents the particular concept. TFICF is calculated as follows:

- TF (term frequency) is defined as the frequency of a term used to describe a concept:
 where TF_{ij} = frequency of keyword j identified in concept C_i ;
 t_j = number of times a keyword j appeared in concept C_i ;
 t_{all} = number of all meaningful words in concept i ;
- ICF (inverse concept frequency) is defined as an inverse of a word that emerges in all concepts. Because a unique, distinguishable term should not appear in every concept, the smaller ICF represents a more distinguishable term:
 where ICF_j = inverse of a word j that appears in all concepts;
 N = number of all available concepts;
 cf_j = number of times the word j appears;
- WEIGHT (the weight assigned to a word that emerges in the vector space of concepts):
 $WEIGHT_{ij} = TF_{ij} \times ICF_j$.

After the initial mapping of textual data using the knowledge structure, the system constructs the space vector for each sentence. We followed the initial mapping with sentences and terms extracted from another set of 20 annual reports to gradually train the software to achieve automation. When new sentences are added to the system, they are compared with the existing concept feature vector. The degree of similarity is calculated using cosine similarity, which is a measure of the cosine of the

Table 1
Accuracy rates for semi-automatic mapping and automatic testing.

Panel A: 40 sample annual reports from 2003 to 2005						
Hierarchy level	1	2	3	4	5	
Cumulated number of elements	2	10	36	69	92	
Accuracy rate						
	Semi-automatic mapping training 1	77.83%	58.36%	46.77%	39.94%	37.42%
	Semi-automatic mapping	87.12%	69.94%	60.11%	52.66%	51.30%
	Automatic testing training 2	87.45%	74.71%	65.21%	58.37%	56.46%
Panel B: 20 sample annual reports from 2009 to 2012						
Hierarchy level	1	2	3	4	5	
Cumulated number of elements	2	10	36	69	92	
Accuracy rate						
	Automatic testing	90.08%	89.13%	88.67%	88.35%	72.04%

```

01 <!DOCTYPE rdf:RDF [
02   <!ENTITY ishbis "http://ishbis.org#" >
03   <!ENTITY tw-gaap "http://ishbis.org#tw-gaap" >
04   <!ENTITY owl "http://www.w3.org/2002/07/owl#" >
05   <!ENTITY xsd "http://www.w3.org/2001/XMLSchema#" >
06   <!ENTITY owl2xml "http://www.w3.org/2006/12/owl2-xml#" >
07   <!ENTITY skos "http://www.w3.org/2004/02/skos/core#" >
08   <!ENTITY rdfs "http://www.w3.org/2000/01/rdf-schema#" >
09   <!ENTITY rdf "http://www.w3.org/1999/02/22-rdf-syntax-ns#" >
10 ]>
11 <rdf:RDF
12   xml:base="urn:cgi:classiferScheme:CGI:TW-GAAP-XBRL:201312"
13   xmlns="urn:cgi:classiferScheme:CGI:TW-GAAP-XBRL:201312#"
14   xmlns:owl="&owl;"
15   xmlns:owl2xml="&owl2-xml;"
16   xmlns:rdf="&rdf;"
17   xmlns:rdfs="&rdfs;"
18   xmlns:skos="&skos;"
19   xmlns:xsd="&xsd;"
20   xmlns:xlink="http://www.w3.org/1999/xlink">
21   <skos:ConceptScheme rdf:about="&tw-gaap;">
22     <rdf:type rdf:resource="&owl;Thing"/>
23     <skos:notation>tw-gaap</skos:notation>
24   </skos:ConceptScheme>
25   .....
26   <skos:Concept rdf:about="&tw-gaap;AvailableSaleFinancialAssets-Current">
27     <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Thing"/>
28     <skos:notation>tw-gaap:AvailableSaleFinancialAssets-Current</skos:notation>
29     <skos:definition>
30
31     <!-- in traditional Chinese -->
32     凡該資產取得之主要目的為短期內出售，於原始認列時即屬合併管理之可辨認金融工具組合之一部分，
33     且有證據顯示近期該組合為短期獲利之操作模式及除財務保證合約或被指定且為有效避險工具外之衍
34     生金融資產屬之。
35
36     <!-- in English -->
37     Available-for-sale financial assets classified as current are those nonderivative financial assets that
38     are designated as available for sale or are not classified as loans and receivables, held-to-maturity
39     investments, or financial assets at fair value through profit or loss and will be sold in a short term.
40     They could be one part of a portfolio that are planned to trade shortly or are assigned as
41     additional derivatives of one financial commitment or hedge instruments.
42
43     </skos:definition>
44     <skos:prefLabel>備供出售金融資產-流動 (Available-for-Sale Financial Assets -
45     Current)</skos:prefLabel>
46     <skos:inScheme rdf:resource="&tw-gaap;" />
47   </skos:Concept>
48   ...
49 </rdf:RDF>
50

