UNIVERSITY OF TURKU

ABSTRACT

	Bachelor's thesis
x	Master's thesis
	Licentiate's thesis
	Doctoral dissertation

Subject	Information System Science	Date	4th April, 2022
Author(s)	Lanmei Du	Number of pages	61
Title	Feature Extraction of Format of Corporate Social Responsibility Reports		
Supervisor(s)	Prof. Reima Suomi		

As an important bridge that transmits signals to the market, the corporate social responsibility report has been an important theme among stakeholders in evaluating organizational efficiency and performance. Its quality will affect the decision-making and judgment of investors, thereby affecting the response of the capital market.

In recent years, social responsibility reports have shown a trend of rich pictures and diverse text designs. On the one hand, social responsibility reports composed of black and white text can hardly meet the information needs of different readers. On the other hand, reports with rich pictures and various text designs can give information users a better reading experience.

However, due to the lack of uniform standards on the format of social responsibility reports, there is still a lot of space for the research on the format of social responsibility reports. This thesis explores how to effectively extract the features of social responsibility reports, so as to further explore the feature dimension, set feature indicators and analyse feature data.

Firstly, existing studies have proved that the format of the social responsibility report has an impact on the stakeholder's impression management, and the way of impression management is mostly based on the design of picture arrangement, text arrangement and page structure.

Secondly, the guidelines of the Shanghai Stock Exchange (the guidelines on social responsibility reports issued by the Shanghai Stock Exchange), the GRI (Global Reporting Initiative) standards take the format as an important criterion for evaluating the quality of social responsibility reports. Based on this, interviews were conducted with five researchers in the field of social responsibility.

Then, a social responsibility report was taken as an example to conceptualize the feature annotation to get the two-dimensional data. Since CSR has different year data, different page data and different feature data, therefore, this thesis used Seaborn graph software to describe the three-dimension feature data.

Finally, this thesis takes a report of Vanke as an example to compare the social responsibility reports of Vanke in different years, and the comparison of the layout format of the real estate industry by Vanke and the financial industry by PingAn Bank.

Key words	Corporate social responsibility report; Format; Feature extraction

FEATURE EXTRACTION OF FORMAT ON CORPORATE SOCIAL RESPONSIBILITY REPORTS

Master's Thesis
in Information Systems Science

Author(s):
Lanmei Du

Supervisor(s):
Prof. Reima Suomi

04.04.2022
Turku

The originality of this thesis has been checked in accordance with the University of Turku quality assurance system using the Turnitin Originality Check service.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

1.1 Research background

After nearly a century of research and discussion, corporate social responsibility (CSR) has become an important topic of study in economics, management, law and other disciplines today. Good corporate social responsibility is the basis for sustainable development, and the fulfilment of CSR by companies not only improves their corporate image but also creates value for them. Inspired by economic interests and policies, CSR has become an important strategy for companies to attain better economic interests and corporate image, and the number of social responsibility reports has been rising year by year. In China, 2,030 CSR reports were released in the year of 2019, compared to 32 in 2006, which shows an increase of about 60 times (Zhao, 2020). With the number of CSR reports rising each year, the presentation of social responsibility reports has been characterized by diversification, visual appeal and better production.

At the early stage of social responsibility reports development, the length of social responsibility reports varied greatly from company to company, and the content of the reports was mainly text-based, with less background images, font settings and picture charts in the reports (Gao et al., 2012). Progress toward the integration of scattered CSR reports is necessary, because stakeholders need to retrieve information from reports. In recent year, it has been difficult for a black and white format report to meet the information needs of readers today. The graphical presentation of social responsibility reports can enhance the vividness and readability of social responsibility reports (Zhang, 2016), and since companies pay more attention to the individual needs of different information users, CSR shows a trend of graphics and text, a variety of formats and a variety of picture background settings.

As Nie (2019) points out, to help understand the CSR development and complexity of a certain company, visually pleasing and well-produced report with graphics are necessary, he also agrees that graphics can enhance the reading experience of information users. For the dissemination of information in social responsibility reports, reports with pictures can increase the rate of information dissemination, and for the same social responsibility report, information users from different cultural backgrounds receive information differently. For example, GF Securities has inserted pictures in its social responsibility reports in recent years, and used tables to represent important text and figures. The design of the format reflects the company's culture and philosophy, and also enhances employees' sense of belonging and motivation (Fang et al., 2019).

Unlike the financial information disclosure of enterprises, CSR information disclosure has the problems of not having to follow the rules of standardization and quantification, being flexible in content and form, having less substance, lacking actual case support,

and being easily manipulated by management, all of which leads to a lower level of quality and decision value of CSR information (Song, 2009). According to Bakar et al. (2012), CSR reports disclosure are often considered closely tied but with distinct conceptions. However, both templates and contents converge heavily throughout the years, and such homogenization of contents, also known as “cloned social responsibility reports”, has raised concern. This phenomenon shows that some enterprises publish CSR reports only to cope with the superficial behavior of stakeholders, but do not truly and objectively record and disclose the actual fulfillment of social responsibility by enterprises. Even though some reports include shareholders' rights, employees' rights, community welfare and environmental protection. However, most of the companies only mention them in passing, without in-depth elaboration and without substantive factual examples and data support. This makes the credibility of the report greatly reduced. In response, the Shenzhen Stock Exchange, the Shanghai Stock Exchange, the International Organization for Standardization (ISO), the Global Reporting Initiative (GRI) and other regulatory authorities and securities markets have issued a series of CSR-related standards and guidelines in recent years.

Even with the regulation in principle, however, social responsibility reports are not objective due to the characteristics of the narrative language. The disclosers have the opportunity to manipulate descriptive information through readability, report topics, rhetoric, language forms, pictures, charts and performance attribution (Sun, 2005). Relying on existing standards and guidelines can only prescribe the surface form of the report, and cannot promote the reliability of information. When there is a situation where a third-party audit system (Song, 2009) is absent, the comparison and verification can be very hard, therefore, it is difficult to test the reliability of social responsibility reports.

Impression management is usually a conscious and purposeful act of information management (Zhao, 2007). Companies are motivated to manage their impressions of social responsibility reports in order to maintain their image and to send good signals to the outside world and information users (Shen, 2010). Many existing studies have shown that this purposeful manipulative behavior can be achieved through formatting features such as the length of the report and the disclosure of quantitative information and pictures (Shen, 2007). Building on previous scholarly research on impression management, Huang et al. (2016) extended corporate impression management from financial reporting to social responsibility reporting, arguing that firms adopt social responsibility reporting impression management strategies for different motives, such as improving corporate performance and competing for scarce resources. Companies use self-interest attributions (Jiang et al., 2008) and design charts in social responsibility reports (Wang, 2020) to manage the overall impressions of the company created by the target audience, the media and the public. Some studies have also confirmed that photographs inscribed with company features inserted in company annual reports do influence investors' perceptions of the company (Beattie et al., 2008). A study by scholar on the use of graphics in UK social

responsibility reports found a clear tendency for companies to manage impressions in the use of background images (Giovanna, 2011). As impression management differs between textual formats and different settings of picture backgrounds, studying the features of picture presentation formats can help to avoid differences in decision makers caused by differences in the way information is presented.

In the past, most social responsibility reports were manually marked and evaluated. With the increase in the number of social responsibility report disclosures, the formatting of social responsibility reports has become diversified. Studies have demonstrated that the format of social responsibility reports has an impact on the judgments and decisions of different information users (Yang, 2016). This diversity brings inconvenience to investors' analysis and inter-company comparison (Nie, 2019). A few scholars have already focused on non-financial accounting texts and conducted research on the semantic aspects of disclosure (Giovanna, 2011). She has confirmed that differences in representation can affect the effectiveness of impression management. However, constrained by technical means and the level of theoretical development, there was no way to effectively analyze the features of the dimensional features of social responsibility report presentation.

Based on the existing research, this thesis explores how to effectively measure and extract the features of social responsibility reports. Then further explore the feature dimensions, set feature indicators and analyze feature data. Based on the sample of social responsibility reports issued by Chinese companies, this thesis explores whether the design of the presentation format in social responsibility reports plays a positive role in enterprises' meeting stakeholders' information needs. The study is based on the sample of CSR reports published by companies in China in 2018. The finding of this thesis is also expected to enrich the current study of feature dimensions on CSR report and to contribute to the ongoing automated representation of report presentation and format features. At the same time, the automated approach saves a lot of labor costs in the field of analyse social responsibility and provides new methods and ideas for the effective representation of report presentation format features.

1.2 Research content

This thesis divided a report into page to do research and these reports mainly consists of Chinese. Since the feature data differs on different page, this thesis uses different page data and different feature data to form a two-dimension feature data. Then, since one company will publish CSR every year, this thesis used different year, different page and different feature data to build a three-dimension feature data. After building the three-dimension feature data, the layout format of overall change and year changes can be explored. Due to the technique limitation, after many times of experiments and failure, this thesis takes feature data as text layout features, page layout features and picture layout features, the measured data are picture location, picture type, and text density. There may be another important dimension in CSR, however, technique for now can only do research

in picture location, picture type, and text density, therefore, this research will mainly focus on these feature dimensions.

Firstly, the evaluation of the layout format of the corporate social responsibility report and related literature are reviewed and analyzed to explore the feature dimensions of the layout format of the corporate social responsibility report. Secondly, according to the quality evaluation in the social responsibility report, the feature indicators (picture position, picture type, text density) of the format of the social responsibility report are established, and then the feature extraction method of this thesis is explained. For extraction method, since a report is composed of multiple pages, the two-dimensional data is formed of the text layout features and picture layout features of each page. Since the social responsibility reports of the same company in different years are different, the features of the social responsibility reports of the same company in different years are combined to form three-dimensional data. Finally, for the feature extraction method proposed in this thesis, this study takes Vanke as an example, and proposes the application of the feature extraction method in the same company, the application in the same company in different years, and the application in different companies in the same year, so as to make a horizontal and vertical comparison to explore the industry features and type features of the social responsibility report layout. The full thesis consists of six chapters:

Chapter 1: The introduction mainly clarifies the purpose and significance of this study; introduces the technical route and research methods of this research, finally explains the research innovations of the thesis.

Chapter 2: The theoretical background mainly reviews the related theories. This chapter provides an in-depth understanding of the literature from two relevant theories on impression management in social responsibility report charts and the evaluation of CSR reports formats, and then exploring the relevant literature.

Chapter 3: The format dimension evaluation mainly conducted an analysis on feature dimensions and feature indicators of the layout format. This chapter analyzes the evaluation of the quality of the layout format of the social responsibility report; takes Vanke's social responsibility report as an example, analyzes the layout format of Vanke. The feature dimensions of this study are established by combining examples, theories and practical evaluations.

Chapter 4: The methodology mainly discusses feature extraction method of layout format. This chapter measures text density, picture position and picture size. For the feature points measured by feature indicators, this chapter combines different years and different industries using the heat map method to perform dimensionality reduction comparisons on three-dimensional data.

Chapter 5: The application explores the application of feature extraction method in a report. Vanke and PingAn Bank are used as examples to explore the application of industry layout features. Since CSR has different year feature, three-dimensional feature data were proposed in this chapter.

Chapter 6: The summary mainly explains the deficiencies found in the research of the format and the directions for future exploration.

1.3 Technical route and research methods

1.3.1 Technical route

This thesis proposes a method to effectively extract the format of social responsibility reports with the help of image processing technology in the management field. Since CSR is an important research area in management. At the same time, computer can be a useful tool in management, therefore, this thesis use Python to do research on management field. In order to explore the feature dimensions of the layout format, establish the feature indicators of the layout format and extract the data of the feature dimensions of the layout format, this study starts from the impression management of social responsibility reports, graphic representation methods and the evaluation of the layout format of social responsibility reports.

After extracting the features of the layout format with the help of image processing technology, the feature dimensions of the layout format of social responsibility reports are determined as text layout features, image layout features and page structure features, among which the indicators of the layout format are established as text density, image size and image position; then each social responsibility report is paginated, and the text layout information, image coordinates and image size were feature extracted in each page. The images are then analyzed by the text layout information, image coordinates and image size in the format of each page, and combined with the same report of different page numbers to form two-dimensional data, also, combined with each report of different years to form three-dimensional data.

For the final three-dimensional data with different page numbers and different years, the thesis proposes a method to downscale the data to form a sample for data analysis and comparison, using Vanke as an example. The technical route is shown in Figure 1-1.

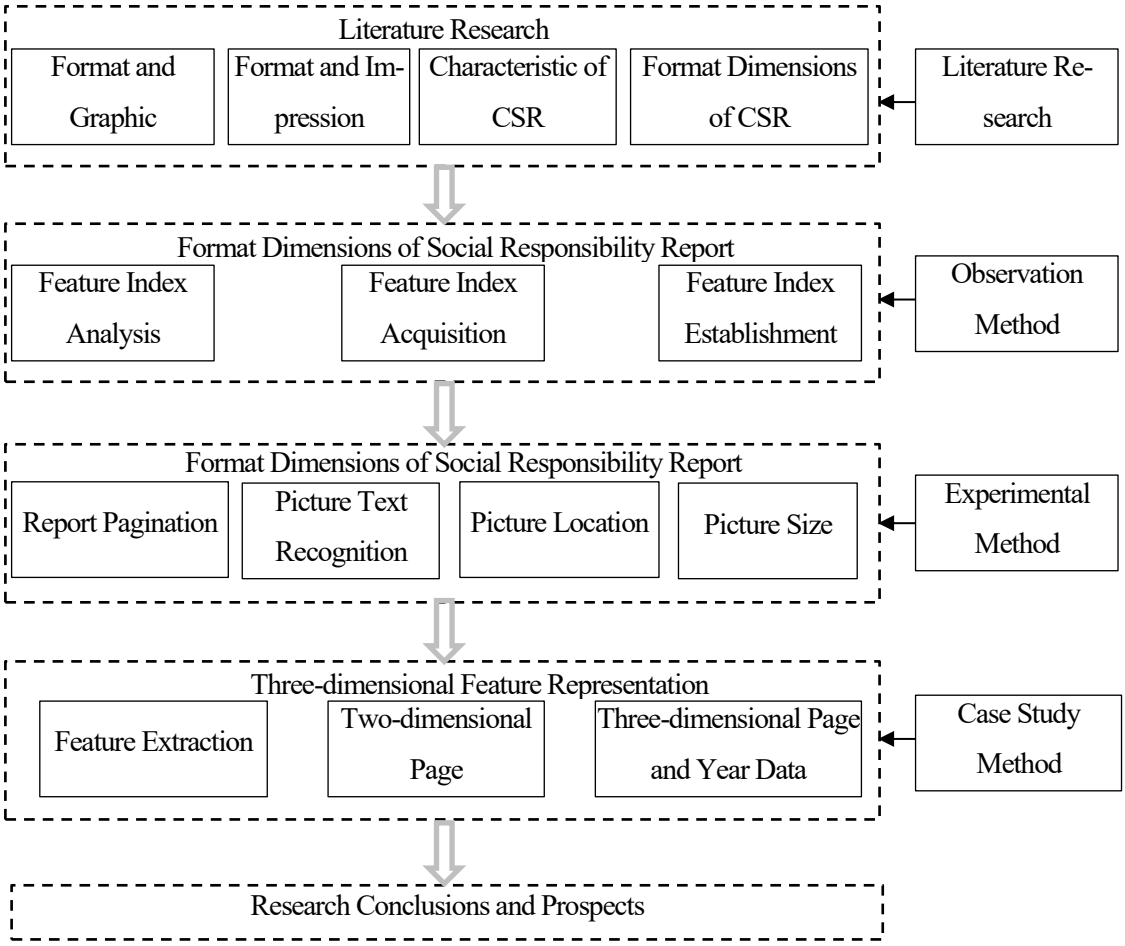

Figure 1–1 Technical Road Map

1.3.2 Research methods

This thesis integrates the relevant knowledge and principles of management, computer science, psychology and other disciplines, and mainly uses literature research method, experimental method and case study method to solve the problems in the research. According to the technology route, the specific research methods applied in this thesis are as follows.

