ISSN: 2688-9323

Research on it Enterprise Annual Report based on Text Mining Technology

Yajun Zhu

Anhui University of Finance and Economics, Bengbu Anhui, 233030, China

2460514562@qq.com

Abstract

Since the 21st century, due to the rapid development of science and technology, the amount of data is blowout growth, data mining is also rising. As a direction of data mining, text mining can sort, extract, filter and understand the natural language of many different documents, which has good research significance and commercial value. The annual report of listed companies is more formal and informative than other public opinions as an authoritative document reflecting the operation of enterprises. At the same time, the annual report is also one of the most accessible corporate information for investors, which can provide reference information for investors to make decisions, it is an important way for stakeholders to understand the profitability, operational capability and future development of each company. Therefore, the text mining of annual reports of listed companies has become an important means for people to understand the operation of listed companies. In this paper, the annual reports of it enterprises are studied, the information contained in the annual reports is extracted by text mining technology, and the extracted information is analyzed, thus better help stakeholders to understand the various aspects of the enterprise, and for the enterprise how to enrich the content of the report made reference to provide policy recommendations.

Keywords

Company Annual Report ;Text Mining; It Enterprises.

1. Introduction

In the information age, a large number of data growth: Enterprise Intranet, Telex News, forums, e-mail and other information are increasing rapidly, directly leading to information overload. While the amount of data has increased, it has been difficult to obtain the available information. In the face of innumerable network information, how to realize the efficient screening of useful information for themselves has become the focus of technology research and development in the information age. The emergence of data mining has solved the problem of discovering useful information from a large amount of data. But the data is not all structured, for the unstructured and semi-structured data, the original data mining algorithm can not be processed, Text Ming came into being, can search and sort multiple different documents, and so on, have Higher commercial value. As a fundamental part of the information technology industry in our country, the stakeholders always regard their annual reports as one of the important references for investment, such as shareholders, investors, potential customers, banks, enterprises need talent, etc. , through the annual report of the relevant enterprises to obtain more information, which often affect their judgment and decision-making. However, as the number of IT enterprises increases, it becomes more difficult to obtain useful information from the annual reports. For investors, much of the information in the annual reports of relevant enterprises is useless, it takes a lot of time and effort to find the information they need. Therefore, how to

ISSN: 2688-9323

improve the efficiency of stakeholders to obtain useful information of annual reports of enterprises has become a hot issue of current research.

2. Overview of the Article

2.1. Text Mining

In the late 1950s, H. Karen. Luhn proposed the method of combining word frequency statistics with automatic classification. Since then, the theory and method of text mining have been put forward In the 1960s, Maron made some achievements on automatic classification on the basis of Luhn's research, and since then, text mining has gradually become one of the hot topics and has been widely used in various fields. Barbosa and others used wavelet analysis to eliminate useless trivia words in social text, which made the accuracy higher ORDENES and others studied the text of the pick-up Service based on customer feedback by text mining method. They subdivided the parking service into stages, and then analyzed each stage separately, in order to realize the whole analysis of the whole parking feedback process.Different from the development of foreign countries, the research on text mining technology in our country has been carried out late, but with the progress of research, some achievements have been made. Wu Tao and others have defined a thesaurus that preserves the typical characteristics of a word, then defined a special storage structure for this thesaurus, and used the Bayesian statistics method to solve the problem of new words, adopting the inverse maximum matching algorithm for word segmentation, therefore, a brand-new word segmentation algorithm is proposed, which makes great progress in part-of-speech recognition and word segmentation accuracy, after preprocessing the text data, the SVM model is used to classify the traffic public opinion automatically, and then Apriori Algorithm is used to analyze the public opinion of the data source through association rules, then the traffic problems reflected by public opinion and their changes with time are mined by co-occurrence network analysis, and the reliability of the method is proved by empirical analysis

