https://doi.org/10.26599/IJCS.2023.9100006 International Journal of Crowd Science

Product Map Analysis from a Crowd of Small- and Medium-Sized E-Commerce Sites: A Bottom-Up Approach

Xin Li[1], Tongda Zhang[1], Xiao Sun[2,3], and Yongsheng Ma[1] ✉

ABSTRACT

The study of product maps in e-commerce has garnered significant attention from academics and practitioners, as they provide insights into the relationship between products, such as complementarity and competition. However, existing studies have focused on the perspectives of large manufacturers and retailers, using data from these central sources. This paper adopts a bottom-up approach based on crowd intelligence, with small- and medium-sized e-commerce (SME) sites serving as independent data providers. This approach allows for the decentralized processing of data and enables the aggregation of diverse perspectives and insights from a large number of independent sources. A graph term frequency-inverse document frequency method is proposed, which can measure the similarities of products and build a product map. The method was employed to find a hierarchical community structure using data from over 90 000 products from 52 SME sites. The results showed that products within the same site tend to be distributed across the same community. Our findings can assist e-commerce sites in making informed decisions about pricing and product offerings, leading to more diversified production.

KEYWORDS

product map; small- and medium-sized e-commerce sites; bottom-up; crowd science

A product map describes how businesses are differentiated and classified according to the type of products they sell (homogeneous/heterogeneous). Analyzing the product map can reflect the extent to which products are considered substitutes or complements, which is fundamental for marketing and pricing[1], especially in the field of e-commerce. When retailers strategize on certain product assortments, they need to determine certain factors, including (1) the type of merchandise, (2) the depth of merchandise, and (3) the amount of inventory (service level) to allocate to each stock keeping unit in order to maintain the best balance of the above three factors[1]. In addition, they must also consider how to customize coupon activities and which products are suitable for simultaneous promotion, among others[2]. These assortment-related decisions affect consumers' choices about a store, thus affecting the store's sales and profits. In the context of the global COVID-19 pandemic, the e-commerce landscape has undergone obvious changes, online retail sales have increased, and the overall share of total retail sales increased from 16% to 19% in 2020[3]. Furthermore, European e-commerce revenue grew by 10% in 2020[3]. Recent research in the field of international entrepreneurship has shown that the adoption of e-commerce can contribute to the further development of small- and medium-sized enterprises through the opportunity to open up new markets and reach untapped customer base; hence more and more SME sites are emerging[4]. However, analyzing product maps from individual SME sites is an ineffective approach as it provides no insight into the relations among products of SME sites. Instead, the entirety of the SME sites should be assessed from the bottom up.

Many studies have already investigated product map generation to explore the relationship between specific products or categories of products. An early study on product maps was Elrod et al.'s inference from customer responses to competing and complementary products[1]. Elrod et al.[1] argued that product map analyses tend to focus on the substitution of narrowly defined product categories due to the limited nature of available data. However, the definition of product categories is not flexible enough, as it excludes the combination of most products and cannot take into account the relationships among certain categories. In other words, this category-centric strategy is not sensitive enough to the product map.

In the recent decade, the development of machine learning (ML) and the availability of high-quality data[5] have made it possible to analyze product maps in numerous ways. For example, Kim et al.[6] provided a descriptive model of how products are searched online for consumer durable goods manufacturers, which can obtain a detailed product-centric visualization of the competitive structure in their industry. Netzer et al.[7] employed a text-mining approach and demonstrated the feasibility of creating a product map based on user-generated content data. More recently, Gabel et al.[5] discussed the product map based on shopping baskets, which invokes a cheap data source that is available to all retailers through their checkout systems. The above mentioned studies are all conducted from the perspectives of manufacturers and large retailers and based on their central data, such as data from Amazon.com[6] and retailers' checkout systems[5]. Thus, past research objects were modeled as supermarkets. In comparison, our research objects comprise a crowd of independent stores.