```

Fig. 3. An Excerpt of Hierarchical Association in SKOS (Taiwan GAAP example). This SKOS document demonstrates how an XBRL concept "AvailableSaleFinancialAssets" is transformed into a semantic link that can be used to connect with knowledge concepts later.

angle between two vectors (Salton and Buckley, 1988). The closer the cosine similarity to 1, the higher the similarity of the new sentence to an existing concept. An expert panel with high domain knowledge verified the classification results. If the results were correct, they were saved to the keyword database, which ensures the accuracy of automatic mapping. Finally, the remaining 10 annual reports were fed to the system without human intervention. Table 1 shows the accuracy rates of our mapping results.

Similar to prior research, we use the whole document as a unit to measure the accuracy (De Bruijn and Martin, 2002; Hui and Yu, 2005; Chi, 2007), and we determine that the mapping is satisfactory if the accuracy rate is greater than 80% at any hierarchical level. As reported in Panel A of Table 1, the accuracy rate at Level 1 is 87.45%, while the classification tends to be ambiguous (not clear-cut) at lower levels.⁹

Additionally, we also use another 20 annual reports between 2009 and 2012 to conduct a robustness test. We apply the same procedures as in the automatic testing process to determine the respective accuracy rates in the five levels. As reported in Panel B of Table 1, the 20 additional annual report samples achieve similar results, which is comparable to the result from the 10 sample reports at the automatic testing process in the first phase. These additional annual reports from 2009 to 2012 provide further evidence that the method we used in analyzing MD&A is generalizable using documents from different companies and with different accounting periods.

⁹ Prior studies indicate that other factors also may contribute to the low accuracy rate, such as insufficient training data or inaccurate initial classification.

The two-phase implementation of our methodology demonstrates that the system evolves as the samples grow. Panel B of Table 1 reports a significantly higher accuracy rate for all levels than that of Panel A of Table 1, particularly at the more granular level-5 concepts.

4.3. Result of step 3: convert XBRL taxonomy into a semantic web-based vocabulary dataset

We use a SKOS editor, SKOSed, to prepare the SKOS document to form a vocabulary dataset for Taiwanese GAAP XBRL taxonomy as we used Taiwanese public firms in our sample. For the sample annual reports from 2003 to 2005, we had to manually create 30 XBRL instance documents using Fujitsu's XWand Instance Editor for firms selected for this study because Taiwanese firms were not required to file XBRL-formatted reports at that time. However, we downloaded XBRL instance documents for the sample annual reports from 2009 to 2012 (directly from the Taiwan Stock Exchange website), since Taiwan's XBRL mandate started in 2009. This dataset is illustrated in Fig. 3.

SKOSed is a plugin for Stanford Protégé 4.0 that allows users to create and edit thesauri (or similar artifacts) represented in SKOS. As Fig. 3 shows, the root element of this SKOS is `rdf:RDF` (line 11), and we use the Commission for the Management and Application of Geoscience Information (CGI) register name space to ensure its publicity, as it is shown as “`urn:cgi:classifierScheme:CGI:TW-GAAP-XBRL:201,312`” in the root element declaration. The “`urn:cgi:classifierScheme:CGI:`” identifies the register of vocabularies that contain terms used to populate property instances in any datasets maintained by CGI model. The aims of CGI are to enable the global exchange of knowledge about geoscience information and systems, which automatically comprise those globally adopted XBRL taxonomies. In lines 26 to 47 we define a SKOS concept using a Taiwanese GAAP Taxonomy element: “AvailableSaleFinancialAssets-Current.” The definition of the concept in Chinese (lines 31 to 34) and in English (lines 36

The screenshot displays the SKOSed application window. On the left, a tree view shows a hierarchy of concepts, with 'FinancialResource' highlighted and enclosed in a red box. A red arrow points from this box to a text box below. The main area on the right shows various tabs for 'FinancialResource', including 'SKOS Annotations', 'SKOS Usage', 'SKOS Object Property Assertions', and 'SKOS Data Property Assertions'. The 'SKOS Data Property Assertions' tab is active, showing a list of assertions with 'prefLabel' and 'altLabel' fields. A red arrow points from this tab to a text box on the right. Another red arrow points from the 'FinancialResource' concept in the tree to a text box at the bottom center.