(1) Literature research method. By making use of the library resources and network resources to query a large number of China high-quality articles from electronic databases such as CNKI, Wanfang and Google Scholar. Secondly, search and download relevant journal articles. Thirdly, explore the impression management, graphical representation methods and social responsibility report evaluation system of the social responsibility report format, so as to fully study the relevant research results.

(2) Observation method: The CSR reports of companies since 2009 were crawled through platforms such as Juchao (a website which collected a lot of CSR reports) Information Network, collecting more than 7,000 reports. With the help of a large sample of

social responsibility reports, this study explores the consistency of the formatting rules and the most distinctive features of the formatting to determine the formatting features.

(3) Experimental method: For the technical route proposed in this thesis, firstly, a page in a report is intercepted to become a picture with the help of Python. Then the features of the picture obtained are annotated as picture position feature data, picture type feature data and text density. Thirdly, combined with multiple pages in a report to form the two-dimensional feature data of this report. Finally, combined with the data of this report in different years to form the three-dimensional data with the help of Seaborn diagram.

(4) Case study method: Through the feature data by Vanke and PingAn bank to explore the application of this method in the report. Then this thesis explored the application of the feature extraction method in same report in different years.

1.4 Innovation

Compared with previous studies, the main innovations of this thesis are the following aspects: At present, Wang (2020) has focused on non-financial accounting narrative texts, confirming that different representation methods will affect the effect of impression management, there is no way to effectively analyze the feature dimensions of the social responsibility report layout. This thesis uses machine learning methods to develop research on the problem. To a certain extent, it provides a certain reference for follow-up research. In the past, most of the research on social responsibility reports was manually marked and evaluated. With the increase in the number of social responsibility reports, the format of social responsibility reports has become diversified. The social responsibility reports of different companies, different years, and different industries have many different types of layout styles, and it is difficult to manually extract the layout features of the social responsibility reports. This thesis proposes that with the help of machine learning, automated methods save a lot of labor costs, and provide new methods and ideas for the effective expression of the features of the social responsibility report format.

Specifically, this thesis has following innovations and contributions: First, this thesis is a methodological study. Although Shen (2010) has confirmed that the difference in the layout format will affect the effect of impression management, but he didn't find an effective way to deal with the problem mentioned above. Therefore, in this thesis, it firstly provides evaluation indicators for layout format features. Second, this thesis proposes a method for extracting feature index data from social responsibility reports, which can transform unstructured information into structured information for analysis. The interdisciplinary approach provides new research on corporate social responsibility reports. The automated method saves a lot of labor costs, and provides a new way to carry out research on social responsibility reporting issues. Third, this thesis introduces the three-dimensional data to the field of social responsibility report. The introduction of three-dimen-

sional data can accurately extract the required feature data and intuitively draw the report's follow-up changes in year and page number. Fourth, this thesis provides the feature data of the 2018 social responsibility report in units of page numbers, which provides a basis for future empirical research on formatting.

2 THEORETICAL BACKGROUND

Corporate social responsibility (CSR) has been a popular topic of academic research since the 1980s. In recent years, in order to maintain a good corporate image, and improve corporate value, enterprises have actively fulfilled their social responsibility through disclosing social responsibility information publicly and communicating with stakeholders (Mackey, 2007). As an important bridge that transmits signals to the market, the CSR report is an important data source and basis for stakeholders to evaluate corporate performance. Its quality will affect the decision and judgment of investors to a certain extent, thereby affecting capital market reaction. In recent years, with the increase in the number of corporate social responsibility reports, there are still problems such as insufficient report credibility, low legibility, and insufficient innovation (Wang, 2020).

To a certain extent, the effectiveness of the capital market and the efficiency of social resource allocation (Lin, 2004), will affects investors' judgment on the sincerity of enterprises' social responsibility (Jiang et al., 2011). However, with the growth of the number of reports, a lot of problems showed up. For example, whether the information disclosed in CSR reports can reveal the true social responsibility performance of enterprises. As a bridge for sending signals to the outside world, CSR reports have become quite important between stakeholders and companies.

However, many existing studies have also shown that this purposeful manipulative behavior can be identified or even quantified by textual features such as report length and average sentence length (Dai, 2014) as well as by disclosing quantitative information and pictures. Impression Management (IM), is usually a conscious and purposeful information management behavior, which will have a direct impact on the decision making of information users (Li, 2009).

Although scholar (Meng, 2012) has put forward the layout format as an important indicator to evaluate the quality of the report, due to the diversity and richness of the layout format, there is neither feature dimension to establish for such unstructured data, nor technical means to extract feature data. Therefore, this chapter will conduct an in-depth understanding of the literature from the impression management and the evaluation of the corporate social responsibility report format. The aim is to explore the feature dimensions of the presentation format therein and to establish feature indicators through a review of the relevant literature.

2.1 Research on the report format composition

2.1.1 Development of CSR

Corporate social responsibility (CSR) has been a hot topic since its introduction. In China, the disclosure of social responsibility information by enterprises started late. For most

companies, the existing laws, regulations and systems do not have mandatory requirements for social responsibility information disclosure. Company can choose the way to disclose the performance, they can disclose social responsibility information, disclose it in the annual report or disclose it in a separate social responsibility report. Also, enterprises can choose the content and form of disclosure reports from the perspective of their own interests. Therefore, CSR information is voluntary, optional and low standardized. However, in terms of the form and content of CSR information disclosure, with the publishment of relevant standards and guidelines in China, the content of CSR reports has been enriched, the form is more standardized than before, and the quality of reports has been improved. But there are still problems, for example, the content of information disclosure is not complete and the credibility of reports is not enough.

2.1.2 Influencing factors of CSR

Scholars generally analyze the research on the influencing factors of CSR information disclosure from both internal and external factors. Generally speaking, internal factors such as company size, company performance, corporate governance structure and industry attributes. External factors such as regulations, the role of government, the existence of pressure groups, economic culture.

The influence of internal factors on CSR information disclosure is mainly in the relationship between firm size and CSR information disclosure. Teoh and Thong used a sample of firms in different countries and eventually found a positive relationship between firm size and CSR information disclosure status through a questionnaire survey (Teoh et al., 1984). Udayasankar (2008) showed that the overall relationship between firm size and CSR information disclosure in a certain range is "U" shaped relationship. Yang (2014) conducted an empirical study on 437 listed companies in Shanghai and Shenzhen in 2010, and found through regression analysis that corporate size was positively related to the level of CSR information disclosure, and the larger the company size, the higher the level of disclosed social responsibility information.

In terms of the relationship between corporate performance and CSR information disclosure, Mills and Gardner's study showed that with great financial performance, companies are willing to disclose CSR information. Through the CSR, company can give signal to the outside world which shows they fulfill their socially responsible and use it to build a good corporate image (Mills, 1984). Dagiliene (2010) argued that a company's operating performance is a prerequisite and guarantee for CSR information disclosure. If the company's performance is poor, the degree of disclosure of social responsibility information is bound to be affected by objective factors such as insufficient funds.

Feng et al. (2016) found through an empirical study that good economic performance can promote companies to improve the level of CSR information disclosure. Based on signaling theory, Tan (2017) conducted an empirical study on the relationship between the degree of industry competition and CSR information disclosure, the results showed

that the more competitive the industry is, the more likely it is that firms will disclose social responsibility information.

Zhang (2014) conducted an empirical study with listed companies that issued social responsibility reports from 2009 to 2011, the results showed that corporate governance as a whole was positively and significantly related to the level of CSR information disclosure. Using the social responsibility information released by SMEs as a sample, Qin et al. (2018) found that the management shareholding had a negative impact on CSR information disclosure, while there was a positive relationship between the chairman having political affiliation, foreign shareholding and CSR information disclosure.

For the influence of external factors on CSR information disclosure, study shows that external groups such as environmental protection organizations affect CSR information disclosure, in addition, media opinion pressure and community supervision also have an impact on CSR information disclosure (Dwyer, 2002). Tao et al., (2013) conducted an empirical analysis on a sample of 251 listed companies in China in 2010, the results of the study showed that the status of the legal institutional environment was significantly and positively related to CSR information disclosure. Their studies also showed the better the status of the economic and legal institutional environment, the higher the level of social responsibility information disclosure of the company, even though the extra-legal institutional environment was negatively related to CSR information disclosure. Using a sample of 31 food industry companies in China from 2010 to 2013, Ying (2015) studied the relationship between media monitoring and social responsibility information disclosure of food companies through empirical analysis, and found that the two were significantly and positively correlated. Wu et al. (2015) analyzed CSR reports issued by companies from 2010 to 2013 and found that differences in both exchanges and regions affect the quantity and quality of CSR reports.

2.2 Research on the report format composition

Since the research on corporate social responsibility reports in China started late, the research on social responsibility reports is mostly on the content and text. The research on the format of social responsibility reports is yet to be explored. Therefore, this chapter will start from impression management and the evaluation of corporate social responsibility report formatting. Through the relevant literature, to explore the feature dimensions of the layout format, and establish the feature dimension indicators of the layout format.

2.2.1 Format and impression management

Impression Management, also known as impression modification or self-presentation, originally belonged to the scope of social psychology. Goffman (1959) draws a conclusion on individuals often shape their roles through different behavioral choices. He mentioned in the process of performance, they are also influenced by the environment and the expectations of others, so as ensure others to form a certain favorable impression. In this

way, people usually choose to display their behaviors purposefully in order to meet the psychological expectations of others and thus obtain a better evaluation of themselves by others.

Subsequently, Jones et al. (1964) extended Goffman's study and the connotation of impression management was enriched, and he considered impression management as the process of attempting to manipulate others to form a certain impression of oneself. The emergence of this idea extended impression management to the psychological level. However, most of the research scholars at that time held a view that impression management had the intention to deceive and manipulate others, and that such behavior was improper.

This view did not change until the 1980s, when psychologists began to revisit impression management and found it to be a universal phenomenon that is common and indispensable in everyday interactions and interpersonal interactions (Rosenfeld, 1995). Leary et al., (1990) proposed a two-factor model of impression management based on a review of previous impression management literature and divided impression management into two parts: impression motivation and impression construction, he also provided the model to study impression management.

With the continuous development and improvement of impression management theory, impression management has intertwined with many different research fields such as management science and organizational behavior, which has promoted the development of cross-discipline to a certain extent (Giovanna, 2011). According to Zhao (2007), impression management is the behavior of corporate management to consciously manipulate the form or content of linguistic disclosures in voluntary disclosures in an attempt to guide investors, government and other stakeholders to form a favorable impression of the company, and then influence stakeholders' decisions. According to Li (2009), impression management can be divided into two categories: active impression management, also called acquired impression management, is the act of establishing a better corporate image to the outside world through some means of information transmission when the company is performing well. Passive impression management, also called protective impression management, is the act of blurring performance by manipulating linguistic information when the company's financial situation is poor, so as to maintain the existing corporate image.

Kahneman et al. (1979) used the "Asian disease problem" to study decision making and found that people are influenced by the way alternatives are described and show preference reversal. He conducted that this phenomenon is universal, which clearly violates the principle of constancy in rational decision making. Kahneman et al. (1981) refer to this phenomenon of making different choices when making decision options for the same outcome due to semantic changes or differences in the form of presentation as the framing effect. In the past 30 years, with the in-depth study of risky decision making, the scope and depth of research on the frame effect as a decision bias has been extended. It includes

the enrichment of the types of frame effect, the extension of the theoretical explanation of the frame effect, and the expansion of the empirical study of the frame effect. The study all shows that people are risk-averse in the benefit zone and risk-seeking in the loss zone through the value function in prospect theory (Kahneman, 1979), and the curve in the loss zone is steeper than the curve in the benefit zone, and the framing effect arises because people place the problem described by the positive frame in the benefit zone and the problem described by the negative frame in the loss zone. However, the limitation is that the reference point and value function are difficult to estimate, some scholars have started to combine cognitive and behavioral related theories to explain the framing effect as a bias. For example, Beattie (2008) argued that people have fuzzy processing preferences and individuals process based on gist, thus preferring risk seeking in the negative frame and risk avoidance in the positive frame.

The better the information readers feel, the more likely they are to agree with the information. There are many ways to manage impressions in corporate information disclosure, such as self-interested attribution, manipulation of readability, manipulation of data, charts and pictures (Cesario et al., 2008). Among impression management methods, setting different quantitative information and pictures are two special methods in CSR, because CSR have greater choice in expression than financial reports.

In the field of cognitive psychology, Vessey (1991) proposed that the formatting method is related to the cognitive level. He also believed that when the decision-making strategy matches the information presentation form, human can make decisions quickly and accurately. Otherwise, cognitive efforts will be required to adjust the information presentation form or decision-making strategy, which will eventually lead to a prolonged decision-making time or a decrease in accuracy. Some scholars conducted research from the perspective of decision-making strategies and task types. For example, Kershaw (1998) studied the relationship between layout format and decision strategy from a cognitive perspective. The study showed that overall strategy is suitable for spatial processing, analysis strategy is suitable for analysis and processing, and graphic format supports overall strategy, the table supports the analysis strategy.

