The Characteristics of SDGs of Internet Enterprises Based on Text Mining

Wanyi Zhang(✉) and Xiang Xie

School of Economics and Management, Beijing Jiaotong University, Haidian District, Beijing, China
zwyya123@foxmail.com, xxie@bjtu.edu.cn

Abstract. Internet enterprises play an important role in promoting the process of global sustainable development. Based on association rules and keyword co-occurrence clustering, this paper studies the sustainable development report issued by Internet enterprises. This paper obtains the sustainable development reports released by Baidu, Alibaba and Tencent, the three largest Internet companies in China from 2016 to 2020, and extracts 919 project text information. After text classification, this paper finds that the implementation of enterprise business is more related to SDG1, 3, 8 and 9. The association rules are used to identify the relevant keywords in the text. It is found that "technology" and "platform" are the keywords of multiple goals. The keyword co-occurrence clustering method is used to carry out visual analysis of the keywords under each sustainable development goal. The research results describe the characteristics, degree and way of implementation of the sustainable development agenda and SDGs in Internet enterprises.

Keywords: SDGs · Internet enterprises · Association rules · Co-occurrence clustering

1 Introduction

In September 2015, the influential document "Transforming our World: The 2030 Agenda for Sustainable Development" was released at the United Nations Development Summit. The document sets out a set of 17 sustainable development goals (SDGs), as shown in Table 1. The agenda aims to solve the problems of hunger, extreme poverty, inequality and injustice, taking into account the three aspects of sustainable development: economy, society and environment, which is linked with the "triple bottom line" of enterprises proposed by Elkington (1998), which means that enterprises should achieve better financial performance, environmental protection objectives and social equity at the same time. Jayaprakash and Pillai (2018) classified the 16 development goals other than the promotion goal partnership (SDG17) as social, economic and environmental development goals respectively. Rosati and Faria (2019) applied the concept of sustainable development to the enterprise level and regarded it as the core element of enterprise sustainable development. The adoption of the sustainable development agenda has further clarified the roles and responsibilities of enterprises in sustainable development.

Z. Zeng et al. (Eds.): ECIT 2022, AHE 11, pp. 348–358, 2023.
https://doi.org/10.2991/978-94-6463-005-3_34

Table 1. SDGs and their contents.

SDG	Content	SDG	Content
1	No Poverty	2	Zero Hunger
3	Good Health and Well-Being	4	Quality Education
5	Gender Equality	6	Clean Water and Sanitation
7	Affordable and Clean Energy	8	Decent Work and Economic Growth
9	Industry, Innovation, and Infrastructure	10	Reduced Inequalities
11	Sustainable Cities and Communities	12	Responsible Consumption and Production
13	Climate Action	14	Life Below Water
15	Life on Land	16	Peace, Justice and Strong Institutions
17	Partnerships		

The proposal of SDGs provides a series of comprehensive, specific and implementable goals for the process of sustainable development. A large number of studies have proposed various suggestions and strategies to support the company to integrate SDGs into strategic management. Jones et al. (2016) outlined SDGs and business participation of enterprises, and put forward some thoughts on many challenges that enterprises will face when making contributions to SDGs. Horne et al. (2020) studied the role of German enterprises in achieving the SDGs in Germany, and proposed a new model to guide decision makers to make the greatest contribution to the SDGs.

As a kind of stakeholders, Internet enterprises play an important role in the process of sustainable development. With the progress of information technology and the rapid development of the Internet industry, a large number of excellent Internet enterprises have emerged in China, such as Sohu, Tencent, Alibaba, Baidu and so on. By setting some key performance indicators, Internet enterprises embed SDGs into their business strategies, and describe their efforts for sustainable development in the sustainable development report according to their completion status. Therefore, the report is an important way to understand the sustainable development status of Internet enterprises. In the research of reporting documents, text mining has been widely used. Wang et al. (2020) applied the method of text mining to study the basic responsibilities and extended responsibilities of the sustainable development goal of the shipping industry and the potential cooperation of the value chain. Long et al. (2020) proposed a bibliometric and text mining method to describe the contribution of water research to the achievement of sustainable development goal 6.