FinancialResource is a narrower term (NT) of FinanceSW in our knowledge structure, while FinancialSW is a broader term (BT) of Financial Resource.

All concepts and terms are labeled in English (using `prefLabel`) and in Chinese (using `altLabel`).

Five XBRL elements have the associate (RT) semantic link to FinancialResource.

Fig. 4. Linking Textual Information with XBRL Elements using SKOS This example shows the software tool used to semantically link XBRL element with concepts defined in the knowledge framework.

to 41), and the label (lines 44 to 45) provide a complete SKOS declaration of this XBRL element. As explained in Section 3.5, this file is later imported to the SKOS semantic link to be integrated with the SKOS file containing textual data from MD&A disclosures.

4.4. Result of step 4: create the textual information vocabulary dataset

Organizing all concepts from our five-level knowledge hierarchy to form an SKOS vocabulary dataset is the main objective in this step. As Fig. 4 illustrates, we first used the “hierarchical” relation (BT and NT) to organize concepts and terms in our five-level knowledge structure. The left panel in Fig. 4 shows an example of this hierarchical relationship between Strengths and Weakness in Finance (Finance SW) and Financial Resource. Financial Resource is a narrower term (NT) of FinanceSW in our knowledge structure, while FinancialSW is a broader term (BT) of Financial Resource. Furthermore, the middle panel in Fig. 4 shows the associate (RT) semantic link: five XBRL elements (FinancialAssetsCarriedCost, CashHand, FinancialLiabilitiesCarriedCost, HeldMaturityFinancialAssets, and AvailableSaleFinancialAssets) are linked to the knowledge structure element “FinancialResource.” Finally, all concepts and terms are labeled in English (using prefLabel) and in Chinese (using altLabel), as demonstrated in the right panel of Fig. 4.

4.5. Result of step 5: establish a semantic link to connect financial and textual vocabulary datasets

Both textual data extracted from annual reports and XBRL instance documents are stored in a relational database. We selected the top 20 terms based on the feature vectors that are associated with financial reporting concepts to build semantic links between hard and soft information. The determination of how to associate related concepts and XBRL elements depends on “key-words matching.” Each concept feature vector is comprised of various related terms. Then we match those terms from textual disclosures to the labels of each XBRL element and the element definition described in the SKOS dataset. For a given feature vector, once its “representative” terms match any terms shown in the labels of any XBRL elements or the SKOS definitions, the association is determined. Table 2 provides a list of four links identified in this study.

For example, the first link in Table 2 shows that when a user reads a paragraph about “revenue” in the annual report, if this paragraph contains any representative terms (such as operation, income, sale, and revenue), the system will provide a SKOS link to related accounts and additional disclosures. These related XBRL elements could be operating revenue, gross and net sales, other

Table 2

Key terms in knowledge tree and related XBRL elements.


Selected tree elements (used in SKOS)	Selected representative terms	Related XBRL elements
Revenue (StrengthAndWeakness\ Finance\Financial Performance\Revenue)	operation, income, sale, revenues, total revenues	<ul style="list-style-type: none"> • Operating revenue (tw-gaap:OperatingRevenue) • Sales (tw-gaap:Sales) • Net sales (tw-gaap:NetSales) • Service revenue (tw-gaap:ServiceRevenue) • Rental revenue (tw-gaap:RentalRevenue) • Repairs and maintenance revenue (tw-gaap:RepairsMaintenanceRevenue) • Notes related to revenue recognition (tw-gaap:NotesAccountingPolicyRevenueRecognition)
Earnings per share (StrengthAndWeakness\ Finance\Financial Performance\Earnings PerShare)	net income, income, net profit, profitability	<ul style="list-style-type: none"> • Primary earnings per share (tw-gaap:PrimaryEarningsPerShare) • Diluted earnings per share (tw-gaap:DilutedEarningsPerShare) • Net income (Loss) (tw-gaap:NetIncomeLoss) • Cash dividends of preferred stock (tw-gaap:CashDividendsPreferredStock) • Number of shares (tw-gaap:NumberShares)
Financial resource (StrengthAndWeakness\ Finance\Financial Resource)	cash flow, financial assets, financial instruments, long term investment	<ul style="list-style-type: none"> • Cash and cash equivalents (tw-gaap:CashCashEquivalents) • Total financial assets measured at fair value through profit or loss – current (tw-gaap:FinancialAssetsMeasuredFairValueProfitLossCurrent) • Total available-for-sale financial assets – current (tw-gaap:AvailableSaleFinancialAssetsCurrent) • Total Held-to-maturity Financial Assets – Current (tw-gaap:HeldMaturityFinancialAssets-Current) • Notes related to financial instruments (tw-gaap:NotesFinancialInstruments)
Proportion of revenue	operation, income, sale, revenues, total revenues	<ul style="list-style-type: none"> • Operating revenue (tw-gaap:OperatingRevenue) • Sales (tw-gaap:Sales) • Net sales (tw-gaap:NetSales) • Service revenue (tw-gaap:ServiceRevenue) • Rental revenue (tw-gaap:RentalRevenue) • Repairs and maintenance revenue (tw-gaap:RepairsMaintenanceRevenue)