Vessey (2006) further expanded the theory of cognitive adaptation, believing that the presentation form of information, task features and whether the decision strategy matches will affect the final decision. Vivien (2000) proposed that compared with professional investors, non-professional investors tend to rely on information in graphical format when evaluating current profits. Beattie et al. (2008) found that the photos with company features inserted in the company's annual report can affect investors' views of the company. From the perspective of the diversification of the layout format and the reader's cognition, some scholars pointed out that the personalized design of the layout format can be used to enhance the reader's cognition of the company's performance (Giovanna, 2011), because pictures and tables will capture and retain the reader's attention. In addition, she thought that the graphs are eye-catching, which play a very important role in

information summarization, extraction and dissemination, and may attract the attention of readers who do not pay attention to the disclosure of the text.

Giovanna (2011) conducted research on the charts in the British CSRs and found that managers prefer to use charts to describe information about air pollution, resource use, and employee activities, and industries that are highly related to pollution also use charts the most. She also found that companies have obvious impression management tendencies in the use of background pictures. When companies decide to use pictures to show or not show certain trends and whether to distort thesis trends, they will show an excessive preference for using pictures to describe good news.

2.2.2 Format and graphical representation

The social responsibility report is an important bridge for companies to send signals to stakeholders. It includes narrative text, charts, pictures, etc. However, both in content and form, financial information disclosure and non-financial information disclosure are different. In terms of content, the company's financial reporting requirements must include statements such as balance sheets, income statements, etc., which can reflect the company's financial status and operating results through chart data. For corporate social responsibility reports, it does not disclose specific information about pictures and charts. Companies can choose to make qualitative descriptions in the form of text, quantitative descriptions in the form of data, or both. In terms of form, the financial reports issued by enterprises are basically standardized and uniform in format. Few companies have included pictures in their financial reports. For corporate social responsibility reports, companies have added pictures related to their social responsibility performance in their social responsibility reports in order to enhance the report's publicity and advertising effectiveness.

Existing studies had proved that the graphical representation in the literature is used in the design of the layout format. Research (Lee, 2016) found that pictures are used to construct, manage and convey the specific image that managers expect. Jones et al. (1994) conducted a study in the United Kingdom and found that the use of pictures and designs in report was still of a secondary importance during the period from 1930 to 1970. Subsequently, the pictures were increasingly used to modify the report. After 1984, the format of the report became a tool to convey the company's image to the outside world.

The number, proportion, and category distribution of the pictures and charts are applied to the layout format to influence the stakeholders of the report, carry out impression management, and achieve the purpose of maximizing benefits. Beattie et al. (2008) used a cross-country comparison method to compare the format of charts in the 1990 annual reports, and found that 92% of American companies used charts in annual reports, while only 80% British companies used charts in annual reports. The charts contain a wealth of information and is an important research object. Teoh (1984) found that there is a significant correlation between whether the chart in the report violates design rules and changes

in corporate performance. Companies with declining performance are more likely to manipulate the chart. John et al. (2004) studied the correlation between the use of charts in the reports of 100 Hong Kong companies and changes in corporate profits, and found that when profits rise or fall, companies will use a lot of charts. Jones et al. (2011) conducted a research on the use of charts in the UK. The research results show that charts are widely used in social responsibility reports, and managers prefer to use charts to describe information about air pollution, waste emissions, resource use, and employees. Industries that receive high attention, use the most graphs.

In general, there are few studies on the format of social responsibility reports, and scholars use cognitive psychology as the theoretical basis to realize interdisciplinary research. Song et al. (2017) conducted statistics on the pictures in the social responsibility report and concluded that the format of the report will produce a certain visual effect. Wang (2014) discussed the impact of the application of management accounting report reporting format on the management accounting report function. There is still a lot of research space for the research on the format of reports. From the review of the literature, it can be seen that the research on the information presentation format is mostly to discuss its classification standards and information content, and there are few special researches on the layout of social responsibility reports.

2.3 Evaluation in the format of the corporate social responsibility report

In China, Shen (2011) pointed out in 2011 that comprehensive reports should be analyzed and explained in terms of format standards, disclosure models, and report effects, and integrate them into social responsibility reports to make up for their shortcomings. The study found that different information presentation formats are positively related to the decision-making, control and evaluation functions of management accounting reports. Some scholars also consider the influence of factors such as personal knowledge, ability and cognition. For example, Mao et al. (2014) considered the factor of personal ability. The study found that decision makers with lower personal ability or professional knowledge use graphs than tables. With the graphs and table, a corporate social responsibility report can be accurate and efficient.

In America, the American Institute of Certified Public Accountants proposed in 1991 that the financial information disclosed in current financial reports did not pay attention to the real needs of information users, such as the understanding of information by different information users (Vessey, 1991). The processing is different, and the chart layout format can be intuitive, and for readers with a lower level of understanding, the chart and other layout formats can be effective. The International Accounting Reporting Expert Group set up an academic research committee to specifically study on how to disclose corporate non-financial information (Reggy, 2000).

In terms of the evaluation of the quality of the social responsibility report, it is mainly based on the corporate social responsibility report guidelines issued. The most global ones

are the GRI standard (GRI, 2016) and the ISO26000 standard (ISO, 2010). The established rating index system includes 5 first-level indicators and 15 second-level indicators of completeness, reliability, comparability, dichotomy, and readability. Some scholars have also compared and analyzed the CSR reports issued in China from the six aspects of completeness, balance, comparability, reliability, accuracy, and readability (Xie, 2015).

However, the most authoritative and widely used rating standard in China is still the rating standard provided by the Chinese Academy of Social Sciences (Zhang, 2020). The center's report rating service has become the most influential third-party evaluation business in China, which was written mostly in Chinese.

Since March 2010, with the profound changes in the field of social responsibility, the report's rating has been revised and upgraded four times over the last eight years, and the latest version of the rating standard was released in 2019, from the perspective of process, materiality, completeness, and balance. Four indicators of comparability, comparability, readability and innovation rate the report's quality.

In March 2019, the China Corporate Social Responsibility Report Rating Expert Committee, based on the "China Corporate Social Responsibility Reporting Guidelines 4.0" (Zhong, 2019), drew on the principles and methods of corporate social responsibility report evaluation at home and abroad, and formulated a corporate social responsibility report rating standard with Chinese features. It points out that the evaluation method of readability is evaluated in terms of report acquisition method, chapter structure, layout design, language, graphics and so on. The innovative evaluation method mainly compares the content and form of the report with the previous CSR to determine whether there is innovation, and whether innovation has improved the quality of the report.

Due to the late start of social responsibility reporting in China, the changes in the versions of the "Chinese Corporate Social Responsibility Compilation Guide" shows the development of social responsibility reporting in China, therefore, it is necessary to record each version of social responsibility reporting in the literature review section.

In January 2014, the general framework of "Chinese Corporate Social Responsibility Compilation Guide 3.0" (Zhong, 2014) and the first sub-industry guide "Chinese Corporate Social Responsibility Report Compilation Guide: General Mining Industry" were released in Beijing. The Chinese Academy of Social Sciences put forward the concept of "full life cycle management of social responsibility reports" when compiling "Chinese Corporate Social Responsibility Compilation Guide 3.0", which was also reflected in the 2014 rating standards. The readability of corporate social responsibility reports includes the fluent, concise, accurate, and easy-to-understanding, and the expressions are intuitive through flowcharts, data tables, and pictures. Innovativeness includes two aspects: innovation in report content and innovation in report form.

Research on social responsibility reports started relatively late. The Chinese Academy of Social Sciences released the "Guidelines for Compiling Chinese Corporate Social Responsibility Report 1.0" (Zhong, 2009) and "Chinese Corporate Social Responsibility

Report 2.0" (Zhong, 2011) in December 2009 and March 2011. The normative study of social responsibility reports in China started very late, and as of 2014, only a few hundred companies had published social responsibility reports, which showed that the overall level of corporate social responsibility in China is still in its infancy. However, the research also showed that seventy percent of enterprises were seriously lacking in social responsibility, and nearly half of state-owned enterprises were on the sidelines, there were even 26 companies that score 0 or even negative points (Dai, 2014).

On October 13, 2020, the Central Corporate Social Responsibility Report meeting was held in Beijing. The meeting proposed that 30% of the corporate social responsibility reports have outstanding innovative performance, and the content format is also showing an increasingly innovative trend. This shows that the format of the layout is rich in pictures and texts, which play a vital role in the quality evaluation of corporate social responsibility reports.

To evaluate the quality of CSR reports, the premise is to determine the index dimensions of report evaluation. By compiling authoritative third-party evaluation agencies and social responsibility report standards and guidelines at home and abroad, this thesis finds that the current evaluation of CSR report content quality mainly revolves around the dimensions of report objectiveness, process, integrity, balance, comparability, readability, innovativeness, accuracy, clearness, reliability, standardization, accessibility and availability as shown in Table 2-1.

Table 2–1 Quality evaluation indicators of CSR reports issued by some authoritative organizations

	China Academy of Social Science	R&L Global	Global Reporting Initiative	WTO Economic Guide	Chinese Industrial Enterprises and Industrial Associations	International Organization for Standardization
Objectivity	√		√	√	√	
Process	√					
Integrity	√	√	√	√	√	√
Balance	√	√	√			√
Comparability	√	√	√	√	√	
Readability	√			√	√	√
Innovativeness	√	√		√		
Accuracy			√			√
Clearness			√			
Reliability		√	√	√	√	
Standardization		√				
Accessibility		√				√
Availability		√	√		√	√

Numerous standard documents and literature show that among the above evaluation indicators, integrity, readability and innovativeness are important indicators for evaluating the quality of reports, and readability is achieved through flowcharts, data tables, and pictures that make the form of expression intuitive as well as easy-to-read typographic design with clear pictures and textual descriptions (Zhong, 2019). The quality evaluation indicators of CSR reports evaluate the readability of a report in terms of how it is obtained, chapter structure, typographic design, language, graphics, and other aspects. The innovative form is the innovation of the layout format. Therefore, it is necessary to study the format of social responsibility reports.

2.4 Summary

When the information receivers read a report, they can have an impression of how much energy and time the company spends in preparing the social responsibility reports. Also, when a report is full of graph and chart as a supplement of the report, the information receiver can be persuaded by the report easily, further affected them to do the market investment. In terms of the graphical representation of social responsibility reports, in a good way, the charts arranged in the report can enable readers with weak information understanding to understand information intuitively and accurately. However, existing studies have shown that managers prefer to use graphs to describe parts of air pollution, waste discharge, and resource use for impression management.

In order to improve the overall layout quality of the report, scholar (Zhong, 2019) have also explored the quality dimension of the layout format, researching and proposing that the disclosure modes and report effects of comprehensive reports can be integrated into contemporary social responsibility reports. With the release of "China Corporate Social Responsibility Reporting Guidelines 4.0", the dimensions of social responsibility reports have become fewer than "China Corporate Social Responsibility Reporting Guidelines 3.0". Regarding the readability and innovation of the quality dimensions of social responsibility reports, some organizations have incorporated the format into the quality evaluation of social responsibility reports. The content in CSR is mostly the arrangement of pictures and charts, but in practice, research on the format of the social responsibility report has not been conducted in-depth.

Based on the above points, in order to solve the problem of impression management in the social responsibility layout format, in the next chapter, this thesis will start with the quality evaluation criteria of the layout format, the content analysis of the layout format, and the feature dimensions of the layout format.

3 EVALUATION DIMENSIONS OF FORMAT IN CORPORATE SOCIAL RESPONSIBILITY REPORT

After conducting literature research on the impression management of the format of the social responsibility report and the evaluation of the format of the social responsibility report, this chapter will conduct a qualitative analysis of the dimensions of the format of the social responsibility report.

Existing studies have demonstrated that social responsibility reports manage impressions through formatting features such as report length (Fang et al., 2015) and the disclosure of quantitative information and images (Guo et al., 2016) quantitatively, but there were no valid criteria to judge how graphic formatting has an impact, and it is difficult to go into data-based objective evaluation.

In the construction of the social responsibility report quality evaluation system, scholars pointed out that the readability indicators are the length of the report, the types of the graphs, the layout design, and the legibility (Ji, 2016). However, the scoring standards for the layout design are only 0 points and 100 points. For content richness, there are 0 points, 20 points, 40 points, 60 points, 80 points, and 100 points. Because of the lack of a unified standard, these report formats are different and diverse. How to effectively extract the formatting features of the report, to extract the formatting format of the report, such as the layout of the text and the layout of the pictures of the report, to explore the changes has become the main issue to be studied in this thesis.

3.1 Analysis of evaluation of the report format

At the same time, as the development of corporate social responsibility reports, many related compilation guidelines and guidelines have emerged, which have continuously promoted the improvement of the quality of social responsibility reports. To evaluate the quality of corporate social responsibility reports, the prerequisite is to determine the index dimensions of the report layout format. After collating the existing corporate social responsibility report compilation standards, this thesis found that the Shanghai Stock Exchange guidelines, GRI standards, and Chinese Academy of Social Sciences guidelines have become three major mainstream reporting standards. Figure 3-1 shows the number of companies implementing the standard.

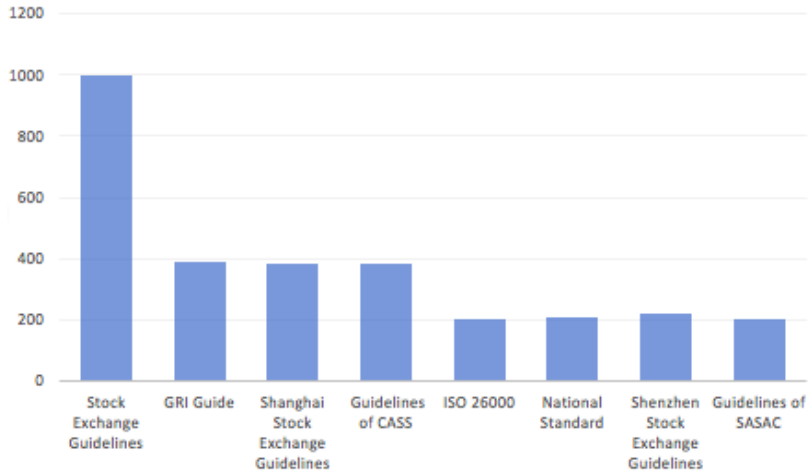

Figure 3–1 The number of companies implement the standard (China News, 2020)

From figure 3-1, it is not difficult to see that the Shanghai Stock Exchange guidelines, GRI standards, Shanghai Stock Exchange guidelines and Chinese Academy of Social Sciences (CASS) guidelines have become the four mainstream reporting standards.