2.2. Research on Annual Reports of Enterprises

Antonina Kloptchenko, et Al. used both data mining and text mining methods to compare the results. The researchers found that the financial ratios contained much less information than the text, furthermore, the researchers conclude that the future development of the enterprise can be obtained from the text information of the annual report, while the financial ratio of the annual report can only be used to see the development of the enterprise in the past, according to the high-frequency statements in the annual report of the enterprise to quantify the tendency to make a prediction of bankruptcy. The researchers used the annual reports of 90 listed companies in Japan as data sources and found that there were some differences in the high-frequency words and sentences between bankrupt companies and normal companies, words such as”Development”, “Exploration”, “Investment”, “New business” and words such as ”Dividends” and”Retained earnings” appear more frequently in the annual report of a normal enterprise, which means that the enterprise has sufficient funds, enough for the next step of Enterprise Development and growth to make financial support, and this is very different from the bankruptcy enterprise. Lin Zhonggao and other researchers first studied the relationship between risk cue information in annual reports of a-share listed companies and bank credit decision-making by text mining method, the regression model between word frequency and the net increase of bank loans in the current period is established. The results show that the more information about risk in the annual reports, the less bank loans the listed companies get, banks tend to pay more attention to non-financial information, such as risk, rather than financial information, including audit opinion and internal control opinion, when making credit decision, and this kind of attention mainly focuses on long-term loan decision Li Changqing, et Al, used regression analysis to study the disclosure quality of the information in the part

ISSN: 2688-9323

of"Management discussion and analysis" in the annual reports of enterprises, mainly by questionnaire, and set up appropriate scoring standards for it. The results show that there are some problems such as poor text comprehensibility and low information quality in the part of"Management discussion and analysis" of annual reports of listed enterprises in our country.

3. Research Methodology

3.1. Text Segmentation

Chinese word segmentation is a basic step of Chinese text processing and a basic module of Chinese man-machine natural language interaction. Different from English, there is no word boundary in Chinese sentences, so when processing Chinese natural language, it is necessary to segment words first, and the effect of segmentation will directly affect the effect of parts of speech, syntax tree and other modules. Of course participle is just a tool, different scene, different requirements. In the man-machine natural language interaction, the mature Chinese word segmentation algorithm can achieve better natural language processing effect, help the computer to understand the complex Chinese language. In the process of constructing Chinese natural language dialogue system, Bamboo inter-intelligence trains a set of algorithm model with better segmentation effect, which lays a foundation for the machine to better understand Chinese natural language.In this paper, Python software for enterprise annual report text word segmentation, word segmentation of the data for the next step.

3.2. Data Visualization

Data visualization is a scientific and technical study of the visual representation of data. The visual representation of the data is defined as the information extracted in a summary form, including the attributes and variables of the corresponding information units.In this paper, the data after word segmentation are presented and analyzed by data visualization software.

4. Empirical Analysis

4.1. Data Collection and Collation

This part is based on the text data of Huawei Technologies Co. Ltd. Annual reports from 2006 to 2019. First, get the text of Huawei's annual report from Huawei's website. Secondly, the content analysis framework of the annual report of IT enterprises constructed in the third part is used to analyze the annual report text of Huawei. Finally, the Framework Transforms Huawei corporate annual report text from unstructured data to structured data to derive the information needed by Huawei stakeholders.

4.2. Text Segmentation

Table 1. High-frequency words in the management speech

2006	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
customer	customer	Global	business	customer	Huawei	join	customer	customer	Huawei	intelligence	intelligence
service	telecom	business	customer	join	ICT	network	operate	business	customer	customer	tech
operate	network	customer	network	network	join	industry	trans	enterprise	business	world	customer
business	terminal	strategy	Global	Huawei	world	service	number	Global	number	Huawei	industry
partner	service	Global	operate	world	Global	business	Huawei	number	Global	industry	partner
network	trans	telecom	business	Global	RMB	product	Global	intelligence	standard	number	Huawei
supply	cooperation	increase	industry	Conduit	strategy	business	trans	consumer	world	ability	number
demand	environment	environment	manufacturer	enterprise	enterprise	number	network	platform	number	Global	country
org	business	lead	broadband	turbulence	Foam	staff	terminal	operate	supply	trans	enterprise
sale	programme	cooperation	steady	terminal	country	industry	Huawei	join	society	company	trans
times	leader	programme	success	partner	company	ideal	human	terminal	world	calculation	Basics

ISSN: 2688-9323

This paper divides the words of management speech and corporate social responsibility respectively, and presents the high-frequency words after the word segmentation with Excel data.