A) Unstructured textual disclosures



年報內文 (ANNUAL REPORT)		年報SWOT分析 (SWOT ANALYSIS)		回首頁(TOP)	
十六、每股盈餘					
民國九十二年度及九十一年度本公司基本每股盈餘及稀釋每股盈餘計算如下：					
	92年度		91年度		
	稅 前	稅 後	稅 前	稅 後	
基本每股盈餘：					
屬於普通股股東之本國淨利	\$ 15,573,259	15,659,928	6,022,697	6,022,669	
加權平均流通在外股數(千股)					
期初普通股	4,015,255	4,015,255	2,970,582	2,970,582	
加：現金增資	-	-	291,667	291,667	
買回庫藏股	(2,926)	(2,926)	(377)	(377)	
溢餘轉增資	244,055	244,055	-	-	
發行債券換取權利證書	33,044	33,044	377,927	377,927	
本期加權平均流通在外股數	4,289,428	4,289,428	3,639,799	3,639,799	
每股盈餘(元)	\$ 3.63	3.65	1.65	1.65	
稀釋每股盈餘：					
屬於普通股股東之本國淨利	\$ 15,573,259	15,659,928	6,022,697	6,022,669	
可轉換公司債之影響	34,301	25,726	170,825	128,119	
計算稀釋每股盈餘之本國淨利	\$ 15,607,560	15,685,654	6,193,522	6,150,788	
加權平均流通在外股數(千股)					
期初普通股	4,015,255	4,015,255	2,970,582	2,970,582	
加：潛在普通股					
合計流通在外股數	81,953	81,953	632,910	632,910	

B) Soft information organized in knowledge structure



年報內文 (ANNUAL REPORT)		年報SWOT分析 (SWOT ANALYSIS)		回首頁(TOP)	
<p>年報SWOT分析 (SWOT Analysis of Annual Report)</p> <p>優勢與弱勢 (Strength and Weakness)</p> <p>財務 (Finance)</p> <p>財務表現 (Financial Performance)</p> <p>收入表現 (Revenue)</p> <p>財務資源 (Financial Resource)</p> <p>顧客 (Customer)</p> <p>顧客市場區隔 (Market Segmentation)</p> <p>主要業務內容 (Business Overview)</p> <p>營收比重 (Proportion of Revenue)</p> <p>市場佔有率 (Market Share)</p> <p>銷售服務對象 (Target Buyer)</p> <p>主要客戶銷售金額 (Sales to Major Customers)</p> <p>主要客戶銷售金額比重 (Percentage of Sales Revenue from Major Customers)</p> <p>主要客戶銷售增加原因 (Reason for the Increase of Sales to Major Customers)</p> <p>銷售與業務發展策略 (Sales Promotion and Business Development Strategy)</p> <p>顧客關係管理 (Customer Relationship Management)</p> <p>顧客服務 (Customer Service)</p> <p>顧客滿意度 (Customer Satisfaction)</p> <p>品質管理 (Quality Control Management)</p> <p>品牌策略與發展策略 (Brand Image and Development Strategy)</p> <p>內部程序 (Internal Business Process)</p>					
<p>年度(year) 92 公司(company) 友達 - 營收表現</p> <p>合併營收不僅首度突破新台幣1000億元，更創新高達新台幣156.6億元，居台灣TFT-LCD同業之冠。</p> <p>全年獲利超過新台幣60億元。</p> <p>友達2002年營業額新台幣755億，相較2001年成長一倍。</p> <p>第一季的營運績效較2002年第四季改善。</p>					
<p>English translation:</p> <ul style="list-style-type: none"> The combined revenue exceeds one trillion NTD for the first time, with gross profit of 156 billion NTD, which is the industry leader in the country. Annual gross profit exceeds 60 billion NTD. AUO achieved 755 billion NTD in revenues in 2002, which was double the 2001 result. The first quarter revenue improved as compared to the fourth quarter of 2002. 					