Although the volume of a single store is relatively small, the

1 Department of Mechanical and Energy Engineering, Southern University of Science and Technology, Shenzhen 518055, China
2 Department of Automation, Tsinghua University, Beijing 100084, China
3 National Engineering Laboratory for E-commerce Technologies, Tsinghua University, Beijing 100084, China
Address correspondence to Yongsheng Ma, mays@sustech.edu.cn

aggregated volume cannot be ignored. Another area of research that deserves mention is the investigation of user and product portraits[8] by analyzing and integrating various data sources, such as user behavior data, product feature data, etc., to extract user preferences, behavior habits, consumption ability, and so on. However, this approach may involve issues with user privacy and sensitive information. The emphasis lies on the accessibility and public nature data used in our study, which are obtained from the textual product descriptions found on SME sites. Thus, in this study, we aim to construct a product map by analyzing the relationships between products.

Inspired by crowd intelligence, a decentralized approach is adopted in this paper, in which each SME site can be regarded as a tiny and independent data source. This means product information invokes a vast set of data sources that are available to all sites through their web pages. Instead of relying on predefined information of item attributes on large companies' sites, the proposed method derives insights into certain product maps solely from publicly available data on web pages. Thus, it can analyze product maps in situations wherein other data sources might be expensive or difficult to acquire.

With this work, we aimed to (1) develop a graph term frequency-inverse document frequency (TF-IDF) for analyzing product maps that meet the needs of SME e-commerce sites and (2) apply our approach to the collected data from 52 SME sites. The proposed graph TF-IDF is based on the TF-IDF method with natural language processing (NLP) and can build a visual network through a graph structure to reflect a product map[9,10]. This visualized network is implemented through Gephi, an open-source network analysis and visualization software based on the NetBeans platform and Java. The innovative points of this paper are as follows. First, we extend the TF-IDF model to allow for modeling product information from SME sites. Then, we use this to build a graph network structure in which we directly quantify the competitive relationship between products. Second, we indirectly quantify the competitive relationship between SME sites. Third, after feeding this network structure to Gephi, we then establish a visualized product map.

The remainder of this paper is organized as follows. Section 1 formally introduces the method and dataset, while Section 2 discusses the procedure and results. Finally, the conclusion is presented in Section 3.

1 Method and Dataset

1.1 TF-IDF

The TF-IDF is a commonly used technique in NLP and information retrieval. It is used to measure the importance of a word in a document relative to a corpus of documents and often employed as a weighting factor in information retrieval and text mining.

TF-IDF is based on the idea that words frequently occurring in a document are likely to be important and that the importance of a word is also related to its rarity across the entire corpus of documents. For example, in this system, a word that frequently occurs in a document but rarely in the overall corpus is given a higher weight than a word that frequently occurs in both the document and the corpus.

TF is used to measure how many times a term appears in a document; obviously, however, documents can be long or short, and the term is more likely to appear in a long document. Hence, to rectify this problem, the "term frequency" must be standardized. As shown in Eq. (1), the TF of term t in document d is determined by the number of occurrences of the term and the number of terms in the document.

$$\mathrm{tf}_{(t,d)} = \frac{f_{(t,d)}}{\sum_{t' \in d} f_{t',d}} \tag{1}$$

For example, document "D_1" has a total of 100 terms, and the term "jeans" appears 5 times; thus, we can obtain the TF value as follows:

$$\mathrm{tf}_{(\text{"jeans"}, D_1)} = \frac{5}{100} = 0.05.$$

Meanwhile, IDF is used to filter some common terms. When calculating the TF of a document, the algorithm treats all terms equally, even though stop words like "a" and "the" are not counted. However, some common terms, such as "and", which cannot be distinguished in documents, obtain large term frequencies, even though they are not important terms. To rectify this problem, IDF assigns lower weights to these common terms and larger weights to rare terms.

$$\mathrm{idf}_{(t,D)} = \ln \frac{N}{|d \in D : t \in d|} \tag{2}$$

where N is the total number of documents and D represents a specific document.