In 2019, the guidelines on social responsibility reports issued by the Stock Exchange mention that the current social responsibility reports have data measurement or inaccurate measurement problems. There have been studies on the text in social responsibility reports.

For text processing and scoring, although there is a large amount of data for analysis in the chart part of the social responsibility report, there is a lack of effective technical means to extract the chart data from the report. The Chinese Academy of Social Sciences put forward the question about how to identify the impression management when compiling CASS-CSR3.0 (Zhong, 2009), which was also reflected in the 2014 rating standards. The readability of corporate social responsibility reports can be embodied in flow charts, data sheets, pictures, etc. to make the form of expression intuitive.

Innovativeness includes two aspects: innovation in report content and innovation in report form. In the "China Corporate Social Responsibility Reporting Guidelines 4.0" issued by the Chinese Academy in Social Sciences, it also emphasizes the promotion of value management of social responsibility reports (Zhong, 2017). The G4 guidelines issued by GRI put forward six principles to ensure the quality of social responsibility reports-clarity, balance, comparability, accuracy, timeliness and reliability.

Clarity requires companies to disclose social responsibility information in a way that is convenient for stakeholders to obtain and understand. Images and comprehensive data sheets can make the information in the report easy to obtain and understand. Table 3-1 shows the definitions of some authoritative organizations on the feature dimensions of the format:

Table 3–1 Layout format in CSR

"Environmental, Social and Governance Reporting Guidelines" of the Shanghai Stock Exchange	The data in the charts in the released social responsibility report chart data have inaccurate problems.
"Chinese Corporate Social Responsibility Guidelines 4.0"	The readability of corporate social responsibility reports can be included in the fluent, concise, accurate, and easy-to-understanding. The expressions are intuitive through flowcharts, data tables, and pictures. The innovation includes two aspects: innovation in report content and innovation in report form.
Global Reporting Initiative (GRI) "Sustainability Reporting Guidelines G4"	Clarity requires companies to disclose social responsibility information in a way that is convenient for stakeholders to obtain and understand the information. Images and comprehensive data sheets can make the information in the report easy to obtain and understand.

In general, the standards are based on the text layout features, picture layout features, and page structure features of the social responsibility report. When focus on the standard definition of the layout format of the social responsibility report, although the social responsibility report quality standards stay at the text layout features, picture layout features and page structure features, but they are not refined to the specific dimension indicators of the layout format standard. The content analysis starts to explore the three-dimensional layout format of text layout features, picture layout features, and page structure features.

3.2 Analysis of the content of the report format

Many companies use pictures to supplement descriptive and analytical text in their social responsibility reports. Yang (2016) believes that pictures in the report not only modify the text information but also have the function of carrying information. It can be seen that the text in the picture supplements the text of the report. This section takes two pages in the 2018 China Vanke Social Responsibility Report as an example, as shown in the figure 3-2. Since different companies has different format. For Vanke company, Python read two pages as one page, therefore, figure 3-2 contains 2 pages.

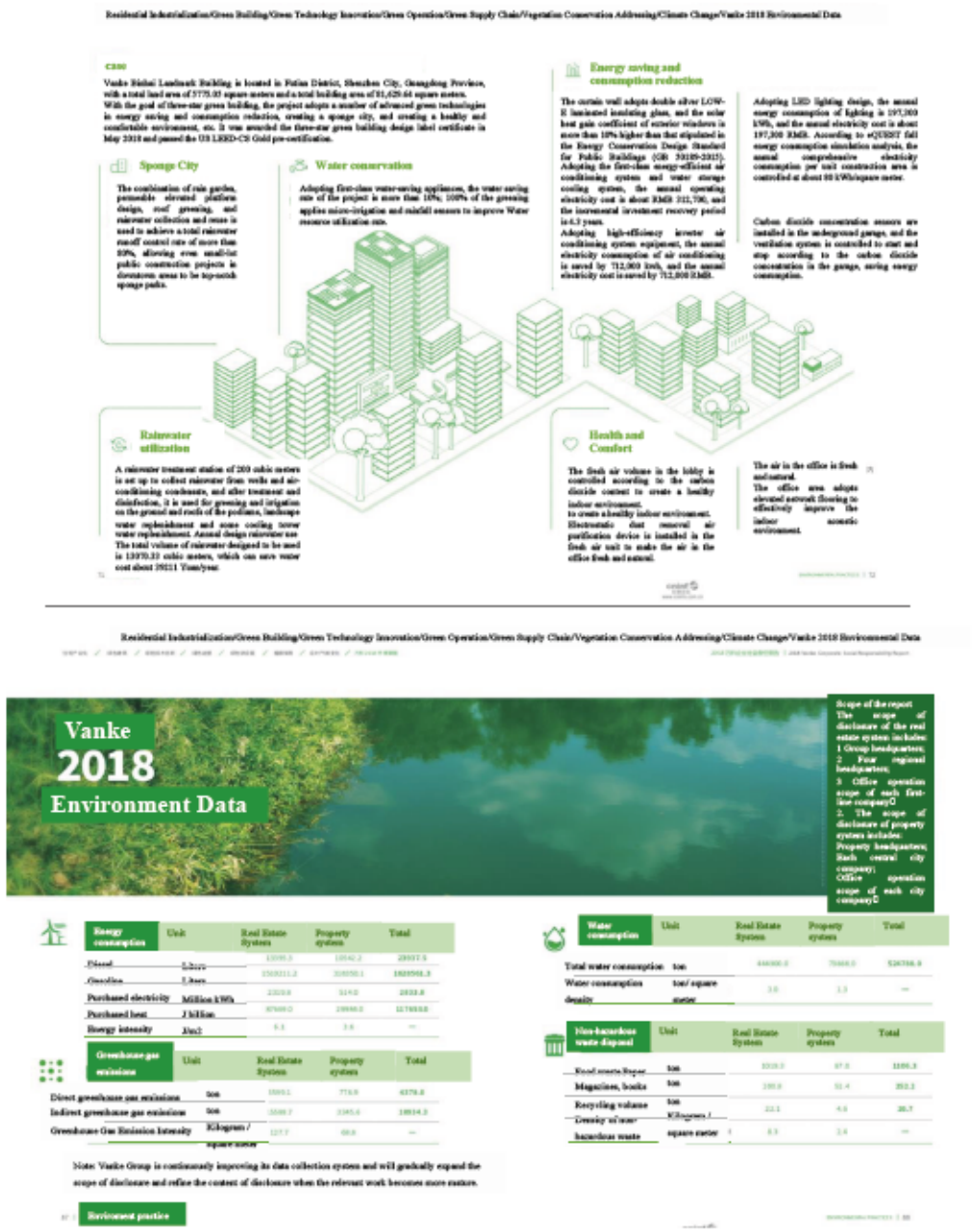

Figure 3–2 Two example pages in the 2018 Vanke CSR

In the design of the layout format, compared with the left page and right page, both consist of pictures and text, but the position settings of the pictures are different; in the top picture, the text is arranged to illustrate the content in the picture; and the figure on the right is the data and text description written in the chart. Relatively speaking, in terms of picture layout features, the second picture has less text and the proportion of the picture is larger, while the picture on top is designed with seven boxes, and the proportion of the text is larger than the proportion of the picture. In terms of page structure, the top picture discloses less quantitative information; while the second picture discloses quantitative information, and a picture is inserted above the page number. The layout format of the

picture and text combination may be due to the attractive picture, which can reduce the complexity of the report, increase the reader's interest in reading, and enhance the social responsibility report readability (Zhong, 2019). Improvements of the readability of the report can act on the reader's perception and have a positive impact on the reader's judgment and evaluation.

For these two pictures, the method of impression management is different. For the top picture, the design of the combination of pictures and narrative text and the larger proportion of pictures on the page can provide vivid information, thereby making reading "feel good" and having a positive impact on the evaluation of their corporate social responsibility performance (Dai, 2014). Through the information modification and information transmission functions of the picture, readers may have the feeling of the persuasiven information; and the relevant data is combined in the second picture, which can give a digitized information. Starting from the feature dimensions of the layout format of the report content (picture layout features, text layout features, and page structure features), the direct indicators that can be measured are the number of words in the text, the location of pictures, and the size of the picture.

3.3 Analysis of the dimensions of report format

This thesis adopted the methods of consulting a large number of documents, summarizing and sorting out the format dimension and expert opinion in the quality evaluation of social responsibility report to establish the feature data of the format.

According to G4 guidelines and the social responsibility report rating standard of Chinese Academy of Social Sciences, quantitative information disclosed in the social responsibility report can improve the accuracy of the report, while the use of pictures in the social responsibility report can enhance the clarity and readability of the report. The information with rich color and rich picture can improve the understanding of information users. From the perspective of enterprises, they try to convey perfect social responsibility information to the readers of social responsibility reports through thesis impression management methods, so that the readers can evaluate the performance of corporate social responsibility, so as to achieve the purpose of obtaining legitimacy. Based on the relevant research of China corporate social responsibility preparation guide 4.0, global reporting initiative organization (GRI) sustainable development report guide G4 and scholars' relevant research (Zhong, 2019), three indicators of the picture feature dimension are constructed, and the format of corporate social responsibility report is standardized later. Because the lack of standard formats, it is hard for companies to compare the quality with the others. Therefore, when there are standard formats, Chinese companies can improve their CSR quality to communicate with their stakeholders.

The first step in extracting the characteristics of the social responsibility report format is to explore the characteristic dimensions of the social responsibility report format,

which are generally obtained from various authoritative institutions that publish evaluation standards for social responsibility reports. This study adopts methods such as reviewing a large amount of literature, summarizing the dimensions of presentation format in the quality evaluation of social responsibility reports, and employing the interview method to establish the characteristic indicators of presentation format.

In response to the characteristic indicators in the presentation format of CSR reports, the Expert Committee on CSR Report Rating in China has developed CSR report rating standards with Chinese characteristics, based on the China CSR Report Guidelines 4.0 (Zhong, 2019) and the principles of CSR report evaluation at home and abroad. Also, on the China CSR Reporting Guide 3.0 (Zhong, 2014), the China CSR Reporting Guide 4.0 and the GRI Sustainability Reporting Guide G4, both of them points out that the readability of a CSR report is reflected in the layout design that makes the expression easy to read through flowcharts, data tables and pictures, with clear pictures and textual explanations. It evaluates the report in terms of chapter structure, typography, language, graphics and other aspects (Zhong, 2017).

For the establishment of the dimension of the features of the format of social responsibility report, this thesis, based on the analysis of the above-mentioned scholars' index of the format, combined with "China corporate social responsibility writing guide 3.0", "China corporate social responsibility writing guide 4.0" and "sustainable development report guide G4", uses the interview method to interview five researchers, evaluators and readers. Combined with the content analysis of Vanke's format in the previous section, this thesis extracts three feature indicators for the format features of social responsibility report. The feature indicators of the format of the social responsibility report are established, as shown in Table 3-2. The measurement dimensions of the social responsibility report layout are picture layout features, word layout features, and page structure features, and the indicators that can be measured after the machine learning test are Picture size, Word density, and Picture location.

Table 3–2 The feature index of format

First level index	**Second level index**	**Third level index**
Picture layout features	Picture size	Height
		Width
Word layout features	Word density	Number of words
Page structure features	Picture location	Picture color
		Top
		Left

The specific feature dimensions are as follows:

(1) The character layout of each page in the CSR report. The most direct way to consider the features of text layout is word density. The attractive picture can reduce the

complexity of the report, and increase the reader's interest in reading. Compared with traditional financial reports. A notable feature of social responsibility reports is that text narratives are the main body, and readers of the report may become bored with the text narratives throughout the reading process. Therefore, the first feature dimension of this study is the density of words in the picture.

(2) The features of picture arrangement in CSR report. The most direct way to consider the features of picture layout is the size of the picture. A picture can strengthen the reading impression to help understanding. A picture can provide the readers of the social responsibility report with clear and vivid information, meet the reader's information needs, and make the reader "feel good." This "feel good" will be transformed into the persuasiveness of the information that, and will have a positive impact on the evaluation of their corporate social responsibility performance. Therefore, the second feature dimension of this thesis is the image size (left and top in the chart).

(3) The page structure features of the CSR report. The most direct way to consider the features of the picture layout is the position of the picture. Exploring the position features of the picture on a page helps to further explore the proportion of the page, so as to find the social responsibility report that satisfies the stakeholders. The coordinate position of the picture can be investigated in two points which are the distance to the left of the page and the distance to the top of the page, by measuring these two positions to get the coordinates of an image in the pages.

3.4 Summary

Firstly, this chapter analyzed the feature dimensions of the layout format from the evaluation criteria of the social responsibility report. Secondly, this thesis explored the layout dimensions of the social responsibility report as picture layout features, text layout features, and pages structural features, the content of the social responsibility report was analyzed in terms of layout dimensions. Finally, combined with the evaluation criteria, the indicators for CSR reports were established as text density, picture size, and picture location. Because the format feature in the social responsibility report is unstructured data, with the help of the method this thesis proposed, the feature data in the format can be extracted to provide structured data for analysis, which turns the unstructured data into structured data. Also, the features of single-page reports were extracted according to the features of text layout, picture layout and page structure features.

4 METHODOLOGY

In the previous chapter, this thesis explored the three feature dimensions, which are text layout features, image layout features, and page structure features, as well as the measurable feature indicators in the feature dimensions, namely text density, image size, and image location. Based on the fact that data is always the manifestation of information, it is hard to discuss data and information separately and abstractly. On the other hand, machine learning also requires a large amount of basic data, and a large number of sample libraries can help the machine to achieve high accuracy in continuous training. Therefore, in this chapter, the research explores machine learning to measure the feature indicators proposed above.

4.1 Study design

As shown in Figure 1-1, firstly, a literature review of graphic representation and impression management in social responsibility reports was conducted. Followed by an introduction to the characteristics of social responsibility reports and the dimensions in the evaluation of social responsibility reports. Thirdly, this thesis analyzes the feature dimensions of the format of social responsibility reports with the observation method, explore the characteristic indicators of social responsibility reports, and extract the characteristic data. Finally, with the help of an experimental method, the social responsibility report was paginated by machine learning, then forming two-dimensional feature data and three-dimensional feature data, which are different page numbers, different years and different feature data respectively.

