How to promote value co-creation through online communities: An exploration of lead user identification from the users participation perspective

DONG Zheng-nan, DAI Wei-wen

(School of Management & Qian Wei College, Shanghai University, Shanghai 200072, China)

Abstract: The emergence of web 3.0 and social platforms provide a convenient way for interactions between users and providers. Focusing on the roles of users in the value co-creation, users' participation becomes an important and prolific source of innovation and a main source of ideas for enterprises. However, not all users are equally likely to innovate but innovative activity occurs especially among lead users. Unlike with the product features of industrial manufacturing, how to identify the lead users from online communities for the industry of mass consumer products is little studied. This paper explores a new method to identify the lead users from the online communities based on index evaluation and biclustering analysis. Through data collected with 11250 community users of 54 sections from MIUI platform, an empirical study of a fast-growing Chinese handset-maker Xiaomi Tech, illustrates the proposed methods. Analysis and calculation of the practical example show that lead users including the core lead user groups and the potential lead user groups can be effectively obtained by these two kinds of methods' similar results matching. And realizing the full participation of lead users with the goal of value co-creation in the end.

Key words: Value co-creation; lead user identification; online community; index evaluation; biclustering; text mining

通信作者：董政男，女，上海大学钱伟长学院，数学与应用数学专业. E-mail：DZN12345@shu.edu.cn

1. Introduction

Information and communication technologies revolutionize the way business has traditionally been done, and create more opportunities for high level of consumer involvement and interaction in the product development Testa et al. (2020). Users are not only playing a role as consumers but also creators Schweisfurth, (2017); Stockstrom et al. (2016). With the transformation from product-centered logic to customer-centric logic, value co-creation, as a cutting- edge commercial idea and a scientific way of innovation, has attracted great attentions from both enterprises and academic researchers Chatterji and Fabrizio (2015); Alves et al. (2016); Fuller et al. (2009); Aghdam et al. (2020).

The concept of value co-creation focuses on the customers' participation, meaning that it not only be limit to the sales satge (value in-exchange) or actual usage (value-in-use) Gronroos (2011). The shared interests can be a driven factor. So, in actual operation, what stage of design for customers to participate in and what requirements of customers should be emphasized on are two issues, which must be taken into account by enterprises. In the long term, to create a win-win situation is an eternal pursuit for both customers and enterprises. For the enterprises, achieving the shift from "back-end treatment" to "early prevention" in manufacturing is critical for both promoting product cost controlling and improving enterprise competitiveness. As the theory goes, the development cost will increase exponentially with the in-depth development process. For customers, customer satisfaction is a key factor for success. On the basis of the classical Kano model, the performance requirements and

excitement requirements, exists more positive impacts on customer satisfaction with the product, especially for the latter Kano et al. (20). Therefore, given all these, the earlier the enterprises capture the beneficial requirements and get the users involved, the better for mutually-beneficial and win-win cooperation. Research on innovation and marketing also shows that this may best be achieved by establishing processes to listen effectively and efficiently to the "voice of the customer" Tontini (2007), since this goal can not be met through the efforts made solely by enterprises on quality improvement.

Fortunately, multimedia-rich interaction opportunities offered by the Internet have made it possible for customers to give feedback to business owners in the shortest possible time on one hand, and for business owners to listen to their customers' interests and opinions Hamidi et al. (2019). Relying on the agglomeration effect, online communities break through the limitations of time and space to provide a data mining platform for enterprises to obtain the customer requirements. Obviously, online communities have become an important source for knowledge and new ideas Martinez (2017); Marchi et al. (2011). Nonetheless, in front of numerous, complicated and diverse customer requests, it becomes impossible for business owners to realize total involvement since it is a costly and time-wasting activity. Hence, it is quite vital to know how to identify the effective demands in the development of new products. With strong evidence, the importance of users' collaboration was verified in this area. In 1986, the lead user, as a new concept, was introduced and refined by Professor Hippel Von Hippel (1977, 1986), which has brought real benefits to cost-efficient and in- depth exchange during a collaborative innovation process. In the course of enterprise's practice, the identification of the lead user becomes the primary issue. According to the two basic characteristics of the lead user Urban and von Hippel (1988), the lead user evaluation index system was gradually raised and continuously improved in the industry manufacture. However, comparing to the industrial manufacturing industry with significant business characteristics, it is not applicable for the mass consumer products industry as it has the most frequent contacts with customers. If the exiting unified standard of the lead user characteristics or the traditional investigation methods like questionnaire and interviews is used, it is unavoidably to cause an incomplete and inefficient identification result.