C) Semantically linked financial data

年報內文 (ANNUAL REPORT)

年報SWOT分析 (SWOT Analysis of Annual Report)

優勢與弱勢 (Strength and Weakness)

財務 (Finance)

財務表現 (Financial Performance)

營收表現 (Revenue)

每股盈餘 (Earnings Per Share)

財務資源 (Financial Resource)

顧客 (Customer)

顧客市場區隔 (Market Segmentation)

主要業務內容 (Business Overview)

營收比重 (Proportion of Revenue)

市場佔有率 (Market Share)

銷售服務對象 (Target Buyer)

主要客戶銷售金額 (Sales to Major Customers)

主要客戶銷售金額比重 (Percentage of Sales Revenue from Major Customers)

主要客戶銷售增加原因 (Reason for the Increase of Sales to Major Customers)

銷售與業務發展策略 (Sales Promotion and Business Development Strategy)

顧客關係管理 (Customer Relationship Management)

顧客服務 (Customer Service)

顧客滿意度 (Customer Satisfaction)

品質管理 (Quality Control Management)

品牌策略與發展策略 (Brand Image and Development Strategy)

內部程序 (Internal Business Process)

年報SWOT分析 (SWOT ANALYSIS)

回首頁(TOP)

公司：友達 關鍵字：營收 對應項目：營業收入

	2003	2004	2005
營業收入	104860642000	168111569000	217388368000
相關科目			
營業成本	81398889000	128468264000	187540389000
營業毛利(毛損)	23461753000	39643305000	29847999000
繼續營業單位稅前淨利(淨損)	15573259000	28024198000	16094568000
應收帳款淨額	11479871000	15297817000	34848588000
應收帳款淨額-應付人	5481108000	5420358000	7766800000
相關財務比率			
應收帳款週轉率	618.25	811.43	510.12
平均收現日數	0.59	0.45	0.72
毛利率	22.37	23.58	13.73
相關政策與策略			

types of revenue, and notes on significant accounting policies related to revenue recognition. This list of semantic links can be extended to include new terms, which allows flexibility of the system and better supports decision-making.

4.6. The system demonstration

After we applied the 5-step processes described in earlier sections, the test of the validity of our design (Gregor and Hevner, 2013) is demonstrated in this section. When users read the description of a firm's strategies in the annual report, they need to manually look for supporting financial data to verify the success of operating strategies. This process usually takes time and different file formats to create additional information processing barriers. The methodology proposed in this study solves this information processing problem by creating hyperlinks from classified textual information that takes users directly to related financial reporting line items stored in the associated XBRL instance document. This design allows efficient information processing.

Fig. 5 provides an example of AU Optronics Corporation's (NYSE: AUO) 2003 annual report to demonstrate how the integrated system works. Panel A in Fig. 5 shows the unstructured textual data from the annual report in the original PDF format. The classification of textual data to our knowledge structure provides soft information that is easier to access for users (Panel B in Fig. 5, left frame). When a user clicks the "Revenue" hyperlink, the system pulls related textual disclosures that are scattered across the "letter to shareholders" and "Management Discussions and Analysis" sections and displays the content in the right frame (Panel B in Fig. 5, right frame). These originally scattered pieces of information can now be found under the hierarchy of Strength and Weakness > Financial Performance > Revenue. Further, these textual disclosures are semantically linked to actual financial data. When users click the hyperlinked terms "revenue" in the right frame of Panel B in Fig. 5, the system responds with information about the 2003–2005 net income and related account balances, such as earnings per share, results of financial ratio analysis, ROA, and ROE (Panel C in Fig. 5).

5. Conclusions and discussions

This study adopts a DSR approach to propose a new way to integrate soft information with hard information using text analytics and semantic web techniques to address the inefficient manual processing of hard/soft information in traditional annual reports. The automatic process that links information from different locations in the annual report potentially leads to the detection of new associations between existing disclosures previously not found by users. The linking of hard/soft information has multiple implications for the users of financial reports.