4.1.1 Operational plan for research

To analyze the characteristic dimensions of the format of social responsibility reports, this thesis uses interview to do the research on the feature index in CSR. Including people in the field of social responsibility reporting, students in the field of social responsibility reporting, graduate students in the field of computer science, staff members who evaluate the reports, and people who had never read the reports before. To explore the formatting dimensions of social responsibility reports, five people were interviewed, including researchers in the field of social responsibility reports, and a group of students in the field of computer science, in order to explore the formatting dimensions of social responsibility reports, and researchers in the field of social responsibility reports.

Firstly, they were contacted by email and asked if it was convenient for them to conduct the interviews. The interviews were recorded using a tape recorder and, with the consent of the interviewees, the discussions were recorded to ensure the accuracy of the interviews. The interviews lasted 30-40 minutes.

Secondly, the interviews were preceded by an introduction to the social responsibility report, which explained the process and purpose of the study, along with an introduction to impression management and impression management in the field of social responsibility reporting. The main questions are the important orchestration dimensions of the report and the indicators of orchestration format features that can be extracted. Considering that the stakeholders of the social responsibility report include equity holders who have not read Vanke corporate social responsibility report before, one person who has never read the report was interviewed. For the interviewer who had not read the social responsibility report, after explaining the source and purpose of the social responsibility report, certain specialized vocabulary for the interview questions was explained, which took an interview time of 37 minutes, but the results obtained were meaningful. The dataset was analyzed using a specific-to-general approach and it was coded and produced with representative quotes (Table 4-1).

Table 4–1 Operationalization plan for research questions

Research question	**Sub questions**	**Theoretical background**	**Main interview themes/questions**	**Extra questions**
The format of corporate social responsibility report?	The dimensions that can be extracted in the format?	A method of feature extracted for academic literature (Yu,2020).	Dimension indicators that can be established?	Is the technology feasible？
	The format of social responsibility report and the dimension of evaluation?	Chinese Corporate Social Responsibility Reporting Guide 4.0 (Zhong,2017).	Feature index extraction in Text feature, picture layout feature，page structure feature?	Are there any other dimensions？

Since each person's field is different, for questions that need CSR knowledge, this thesis used questions from the interviewee's familiar field as an alternative to prevent the interviewer from not being able to answer for the interview questions. All interviews were conducted in complete confidentiality and the interview data was kept confidential. After the interviews, respondents were also asked for relevant feedback and whether there was any data that could not be included in the research. The study of this thesis focuses on the choreographic formatting dimensions worth studying and the choreographic formatting indicators available for extraction. After exploring the similarities and differences in the interview responses, we labeled the words that recur in the responses to the same question. The encoding process was completed by selective encoding, and the repeated words in the interview were integrated and refined into specific concepts (picture layout features,

text layout features, word layout features). This interview used MAXQDA2020 software to code.

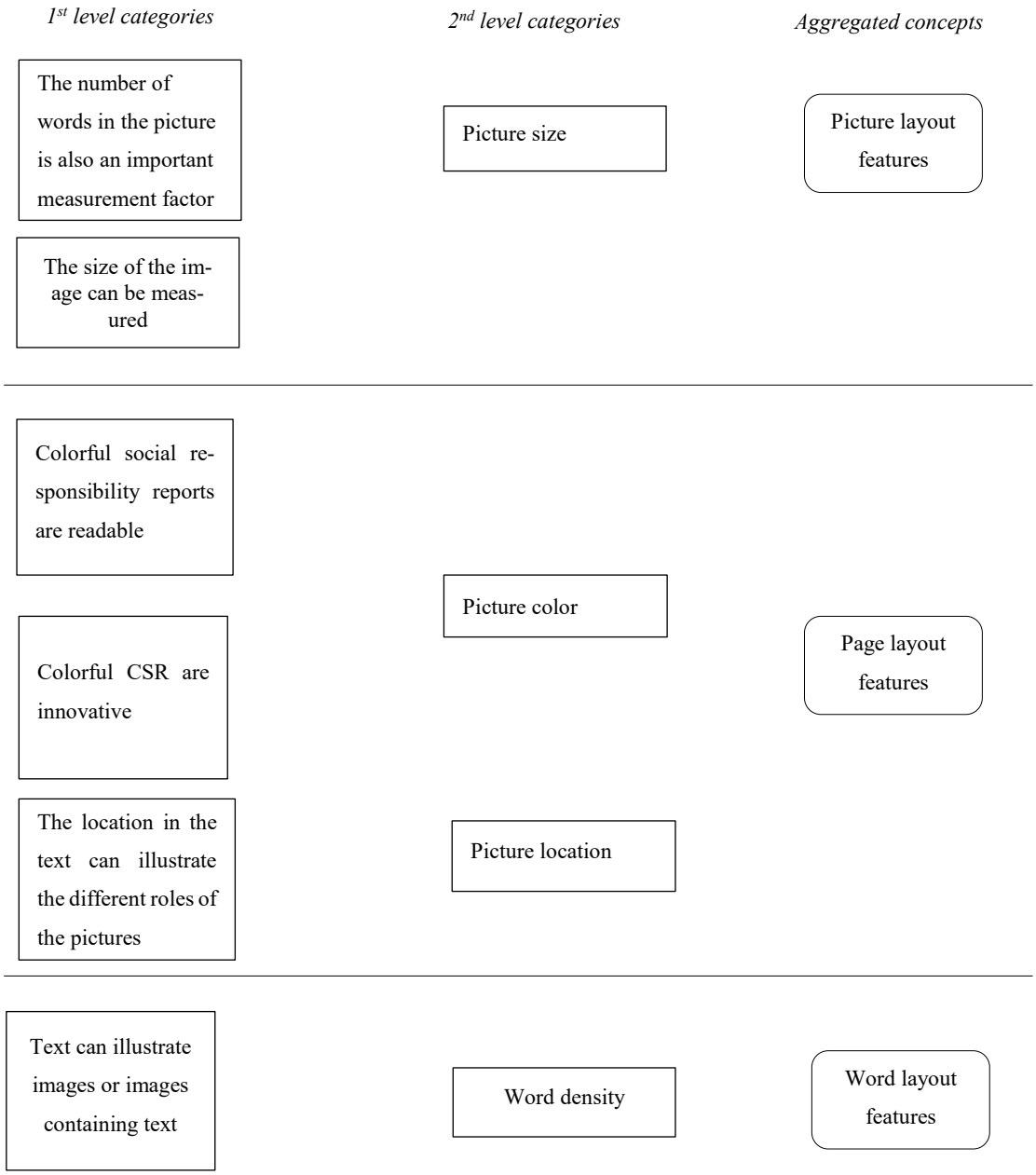

Figure 4–1 Selective coding of CSR feature

4.1.2 Experimental method for research

The data set in this thesis is mainly composed of pictures. First, with the help of Python, a page of a "PDF" format report is intercepted into a "PNG" format picture. The labeling of the corporate social responsibility report mainly includes manual labeling and machine labeling. Among them, manual labeling is the most accurate method. Compared with the

machine labeling, the comprehensive rate and accuracy of the data are high, and the problems in the data can be found in time, but it takes a lot of manpower and time costs, and repetitive data labeling work, in the same time, a lot of time was wasted on initial data collection, and the research progress was slow. Machine labeling has certain advantages in terms of labor cost and time cost. It can become accurate through continuous learning. Some scholars use word vectors to extract features from reports. After several iterations of learning, the model can be used to label the same type of reports (Zech, 2018).

For these two mainstream methods of feature extraction, the time and labor cost of manual labeling are difficult to accept. Therefore, this thesis adopts the idea of machine labeling and proposes a method of automatically labeling the format feature dimensions of social responsibility reports. Because the format feature data in the social responsibility report is unstructured data, with the help of this method, the feature data in the format can be extracted to provide structured data for analysis. Among them, the features of single-page reports are extracted according to the features of text layout and picture layout, and the features of different features, different page numbers, and different years of reports are described to explore the overall change and year of the layout format in the social responsibility information disclosure.

The method of measuring the feature indicators of the format of the social responsibility report is shown in the figure 4-2.

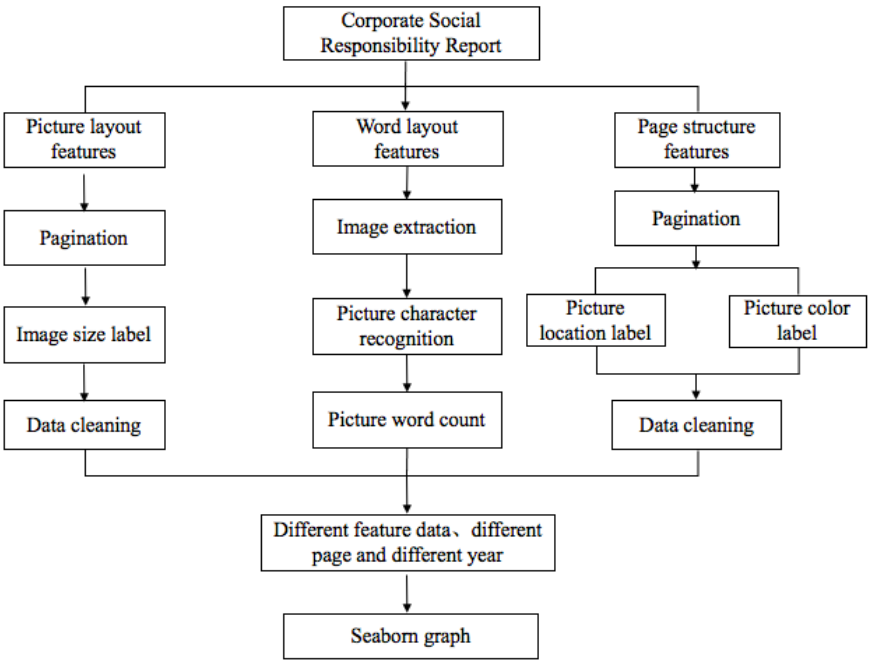

Figure 4–2 Flow chart of feature extraction method

4.2 Tools used in the research

This thesis integrates the relevant knowledge and principles of management, computer science, psychology and other disciplines, and mainly uses literature research methods, experimental methods and case study methods to solve the problems in the research.

For the experimental method, the research in this thesis uses Baidu AI, which includes functions such as image technology, text recognition, picture recognition, video technology, and knowledge graphs. The Baidu AI Open Platform is a platform for deep learning, and is based on a large-scale open data set of real samples, combined with a variety of algorithms to continuously train the data to obtain a relatively complete image subject recognition. Its functions include locating the key subject position of the image, and intercepting the subject position in the picture for image processing.

With the help of Python, the Seaborn plot can be expressed in two colors: red and white. In terms of data visualization, there are two useful packages for Python: Matplotlib and Seaborn. The Matplotlib and Seaborn used in this thesis have the following advantages: Matplotlib is an extended library of Python, so it inherits the advantages of Python syntax, being object-oriented, easy to read, easy to maintain, and the code is concise. Seaborn is an advanced visualization library based on Matplotlib, which is biased towards statistical drawing. Therefore, the points targeted are mainly variable feature selection in data mining and machine learning. Compared with Matplotlib, its syntax is relatively simplified, and the graphs drawn do not require a lot of effort to modify. Since this chapter mainly focuses on the technical route proposed in this research, the application of the three-dimensional feature map dimensionality reduction proposed in this research in the social responsibility report will be explained in detail in the next chapter.

4.3 Feature extraction method of layout format

In recent years, with the further in-depth and expansion of related technologies and applications, some scholars have begun to try to introduce machine learning into the text analysis of corporate social responsibility reports. For example, Apostolova et al. (2013) pointed out the accurate positioning and indexing of pictures in the literature and their retrieval of pictures. Correctness and efficiency have an important impact. Splendiani (2014) used a decision tree to match the text in the medical literature with the pictures in the literature, and add text descriptions to the pictures. In general, this is based on accurate extraction and positioning of pictures. Based on this, in order to extract the feature data of the layout format of the social responsibility report, this thesis chose machine learning to extract the feature data of the layout format of the social responsibility report.

The schematic diagram of the paging method is shown in figure 4-1. The social responsibility report is processed and output as a picture through the pagination of “PDF” format by Python. It is worth noting that since the report itself contains page number

annotations, the way the machine opens the reports may be different. However, the extraction of feature data will not cause omissions because of the difference between the page number of the report itself and the opened format.

Because there are many layout formats in social responsibility reports, such as the design of the entire report and the design of the single-page CSR format, and the single-page report only inserts charts with no pictures. Because the "PDF" format file analytical program is paged according to the page number, when the layout design of the first page and the layout design of the second page are designed together, especially the line drawings with complex elements, it will cause recognition errors. Although the effect of picture paging is better, due to the diversity of picture data, it is inevitable that there will be a small number of format errors.

4.3.1 Feature extraction of text layout

Firstly, this thesis intercepted the pictures in a PDF report. In this step, the pictures are cropped with the help of Pillow and OpenCV. Both Pillow and OpenCV have excellent image processing performance and are widely used in the development of advanced image processing and computer vision fields. The second step is to use the PIL package for extracting text density features. It is a Python image processing package. Basic operations on images are basically included in this module. For functions such as opening, saving, conversion, etc., in this research, PIL is mainly used to open pictures in a way required by machine learning. The third step requires the help of the pytesseract package, which is currently recognized as the best and most accurate OCR (Optical Character Recognition) system. In addition to extremely high accuracy, the Tesseract package also has high flexibility. It can recognize any font through training, and it can also recognize any Unicode character. This study uses the Tesseract package to identify Chinese characters and letters in the pictures to count the text density.

To check the labeling results manually, one has to check whether the positioning of the picture in the text is accurate, such as whether the inserted picture is completely and comprehensively detected, and the other is to check the accuracy and accuracy of text recognition. For the text layout feature in the feature dimension of the social responsibility report, this thesis uses Tesseract, developed by Google, to recognize the text of the image. It recognizes the text density of the image output after the corporate social responsibility report is paginated. The balance of text and pictures is important, the reader can easily feel tired if the page is full of text. When the word count of each page of the social responsibility report is concise, the reader can read the information intuitively.

