A Visual Comparison of the Annual Reports of an IT Company Listed in Shanghai Stock Exchange Using Collocational Networks

Yuanqin Li
Shanghai Jiaotong University
Rancy@china.com

Guoqi Cheng
Industrial and Commercial Bank of China
Chengguoqi@china.com

Abstract

Traditionally, investors can only rely on a company’s annual financial statements (balance sheet and income statement, etc) to know whether the company is financially sound and effectively managed. However, annual financial statements may not be able to present an accurate picture about a company financial soundness, and sometimes, they may even mislead investors. Therefore, some other new methods are needed to supplement the current method of largely relying on annual financial statements. This paper uses the collocational networks, a method originating in corpus linguistics, as a tool for visualizing sequences of textual contents of annual reports of an IT company listed in Shanghai Stock Exchange of China, which is the first time to use such a new method to do a qualitative data analysis for an IT company in China. It compares the derived networks to identify the most significant differences between the documents in order to describe key issues that may not be revealed in a company's annual financial statements. The new methodology is introduced, its difficulties in applying it to Chinese language analysis are discussed and an example is given to show how this method works. Moreover, future studies are proposed for this new method to be effectively used in reality.

Keywords: Collocational Networks Analysis Method, Analysis of IT company listed in stock exchange.

1. Introduction
The annual reports of listed companies communicate financial and operational information to the outside. Traditionally, the financial data in these reports have received a lot of attention from investors, analysts, and regulators etc. However, the analysis of financial statements couldn’t completely reveal the real financial and operational status of a listed company because some listed companies prettified their quantitative data and mislead investors in the stock market. Even in USA, where one of the most developed stock market and the most perfect regulations have been built, financial scandals broke out now and then such as Enron scandal and the Worldcom scandal. In China with an emerging stock market of less than 20 years' history, financial scandals were found out almost every year such as Yinguangxia company in 2001, Lantian company in 2002, and Dongfang Electron in 2003. The annual reports of listed companies include both qualitative and quantitative data. Qualitative data “are management’s summary of the financial performance for the year, and as such, attempt to put in perspective the success or failure of the various initiatives of the company…..embody the corporate- speak representative of the top management of the company.” (Thomas 1997). Compared to the analysis of quantitative data, the analysis of qualitative data has been overlooked.
Therefore, in order to protect most investors’ interests and build a healthier and more effective stock market, the traditional analysis method, the financial statements method, is not enough. A new method is needed to supplement the financial data analysis, that is to analyze the language of the reports. The collocational networks analysis method is one of such new methods.

The collocational networks method originates in corpus linguistics. This method can examine which concepts are emphasized by the company in a particular report, and which concepts are most frequently linked to each other by revealing which words regularly appear within a certain window size to each other. (Magnusson Camilla et al. 2003)

There are a few English papers published overseas on this language analysis issue. Magnusson Camilla et al. (2003) visualized the sequences of texts in quarterly reports using collocational networks. Thomas Jane (1997) analyzed the meaning of the language in annual reports. Kendall J.E.(1993) classified the language of the chairman’s letters in annual reports by using the concepts of drama. However, no such a research has been done in China's listed companies so far. This paper will be the first one in China using collocational networks method to analyze annual reports of an IT company over a period of five years.

The main research issue for this study is: Using the collocational networks method to work out the networks for a sequence of annual reports from the Clever company, an IT company listed in Shanghai Stock Exchange, China, and then compare these networks to identify the most significant differences between the documents in order to provide a more accurate picture about the company's status.

The specific research questions are: Can the new collocational networks method disclose the most significant differences of the textual contents in annual reports of the company in China? If so, what specific implications of this research findings to China's financial market's regulation and management?

2. Literature Review

Magnusson et al.(2003) researched whether there was a systematic link between the financial data and the accompanying textual data in quarterly reports. Compared to other articles, this paper analyzed the quantitative data and the qualitative data at the same time using different methods, then found some patterns. Quarterly reports of telecommunications companies Nokia, Motorola and Ericsson from a period of two years were chosen as research material. The textual contents were turned into collocational networks, and the financial data of the reports were visualized using the self-organizing map. The authors concluded that the textual information of quarterly reports appears based on the selected companies to be a good predictor of the financial performance of the companies, with a time lag of one quarter between the change in the collocational network representing the text and a shift in the SOM model. The major change in the collocational network is in turn anticipated by a minor change in the preceding collocational network( Magnusson et al. 2003). The weakness of this paper is the material period is only 2 years, and It would be meaningful to study these companies over a longer period of time.

