A TEXT MINING ANALYSIS FOR THE EFFECTIVENESS OF GREEN INFORMATION TECHNOLOGY IN THE TAIWANESE CORPORATE SOCIAL RESPONSIBILITY REPORT

Shi-Ming Huang, Department of Accounting and Information Technology, National Chung Cheng University, No.168, University Road, Min-Hsiung, Chia-Yi Taiwan, R.O.C., +886-272-0411#16812, smhuang@mis.ccu.edu.tw

Jhih-Wei Liao, Department of Information and Management, National Chung Cheng University, No.168, University Road, Min-Hsiung, Chia-Yi Taiwan, R.O.C., +886-272-0411#16812, steven789632145@hotmail.com

Ling-Yi Chou, Department of Accounting and Information Technology, National Chung Cheng University, No.168, University Road, Min-Hsiung, Chia-Yi Taiwan, R.O.C., +886-272-0411#16812, choulinyi@gmail.com

ABSTRACT

With the increasing importance of the environment protection, socially responsible investing growth potential is extraordinary. Many companies incorporate green information technology into the report of corporate social responsibility to consider both financial return and social good. However, it also brings various voluntary disclosure and review problems that lie on the path toward success. One of the problem is the stakeholders generally lack of sound assurance that whether their corporate social report are effective and efficient, or information announced to stakeholders are reliable and integrity. Thus, for the sake of disclosure and assurance, a lot of design and implement integrative platforms have been developed in the field of green information technology assurance.

Text mining is a statistical pattern learning method, and has been adopt to verify text categorization and clustering, sentiment analysis, and document summarization, and entity relation modeling. In this research, we try to develop an automatic mechanism to detect the structure and content of the corporate social responsibility reports, and the inconsistency between actual corporate social responsibility with green information technology and announced content of corporate social responsibility reports. The mechanism adopts text mining as verification tool, and we apply this mechanism to audit the content of corporate social responsibility to reach automatic analyzing the content of corporate social responsibility report. A prototype has been developed to evaluate feasibility of mechanism, and we evaluated the verification and validation through a real case. The result shows that it is useful for assurance of the case company’s corporate social responsibility report, and the system can provide information effectiveness and efficient for them to detect possible difference.

Keywords: Text mining, TF-IDF, Zipf's Law, ontology, design science research method, corporate social responsibility, green technology

1. INSTRUCTION

In the recent years, environment protection has become significant force in the business world, it not only speeds up the business process with green information technology, but also affects the ways that business running their daily operations in. With more and more organizations attach great importance to green information technology related information, the number of corporate social responsibility (CSR) reports grow rapidly. According to the forecast and estimation of the report CR perspectives 2013 conducted by CorporateRegister.com, the integrated CSR reports are increasing and the global CSR reports would be equivalent to about 52,000 reports from the world's largest corporate responsibility reporting database in 2013. Although CSR reports that tend to give rise to various perspectives, it is also brings a lot problems about the voluntary disclosure and review difficulty that cause damages to the stakeholders. These damages can lead to errors, undetected fraud, impairment of reputation, and eventually cause business lost revenue, lost data, lost customer trust, and increased costs.

Great deals of corporate responsibility problems above mentioned are about automatic analysis over the content of CSR. However, stakeholders generally lack of sound assurance that whether their corporate social report are effective and efficient, or information announced to customers are reliable and integrity. In other words, many stakeholders face review problems caused by various voluntary disclosure information of CSR. Moreover, because it is directly visible to the stakeholders, the assurance over business operations in the CSR reports are particular important. In addition, CSR reports were involved maximize business benefits [1], and even it is more likely to affect firm financial performance in the more stakeholder-oriented countries. [2] Therefore, there is a criterion which is needed to implement some solutions for CSR reports, especially real-time analysis and assurance, in order to help stakeholders to verify the corporate social report are effective and efficient, or information announced to stakeholders are reliable and integrity. We aim at developing an automatic mechanism for corporate social responsibility analysis and assurance.