First, the text analytics techniques applied in this study allow previously unstructured textual data to be presented in a knowledge structure, which removes the need to manually review and search for information presented in textual format. This information retrieval process decreases the amount of time needed to gather information and increases the relevance of information used to support decision-making. Second, we integrate information from financial reporting line items, footnotes, and textual descriptions that were previously scattered in an annual report. By doing so, users can extract the most relevant information to support their decision-making and to create different business reports to meet their specific needs. Third, one reason that nonfinancial information has not played the critical role in decision making that it deserves is its high information-processing cost. The use of technologies such as text analytics and XBRL can reduce processing cost significantly, which encourages the financial reporting community to consider how business reporting should be produced and disseminated.

The new reporting framework considers the relevance of soft business information and how it can be used to improve efficiency in providing useful information for decision-making. Although the PDF format of annual reports makes these reports readable to human eyes, it creates an information-processing barrier for machines. The XBRL detailed footnote tagging of textual information promotes the usability of soft information. However, the supplementing and complementing of soft information cannot be fully utilized until it is cross-referenced with financial report line items. The system presented in this study is an extendable prototype that is designed to link soft business information with financial data. This system can be extended to other forecasting models, such as earnings quality analysis or stock pricing forecasts. The classification schemes also can be modified to include other classification hierarchies to analyze unstructured textual data in different domains.

This study has the following limitations. First of all, this paper is descriptive in nature and our concept-testing sample size is very small. As we demonstrated in our results section, higher accuracy can be achieved by increasing the sample size. However, one technical barrier worth noting is that as the number of words increases, the system performance reduces dramatically since the TFIDF vectors will become larger and larger. One solution could be to migrate the system to the cloud to mitigate the technical risks caused by the limited computing capacity in our premise system. In addition, the knowledge structure used in this study to classify soft information is based on the domain knowledge of the authors. By definition, the prior knowledge comes from professional training and experiences may not be replicable (Roberts, 1991). Although the result of this research may not be generalized, the framework and approach proposed can be used to create other knowledge structures as well as illustrate the capability of our model to link soft business information and numerical financial data.

Fig. 5. Integrated Financial and Nonfinancial Information. Panel A shows the original annual report in PDF format. When an user clicks a concept "Revenue" (Panel B) the system pulls related textual disclosures (right frame of the Panel B). Hyperlinks in textual disclosures take users to related financial data previously found in XBRL instances (Panel C).