Thomas (1997) studied the development of one company over five years to look at the differences between “good news” and “bad news” annual reports as communicated through the manager’s message to the stockholders. During this period, as the company started to
experience more and more severe difficulties, the structure of the language used in the reports also changed. The result is that, as the news becomes more negative, linguistic structures suggest a factual, “objective” situation caused by circumstances not attributable to any persons who might otherwise be thought responsible (Thomas 1997). In this paper, Thomas only concentrated on one company. A study of a larger body of annual report managers’ messages will be able to supplement this weakness.

Stubbs(1995) studied collocations of words in corpora of data. Using data from corpora of up to 120 million words, the author have defined a method for identifying collocations, and shown that related methods discussed in the literature can be simplified in various ways. This paper presented three different formulae for studying collocations: O/E, I-value, and T-value. I-value, that is the frequency of co-occurrence of node word and collocate with the frequency of their occurrence independently of each other, was called MI score in Magnusson and Vanharanta(2003)’s article.

Magnusson and Vanharanta(2003) presented the collocational networks, a method originating in corpus linguistics, as a tool for visualizing sequences of texts. Unlike Stubbs(1995)’ paper, this paper explained clearly how to create collocational networks. Then, as a case study, a sequence of quarterly reports of Ericsson was turned into collocational networks for visual comparison over different years.

Kendall(1993) analyzed the chairmen’s message in annual reports using dramatism method. God terms and devil terms were classified and used to discover and interpret corporate dramas existing in the language of the message of the Dow Jones Industrials. Words belonging to God terms represented good news in the eyes of the company. Words in Devil terms, on the other hand, brought bad news. The drama method used by Kendall was interesting.

3. Research Data and Methodology

3.1 Data
Annual reports from an IT company were chosen as research material, and the period span was from 1999 to 2003. The company is Heilongjiang Clever Net Corp., Ltd. (abbreviated to Clever). Clever is listed on Shanghai Stock Exchange, making it one of only a handful of private high-tech enterprises to go public in China. E-commerce and web-based education, two of its core businesses, helped it become one of the leading software companies in China. Our research will construct the collocational networks for Clever at the basis of selected data, then compare these networks and find out whether this new collocational networks method can identify the most significant differences of sequences of textual contents in annual reports of the company in China.

3.2 Collocational networks
“Collocational networks are two-dimensional networks which contain interlinked collocations, i.e.” words which occur together in a text.” (Magnusson and Vanharanta 2003). A collocation is the occurrence of two or more words within a certain window in a text. The window size in analyzing textual contents of annual/quarterly reports is usually 4:4, i.e. four words to left or right. Significant collocation is an important concept in this method, which is
measured using the Mutual Information or MI score. The MI score, an information theoretic concept introduced in linguistics by Church & Hanks (1990), compares the frequency of co-occurrence of node and collocate with the frequency of their occurrence independently of each other. (Magnusson and Vanharanta 2003).

\[
MI(n,c) = \log_2 \frac{f(n,c)N}{f(n)f(c)}
\]

Where \( n \) stands for node, \( c \) for collocate and \( N \) for the size of the text; \( f(n,c) \) stands for joint frequency of node and collocate, \( f(n) \) and \( f(c) \) for their independent frequencies. The MI score is symmetric, that is the calculation of the score do not take into account the order in which the words occur in the text, since \( f(n,c) \) has the same value as \( f(c,n) \).

The text size in this study consists of approximately 1500 words each, and an MI score of 4.00 is suitable for these texts in order to produce the networks, that is all collocates belonging to a nuclear node with an MI score equal to or larger than 4.00 should been placed on the network. In addition, in the network the frequency of a nuclear node (i.e., \( f(n) \)) should equal to or larger than 9, and the joint frequency of a nuclear node and a collocate (i.e., \( f(n,c) \)) should equal to or larger than 3. Therefore, the necessary conditions for a collocational network includes:

1. \( MI(n,c) \geq 4.00 \);
2. \( f(n) \geq 9.00 \);
3. \( f(n,c) \geq 3.00 \).