Baidu, Alibaba and Tencent, three representative enterprises in China's Internet industry with large scale and high proportion of operating profits, are selected to study the sustainable development report by using the method of text mining, analyze the implementation of sustainable development of Internet enterprises, and explore what practical measures and plans for sustainable development are adopted by Internet enterprises, provide specific support for Internet enterprises to implement SDGs.

2 Materials and Methods

2.1 Materials

The research selected the sustainable development reports disclosed by Baidu, Alibaba, Tencent as the research goal, because these companies are large-scale Internet enterprises in China, their business income accounts for more than 25% of the total income of China's top 100 Internet enterprises, and their operating profit accounts for nearly 60% of the total. They are the most influential participants in the industry. The reports obtained in this paper are the reports related to the SDGs disclosed by these enterprises from 2016 to 2020, which are from the official websites of various enterprises. After obtaining the report, further analysis is carried out through data preprocessing and text mining.

Through optical character recognition (OCR), the report in PDF format is converted to TXT format for subsequent processing. The information obtained includes a large number of text paragraphs, picture information, tables and other contents. According to the needs of research, delete lines with less than 10 words to remove the information such as tables and titles in the text, and correct a small amount of errors such as garbled code and wrong characters in the text.

In the process of text classification, the research takes an activity or item mentioned in the report as an analysis unit, extracts 919 item texts from the report, and forms a further treatable text corpus data set. Next, refer to the official website of the United Nations sustainable development goals (https://sustainabledevelopment.UN.org), identify the key words related to each SDG, use this as an indicator to classify the text, and assign groups of SDGs to each analysis unit. For example, activities that include keywords such as poverty, and low income are assigned to SDG1; Assign activities with keywords such as education and students to SDG4. According to the results of text classification, the research can describe the overall implementation of the SDGs of Internet enterprises.

2.2 Methods

2.2.1 Chinese Word Segmentation

Word segmentation is the first step in the semantic processing of Chinese text. Research and use the Jieba of Python to realize Chinese word segmentation and delete stop words, and add custom words such as "cloud computing", "government cloud", "globalization" and so on; The meaningless words in text are removed in combination with the deactivation vocabulary. On the basis of the general deactivation vocabulary and according to the results of multiple word segmentation, the original deactivation vocabulary is expanded and meaningless high-frequency words are added, such as "Baidu", "Alibaba", "Alibaba", "Tencent" and so on. After word segmentation and removing stop words, a text matrix is formed. The rows of the matrix represent events related to the SDGs. The column contents of the matrix are words. The word segmentation results are obtained for the next keyword recognition of each sustainable development goal based on association rules.

2.2.2 Association Rules

Association rules are used to mine valuable relationships between data items in a large amount of data. The implication of P ⇒ Q is used to represent association rules, and D represents the set of transactions. Support represents the support degree of association rules. Its calculation is shown in formula (1), which represents the probability that P and Q appear in transaction set D at the same time. The minimum support threshold of transaction set is called the minimum support degree, which is expressed by minSup, which reflects the minimum importance of Association rules. Confidence is used to measure the reliability of association rules, that is, the conditional probability of data. Specifically, it refers to the probability that one data p appears and another data Q appears. Formula (2) is used to calculate the confidence of association rules. The lowest confidence threshold of transaction set is called the minimum confidence, which is expressed by minConf, which reflects the lowest credibility of association rules. Association rules with support not less than minSup and confidence not less than minConf are called strong association rules. The values of minSup and minConf are set by a priori knowledge judgment.

$$\text{Support}(\text{P} \Rightarrow \text{Q}) = \text{count}(\text{P} \cup \text{Q})/\text{count}(\text{D}) \tag{1}$$

$$\text{Confidence}(\text{P} \Rightarrow \text{Q}) = \text{count}(\text{P} \cup \text{Q})/\text{count}(\text{Q}) \tag{2}$$

In the actual mining, the data may be sparse, that is, the number of items in the item set is more, and each transaction in the transaction set contains less data. At this time, transforming the data into sparse matrix can improve the efficiency of the algorithm.