References

- Abrahamson, E., Amir, E., 1996. The information content of the president's letter to shareholders. *J. Bus. Finance Account.* 23 (8), 1157–1182.
- Alles, M., Debreceeny, R., 2012. The evolution and future of XBRL research. *Int. J. Account. Inf. Syst.* 13, 83–90.
- Antweiler, W., Frank, M.Z., 2004. Is all that talk just noise? The information content of internet stock message boards. *J. Financ.* 59 (3), 1259–1294.
- Blankespoor, E., 2012. The Impact of Investor Information Processing Costs on Firm Disclosure Choice: Evidence From the XBRL Mandate. Ph.D. dissertation at University of Michigan. Available at http://www.stern.nyu.edu/sites/default/files/assets/documents/con_034218.pdf (accessed on January 24, 2016).
- Blankespoor, E., Miller, B.P., White, H.D., 2014. Initial evidence on the market impact of the XBRL mandate. *Rev. Acc. Stud.* 19, 1468–1503.
- Boritz, E., Hayes, L., Lim, J.H., 2013. A content analysis of auditors' reports on IT internal control weaknesses: the comparative advantages of an automated approach to control weakness identification. *Int. J. Account. Inf. Syst.* 14 (2), 138–163.
- Brown, S.V., Tucker, J.U., 2011. Large-sample evidence on Firms' year-over-year MD&A modifications. *J. Account. Res.* 49 (2), 309–346.
- Bryan, S.H., 1997. Incremental information content of required disclosures contained in management discussion and analysis. *Account. Rev.* 72 (2), 285–301.
- Chi, Y.-L., 2007. Elicitation synergy of extracting conceptual tags and hierarchies in textual document. *Expert Syst. Appl.* 32 (2), 349–357.
- Cimiano, P., 2006. *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications*. Springer-Verlag, New York.
- Davis, A., Tama-Sweet, I., 2012. Managers' use of language across alternative disclosure outlets: earnings press releases versus MD&A. *Contemp. Account. Res.* 29, 804–837.
- De Bruijn, B., Martin, J., 2002. Getting to the core of knowledge: mining biomedical literature. *Int. J. Med. Inform.* 67 (1–3), 7–18.
- Debreceeny, R., Gray, G.L., 2001. The production and use of semantically rich accounting reports on the Internet: XML and XBRL. *Int. J. Account. Inf. Syst.* 2, 47–74.
- Debreceeny, R.S., Chandra, A., Cheh, J.J., Guithues-Amrhein, D., Hannon, N.J., Hutchison, P.D., Janvrin, D., Jones, R.A., Lamberton, B., Lymer, A., Mascha, M., Nehmer, R., Roohani, S., Srivastava, R.P., Trabelsi, S., Tribunella, T., Trites, G., Vasarhelyi, M.A., 2005. Financial reporting in XBRL on the SEC's EDGAR system: a critique and evaluation. *J. Inf. Syst.* 19 (2), 191–210.
- Eccles, R., Herz, R., Keegan, E., Phillips, D., 2001. *The Value Reporting Revolution: Moving Beyond the Earnings Game*. John Wiley & Sons, Inc., New York.
- Engelberg, J., 2008. Costly information processing: evidence from earnings announcements. American Finance Association Annual Meeting, San Francisco. Available at SSRN: <http://ssrn.com/abstract=1107998>
- Fan, W., Wallace, L., Rich, S., Zhang, Z., 2006. Tapping the power of text mining. *Commun. ACM* 49 (9), 77–82.
- FASB, 2012. *XBRL-US GAAP Taxonomy in SKOS* (not published, retrievable by request).
- Fisher, I.E., Garnsey, M.R., Goel, S., Tam, K., 2010. The role of text analytics and information retrieval in the accounting domain. *J. Emerg. Technol. Account.* 7, 1–24.
- Forst, M., Fang, J., 2009. TBI-improved non-deterministic segmentation and POS tagging for a Chinese parser. 1Proceedings of the 12th Conference of the European Chapter of the ACL, Athens, Greece, pp. 264–272.
- Garnsey, M.R., 2006. Automatic classification of financial accounting concepts. *J. Emerg. Technol. Account.* 3, 21–39.
- Geerts, G.L., 2011. A design science research methodology and its application to accounting information systems research. *Int. J. Account. Inf. Syst.* 12, 142–151.
- Ghani, E., Laswad, F., Tooley, S., 2009. Digital reporting formats: users' perceptions, preferences, and performances. *Int. J. Digit. Account. Res.* 9, 45–98.
- Gómez-Pérez, A., Fernández-López, M., Corcho, O., 2004. *Ontological Engineering: With Examples From the Areas of Knowledge Management, E-commerce and the Semantic web*. Springer-Verlag, New York.
- Gregor, S., Hevner, A.R., 2013. Positioning and presenting design science research for maximum impact. *MIS Q.* 37 (2), 337–355.
- Hevner, A.R., March, S.T., Park, J., Ram, S., 2004. Design science in information systems research. *MIS Q.* 28 (1), 75–105.
- Hirst, D.E., Hopkins, P., Wahlen, J., 2004. Fair values, income measurement, and bank analysts' risk and valuation judgments. *Account. Rev.* 79 (2), 454–472.
- Hodge, F.D., Kennedy, J.J., Maines, L.A., 2004. Does search-facilitating technology improve the transparency of financial reporting? *Account. Rev.* 79 (3), 687–703.
- Huang, C., Kuo, C., 2003. The transformation and search of semi-structured knowledge in organizations. *J. Knowl. Manag.* 7 (4), 106–123.
- Humpherys, S.L., Moffitt, K.C., Burns, M.B., Burgoon, J.K., Felix, W.F., 2011. Identification of fraudulent financial statements using linguistic credibility analysis. *Decis. Support. Syst.* 50 (3), 585–594.