This study attempts to construct an automatic mechanism to detect the content of CSR reports, and analyze the consistency between actual corporate social responsibility with green information technology and announced content of corporate social responsibility report. In the current paper, our main focus is on how use a Chinese and an English document collection to evaluate the ontology for knowledge of green IT and why the optimal strategy of CSR disclosure therefore depends on the organizational forms. We extend Murugesan [3] by introducing approach of green IT into the CSR processes employed by investors and stakeholders. In particular, CSR reports are now produced using various forms. But whereas disclosures of CSR are controlled by the entrepreneur, the voluntary disclosure of CSR is able to influence the stakeholders’ advocacy behavior and corporate image. We apply a text mining mechanism to build a framework that identify CSR items for green IT technology, and compare the CSR disclosure degree between different organizations.

2. LITERATURE REVIEW

2.1 TF-IDF

Term frequency–inverse document frequency (TF-IDF) is one of statistic methods which reflect the importance of a word to a document. According the definition, TF is the term frequency in a document, and term is likes words or phrases. The frequency of each term in the document may vary widely. Thus, term frequency becomes an important attribute to discriminate itself from other terms [4].

The TF algorithm calculates terms weight only based on their frequency. The IDF algorithm is to measure whether the terms are common or rare across all documents. In addition, TF-IDF term weight algorithm that tends to filter out common terms and to preserve important words. For example, high weight of TF-IDF is achieved by a high term frequency and low document frequency of the term in the all collection of documents. Hence, high TF-IDF value tends to preserve important words.

2.2 Zipf's law

The concept of Zipf’s law proposed by [5] that given many types of data could follow a Zipf-like distribution [6]. The frequency of the pattern is inversely proportional to its rank in the frequency list [7]. The occurrence of frequency for most frequent word is twice as high as the second most frequent word, and the occurrence of frequency for most frequent word is three times as high as the third most frequent word, and so on. Zipf’s law has used in the area of information retrieval [7]. Information retrieval is an activity that obtains information resources from a corpus, providing functionalities for users to find a particular subset of it by constructing queries. Zipf’s distribution would impact on the information retrieval process, especially preserving the medium frequency terms that have content-bearing [7].

2.3 Latent semantic analysis

Latent semantic analysis (LSA) is a method for natural language processing can analyze the relationship between articles and terms. In specific, LSA can apply the statistical computation to extract and represent the contextual-usage meaning of words from a large corpus of text [8]. The underlying idea of LSA is that all the word contexts have a set of mutual constrains that largely determine the similar meaning of words and sets [9]. Besides, LSA also can be regarded as an extensive method on implementing vector space model (VSM). There are various ways to apply LSA to reflect of human knowledge that has been established such as information retrieval, the construction of allo-graphic synonyms, search engine optimization.

3. RESEARCH METHODOLOGY

In this section, we present the mechanism of text mining analysis for effectiveness of green IT in CSR reports. Our current research is exploring alternative structure CSR disclosure analysis, to also identify cases where CSR disclosure should be address the green IT issues. We are also building a general CSR disclosure model with a text mining analysis for the effectiveness of green information technology in Taiwan. According the a methodology for design science research in information systems, this study involves the design of innovative analysis of CSR report of the use of text mining technique to understand the degree of CSR disclosure of firms for green IT. Figure 1 demonstrates the complete mechanism. In order to achieve the goal, we need to define the audit object first, and then follow the three phases below:

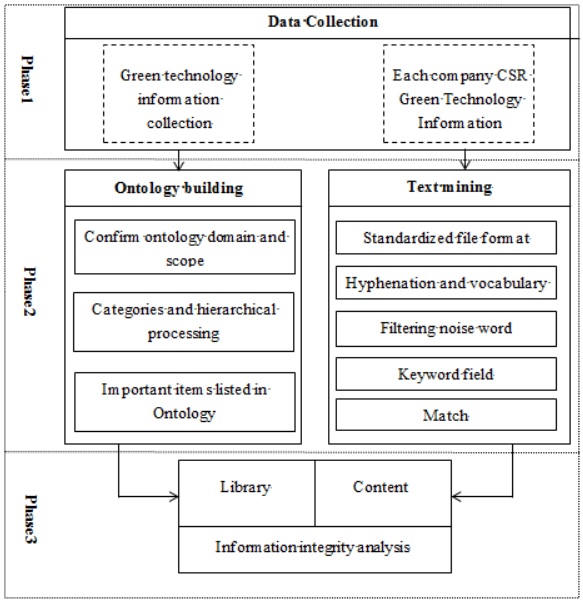

Figure1. The research mechanism of this study

Phase1: Data Collection

Step1. Identify the green IT related issues from literature, CSR reports and websites.