Table 2. Corporate social responsibility part of high-frequency words

2006	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
Huawei	Huawei	Huawei	Huawei	Huawei	Huawei	Huawei	Huawei	Huawei	Huawei	Huawei	Huawei
staff	development	network	green	Global	network	ICT	Global	supply	number	Global	staff
Africa	development	signal	network	development	development	development	development	customer	development	supply	Global
Global	enterprise	green	customer	supply	supply	staff	staff	staff	supply	green	development
green	supply	supply	signal	network	ICT	society	country	network	Administration	staff	supply
RMB	society	Administration	green	staff	staff	programme	supply	energy	healthy	ICT	epidemic
network	country	society	environment	ICT	energy	energy	SDGs	green	network	environment	energy
operate	fund	network	law	operate	environment	world	environment	signal	environment	EHS	ecology
energy	environment	countryside	guarantee	customer	world	environment	signal	resources	innovate	children	ICT
society	number	supply	life	number	operate	join	strategy	Emission	power	number	green
internet	countryside	guarantee	network	network	network	energy	environment	education	risk	number	human
environment	society	society	country	network	medical	environment	CSR	EHS	intelligence	green	environment

4.3. Data Analysis

4.3.1. Analysis of the Management Speech

According to the distribution of high-frequency keywords in each year, there are some identical words in the high-frequency keywords in each year. 'customer' occupies the top five places in the high-frequency keyword list in all years except 2005, this is closely related to Huawei's corporate goal of “Customer Focus and value creation for customers” and demonstrates Huawei's commitment to the service concept of creating value for customers through innovation. Words such as 'global' and 'world' were among the top 20 most-used keywords in all the years 2006-2020. Globalisation has always been part of Huawei's development strategy, with sales in Huawei reaching RMB65.6 BN, of which overseas sales account for more than 65 per cent, Huawei's globalisation strategy has scored an early victory, with Huawei overtaking Ericsson in 2013 to become the world's largest maker of communications equipment and, in 2020, huawei has overtaken Samsung to become the world's largest smartphone maker. These figures and facts show that Huawei's path to globalisation is already strong and will continue to be so, and that Huawei will be a very reliable partner for other companies and institutions abroad.According to the statistical distribution of high-frequency keywords in each major year, there are also great differences and differences in some keywords of the management speeches in each major year, and from the changes of words existing in each year in these keyword lists, we can extract the required information from it for stakeholder decision-making. From 2006 to 2010, 'operators' were consistently in the top 20 of the high-frequency Keyword list. By 2009, Huawei had served 45 of the world's top 50 telecom operators. Huawei's business was dominated by operators, huawei has always maintained an open mind and actively worked with industry chain partners to create a win-win business ecosystem with operators at its core, continuously creating long-term value for its customers. Since 11 years ago, the term 'operator' has rarely appeared in high-frequency keywords. This is mainly due to the fact that since 10 years ago, the telecom industry has been at a new starting point in the global digital era, huawei's shift from a carrier-focused business to one focused on operators, businesses and consumers is growing at a time when there are huge opportunities for the telecoms industry to grow by connecting broadband and digital, the decision to set up four operations centres -- networks of operators, enterprises, terminals and others -- to make their objectives clearer, their levels of organisation and their management processes simpler is one of the most significant in Huawei's history, has Important Implications for Huawei's development.