For example, we have 1000 documents in the corpus, and "jeans" appears in 101 of them. Thus, according to Eq. (2), we can get the IDF value as follows:

$$\mathrm{idf}_{(\text{"jeans"}, D)} = \ln \frac{1000}{101} = 2.293.$$

In practice, the TF-IDF weight of a word in a document is calculated as the product of its TF and IDF, as shown in Eq. (3). The TF is simply the number of times the word occurs in the document, while the IDF is calculated as the logarithm of the total number of documents divided by the number of documents containing the word.

$$\mathrm{tfidf}(t,d,D) = \mathrm{tf}(t,d) \cdot \mathrm{idf}(t,D) \tag{3}$$

In this way, the TF-IDF of keywords in each document can be obtained, and then these TF-IDF values can be used to convert the document into a word frequency vector. Next, the cosine similarity of the vector is calculated using Eq. (4) to determine the document similarity[11].

$$\mathrm{similarity}(d_i, d_j) = \frac{\sum_{i=1}^{n} (D_i \times D_j)}{\sqrt{\sum_{i=1}^{n} (D_i)^2} \times \sqrt{\sum_{i=1}^{n} (D_j)^2}} \tag{4}$$

In the context of this paper, we consider each product to be a document and the words in its title and description to be terms in that document. Then, TF-IDF can be used to measure the importance of a word in a product's title and description relative to the entire corpus of products. For example, if a certain word occurs more frequently in the title and description information of a product but rarely in that of other products, then this word will be assigned a higher weight based on the TF-IDF algorithm. In this way, the similarity measure can help us identify the unique characteristics of a product and understand its relationship to other products in the corpus.

Notably, TF-IDF has two limitations. The first is that the

extraction of keywords is heavily dependent on the quality of the corpus used for training. One way to resolve this issue is to use a larger and more diverse corpus to train the keyword extraction model. Doing so can help ensure that the model learns to identify relevant and meaningful keywords that are representative of the target domain. Additionally, using domain-specific dictionaries and ontologies can facilitate keyword selection and improve the training corpus quality. The second limitation relates to how the degree of differentiation of text can be affected by the location of the feature words. One approach is to use N-grams[12], which is a combination of n consecutive words. By using N-grams, the context of the feature words can be considered, thereby improving the accuracy of text discrimination. For example, instead of considering only single words, we can consider two- or three-word combinations that capture more of the context of the text.

1.2 Graph TF-IDF design

In this paper, we propose that the similarities of textual descriptions of products, as determined by Eq. (4), indicate the relevance between the products. To provide a comprehensive view of the product landscape across independent SME sites, we introduce the graph TF-IDF network, which depicts the relationship between products and the SME sites themselves.

We define the product network as a weighted undirected graph $G = \langle V, E \rangle$, where V and E represent the node set of products and the set of edges, respectively. An edge between two nodes (i.e., two independent products) exists if the similarity score (defined in Eq. (4)) is positive. The edge weight between nodes i and j is represented by A_{ij}, which can be calculated as follows, using the similarity (n_i, n_j) obtained from Eq. (4):

$$A_{ij} = \begin{cases} \text{similarity}(n_i, n_j), & i \neq j; \\ 1, & i = j \end{cases} \tag{5}$$

There are two levels of network structures in our approach. First, we can construct a graph network on the product layer based on the aforementioned method. Second, we can treat each site as a large collection of documents and build a site graph network that utilizes the same approach, given that the products belong to their respective SME sites. This paper focuses on analyzing the product map on the product layer.

1.3 Dataset

The dataset was collected from 52 different SME sites that sell clothing. These sites feature information on more than 90 000 products, including product titles, body_html (description information), production types, tags, and other information.

After reviewing the dataset, some statistical results show that there are 83 698 non-repeating tags, indicating that almost every product has a unique tag. Furthermore, most products have more than one tag, suggesting that the tag has a high degree of granularity. Although there are 116 non-repeating production types, their texts are short, and the distribution is extremely unbalanced. The statistics of each product type are shown in Table S1, which is in the Electronic Supplementary Material (ESM) of the online version of this article, with three significant figures reserved and in descending order. On one hand, the product type "Dresses" has more than twice the number of "Knit tops", but the number of many product types is so small that the proportion is close to 0. On the other hand, the title and body_html of product data contain information that can better characterize product attributes. Based on the above three points, we mainly use the title and body_html text information of the product in this paper to represent its attributes.