4.3.2 Feature extraction of picture layout

This thesis marked the features of the social responsibility report page. Firstly, this thesis intercepted the form of pictures, and then this thesis marked the picture size and position of the pictures. The output result of Python is shown in the figure 4-3. Because the output format is changed, in order to ensure the standardization of the data, they data needed to

be recollected in order. For example, in the output of coordinates and size, the feature data of the picture size and picture position shows in different format. In order to ensure the standardization of data processing, the data needs to be cleaned up after data extraction. The cleaned format is “top” and “left”, “height” and “width” are output in the order. They have a maximum value and a minimum value, assuming 500 as the page divider, the value below 500 is located on the left side of the page, the value above 500, the location on the right side of the page. While the width and height are the size of the picture. For a page in the social responsibility report, a picture is inserted in the layout format of the page, and the result of the feature annotation is shown in the picture, the coordinates are (4, 10), and the size of the picture is (1157, 822). The experimental process is as follows, and the figure 4-3 shows the schematic diagram of the picture position and picture size labeling method.

```
pdf/venv/bin/python /Users/dulanmei/Desktop/pdf/pic/图片坐标和大小
'result': {'left': 10, 'width': 1157, 'top': 4, 'height': 822}}
'result': {'left': 3, 'width': 819, 'top': 3, 'height': 563}}
'result': {'left': 47, 'width': 496, 'top': 43, 'height': 702}}
'result': {'left': 16, 'width': 539, 'top': 40, 'height': 759}}
```

Figure 4–3 Schematic diagram of picture position and picture size labeling method

Regarding the features of picture arrangement and page structure, the first step of this research is to use the json package, which is mainly used to convert the data format. It uses its own defined standards to exchange data between the server and the client's browser. The second step used the module of AIP image classify, which is a Python client for image recognition and provides a series of interactive methods for image recognition developers. The research in this thesis uses Baidu AI, which includes functions such as image technology, text recognition, picture recognition, and knowledge graphs. The Baidu AI Open Platform is a platform for deep learning, and is based on a large-scale open data set of real samples, combined with a variety of algorithms to continuously train data to obtain a relatively complete image subject recognition. Its functions include locating the key subject position of the image, and intercepting the subject position in the picture for image processing.

4.4 Three-dimensional data construction of feature extraction in report format

This chapter takes Vanke’s 2018 social responsibility report as an example. It uses image processing technology in the computer field to conduct related experiments. Finally, a method for extracting the features of the format of the social responsibility report is proposed to define the dimensions of the format of the social responsibility report. The Vanke report is suitable as a sample to illustrate the application of the method proposed in this thesis in the social responsibility report. On the other hand, the layout format of the social

responsibility report released by China Vanke in March 2019 has abundant elements for analysis compared with the report published in the year. The data set in this thesis is mainly composed of pictures. Firstly, with the help of the technical route proposed in the previous chapter, the 2018 Vanke Social Responsibility Report is paginated and 65 pictures are obtained. The feature data of the report format is organized into folders.

4.4.1 Two-dimensional feature extraction of report format

The two-dimensional feature of the layout format is composed of the page number and the feature data of each page. Because there are values for top, left, height, width, and word, there are 65 feature data for each feature indicator in the 2018 Vanke Social Responsibility Report, resulting in a total of 325 feature data. The feature data of 65 pictures was extracted into 325 feature data with page number as the unit, as shown in the table 4-2. All features are obtained with the help of the technical route described in the previous chapter. The feature data consists of six feature data in the social responsibility report on each page. The two-dimensional layout format feature data in this study refers to a data collection composed of single-page social responsibility report layout features (text density, picture location, and picture size). Since a report is composed of multiple pages, the two-dimensional data is formed by combining the text and picture layout features of each page. As shown in table 4-2, the ordinate is the page number, and the abscissa is the feature data.

For left, they have a maximum value and a minimum value. The maximum value for left is 1000, the minimum value for left is 0. Assuming 500 as the page divider, therefore, the value below 500 is located on the left side of the page, and the value above 500, the location on the right side of the page. For top, they have a maximum value and a minimum value, the maximum value for top is 600, the minimum value for top is 0. Assuming 300 as the page divider, therefore, the value below 300 is located on the top side of the page, the value above 300, the location on the bottom side of the page. For height, the maximum value for height is 1500, the minimum value for top is 0. Assuming 750 as the page divider, therefore, the value below 750 is relatively small picture, the value above 750 is relatively large picture. For width, the maximum value for height is 2000, the minimum value for height is 0. Assuming 1000 as the page divider, therefore, the value below 1000 is relatively small picture, the value above 1000 is relatively large picture. For the word, the maximum value for the word is 4000, the minimum value for word is 0. As was shown in table 4-2, by using these data, this thesis can transfer un-structure data (picture) into structure data, in order to illustrate the method step by step, table 4-2 take 5 pages as an example.

Table 4–2 Two-dimensional picture page number feature data

Page	Top	Left	Height	Width	Word

1	127	4	337	502	0
2	120	273	577	842	58
3	59	635	676	509	430
4	5	14	784	1142	128
5	4	9	785	1154	53

Due to the rich content and diverse forms of the social responsibility report, the feature data of the actual format of the social responsibility report is complicated. In order to further analyze the page number features and year features of a social responsibility report, this thesis provides the feature values of the five dimensions. The reason for setting the average number for feature description is that the feature data of the social responsibility report has geometric spatial features, and the three-dimensional data can better describe the formatting features of each social responsibility report in different industries and different years.

4.4.2 Three-dimensional feature description of report format

For the three-dimensional data formed by combining industry and page number, this section proposes the method of three-dimensional data feature description to form the feature data of a sample, so as to form a comparison between the sample and the sample. For the processing of three-dimensional data, this thesis proposes a method of describing the feature data, which is to describe the feature data in 0 and 1. By using the number of 0 and 1, CSR feature can be seen clearly, no matter for a report or the reports in the whole industry. First, the division of 1 and 0 is technically easier to achieve, and later clustering or classification to analyze the consistency features of it is easier on the technical route. Second, the division of 1 and 0 has high reliability. Only two numbers are difficult to make mistakes on during data analysis and processing. The choice of two numbers can also be described simply and intuitively. Third, the numbers 1 and number 0 are simpler in terms of rules, which is conducive to extracting feature dimensions conveniently, quickly and intuitively to analyze feature data.

For the feature data that has been extracted, in the text layout feature, if the feature data is 0, and there is no text in the picture on the page, it was marked as 0, and it was indicated in red. If the feature data has a corresponding value, it will be the picture on the page. If the text was designed, it was marked as 1, and it was indicated in white. In terms of picture layout features, if the picture value was 0, there was no picture inserted on the page, and it was marked as 0, otherwise it was marked 1, as was shown in the table 4-3.

Table 4–3 Annotation diagram of two-dimensional feature data

Page	Top	Left	Height	Width	Word

1	0	0	0	0	0
2	0	0	1	0	1
3	0	1	1	0	1
4	0	0	1	1	1
5	0	0	1	1	1

As shown in the table 4-3, the feature data of the first 10 pages of the 2018 Vanke social responsibility report has been marked with the 0 and 1 method described above.

4.5 Summary

This thesis using Python to extract the feature data, which includes text density, picture location and picture size. And this chapter proposes a method for describing a report with three-dimensional data. Firstly, the feature data of each page is obtained with the help of the technical route proposed in this thesis. Because the feature data in the format of the social responsibility report has spatial features, it is combined with page number and year to form one-dimensional data of a sample. For the formed three-dimensional data, this chapter proposes a Seaborn diagram to describe the three-dimensional feature data.

In essence, this chapter used the image recognition model to extract the feature dimensions in the social responsibility report. At the same time, it is found that the feature dimensions of the layout format are difficult to extract. Even if the report has the corresponding chart layout format design, there is no corresponding feature dimension to judge whether the personalized settings of the company in the report have produced the effect of impression management, thereby affecting investment judgment. In the face of this complex and abstract problem, this chapter started from using machine learning to effectively express the features of the format of the corporate social responsibility report.

5 APPLICATION OF FEATURE EXTRACTION METHODS IN CORPORATE SOCIAL RESPONSIBILITY REPORTS

Social responsibility reports have a layout format with colorful pictures and texts. Just like other reports, but the literature review demonstrates that these reports have colorful features, but it is difficult to evaluate these reports objectively, or in terms of data. For CSR, there is no way to evaluate on the report objectively, which gives the report discloser (individual company) the opportunity to disclose the report selectively, such as disclosing only the good parts, and choosing not to disclose the crucial part, for example, the corporate sales data. This is obviously unfavorable for stakeholders making investments (Meng, 2012). This thesis provides a relatively effective method for the feature extraction method for the feature data extraction of social responsibility reports. This chapter takes a sample of Vanke as an example, compares the social responsibility reports of the sample in different years, and compares the differences in the layout format of the real estate industry represented by Vanke and other industries in the same year, to demonstrate the use of the methodology mentioned in this thesis in the field of reports.

5.1 Application in the same report in the same year

The third chapter of this thesis established the feature dimensions of the layout format as text layout features, picture layout features, and page layout features. After paginating the 2018 Vanke Social Responsibility Report, 65 pictures were obtained, and feature data extraction was performed on all 65 pictures. There are 325 feature data in a report. With the page number as the horizontal axis and the feature value as the vertical axis, in the Vanke's existing catalog (which was provided by Vanke), it was divided into company profile, company operations, environmental protection, and corporate governance, and a sample is used to compare different area in different descriptions. The arrangement method adopted at the time to compare the differences in the arrangement of different content. As shown in figure 5-1, all features are obtained with the help of the technical route described in the previous chapter. The feature data consists of six feature data in the social responsibility report on each page. The two-dimensional layout format feature data in this study refers to a data set composed of single-page social responsibility report layout features (text density, picture location, and picture size). Since a report is composed of multiple pages, the two-dimensional data is formed by combining the text and picture layout features of each page.

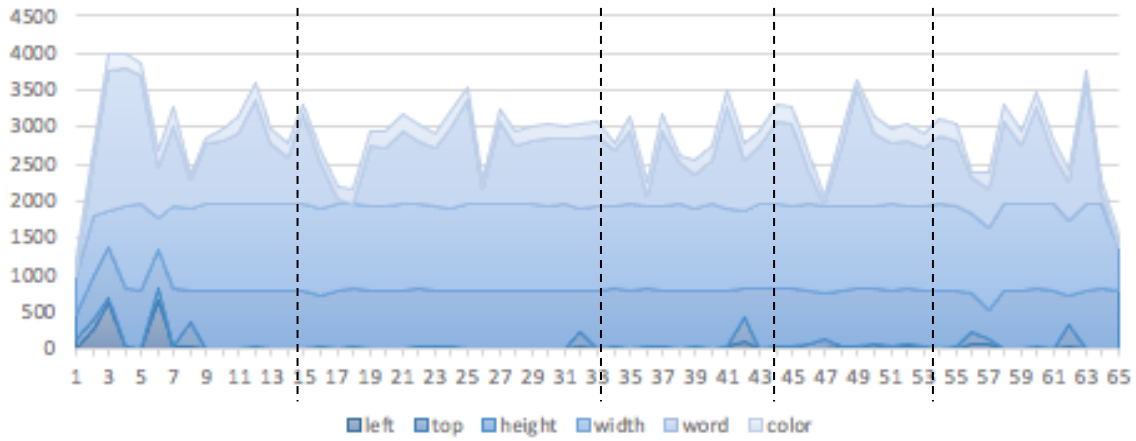

Figure 5–1 Feature data of Vanke's layout format in 2018

Among them, the values of left, top, height, width and text are used as vertical coordinates in the figure, from darker to lighter colors, respectively, where each color has a measured characteristic value shown in the figure, and the page number of the 2018 Vanke Social Responsibility Report is used as the horizontal coordinate. For top, they have a maximum value and a minimum value. The maximum value for top is 600. The minimum value for top is 0. Assuming 300 as the page divider, therefore, the value below 300 is located on the top side of the page, the value above 300, the location on the bottom side of the page. For left, they have a maximum value and a minimum value. The maximum value for left is 1000. The minimum value for left is 0. Assuming 500 as the page divider, therefore, the value below 500 is located on the left side of the page, the value above 500, the location on the right side of the page. For height, the maximum value for height is 1500, the minimum value for top is 0. Assuming 750 as the page divider, therefore, the value below 750 is relatively small picture, the value above 750 is a relatively large picture. For width, the maximum value for height is 2000, and the minimum value for height is 0. Assuming 1000 as the page divider, therefore, the value below 1000 is relatively small picture, the value above 1000 is relatively large picture. For the word, the maximum value for the word is 4000, the minimum value for word is 0.

Since 2018, Vanke's catalog has five sections: company overview, environmental protection, social responsibility, corporate governance, and employee care, so the black line is used to divide on the corresponding catalog into different sections, which can visualize the input of Vanke as a company in different sections. Five areas are divided into company profile, company operation, environmental protection, and corporate governance. The value range of "top" is 0-350, where the smaller the value, the closer to the top of the page number; the value range of "left" is 0-600, the smaller the value, the closer to the left; the value range of "height" is 0-1500, "width" is 0-2000.

However, due to the design of flower patterns and artistic fonts on some pages, the positioning of that part of the picture is the entire picture, and the largest page number is the layout design of the entire picture. From the feature data point of view, the same layout

format design is carried out at the beginning and end of each part. In the company introduction areas, there is not much texts and larger pictures. In terms of company operations and corporate governance, Vanke has designed a lot of pictures with small sizes, which may be photos of the company's daily operations, to reduce the complexity of the report and increase readers' interest in reading. In terms of environmental protection, Vanke has inserted large-scale pictures, which show that Vanke held a lot of environmental protection activities and took a lot of photos. The data can show the company invests a lot of manpower and material resources into collecting statistics, which means the company attaches great importance on the CSR. It is will not choose to use all text description to "fuzzy" the information.

5.2 Application in the same report in different years

Based on its unique advantages, Python can not only process the data of social responsibility reports in batches, but also extract the required data according to needs, which can greatly improve the efficiency of batch data analysis and processing. First, this thesis paginated China Vanke's social responsibility report from 2015 to 2019 and got 304 pictures. Secondly, this thesis used the aforementioned methods to characterize the text layout features, picture layout features, and page structure features of the image data and get 1520 feature values. Among them, the smaller the value of "top", the closer to the top of the page number; the smaller the value of "left", the closer to the left; the two feature values of "height" and "width" are larger. Therefore, the feature dimensions of the layout format of Vanke in different years are marked, the most obvious features are extracted, and a set of layout format feature data for Vanke is formed, as shown in the figure 5-2. They are represented as "word", "top", "left", "height", and "width" from top to bottom.

In the theoretical background, it is mentioned that impression management behavior occurs in the presentation format of social responsibility reports, but there is a lack of an effective means to transform unstructured data into structured data. Red and white are an effective way to transform unstructured data into structured data, which is similar to the 0 and 1 inside the computer. In fact, the color designation is only a means to transform unstructured data into structured data. Similar to the 0 and 1 in the computer, 0 represents the smaller value, located on the right and top of the page, 1 represents the larger value, located on the left and bottom of the page. With the help of red and white colors, this thesis can draw the color chart, so as to visualize the difference between the two reports. With the help of machine learning, this thesis can classify the picture. This is the post-processing of structured data.