ISSN: 2688-9323

4.3.2. Analysis of Corporate Social Responsibility

Looking at the distribution of the top 20 high-frequency words in the social responsibility (sustainable development) section of Huawei's annual report for 2006-20, the proportion of the same words is relatively large, and some of the words remain among the high-frequency words, huawei Huawei has always adhered to and implemented certain ideas or measures on corporate social responsibility and sustainable development. From the high-frequency word distribution word list, always maintain in the top or even the first three words are 'staff' , 'sustainable development' , 'green' and so on. The frequent use of the word “Employee” reflects Huawei's 'striver-oriented' corporate culture, which means that Huawei attaches great importance to the physical and mental health of its employees and the harmony of the internal organizational atmosphere, huawei insists on the core idea that a motivated, motivated workforce is the company's most valuable asset. The term 'sustainable development' first appeared in international documents in 1980, when the International Union for Conservation of Nature (IUCN) formulated the World Programme for Conservation of nature, the formal definition of the model and concept was put forward in the report”Our Common Future” published by the World Commission on Environment and Development in 1987. The frequent use of the term “Sustainable development” shows that Huawei, as a company, has always adhered to the strategy of “Sustainable development” in the face of environmental degradation and a relative shortage of resources per capita, and adhere to the principle of equity, sustainability, the principle of commonality, and strive to achieve sustainable economic development, ecological sustainable development, social sustainable development. The term “Green” corresponds to Huawei's “Green Huawei, green communications, Green World” environmental agenda, which means that Huawei takes environmental requirements into account in its product development and commercial activities, huawei is trying to achieve environmental protection and energy conservation through technological innovation by complying with international environmental regulations and incorporating words such as “Energy conservation” and “Emission reduction” from the glossary.In 2010, the word 'female' appeared on the high-frequency list. This year, Huawei held its first conference on the theme of “Growth, enterprise and excellence” for female employees, which indicated that Huawei will continue to strengthen the vocational training for female employees, in 2013, 'mobile phone' appeared in the glossary, a year when the French team in Huawei and two French companies collaborated to set up a mobile phone recycling platform in France, encouraging consumers to replace their old phones with new smartphones, which will be recycled, a program that has survived to this day and helped reduce Electronic waste, in 2015, there was an 'earthquake' on the high-frequency list, the year of the 8.1-magnitude earthquake in Huawei, heavy losses were caused to the lives and property of the local people, and the communications infrastructure in the affected areas was severely damaged. In the face of the earthquake disaster, Huawei took emergency action to ensure the operation of the network equipment in the disaster-stricken areas through a series of measures, and provided security for the communication in the disaster-stricken areas, it shows Huawei's ability to do business reliably and to respond to emergencies. In 2020, there was an 'outbreak' in the high-frequency list. In 2020, the global covid-19 pandemic posed a major threat to people's physical and mental health, huawei moved quickly to develop a special epidemic prevention programme for countries and subsidiaries hit hard by the outbreak, demonstrating Huawei's determination to stand firm in the face of the epidemic and weather the storm together.

5. Conclusion

According to the results of this study, Huawei companies generally maintain a good balance of development trends, and the company has a good response capacity in the face of the covid-19

ISSN: 2688-9323

pandemic and the Wenchuan County earthquake and other major disasters, so Huawei is a big IT company that investors can trust.

Acknowledgments

Anhui University of Finance and Economics Postgraduate Research and Innovation Fund Project "Research on IT enterprise annual report based on text mining technology" (Project Approval Number: ACYC2021437).

References

[1] Luhn H P. Auto-encoding of documents for information retrievalsystems[M].Modern Trendsin Documentation. New York:PergamumPress,1959.

[2] Ordenes F V, Theodoulidis B, Burton J, et al. Analyzing customer experience feedback using text mining: a linguistics-based approach[J]. Journal of Service Research, 2014, 17(3):278-295.

[3] Sun Y, Wang Z, Zhang B, et al. Residents' sentiments towards electricity price policy: Evidence from text mining in social media[J]. Resources Conservation and Recycling, 2020, 160:104903.

[4] DE Brown. Text Mining the Contributors to Rail Accidents[J]. IEEE Transactions on Intelligent Transportation Systems, 2016, 17(2):346-355.

[5] Feldman R, Hirsh H. Finding Associations in Collections of Text[J]. machine learning & data mining, 1997.

[6] Ponomariov B. Government-sponsored university-industry collaboration and the production of nanotechnology patents in US universities[J]. Journal of Technology Transfer, 2013, 38(6):749-767.