Phase 2-1: Ontology building

Step2. Confirm ontology domain and scope

Step3. Identify important items to the hierarchical processes

Phase 2-2: Text mining

Step 4 Standardized text format and construct hyphenation and vocabulary

Step 5 Filter noise words

Step 6 Extract key words for matching

Phase 3: Verification

Step 7 Evaluate the verification and validation

3.1 Data collection

3.1.1 Identify the green IT related issues from literature, CSR reports and websites.

The first step identifies the green IT related information of corporate social reports announced by prospects manufacturing industry market observation platform, which contains the corporate social responsibility information in high-tech industry in Taiwan. To achieve this goal, we need to browse the web site to find out the published information or obtain the original designed documents about the target process. Then, we depict the green indicators of carbon disclosure information in CSR reports. To build ontology later, we need to identify the green IT information in the platform website.

3.1.2 Confirm ontology domain and scope

Previous study of [12] conducts 27 measure indicators of green IT to examine four product life

cycles from CSR report by using ground theory. In this section, this study adopts 27 green IT indicators [12], which contain the multiple concepts for green IT, and defines green IT items based on these indicators for confirming ontology domain and scope. The similar meaning of green IT items is composed of a category for the process of ontology building.

3.1.3 Identify important items to the hierarchical processes

The third step identifies the relationship between different entities, and defines the each category of dimensions. First, we use the Chinese knowledge and information processing (CKIP) system which was developed by Academia Sinica of Taiwan to produce the minimum analysis unit in the sentence. Then, we use the TF-IDF weight technique to compute the score of each minimum analysis word. If the score of each word is above the average value of all words, we preserve the word as regard as the key entity of green IT. And then, point out the links of the similar entities as a category. Moreover, we identify the individual relationships with each category and find out hierarchy based on the green IT indicators conducted by [12]. Finally, we establish the “iskeywordof” and the “isParentOf” relationships based on the entities were extracted from TF-IDF weight technique.

3.1.4 Standardized text format and construct hyphenation and vocabulary

This section focus on capturing the green IT information from the CSR reports of Taiwan. In order to achieve goal, we need to parse the source from prospects manufacturing industry market observation platform. And then, we download CSR reports and other related green IT information, and transform their format into TXT format of each file. Third, we employ the CKIP system from Academia Sinica to identify the smallest unit in each sentence, and store them in our vocabulary database.

3.1.5 Filter noise words

This study refers to the Zipf’s law mechanism proposed by [11], getting the terms frequency and distribution, finding out the upper cut-off point and the lower cut-off point, and then we eliminate noisy words and get the significant words. We then use Excel VBA model remove some unnecessary parts of word such as adverb, unit, conjunction, interjection from those significant words. The rest words will be candidate key words and saved into candidate words database.

3.1.6 Extract key words for matching

This study use TF-IDF algorithm to compute the weight of terms to filter out common words and to preserve important words. And then, we the important words match the important items that extract from green IT appropriate key words through ontology building to get the appropriate words in the phase 2.

3.1.7 Evaluate the verification and validation

After identifying the appropriate words of the CSR reports, we need to compare them with the results of [12]. The first one is to confirm related data was extracted to our sample. The second one is to identify rule, and we compute the score of each key words for doing similarity analysis. The matching formula is: α∩β/β, α = the number of terms in a CSR report of a company through Zipf's Law. β = the total number of keywords which were included in an indicator. Third, we produce the green IT integrity score reports.