In the 2018 Vanke Social Responsibility Report, Vanke's five-year samples were counted. For top, they have a maximum value and a minimum value. The maximum value for top is 600. The minimum value for top is 0. Assuming 300 as the page divider, therefore, the value below 300 is located on the top side of the page, marked as 0 and red. The value is above 300, at the location on the bottom side of the page, marked as 1 and white.

For left, they have a maximum value and a minimum value, the maximum value for left is 1000. The minimum value for left is 0. Assuming 500 as the page divider, therefore, the value below 500 is located on the left side of the page, marked as 0 and red, the value above 500, the location on the right side of the page, marked as 1 and white. For height, the maximum value for height is 1500, the minimum value for top is 0. Assuming 750 as the page divider, therefore, the value below 750 is relatively small picture, marked as 0 and red, the value above 750 is relatively large picture, marked 1 and white color. For width, the maximum value for height is 2000, the minimum value for height is 0. Assuming 1000 as the page divider, therefore, the value below 1000 is relatively small picture, marked as 0 and red, the value above 1000 is relatively large picture, marked as 1 and white. For the word, the maximum value for word is 4000, the minimum value for word is 0. Assuming 2000 as the page divider, therefore, the value below 2000 is relatively small picture, marked as 0 and red, the value above 2000 is relatively large picture, marked 1 and white. The reason why the values are recorded is that a picture is visually unstructured and cannot be objectively compared and analyzed. With the help of these values the picture can be transformed into a comparable value for next step analysis.

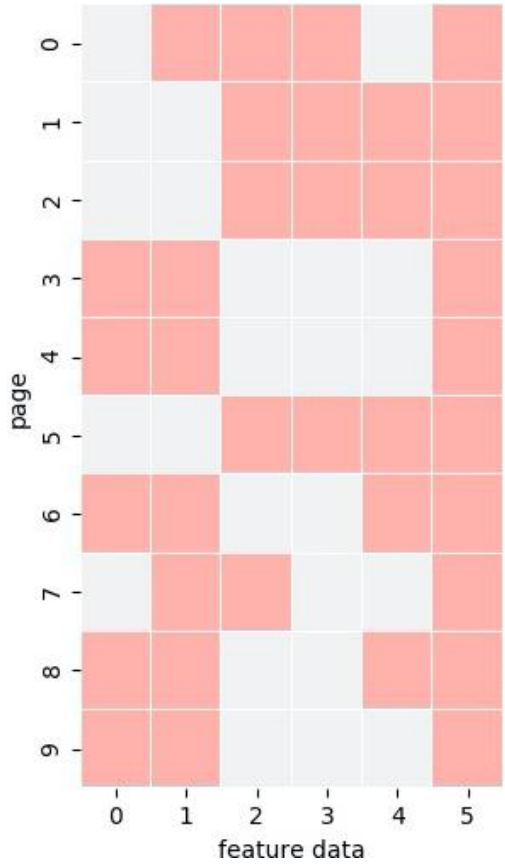

Figure 5–2 Page number feature data

The 1520 feature values are processed with the help of the Seaborn two-color diagram, and the feature data composed of each report from 2015 to 2019 is counted, and the dimensionality is reduced to form Vanke's feature data from 2015 to 2019. At present,

for the feature extraction of the format of the social responsibility report, the traditional manual annotation is time-consuming and labor-intensive.

Feature extraction is a key point in accurately describing the formatting of a CSR report. It is the process of extracting those points that represent the image features or text features of a page of a CSR report from the image features of a single page of a CSR report, and then generating feature lines and feature surfaces based on those feature points.

The goal of feature extraction is thesis feature points, which will help to quickly and easily understand the important information of the object. Figure 5-2 shows the characteristic data of the first ten pages of Vanke's social responsibility report, which are shown in white and red. Taking the first ten pages of Vanke's social responsibility report in 2018 as an example, in terms of picture position, the pictures in Vanke's social responsibility report are mostly located on the left and above. In terms of picture size, larger pictures are inserted in the first three sides. In terms of text, the corresponding pictures are larger and the text is smaller. In terms of color, light-colored pictures are used. The red and white colors are to visualize the use of pictures in a social responsibility report.

This thesis is to investigate whether the use of pictures in a social responsibility report has an impression management behavior, but since there is no previous study about pictures in the report, this thesis proposes a method to do research on the impression management in social responsibility reports. In this section, after extracting the feature data of Vanke from 2015 to 2019, the feature data in the unit of year is formed, and the feature points of Vanke are extracted by combining the feature data extraction method proposed in the previous chapter.

5.3 Application in different industries in the same year

With the help of the research methods mentioned above, feature extraction was performed on the 2018 PingAn Bank Social Responsibility Report and the 2018 Vanke Social Responsibility Report, a feature data map was formed for each report, and the similarity and correlation between the maps were used to judge the difference in industry features. The reason why Vanke and PingAn Bank are selected as samples in this thesis is that the ratings on Hexun.com have a relatively stable difference between Vanke and PingAn Bank. Secondly, after processing 2000 CSR in 2018, it is concluded that compared with other reports in the same industry (in the industry of financial and real estate), Vanke and PingAn Bank are richer in format. In terms of layout, the social responsibility reports of Vanke and PingAn Bank have a specific compilation basis. Figure 5-3 shows the comparison of feature maps between PingAn Bank and Vanke.

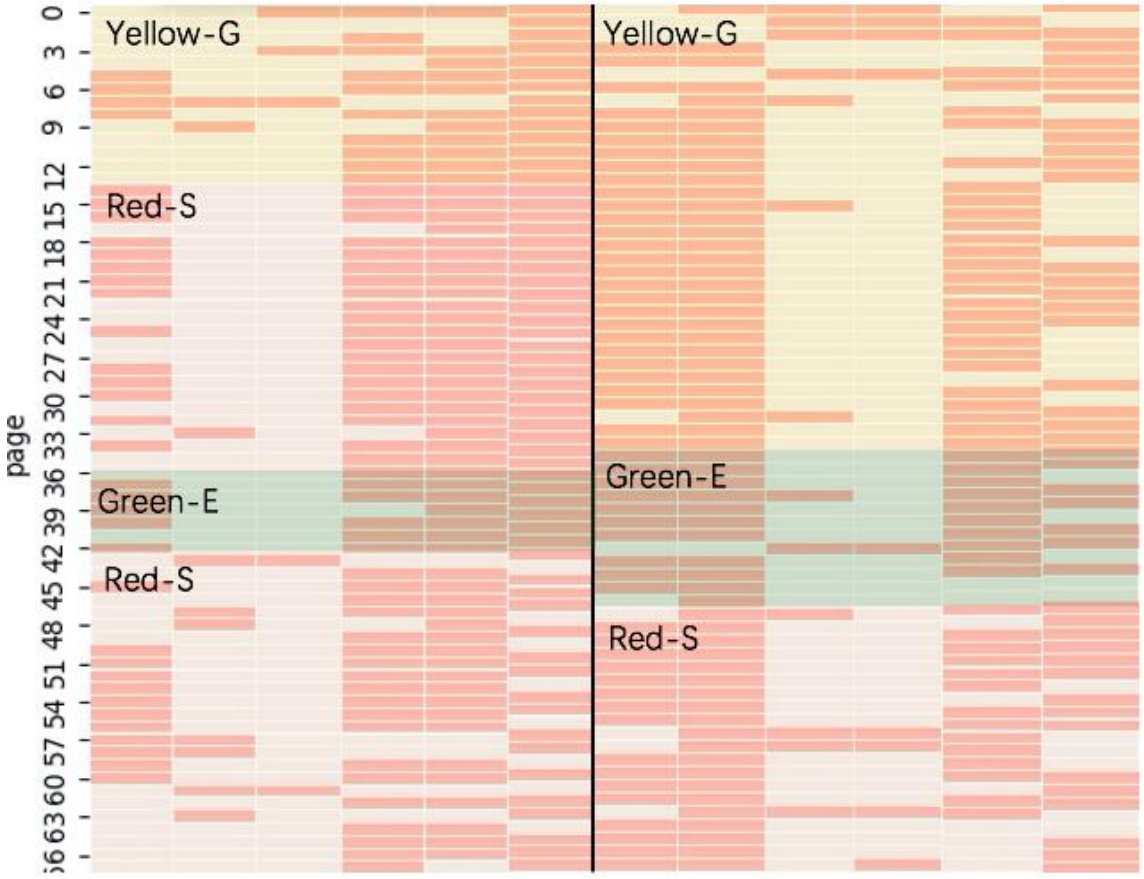

Figure 5–3 Comparison of feature data between PingAn Bank and Vanke

As shown in the figure above, with the page number as the vertical axis and the feature data as the horizontal axis, ESG is used as the color division of the page numbers of the two social responsibility reports. From the perspective of social governance (ESG) and corporate value, it also refers to the relationship between corporate governance and social value. Among them, green represents the environment (E), which mainly evaluates the green investment in the production and operation activities of enterprises, the recycling and sustainable utilization of natural resources and energy, and the treatment of hazardous waste; red represents society (S), which mainly investigates the relationship between enterprises and the government, employees, customers, creditors, as well as relevant stakeholders inside and outside the community. Yellow represents corporate governance (G), which mainly includes the structure of the board of directors, ownership structure, management compensation and corporate ethics, such as the interests and responsibilities of shareholders and management, avoiding corruption and financial fraud, and improving the transparency, independence and professionalism of the board of directors (Yang, 2020).

For top, they have a maximum value and a minimum value. The maximum value for top is 600. The minimum value for top is 0. Assuming 300 as the page divider, therefore, the value below 300 is located on the top side of the page, marked as 0 and red. The value above 300, at the location on the bottom side of the page marked as 1 and white. For left,

they have a maximum value and a minimum value. The maximum value for left is 1000. The minimum value for left is 0. Assuming 500 as the page divider, therefore, the value below 500 is located on the left side of the page, marked as 0 and red. The value above 500 is located on the right side of the page, marked as 1 and white. For height, the maximum value for height is 1500, the minimum value for top is 0. Assuming 750 as the page divider, therefore, the value below 750 is relatively small picture, marked as 0 and red, the value above 750 is relatively large picture, marked 1 and white color. For width, the maximum value for height is 2000, and the minimum value for height is 0. Assuming 1000 as the page divider, therefore, the value below 1000 is relatively small picture, marked as 0 and red, the value above 1000 is relatively large picture, marked as 1 and white. For the word, the maximum value for word is 4000, the minimum value for word is 0. Assuming 2000 as the page divider, therefore, the value below 2000 is a relatively small picture, marked as 0 and red, while the value above 2000 is a relatively large picture, marked 1 and white.

The first part of PingAn Bank is corporate governance, which is mainly composed of society and relevant stakeholders, and a small part describes the environment. From the similarity point of view, PingAn Bank and Vanke have carried out design in the format, and the image size distribution in each part of ESG is uniform. Although the layout design of PingAn Bank is relatively small, the inserted pictures are larger. Combined with the text density, it may be table design, expressed in the form of data, reducing the text.

5.4 Summary

In this chapter, the application of the feature extraction method proposed in this research in the field of social responsibility is used, including the feature description in a report, the feature description of different industry reports, and the feature data of different years. In the company profile, company operations, environmental protection, and corporate governance sections of a report, the same layout format is designed at the beginning and end of each chapter, often with less text and larger pictures.

In the company profile, large pictures are designed to help understanding. In terms of company operations and corporate governance, Vanke has designed pictures of small sizes to reduce the complexity of the report. In different industries, PingAn Bank and Vanke have made designs in the layout format compared with other reports in the financial and real estate industry. Most of the pictures inserted by Vanke are photos, because the size of the pictures is small, but the position of each page is relatively small in the middle of the page. Although PingAn Bank's layout format design is relatively small, it has inserted tables and presented it in a data-based form, reducing the text.

6 DISCUSSION AND CONCLUSION

6.1 Research conclusion

As a bridge for companies to send important signals to the outside world, social responsibility reports are of great importance to a company. Regarding the layout format, the layout format of social responsibility reports in recent years has been shown with pictures and texts. It is difficult to define the feature dimensions of the format of the social responsibility report. Even if the definition problem can be solved, the feature data is difficult to get. This research proposes a method of extracting formatted feature data from social responsibility reports.

Firstly, from the perspective of impression management and graphical representation effects, this thesis explored the feature dimensions of the format of the social responsibility reports. Studies have demonstrated the impact of the presentation of social responsibility reports on stakeholder perception management, and have confirmed a correlation between reporting effectiveness and presentation, but there is a lack of effective methods for extracting characteristic dimensions.

Secondly, after studying the quality evaluation of the existing social responsibility report layout format, and combining the layout formats of social responsibility reports in recent years, the feature dimensions available for analysis and measurement are image layout features, text layout features and page structure features. After examining the technical routes that can be implemented, the feature indicators that can be measured in this study's feature dimensions are text density, image location, image color and image size. This study proposes a Seaborn diagram for feature extraction for two-dimensional data with different page numbers and three-dimensional data with different page numbers and years.

Finally, for the methodology proposed in this study, how it is applied in social responsibility reports is explored in a chapter, which includes the characterization of one social responsibility report by catalogue classification. The application and characterization of data on the characteristics of Vanke's social responsibility report as a sample in different years in the same industry; and the comparison of the characteristics of the financial industry represented by PingAn Bank and Vanke's social responsibility report in different industries in the same year. The thesis found out that compared with black and white background, which was full of text, some companies have started to design large-sized pictures in the company profile these years.

From the perspective of different industries, comparing the financial industry with PingAn Bank as the sample with the real estate industry with Vanke, Vanke’s layout format elements are abundant. In comparison, PingAn Bank inserted a lot of table data. Looking at the same report year wise, the feature description method can have a look at

the feature data dimension and page dimension, then use the image classification to form a feature description of the sample.

This research proposes a method for effectively extracting the feature data from the format of the social responsibility report. The unstructured data is extracted as structured data for a single sample analysis and comparison.

6.2 Limitation and further suggestions

Compared with the existing research, the feature dimension model of layout format proposed in this thesis broadens the connotation of the evaluation of the layout format of social responsibility report, and introduces it from computer graphics to the research field of management, and the extraction of data feature dimensions can also be effective. It breaks through the difficulty of establishing the feature index of the past layout format, the difficulty of extracting the feature data, and the complex and diverse data that is conducive to intuitively obtaining the required data.