2.2.3 Keywords Co-occurrence Clustering

Keyword co-occurrence is an analysis method of the common phenomenon of keywords and other information in an article. This method holds that the keywords in a piece of text information centrally reflect its main idea, and the simultaneous appearance of a group of keywords reflects its relevance. Therefore, the keyword co-occurrence network can be constructed according to the co word frequency of keywords in the text information. A co-occurrence matrix M containing keywords can be defined as shown in Eq. (3), where $\text{C}(\text{w}_\text{i}, \text{w}_\text{j})$ represents the co-occurrence frequency of keywords w_i and w_j.

$$M = \begin{bmatrix} C(w_1, w_1) & C(w_1, w_2) \cdots & C(w_1, w_n) \\ C(w_2, w_1) & C(w_2, w_2) \cdots & C(w_2, w_n) \\ \vdots & \vdots & \vdots \\ C(w_n, w_1) & C(w_n, w_2) \cdots & C(w_n, w_n) \end{bmatrix} \tag{3}$$

According to the keyword co-occurrence matrix, taking the keyword w_i (i = 1, 2, 3…n) as the node and the co-occurrence frequency of each group of keywords as the weight of the edge, the co-occurrence network composed of keywords and their correlation can be drawn. The network is divided by clustering method to form a cluster with relatively independent concepts and obvious characteristics, and the characteristics of each cluster can be analyzed. The hierarchical clustering algorithm is used to divide the cluster of co-occurrence network. Through the input data set and the number of clusters, the diameter and average distance of classes are used to measure.

3 Results and Discussion

3.1 The Overall Characteristics of the Implementation of SDGs

By assigning the 919 project texts obtained in the sustainable development report to the SDGs, the proportion of SDGs can be used to describe the implementation of the SDGs of Internet enterprises, as shown in Fig. 1. The Internet industry has the main contribution to SDG3 (Good Health and Well-Being, 18%) and SDG8 (Decent Work and Economic Growth, 13%), followed by SDG1 (No Poverty, 11%) and SDGs9 (Industry, Innovation, and Infrastructure, 10%). Internet companies have relatively little sustainability work in areas related to SDG16 (Peace, Justice and Strong Institutions, 9%) and SDG4 (Quality Education, 7%). The current status of implementation seems to indicate that the remaining goals are less relevant to Internet enterprises, and the content of each SDG is less than 5%.

In addition, it studies the text information related to SDG17 (Partnerships) and analyzes the specific implementation of goal partnerships promoted by Internet enterprises. As shown in the secondary pie chart in Fig. 1, 42% of partnerships are related to SDG 8 (Decent Work and Economic Growth), 27% of the content describes cooperation related to SDG3 (Good Health and Well-Being), while Internet enterprises also actively promote the establishment of industrialization, innovation and infrastructure construction (SDG9, 15%).

3.2 Keyword Recognition of SDGs Based on Association Rules

Use FP_Growth algorithm in association rules, mining rules by constructing FP Tree and recursively mining frequent itemsets. Specifically, the text matrix under each sustainable development goal is transformed, and the keyword are taken as column attributes. If the keyword appears in the row, the corresponding position in the matrix is marked with 1, and if not, it is marked with 0. Through the transformation, the text matrix becomes

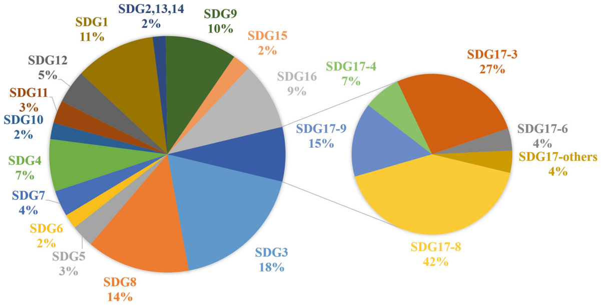

Fig. 1. Proportion of SDGs.

01 matrix, the generated matrix is used to analyze association rules after attribute type conversion and data sparsity. The analysis results are generally given in the form of rules. In order to ensure the reliability of the mined association rules, the threshold of minimum support is set at 0.1 and the threshold of minimum confidence is set at 0.8. Due to the large length of strong association rules, only the association rules in the first three confidence levels are displayed, as shown in Table 2. Each line represents a sustainable development goal. The keyword in a rule is placed in a cell, number in parentheses indicate confidence, and it does not distinguish whether the keyword is in the front or rear of the association rules, that is, the directionality of the keyword is not considered in this study. This paper studies the sustainable development goal text set with the number of texts greater than 5. The number of texts of SDG2 (Zero Hunger) is too small, so the keyword recognition based on association rules is not carried out.