- Hui, B., Yu, E., 2005. Extracting conceptual relationships from specialized documents. *Data Knowl. Eng.* 54 (1), 29–55.
- Kim, J.W., Lim, J.H., No, W.G., 2012. The effect of mandatory XBRL reporting across the financial information environment: evidence in the first waves of mandated U.S. filers. *J. Inf. Syst.* 26 (1), 1–27.
- Li, F., 2010. The information content of forward-looking statements corporate filings - a naïve Bayesian machine learning approach. *J. Account. Res.* 48 (5), 1049–1102.
- Li, F., Lundholm, R., Minnis, M., 2013. A measure of competition based on 10-K filings. *J. Account. Res.* 51, 399–436.
- Liedtka, S.L., 1999. *Nonfinancial and Financial Performance Measures: An Empirical Analysis of the Airline Industry*. University of Maryland Ph.D. dissertation.
- Maines, L.A., McDaniel, L.S., 2000. Effects of comprehensive-income characteristics on nonprofessional investors' judgments: the role of financial-statement presentation format. *Account. Rev.* 75, 179–207.
- Maines, L.A., Bartov, E., Fairfield, P.M., Hirst, D.E., Iannaconi, T.E., Mallett, R., Schrand, C.M., Skinner, D.J., Vincent, L., 2002. Recommendations on disclosure of nonfinancial performance measures. *Account. Horiz.* 16 (4), 353–362.
- Nelson, M.W., Tayler, W.B., 2007. Information pursuit in financial statement analysis: effects of choice, effort, and disaggregation. *Account. Rev.* 82 (3), 731–758.
- O'Riain, S., Curry, E., Harth, A., 2012. XBRL and open data for global financial ecosystems: a linked data approach. *Int. J. Account. Inf. Syst.* 13, 141–162.
- Peng, F., Feng, F., McCallum, A., 2004. Chinese segmentation and new word detection using conditional random fields. Proceedings of the 20th International Conference on Computational Linguistics.
- Plumlee, M.A., 2003. The effect of information complexity on analysts' use of that information. *Account. Rev.* 78 (1), 275–296.
- Previts, G., Bricker, R., Robinson, T., Young, S., 1994. A content analysis of sell-side financial analyst company reports. *Account. Horiz.* 8 (2), 55–70.
- Roberts, E.B., 1991. *Entrepreneurs in High Technology: Lesson From MIT and Beyond*. Oxford University Press, New York.
- Salton, G., Buckley, C., 1988. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manag.* 24 (5), 513–523.
- Securities and Exchange Commission (SEC), 2003. Interpretation: Commission Guidance Regarding Management's Discussion and Analysis of Financial Condition and Results of Operations. Release Nos. 33-8350; 34-48960; FR-72. SEC, Washington, DC.
- Securities and Exchange Commission (SEC), 2009. Interactive Data to Improve Financial Reporting Release Nos. 33-9002; 34-59324; 39-2461; IC-28609; SEC, Washington, DC.
- Securities and Exchange Commission (SEC), 2010. Interpretation: Commission Guidance on Presentation of Liquidity and Capital Resources Disclosures in Management's Discussion and Analysis Release Nos. 33-9144; 34-62934; FR-83. SEC, Washington, DC.
- Sedbrook, T., Newmark, R., 2008. Automating REA policy level specifications with semantic web technologies. *J. Inf. Syst.* 22 (2), 249–277.
- Shirata, C.Y., Takeuchi, H., Ogino, S., Watanabe, H., 2011. Extracting key phrases as predictors of corporate bankruptcy empirical analysis of annual reports by text mining. *J. Emerg. Technol. Account.* 8, 31–44.
- Spies, M., 2010. An ontology modelling perspective on business reporting. *Inf. Syst.* 35 (4), 404–416.
- Sproat, R., Emerson, T., 2003. The first international Chinese word segmentation bakeoff. Proceedings of the Second SIGHAN Workshop on Chinese Language Processing, pp. 133–143.
- Sullivan, D., 2001. *Document Warehousing and Text Mining: Techniques for Improving Business Operations, Marketing, and Sales*. John Wiley & Sons, Inc., New York.
- Sun, Y., 2010. Do MD&A disclosures help users interpret disproportionate inventory increases? *Account. Rev.* 85 (4), 1411–1440.
- Sutton, S.G., Arnold, V., Bedard, J.C., Phillips, J.R., 2012. Enhancing and structuring the MD&A to aid investors when using interactive data. *J. Info. Sys.* 26 (2), 167–188.
- Tetlock, P.C., 2007. Giving content to investor sentiment: the role of media in the stock market. *J. Financ.* 62 (3), 1139–1168.
- Vasarhelyi, M.A., Chan, D.Y., Krahel, J.P., 2012. Consequences of XBRL standardization on financial statement data. *J. Inf. Syst.* 26 (1), 155–167.
- Vincent, L., 1999. The information content of funds from operations (FFO) for real estate investment trusts (REITs). *J. Account. Econ.* 26, 69–104.
- Visa, A., Toivonen, J., Ruokonen, P., Vanharanta, H., Back, B., 2000. Knowledge discovery from text documents based on paragraph maps. Proceedings of the 33rd Hawaii International Conference on System Sciences, 2, pp. 1–9.
- Yang, C.C., Luk, J.W.K., Yung, S.K., Yen, J., 2000. Combination and boundary detection approaches on Chinese indexing. *J. Am. Soc. Inf. Sci.* 51 (4), 340–351.