The body_html contains the description information of the product, which is useful for characterizing data features. First, the text information in body_html must be extracted as body_text. Second, the data also contained many stop words that are neither meaningful nor useful in this context, as previously explained in Section 1. Therefore, the data must be filtered by removing stop words. In addition, we also performed operations, such as punctuation regularization and lowercase conversation. After these preprocessing processes, the word length distribution of the product title and body_text is counted, as shown in Table 1. Note that the minimum length of body_text is 0; however, this does not mean that its body_html was originally empty. In addition, we concatenated the title and body_text of the product as the text feature so that the text will not be empty.

2 Result and Discussion

Two components of the implementation process are presented in this section. The first is to establish a quantitative measure of similarity between products using the TF-IDF algorithm. Second, a product network visualization is constructed to provide a graphical representation of the relationships among the products.

2.1 Product relationship quantification

As discussed in Section 1, the TF-IDF algorithm is used to analyze each product as a distinct document. The TF-IDF values of the terms within each document can be calculated and subsequently utilized to transform the document into a word frequency vector. Through this process, the cosine similarity metric can be used to determine the similarity of documents. Given that our methodology is corpus-based, it maintains applicability in scenarios wherein new documents are introduced.

This paper gives an example of the process of quantifying the correlation between products. First, we randomly select a product as the test data, titled "classic high waist skinny jeans dark denim". The other content of the product is to explain some attributes of the jeans, including how many pockets it has, the proportion of textile materials, etc. Then, we calculate the n data with the largest similarity score to the test sample.

Table 2 shows the eight samples that are the most similar to "classic high waist skinny jeans dark denim", along with the similarity scores and the site indexes where these samples come from. Due to the length of the text, only the title information is shown here. As can be seen, on one hand, these products all express the same style of jeans—"classic high waist skinny jeans"—with color being the only main difference. On the other hand, these eight samples all come from sites 1 and 52, which indicates that these two sites have a strong competitive relationship. Such a relationship is determined according to the similarity scores between the products among them.

2.2 Product map revealed by communities

In Section 1.2, we presented a strategy to construct the product network mainly based on the similarity scores between products. To conduct a more in-depth analysis of the product map in the entire product pool, we continue to concentrate on identifying

Table 1 Word length statistics of body_text and title_text.

Text	Minimum length	Maximum length	Average length
Body_text	0	571	24.33
Title_text	2	14	5.98

Table 2 Eight samples that are the most similar to the test data.

Series No.	Shop No.	Title	Similarity score
1	1	Classic high waist skinny jeans light blue wash	0.989 217
2	1	Classic high waist skinny jeans white	0.966 868
3	52	Classic high waist skinny jeans light blue	0.942 952
4	52	Classic high waist skinny jeans wine	0.924 600
5	52	Classic high waist skinny jeans khaki	0.923 623
6	52	Classic high waist skinny jeans rust	0.921 935
7	52	Classic high waist skinny jeans light chocolate	0.920 836
8	52	Classic high waist skinny jeans olive	0.915 092

and characterizing communities and the product map in the network.

The original graph network without communalization is shown in Fig. 1, in which we randomly selected 18 sites from all the SME sites, each sampling 10 samples for a total of 180 nodes (the graph network of more nodes or sites has similar structures). Each node represents a product, and different colors are used to distinguish the SME sites from one another. We denote the node size by the weighted degree, which is the sum of the weights of the edges incident with the node. Here, the bigger the node is, the more closely connected it is with other nodes and the more important it is. Moreover, the thickness level of the edge is proportional to its weight. From Fig. 1, it is difficult to find the correlation between nodes for two main reasons. First, the original graph has a "fully connected" shape, which makes the number of edges so large that the structure of the graph is not clear enough. Second, the original graph only reflects the connection relationships among nodes; it cannot reflect the spatial distribution of nodes.