As can be seen from Table 2, "platform" is the key word of SDG5, 11, 12, 15 and 17, which shows that for Internet enterprises, most of the goals are promoted through the construction of Internet platform. "Technology" is the key word of SDG 9, 10, 11, 16 and 17, which shows that the implementation of various businesses and projects of Internet enterprises is inseparable from the support of technology. For SDG 1 (No Poverty), Internet enterprises mainly strive to fight poverty and rural revitalization. Key words such as "poverty alleviation" and "dilemma" appear in SDG 4 (Quality Education) and 5 (Gender Equality), indicating that Internet enterprises also promote the realization of sustainable development goal 1 (No Poverty) when implementing goal 4 and related businesses of goal 5. The launch of environmental protection projects and the support of the foundation are the main ways for Internet enterprises to commit to sustainable development goal 6 (Clean Water and Sanitation). As the basic carrier of Internet enterprises, the data center has achieved the goal of energy conservation and consumption reduction (SDG 7). For sustainable development goal 8 (Decent Work and Economic Growth), in the report, Internet companies focused on describing how to promote the implementation of decent work. For sustainable development goal 10 (Reduced Inequalities), Internet enterprises achieve it through information accessibility action.

3.3 Specific Goal Implementation Characteristics Based on Co-occurrence Clustering

Based on the co-occurrence clustering network, this section analyzes the specific implementation characteristics of SDG1, 3, 8 and 9, which account for a relatively high proportion. The identified keywords are used to construct the co-occurrence matrix and draw the co-occurrence network. The hierarchical clustering method is used to divide the clusters in the network and present the visual co-occurrence clustering results. According to the results of the mined association rules, the keywords appearing in the same association rule are connected in pairs. Taking the keyword as the node and the co-occurrence frequency of each group of keywords as the weight of the edge, a keyword co-occurrence network is constructed. The size of the keyword in the network reflects the number of connections between the keyword and other keywords, that is, the influence of the keyword, the thickness of the edge reflects the link strength between keywords.

Table 2. Partial association rule results.

SDG	Keywords		
1	Country, revitalization (1)	Poverty alleviation, Country, development (1)	Society, poverty alleviation (0.88)
3	Employees, public welfare (0.94)	Initiation, public welfare (0.83)	Organization, public welfare (0.81)
4	Digital, education (1)	Poverty alleviation, education (1)	Education, innovation (0.88)
5	Rural, female (1)	Security, female (1)	Dilemma, women, platform (0.86)
6	Environmental protection, data (1)	Public, environmental protection, project (0.86)	Public welfare, foundation, environmental protection (0.8)
7	Emissions, data center, carbon (1)	Efficiency, data center (0.88)	Energy, green (0.82)
8	Company, management, employees (1)	Culture, employees (0.9)	Employees, system (0.83)
9	Services, Internet, technology (1)	Data, infrastructure, technology (0.89)	Data, artificial intelligence, calculation (0.88)
10	Product, information accessibility, crowd (1)	Information accessibility, technology (0.86)	Technology, crowd, Internet (0.83)
11	City, technology, data (1)	Resources, integration, platform (0.83)	Platform, participation, public welfare (0.83)
12	Consumption, green (1)	Platform, carbon, green (1)	Internet, platform (0.91)
13	Public, environment (1)	Environmental protection, development (1)	Environment, public, Internet (1)
14	Nature, conservation (1)	Guardian, public welfare (1)	Public welfare, activity and protection (1)
15	Ant forest, green, region (1)	Protection, platform, ecology (0.88)	User, online, protection (0.88)
16	Technology, enterprise, service (0.9)	Service, capability, technology (0.9)	Privacy, protection (0.83)
17	Cooperation, platform, China (1)	Digital, economic (0.9)	Provide, technology, technology (0.88)

3.3.1 The Implementation Characteristics of SDG1

Under SDG1 (No Poverty), a total of 17 strong association rules meet the threshold requirement with a minimum confidence of 0.8, including 14 keywords, and the connection between 20 keywords is obtained. The visualization results of SDG1 are shown in Fig. 2. The blue cluster shows that Internet enterprises are committed to continuously promoting the development of poor areas. The green cluster illustrates the direct method