Due to time and space constraints, this thesis extracts three feature dimensions and six feature indicators. While the three dimensions of the layout format as an artifact are the focus of the research, future research can focus on the other dimensions of the layout format, such as color. Will the company use colors in the report to form its own corporate identity? The exhaustive content of the CSR report should be considered if future research about format is used to make a shift from facts and tables.

REFERENCES

Adams, C. (2004). The Ethical, Social and Environmental Reporting-Performance Portrayal Gap. *Accounting Auditing & Accountability Journal*, Vol.17(5), 731-757.

Apostolova, E., You, D., Xue, Z. (2013). Image Retrieval from Scientific Publications: Text and Image Content Processing to Separate Multi panel Figures. *Journal of the American Society for Information Science and Technology*, Vol. 64(5), 893-908.

Bakar, A. S. A., Ameer, R. (2012). Readability of Corporate Social Responsibility communication in Malaysia. *Corporate Social Responsibility & Environmental Management*, Vol. 18(1), 50-60.

Beattie, V. A., Jones, M. J. (2008). Corporate reporting using graphs: A review and synthesis. Journal of Accounting Literature, 27.

Cesario J, Higgins, E.T. (2008). Making message recipients "feel right": How nonverbal cues can increase persuasion. Psychological Science, Vol. 19 (5), 415-420.

China News. (2020). Chinese companies are fulfilling their social responsibility at high speed.http://epaper.bjnews.com.cn/html/2019-03/22/127content_749819.Html.div=-1.

Dagiliene, Lina. (2010). The Research of Corporate Social Responsibility Disclosures in Annual Reports. Engineering Economics, Vol. 21(2), 197-204.

Dai, H. T. (2014). Research on the effect of corporate social responsibility report impression management. Jinan University.

Dwyer, B. (2002). Managerial perceptions of corporate social disclosure: An Irish story. *Accounting Auditing Accountability*, Vol.15(3), 406-436.

Fang, H., Shang, Y., Zhang, R. (2019). The reading effect of infographics in data journalism: evidence from eye movements. *Library and Information Service*, Vol. 63 (8), 74-86.

Fang, Y., Xie, J., Dai, G. (2015). 3D deep shape descriptor. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2319-2328.

Feng, L. Y., Xiao, X., Zhao, T. J. (2016). The impact of economic performance on corporate social responsibility disclosure. *Journal of Management,* Vol. 13(7), 1060-1069.

Financial and Investment Institutions Operating Environment and Policy Research Group. (2020). ESG Investment Philosophy and application prospects. Vol. 1, 68-76.

Friedman. (1970). The social responsibility of business is to increase its profits. *New York Times Magazine*, 122-126.

Gao, Y. Q., Chen, Y. J., Zhang, Y. J. (2012). "Red Scarf" or "Green Scarf": Research on the Motivation of Charitable Donations by Private Enterprises. *Management World,* Vol. 8, 106-114.

Giovanna, Michelon. (2011). The Nature, Use and Impression Management of Graphs in Social and Environmental Accounting. *Accounting Forum,* Vol. 2, 75-89.

Global Reporting Initiative. (2016). Gri Standard. https://www.globalreporting.org/standards.

Goffman, E. (1959). *The Presentation of Self in Everyday Life*. New York: Doubleday Anchor.

Guo, H., Wang, J., Gao, Y. (2016). Multi-view 3D object retrieval with deep embedding network. *IEEE Transactions on Image Processing*, Vol.25(12), 5526-5537.

Huang, Y. X., Yao, Z. (2016). Corporate Social Responsibility Report, Impression Management and Corporate Performance. *Economic Management,* Vol. 1, 105-115.

International Original for Standardization. (2010). *Guidance on Social Responsibility*. Switzerland: ISO.

Jiang, Y. J., Xu, Y., Liu, D. (2011). The market response of corporate social responsibility information disclosure: an event study based on social responsibility reports issued by listed companies in China. *Economic and Management Research*, Vol. 8, 123-128.

Jiang, Y. P. (2008). Research on Self-interest Tendency in Information Disclosure of Attribution of Earning Changes in Listed Companies. *Modern Management Science*, Vol. 6, 117-119.

Ji, L. (2016). A research framework for managing the impressions of CSR disclosure: motivations, strategies and economic consequences. *Journal of Zhengzhou Institute of Aviation Industry Management*, Vol. 34(2), 76-84.

John, K. Courtis. (2004). Color as Visual Rhetoric in Financial Reporting. *Accounting Forum*, Vol. 28 (3), 265-281.

Jones, E. E. (1964). *Ingratiation.* New York: Appleton-Century-Crofts, 1964.

Jones, M., A, Shoemaker. P. (1994). Accounting Narratives: A Review of Empirical Studies of Content and Readability. *Journal of Accounting Literature,* Vol. 13, 142.

Kahneman, D., Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, Vol. 47(2):263-291.

Kar, A., Tulsiani, S., Carreira, J. (2015). Category-specific object reconstruction from a single image, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. *Boston*, 1966—1974.

Kershaw, R., Tuttle, B. M. (1998). Information presentation and judgment strategy from a cognitive fit perspective. *Journal of Information Systems*, Vol. 12(1), 1-17.

Lee, P. S., West, J. D., Howe, B. (2016). Vision metrics: Analyzing Visual Information in the Scientific Literature. *IEEE Transactions on Big Data*, Vol. 4(1), 117-129.

Li, H., Wang, J. Q., Tu, J. (2009). Management's impression management behavior and corporate social responsibility disclosure. *Finance and Accounting Newsletter*, Vol. 21, 144-146.

Lin, Z. G., Wu, L. J. (2004). The relevance of corporate governance to the quality of accounting information. *Accounting Research*, Vol. 24(8), 65-71.

Liu, W. L., Song, Y. F. (2017). Construction of the Quality Evaluation System of Corporate Social Responsibility Reporting. *Business Accounting*, Vol. 4, 24-27.

Mackey, A., Mackey, T. B., Barney, J. B. (2007). Corporate social responsibility and firm performance: investor preferences and corporate strategies. *Academy of Management Review,* Vol. 32(3), 817-835.

Mao, H. T., He, X. Q., Su, M. (2014). Reporting format, personal ability and management accounting information decision value: an experimental study of pricing decision. *Accounting Research*, Vol. 7, 67-74+97.

Meng, X. (2012). *Research on the readability of corporate social responsibility reports based on impression management perspective*. Southwest University of Finance and Economics.

Mills, D. L., Gardner, M. J. (1984). Financial profiles and the disclosure of expenditures for socially responsible purposes. *Journal of Business Research*, Vol. 12(4), 407-424.

Nie, W. Q. (2019). Reshape the brand image design of Chinese private enterprises based on CIS theory. Hefei University of Technology.

Qin, X. Z., Wang, Z. S., Zhao, H. (2018). Corporate governance and corporate social responsibility disclosure. Corporate governance and CSR disclosure - a study of SMEs based on GEM [J]. Management Review, Vol. 30(3), 188-200.

Reggy, Hooghiemstra. (2000). Corporate Communication and Impression Management: New Perspectives Why Companies Engage in Corporate Social Reporting. *Journal of Business Ethics,* Vol. 27(1/2), 55-68.

Rosenfeld, R. A., Riordan, C. A. (1995). *Impression management in organizations: Theory, measurement and practice*. London, Rout Ledge.

Shaft, T. M., Vessey. (2006). The role of cognitive fit in the relationship between software comprehension and modification. *Management Information Systems Quarterly*, Vol. 30 (1), 29–55.

Shen, H. T., Li, Y. L. (2010). Analysis on the status quo of environmental information disclosure of listed companies in my country's heavy pollution industries. *Securities Market Guide,* Vol. 6, 51-57.

Shen, H. T. (2007). A review of the main models of foreign corporate social responsibility reporting. *Securities Market Herald,* Vol. 8, 7-13.

Song, X. Z. (2009). Corporate sustainable development and social responsibility and its information disclosure. *Financial Accounting Study*, Vol. 3, 12-13.

Song, X. Z., Gong, M. X. (2009). Quality of social responsibility information and evaluation of decision-making value: content analysis of annual accounting reports of listed companies. *Accounting Research*, Vol. 2, 37-43.

Splendiani, B., Ribera, M. (2014). How to Textually Describe Images in Medical Academic Publications. Proceedings of the XV International Conference on Human Computer Interaction.

Sun, M. L., Jiang, L., Sun, J. (2013). A study on self-interest disclosure and market response of performance attribution: the performance of textile industry under unilateral appreciation of exchange rate. *Accounting Research,* Vol. 4, 46-51.

Sun, M. L. (2005). Research on impression management behavior in company annual reports. Beijing: Renmin University of China Press.

Tan, X. (2017). Industry competition, nature of property rights and CSR information disclosure--an analysis based on signaling theory. Industrial Economics Research, Vol. 3, 19-32.

Tao, Y., Dong, D. Y. (2013). An empirical study on the relationship between institutional environment and corporate social responsibility information disclosure. *China CPA*, Vol. 12, 63-68

Teoh, H. Y., Thong, G. (1984). Another look at corporate social responsibility and reporting: An empirical study in a developing country. *Accounting Organizations & Society*, Vol. 9(2), 189-206.

Tversky, A., Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, Vol. 211(4481), 453-458.

Udayasankar, K. (2008). Corporate social responsibility and firm size. *Journal of Business Ethics*, Vol. 83(2), 167-175.

Vessey, I. (1991). Cognitive fit: A theory-based analysis of the graphs versus tables literature. *Decision Sciences*, Vol. 22, 219–240.

Vivien, Beattie., Michael, John. (2000). Impression Management: The Case of Inter-Country Financial Graphs. *Journal of International Accounting, Auditing and Taxation,* Vol. 9 (2), 159-183.

Vivien, F. (2010). Fuzzy-trace theory and framing effects in choice: Gist extraction, truncation, and conversion. *Journal of Behavioral Decision Making*, Vol. 4(4), 249-262.

Wang, Q., Zhang, J. (2020). Research on the Disclosure of Corporate Social Responsibility Accounting Information. *Modern Auditing and Economy*, Vol. 1, 30-33.

Wang, X. D. (2014). Research on Influencing Factors of Management Accounting Reporting Function. Southwestern University of Finance and Economics.

Wood, D. J. (1991). Corporate Social Performance Revisited. *Academy of Management Review*, Vol.16(4), 691-718.

Yang, H. (2016). The impact of graphic accounting information on investor judgment and decision-making. Shandong University of Finance and Economics.

Yang, H. Y. (2020). A Comparison of Domestic and Foreign ESG Systems. *Enterprise Reform and Management*, Vol. 2, 51-52.

Yang, J. (2013). An empirical study on the factors influencing corporate social responsibility disclosure--based on 2010 CSR reports of listed companies in Shanghai and Shenzhen. *Finance and Accounting Newsletter*, Vol. 3, 85-87.

Yang, S. L., Mao, Y. H. (2014). International Comparison and Enlightenment of Comprehensive Reporting Practice. *Finance and Accounting Monthly*, Vol. 1, 154-157.

Yin, H. (2015). Media monitoring and CSR disclosure: empirical evidence based on the food industry. *Financial Accounting Newsletter,* Vol. 14, 24-27.

Yuan, Z. Q., Shen, H. T. (2011). The practice of Novo Nordisk's comprehensive report and its enlightenment to my country. *Finance and Accounting,* Vol. 4, 73-74.

Yu, F. C., Lu, W. (2020). A method of constructing a data machine for academic literature icon position annotation. *Data Analysis and Knowledge Discovery,* Vol. 6, 35-42.

Zech, J., Pain, M., Titano, J. (2018). Natural Language-based Machine Learning Models for the Annotation of Clinical Radiology Reports. *Radiology*, Vol. 287(2), 570-580.

Zhang, A., Xu, Y. Q., Chen, F. (2014). Research on the Quality Evaluation and Influencing Factors of the Social Responsibility Report of Central Enterprises. *Journal of Capital University of Economics and Business*, Vol. 3, 97-105.

Zhang, B. B., Qi, L. Y., Li, T. F. (2016). Research on the Quality Evaluation of Corporate Social Responsibility Reports. *Scientific Research Management*, Vol. 4, 645-651.

Zhang, B., Zhao, D., Chang, Z. (2020). Public Transport's practice and experience in compiling and publishing corporate social responsibility reports. *Urban Public Transport*, Vol. 2, 26-27.

Zhang, Z. Y., Ji, L., Mao, H. T. (2018) Does corporate governance affect social responsibility information disclosure? --Empirical evidence from social responsibility reports of Chinese listed companies. *Economic Economics,* Vol. 6, 107-111.

Zhao, M. (2007) Analysis of impression management behavior in voluntary information disclosure of listed companies. *Contemporary Finance,* Vol. 3, 117-119.

Zhao, X. H., Wang, Z. J. (2020). Analysis on the compilation and application of corporate social responsibility reports. *International Vision,* Vol. 10, 120-124.

Zhong, H. W. (2009). *Chinese Corporate Social Responsibility Reporting Guide 1.0 (CASS-CSR1.0)*, Economic Management Press.

Zhong, H. W. (2011). *Chinese Corporate Social Responsibility Reporting Guide 2.0 (CASS-CSR2.0)*. Economic Management Press.

Zhong, H. W. (2014). *Chinese Corporate Social Responsibility Reporting Guide 3.0 (CASS-CSR3.0)*. Economic Management Press.

Zhong, H. W. (2019) *Chinese Corporate Social Responsibility Reporting Guide 4.0 (CASS-CSR4.0)*. Economic Management Press.

APPENDICES

Appendix 1 Interview questions:

1. Have you read the social responsibility report before? What's your impression?
2. Do you know the motivation or purpose of the format design of the corporate social responsibility report?
3. When it comes to the innovation of social responsibility reports, what do you think?
4. When it comes to the readability of social responsibility reports, what do you think?
5. What's your impression on the format of the social responsibility report in recent year?
6. What do you think is the standard of social responsibility report format?
7. When it comes to measuring the format of social responsibility reports, what comes to mind?

Picture:

8. What do you think of the graphic features of this page (page 4 in 2018 Vanke CSR)?
9. Which dimensions of a picture layout features do you think are important?
10. The dimensions that can be extracted in the format?
11. What do you think are the Feature index in the picture dimension?
12. What other dimensions do you think in the picture format?

Word:

13. What do you think of the character of text arrangement?
14. What is the feature index in this dimension?
15. How do you think the feature index should be measured?
16. What other indicators are important?

Page:

17. What do you think are the layout features of page structure?
18. If you were asked to analyze the page structure feature, how would you analyze it?
19. What do you think is the measurement of page layout in terms of technical means?
20. What do you think can be improved in the layout of page structure?