The construction of a collocational network needs to locate a nuclear node and its collocates. Before the construction of the networks, a software is needed to calculate MI score. The construction starts with the locating of the most frequent word in the text. It is given the status of central nuclear node in the network. Then, the collocates are traced, and these collocates are then treated as nodes, the collocates of which are traced. This process is repeated until three necessary conditions above cannot be satisfied. In some cases, besides the main networks, there are other networks.

This method can produce collocational networks for the textual contents of annual reports. These networks consist of the most central concepts in the texts and reflect the connections between the concepts, which provide a new method for the public to analyze the annual reports by comparing these networks.

### 3.3 The difference between English words and Chinese characters

According to Magnusson and Vanharanta (2003), a collocational network consists of nuclear nodes and collocates. In the context of English words, a nuclear node is actually a single word. Similarly, a collocate is a single word too, for a single word in English does make sense. However, a single Chinese character usually means nothing. For example, the counterpart for “we” in English is “wo-men”, two words in Chinese. Therefore, compared to the method introduced by Magnusson and Vanharanta (2003), the method in this paper needs one more part: words segmentation, that is to identify words that usually occur together and represent one meaning in the annual reports.

For the purpose of this study, some preliminary measures have to be carried out as follows:

1. Low-content words such as prepositions, articles, conjunctions etc. are moved out.
(2) All tables in annual reports are removed.
(3) Words segmentation:
Characters that always occur together and represent one certain meaning in the annual report are considered as a phrase;
A phrase is separated from another phrase (to the left or/and to the right) by a blank so that the software can identify \( f(n,c) \), \( f(n) \), and \( f(c) \) consistently and precisely. So, a nuclear node or a collocate in a network is a phrase, not a single word.
It should be pointed out that the aim of this method is to find out what is central in these annual reports, so some collocations may be not typical for business language in general.

3.4 an example
The material of this example comes from the annual report of Clever in 2002. After carrying out the preliminary measures described in section 3.3, we get analyzable qualitative data. Then, the construction of the main network starts with the locating of the most frequent phrase in the text, that is “sell” (“xiao-shou” in Chinese). “Sell” is the nuclear node in the main network, and it has three collocates. Moreover, one of its collocates, “mode” (“mo-shi” in Chinese), has two collocates of it own. In addition to the main network, there are four other collocational networks, and the corresponding four nuclear nodes are “software”, “manage”, “education”, and “decrease”. “Decrease” has four collocates, and one of its collocates is “income”. Figure 1 contains the collocational networks for Clever’s annual report 2002, in which each Chinese phrase is translated into English word.

4. Future Research
The current status of this study is to calculate the MI score for all phrases in selected annual reports. The next step is to construct the collocational networks for each annual report. Finally, this study will analyze and compare these networks.
If the hypotheses can be verified, that is the collocational networks method can disclose the most significant differences of the textual contents in annual reports of the company in China, this will provide a new tool for analyzing annual reports of companies. Then, the future research should concentrate on the relationship between changes in the texts and changes in the financial performance. The combination of the analysis to the two kinds of data in an annual report will increase the capability of the stock market to protect investors and will help build a more effective stock market in China. Further, this new method may also provide CIOs (Chief Information Officers) of companies with another means to identify key strategic issues and do a better alignment between business strategies and IT strategies for their companies.

5. References
Kendall, J.E. “Good and Evil in the Chairman’s “Boiler Plate”: an Analysis,” *Organization Studies*, (14), 1993, pp.571-592.
Magnusson, Camilla., and Hannu, Vanharanta. “Visualizing Sequences of Texts Using
Collocational Networks,” University of Helsinki, Finland. Tampere University of Technology, Finland, 2003.


Figure 1

- **Software** (14)
  - **Design(6)**
  - **Product(12)**
    - **Market(11)**
    - **Popularize(5)**
  - **Prospect(5)**

- **Manage** (14)
  - **Income(3)**
  - **Corporation(4)**
  - **Profit(4)**

- **Education** (13)
  - **Families(5)**
  - **Domain(5)**

- **Decrease(10)**
  - **Income(4)**
  - **Operation(5)**
  - **Mainly(5)**

*Independent frequency of word indicated in brackets*

*MI-score indicated on line connecting words*