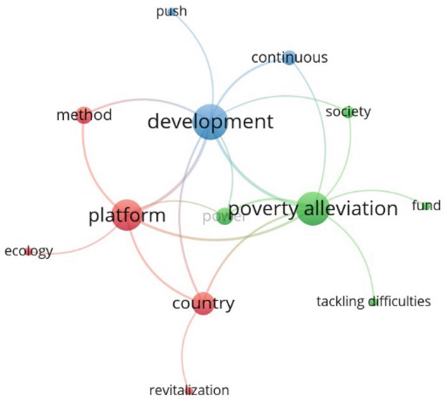

Fig. 2. Keyword co word network of SDG1.

for Internet enterprises to get rid of poverty. The poverty alleviation foundation gathers the Internet enterprises themselves and the forces from all walks of life, which gives economic support to the cause of poverty alleviation. The red cluster reflects the more accurate and transparent poverty alleviation ecosystem of Internet enterprises indirectly through the construction of the platform.

3.3.2 The Implementation Characteristics of SDG3

Through the operation of association rule algorithm, it is found that there are only 3 strong association rules in line with the minimum confidence obtained by SDG3, and there are not many rules with a confidence of more than 0.8, including only 4 keywords. The co-occurrence clustering of the data in the table shows that the results are shown in Fig. 3, "employees", "public welfare", "initiation" and "organization" are all in one category. Although sustainable development goal 3 accounts for a high proportion in the report, its clustering effect is not obvious. The main keyword is "public welfare", which shows that the Internet is involved in the diversification of ways to promote the well-being of people of all ages, such as medical and health, financial well-being, transportation.

3.3.3 The Implementation Characteristics of SDG8

Through the screening of association rules for text information under SDG8 (Decent Work and Economic Growth), a total of 8 strong association rules meet the requirements, including 11 keywords, and a total of 14 connections between keywords are obtained. The visualization results are shown in Fig. 4. When the number of clusters is set to 2, the network achieves the best clustering effect. The visualization results show that the text information in the sustainable development report revolves around two types of subjects: the company and employees. Green cluster focuses on the key word of company, which shows that enterprises are committed to improving their management ability and ensuring the working environment of employees. The red cluster focuses on the keyword of employees. For employees, employees can exercise their ability by

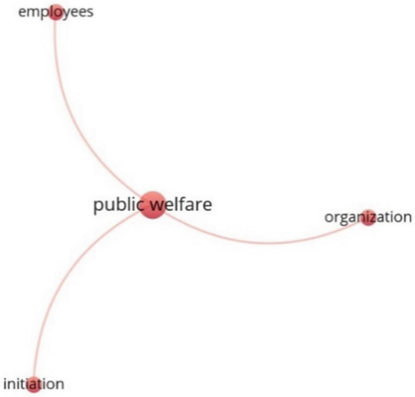

Fig. 3. Keyword co word network of SDG3.

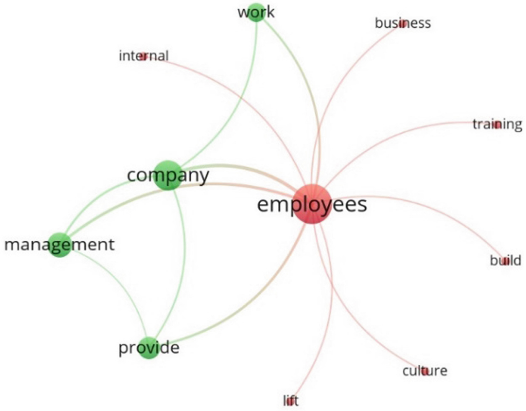

Fig. 4. Keyword co word network of SDG8.

using the resources, support and help provided by the company, and improve the internal vitality of the enterprise.