To obtain an in-depth graph network, we hide edges with comparatively low weights (edges with weights below 0.14 are filtered). Here, the smaller the weights, the weaker the relationship between nodes. Additionally, to achieve better community detection in our product network, we used the Louvain method to extract communities from large networks. This method, created

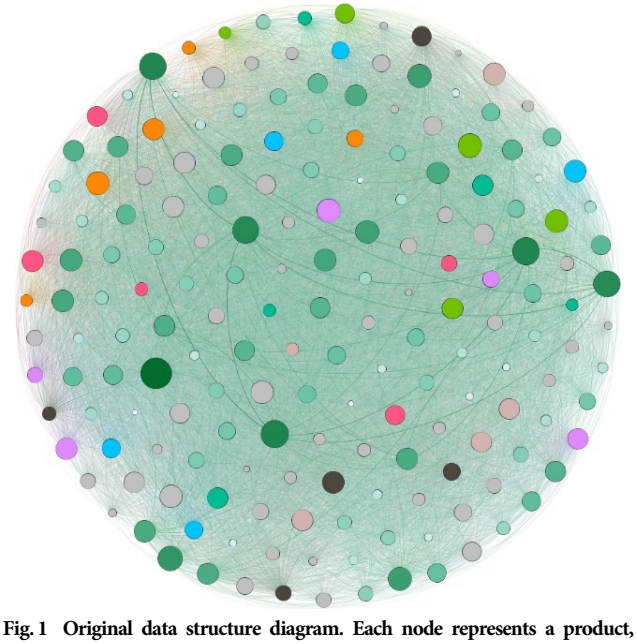


Fig. 1 Original data structure diagram. Each node represents a product, and the color represents the SME site.

by Blondel et al., is based on the assumption that the within-community connections are supposed to be denser than the between-community connections[13,14]. The basic idea is that nodes in the network try to traverse the community labels of all neighbors, after which they choose the community label that maximizes the modularity increment. After maximizing the modularity, each community is regarded as a new node, and the process is repeated until the modularity no longer increases. For our product map network, the modularity is defined in Eq. (6):

$$M = \frac{1}{\sum_{ij} A_{ij}} \sum_{i,j} \left[A_{ij} - \frac{\sum_{i} A_{ij} \sum_{j} A_{ij}}{\sum_{ij} A_{ij}} \right] \delta(c_i, c_j) \tag{6}$$

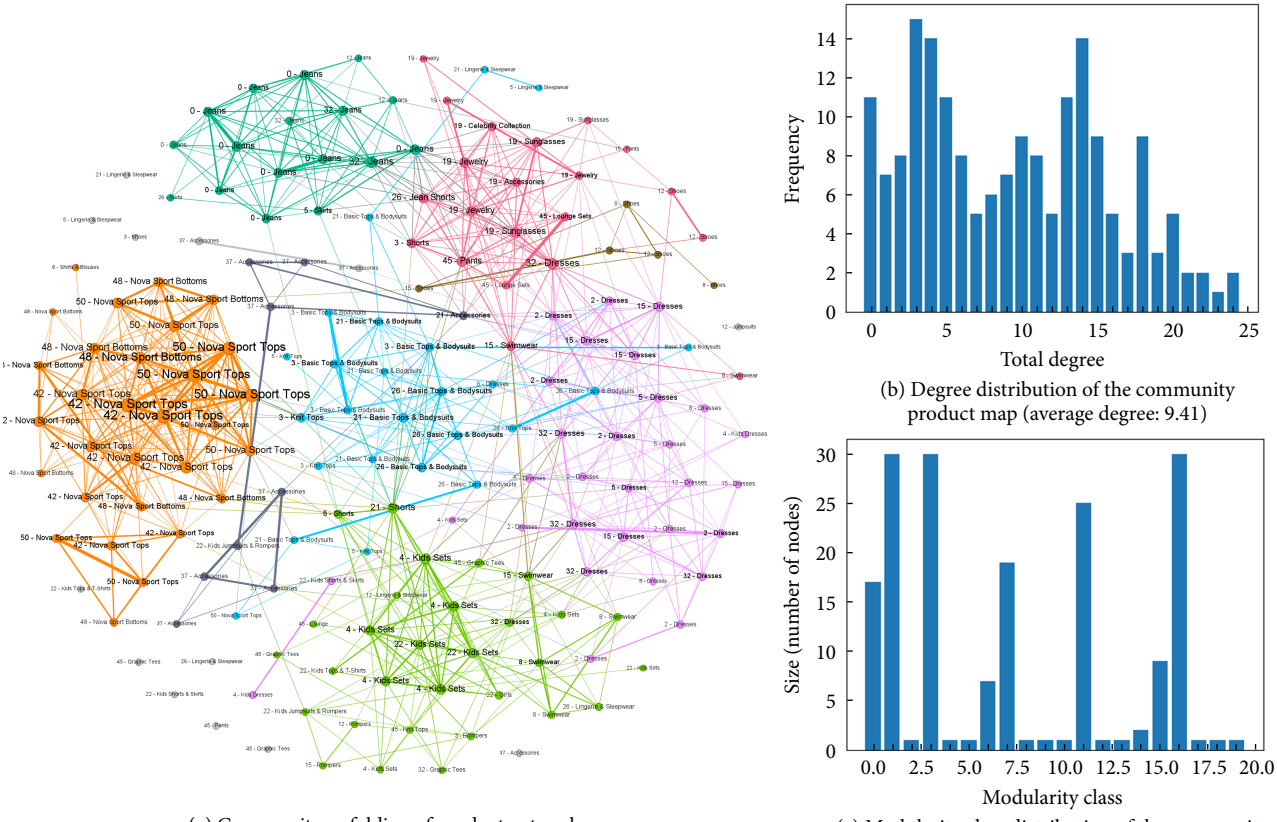
where c_i is the community to which product i is assigned. The δ function is used to determine whether nodes i and j are the same community. If yes, then the value is 1; otherwise, it is 0.