4. SYSTEM IMPLEMENTATION

To verify whether the mechanism can assist stakeholders to analyze and assure the content under the CSR environment, this research implements a prototype system. The prototype system is constructed to implementation our mechanism, and it contains five modules responsible for parsing the website, extracting the CSR report contents, translating into EXCEL, and exporting to PHP file respectively.

5. SYSTEM EVALUATION

5.1.The subjects of case company

We obtain their CSR reports in December 2012, and the data is collected from their CSR reporting website. The case company deals with high-technology industry and implements CSR reports for a long time. The case company has founded the structure and content problems. With CSR report structure and content are continuously being updated, the stakeholders often response too much information contained in the CSR report, such as differs form and content from each CSR reports. The stakeholders want to audit their CSR related information on their website to assure the correctness and completeness of the structure and content of CSR reports. Thus, we assess the confidentiality, integrity and availability.

5.2. System verification and validation

In this study, total 27 subjects to be our sample. According to the questionnaires, we propose the audit results to the case company. The precision rate = $\alpha \cap \beta / \alpha \cup \beta$, α = indicators were determined by the mechanism; β = indicators were determined by human judgment. The results can be summarized below:

Table 1. The result for verification and validation

Total number of subjects	Total indicators ($\alpha \cup \beta$)	Intersection indicators ($\alpha \cap \beta$)	Precision rate
27	729	519	0.71

In the final audit results, we find out the 0.71 precision rate for our mechanism. The participators the cases indicate that the system is useful for monitoring their CSR reports and detecting the potential green IT information.

6. CONCLUSION

In this thesis, we adopted text mining technique to establish an automatic mechanism for detecting the structure and content of the CSR report, and analyze the inconsistency between CSR reports with green IT information and announced related information on the website. We implemented it into a prototype system to test the feasibility of the proposed mechanism through a simplify text mining process, and applied the prototype system to the CSR reports to continuous analyzing CSR related information. The result also shows that the implemented system can correctly detect the possible inconsistency between two building indicators methods. In addition, we evaluated the verification and validation of our system through the real cases. The result indicates the system is useful for the case company's CSR reports and the system can also provide information for them to review CSR reports more effective and efficient.

7. REFERENCE

[1] S. Du, C. Bhattacharya, & S. Sen, Maximizing business returns to corporate social responsibility (CSR): The role of CSR communication, *International Journal of Management Reviews*, 2010, 12(1), 8-19.

[2] D. S. Dhaliwal, S. Radhakrishnan, A. Tsang, & Y. G. Yang, Nonfinancial disclosure and analyst forecast accuracy: International evidence on corporate social responsibility disclosure, *The Accounting Review*, 2012, 87(3), 723-759.

[3] S. Murugesan, Harnessing green IT: Principles and practices, *IT professional*,2008., 10(1), 24-33

[4] T. Xia, & Y. Chai, An improvement to TF-IDF: term distribution based term weight algorithm, *Journal of Software*,2011, 6(3), 413-420

[5] G. K. Zipf, Human behavior and the principle of least effort, Addison-Wesley, 1949.

[6] Q. Jiang, C.-H. Tan, C. W. Phang, J. Sutanto, & K.-K. Wei, Understanding Chinese online users and their visits to websites: Application of Zipf's law, *International Journal of Information Management*, 2013, 33(5), 752-763..

[7] S.-M. Huang, D. C. Yen, L.-W. Yang, & J.-S. Hua, An investigation of Zipf's Law for fraud detection ,*Decision Support Systems*, 2008, 46(1), 70-83.

[8] T. K. Landauer, and S. T. Dumais, A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge, *Psychological review*, 1997, 104(2), 211.

[9] T. K. Landauer, P. W. Foltz, and D. Laham, An introduction to latent semantic analysis, *Discourse processes*, 1998, 25(2-3), 259-284.

[11] Hsueh, H. Y., and Huang, S.-M. Information Retrieval on Socialized Peer-to-Peer Environment, PhD. Thesis, National Chung Cheng Univeristy, 2007

[12] Chu, W. C., and Hong, L. Y. The companies commitment of green information technology implementation in notebook computer industry, Journal of e-business, 2012,14(4): 723-742