3.3.4 The Implementation Characteristics of SDG9

Through the mining of association rules for the project text information of SDG9 (Industry, Innovation, and Infrastructure), 22 strong association rules are obtained, including 14 keywords, such as “technology”, “data” and “cloud computing”. The visualization results are shown in Fig. 5. When the number of clusters is set to 4, clustering achieves the best effect. The analysis shows that Internet enterprises promote industrial innovation through the development of technology to provide better data services for enterprises and consumers. Specifically, infrastructure construction needs strong technical support and guarantee (red cluster). Through the application of technical means, enterprises

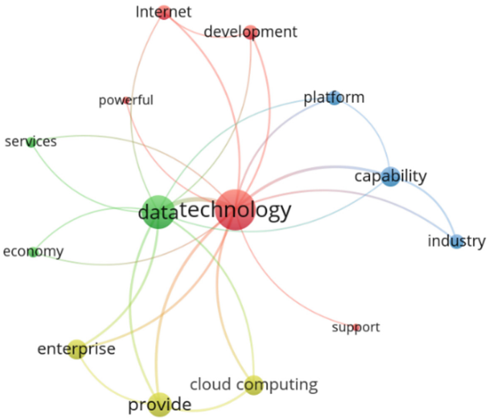

Fig. 5. Keyword co word network of SDG9.

continue to improve Internet infrastructure. At the same time, Internet enterprises attach importance to the value of data (green cluster), transform digital operation ability into economic resources, and improve the convenient service ability for the public. The blue cluster shows that using the influence of the Internet platform, Internet enterprises are promoting the transformation and upgrading of all walks of life. For example, nail helps enterprises build a clear organizational structure, personnel structure and management system. Enterprise Cloud Computing (yellow cluster) is also a recurring theme in the report. Cloud computing is a solution for massive big data computing. It uses a low-cost and commercial model to solve the problem of big data computing.

4 Conclusions

The proposal of sustainable development goals further points out the direction of the implementation of Internet corporate responsibility. The study found that the contribution of Internet enterprises to sustainable development mainly lies in the following goals: Internet enterprises are committed to the sustainable development of economy and the development of goals consistent with their core business, that is, SDG8 (Decent Work and Economic Growth) and SDG9 (Industry, Innovation, and Infrastructure). Furthermore, when implementing businesses related to some other sustainable development goals, it contributes to society and the environment, relies on the power of the Internet to carry out poverty eradication actions (SDG1), and promotes human health and well-being in many ways (SDG3). In this process, Internet enterprises seek common values that can bring economic, environmental and social benefits by cooperating with some international organizations or supply chain members (SDG17).

Using the method of association rules, this paper finds that “technology” and “platform” are the keywords of multiple sustainable development goals. When Internet enterprises implement the related businesses of SDG4 (Quality Education) and SDG5 (Gender Equality), they also promote the realization of SDG1 (No Poverty). Then, the method of keyword co-occurrence clustering is used to carry out visual analysis of SDG1 (No

Poverty), SDG3 (Good Health and Well-Being), SDG8 (Decent Work and Economic Growth) and SDG9 (Industry, Innovation, and Infrastructure). This paper specifically studies the various responsibilities and ways of Internet enterprises to achieve these sustainable development goals in the strategic planning of sustainable development.

The research results of this paper strongly advocate the relevance and importance of sustainable development goals in the sustainable development of Internet enterprises. As the enterprises selected by the research institute are all large-scale Internet enterprises in China, although these enterprises have a very high proportion of revenue, they can not represent all enterprises in the Internet industry. The research will expand the scope of the text for more detailed exploration in the future.

Acknowledgements. This research was financially supported by National Natural Science Foundation of China (Grant No. 71874011). Thanks for the help.

References

Elkington J (1998) Cannibals with forks, vol 8, no 1. Top sustainability books, pp 108–112

Horne J, Recker M, Michelfelder I, Jay J, Kratzer J (2020) Exploring entrepreneurship related to the sustainable development goals - mapping new venture activities with semi-automated content analysis. J Clean Prod 242:118052

Jayaprakash P, Pillai RR (2018) Role of Indian ICT organisations in realising sustainable development goals through corporate social engagement

Jones P, Comfort D, Hillier D (2016) Common ground: the sustainable development goals and the marketing and advertising industry. J Public Affairs

Long TH, Alonso A, Forio M, Vanclooster M, Goethals P (2020) Water research in support of the sustainable development goal 6: a case study in Belgium. J Clean Prod 277:124082

Rosati F, Faria LGD (2019) Addressing the SDGs in sustainability reports: the relationship with institutional factors. J Clean Prod 215:1312–1326

Wang XQ, Yuen KF, Wong YD, Li KX (2020) How can the maritime industry meet sustainable development goals? An analysis of sustainability reports from the social entrepreneurship perspective. Transp Res Part D Transp Environ 78