The representation of SME sites' product map is shown in Fig. 2. In particular, the spatial characteristics of the product network and the relations across different products can be seen in Fig. 2a. Each product represents a node, and the node label consists of two parts: the SME sites index and the product type. The color of the node and edge is the same, which is used to determine the community to which the node belongs. Node degree represents the number of edges associated with the node, modularity class represents the community to which that node is assigned, and the weighted edges capture the relationship between products. The more closely connected part can be considered as a community with relatively close connections between nodes but sparse connections between communities. Within this community, the similarity of products is positively correlated with the weight of edges. The degree and modularity class distribution, shown in Figs. 2b and 2c, have an average degree equal to 9.41 relative products and a maximum of 24, respectively. The modularity score, calculated using Eq. (6), is 0.682. Within the community network, there are 6 larger hidden neighborhoods (having more than 15 nodes) depicted using the following colors: decorations shown in red, knits and bodysuits in blue, jeans in dark green, children's clothing in green, sportswear in orange, and dresses in purple. Products within the same SME sites tend to be distributed in the same community in accordance with the densities of node labels and edges. Moreover, the similarity of products within a community is higher than that between communities, which is consistent with real-world observations.

3 Conclusion

Overall, the proposed decentralized approach for studying product maps in e-commerce can provide valuable insights and assistance to SME e-commerce sites. This method, which uses crowd intelligence and a bottom-up approach, has the potential to enhance our understanding of competitive relationships within the market. It can also help provide valuable guidance to e-commerce sites in making informed decisions about pricing, product offerings, and marketing strategies.

Further research in this area has the potential to yield significant benefits for both academics and practitioners in the field. For example, future works on this topic may include expanding the study to incorporate additional e-commerce sites to further refine the product map analysis and enhance the accuracy of the results. Additionally, further research may be conducted to explore the benefits and challenges of implementing the proposed decentralized approach on a larger scale.



(a) Community unfolding of product network (N = 180 products and E = 847 relationships)

(b) Degree distribution of the community product map (average degree: 9.41)

(c) Modularity class distribution of the community product network (modularity: 0.682)

Fig. 2 Representation of SME sites' product map.

Additionally, future works could focus on incorporating additional data sources, such as customer reviews and ratings, to gain a more comprehensive understanding of product relationships within a market. This could provide valuable insights into consumer preferences and behaviors, which could be used by e-commerce sites to improve their product offerings and marketing strategies.

Another possible avenue for future research involves exploring the potential applications of the proposed bottom-up method for studying product maps in other industries. This approach could be applied to a wide range of industries, including finance, healthcare, and transportation, to help gain a more detailed understanding of the complex relationships between different products and services within these distinct markets.

Aside from these potential research directions, future studies could also focus on developing more advanced algorithms and techniques for analyzing and visualizing product map data. This could include utilizing ML and other advanced data analysis techniques to identify patterns and trends within the data, as well as to develop predictive models for forecasting future market conditions.

Acknowledgment

The authors wish to thank all the respected professors who helped during the experiment, development, and writing phase of this paper.

Electronic Supplementary Material

Supplementary materials including

- Table S1: Proportion of production type.

All the supplementary materials are available are available in the online version of this article at https://doi.org/10.26599/IJCS.2023.9100006.

Dates

Received: 16 February 2023; Revised: 28 March 2023; Accepted: 3 April 2023

References

[1] T. Elrod, G. J. Russell, A. D. Shocker, R. L. Andrews, L. Bacon, B. L. Bayus, J. D. Carroll, R. M. Johnson, W. A. Kamakura, P. Lenk, et al., Inferring market structure from customer response to competing and complementary products, *Mark. Lett.*, vol. 13, no. 3, pp. 221–232, 2002.

[2] R. Venkatesan and P. W. Farris, Measuring and managing returns from retailer-customized coupon campaigns, *J. Mark.*, vol. 76, no. 1, pp. 76–94, 2012.

[3] I. K. Mensah and D. S. Mwakapesa, Cross-border e-commerce diffusion and usage during the period of the COVID-19 pandemic: A literature review, in *Proc. 3rd Africa-Asia Dialogue Network (AADN) Int. Conf. Advances in Business Management and Electronic Commerce Research*, Ganzhou, China, 2021, pp. 59–65.

[4] D. Tolstoy, E. R. Nordman, S. M. Hånell, and N. Özbek, The development of international e-commerce in retail SMEs: An effectuation perspective, *J. World Bus.*, vol. 56, no. 3, p. 101165, 2021.

[5] S. Gabel, D. Guhl, and D. Klapper, P2V-MAP: Mapping market structures for large retail assortments, *Journal of Marketing Research*, vol. 56, no. 4, pp. 557–580, 2019.

[6] J. B. Kim, P. Albuquerque, and B. J. Bronnenberg, Mapping online consumer search, *J. Mark. Res.*, vol. 48, no. 1, pp. 13–27, 2011.

[7] O. Netzer, R. Feldman, J. Goldenberg, and M. Fresko, Mine your

own business: Market-structure surveillance through text mining, *Mark. Sci.*, vol. 31, no. 3, pp. 521–543, 2012.

[8] C. I. Eke, A. A. Norman, L. Shuib, and H. F. Nweke, A survey of user profiling: State-of-the-art, challenges, and solutions, *IEEE Access*, vol. 7, pp. 144907–144924, 2019.

[9] S. Qaiser and R. Ali, Text mining: Use of TF-IDF to examine the relevance of words to documents, *Int. J. Comput. Appl.*, vol. 181, no. 1, pp. 25–29, 2018.

[10] L. Yao, C. Mao, and Y. Luo, Graph convolutional networks for text classification, *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 1, pp. 7370–7377, 2019.

[11] C. H. Huang, J. Yin, and F. Hou, A text similarity measurement combining word semantic information with TF-IDF method, *Chin. J. Comput.*, vol. 34, no. 5, pp. 856–864, 2011.

[12] S. S. M. M. Rahman, K. B. M. B. Biplob, M. H. Rahman, K. Sarker, and T. Islam, An investigation and evaluation of N-Gram, TF-IDF and ensemble methods in sentiment classification, in *Proc. Second EAI International Conference on Cyber Security and Computer Science* (*ICONCS 2020*), Dhaka, Bangladesh, 2020, pp. 391–402.

[13] V. D. Blondel, J. L. Guillaume, R. Lambiotte, and E. Lefebvre, Fast unfolding of communities in large networks, *J. Stat. Mech.*, vol. 2008, no. 10, p. P10008, 2008.

[14] T. Zhang, J. Qian, X. Sun, Y. Yuan, and M. Chen, A glance at people's engagement behavior with fully pay transparency: From the perspective of a crowd collaboration system, in *Proc. 5th Int. Conf. Crowd Science and Engineering*, Jinan, China, 2021, pp. 1–7.