Based on the current study achievements of the lead user theory, therefore, taking the mass consumer products industry as an example, this paper proposes a data-driven approach for lead user identification, which combines the text mining, bi-clustering algorithm and other data mining methods. Moreover, a real case of one domestic smartphone firm's online community is given to validate the effectiveness of the method, in which the comment contents and be- havioral data obtained from the platform are analyzed and the operational process of the method in the real application is demonstrated in detail. Through theoretical analysis and case study, this paper further proposes a co-creation model of lead user participation on the basis of product life cycle and intends to provide certain references for manufacturing enterprises about the customer participation in new product development and marketing. With a view to achieve the ultimate goal of value co-creation.

The paper proceeds as follows. Following the introduction, in Section 2 a literature review on value co-creation, user innovation and attempt to draw forth the lead user theory and extend the exiting methods within online communi- ties. Research methodology about the index evaluation and biclustering methods applied in this paper are presented in Section 3. Subsequently, Section 4 describes more details about the empirical study, including the case presentation, data collection and

processing, and the lead user identification analysis. The implications of the findings are discussed from two aspects based on a novel model in Section 5, and finally, conclusions are given in Section 6.

2. Literature Review

Since lead user theory, focusing on customer engagement in product development and innovation, gives a signifi- cant impetus to value co-creation, in this section, we present a more detailed analysis of the related work that consists of value co-creation and user innovation, the lead user theory and identification, as well as the studying method biclus- tering mostly concerned in the whole paper including the arithmetic investigations and application fields.

2.1. Value co-creation and user innovation

The proliferation of ICTs has changed the way how online community interacts and somehow facilitated co-creation activities in recent years Fu et al. (2017). As a new concept in the business management literature, the co-creation of value first appeared in 2004 by Prahalad and Ramaswamy Prahalad (2004). Following this debut in the literature, large numbers of scholars have studied it from different perspectives Alves, Fernandes and Raposo (2016). Value co-creation is now becoming a dominant trend in service science, particularly in the information management, marketing and service domains Galvagno and Dalli (2014); Xie et al. (2016); Osei-Frimpong et al. (2018). According to Vargo and Lush, the dominant marketing logic is transforming from a Good Dominant (G-D) logic to a Service-Dominant (S-D) logic, and the role of consumers is now extending beyond being passive product and service recipients from being "choosers" to becoming active "makers" Vargo and Lusch (2004); Prahalad and Ramaswamy (2004); Janamian et al. (2016). This framework leads to the assumption that the value creation process is transformed from enterprise and product-centered to individual and experience-centered Xie, Xiao and Hu (2016); Wu and Liu (2018). Value co-creation theory has been the focus of research and discussion among scholars since 2000, with two representative viewpoints, namely, co-creation theory based on "S-D logic" and co-creation theory based on consumer experience. Relying on previous literatures research outcomes on the conception and connotation of value co-creation, some scholars summarized their core ideas in 6 aspects and pointed that despite the differences in research perspective and connotation, both view-points aim to create product and service value through effective interaction between consumers and enterprises Xie et al. (2008); Greer et al. (2016); Hu et al. (20).

As for the output from the interactions of all actors, it is not only an opportunity to co-create value Gronroos and Voima (20), but also a facilitator to the innovation process Souto (20). In this matter, research about co-creation and innovation has been gradually in the spotlight. For instance, Hsieh and Hsieh Hsieh and Hsieh (20) investigated how customer co-creation affects the performance of service innovation through the operant resources. Based on grounded theory, Fu, Wang and Zhao Fu, Wang and Zhao (2017) explored the micro-mechanisms of how three types of platform service innovation-product innovation, process innovation and business model innovation-affect value co-creation activities and the network effect longitudinally. In the context of sharing economy, Casais, Fernandes and Sarmento Casais et al. (20) discussed tourism innovation developed by hosts of sharing accommodation, based on the outcomes of guests' value co-creation. Co-creation offers firms and their network of actors significant opportunities for innovation, as each actor offers access to new resources through a process of resource integration Frow et al. (20). Among numerous researches,

customer engagement has been given a prominent role. Particularly, it has been perceived as a strategic factor to enhance business performance directly Neff (20) or indirectly through supporting product innovation Hoyer et al. (20). Additionally, virtual communities show a high potential as drivers of value co- creation and co-innovation Romero and Molina (20). Given the importance of user participation, recently, scholars are trying to study the motivations of user participation in online communities et al. (20).

2.2. Lead user theory and identification

Ever since it was agreed that users can be a main part of participatory product development, the literature on the role of users during innovation has grown tremendously, which also be known as customer-driven innovation or user-centered innovationBogers et al. (2010). Against the background of a breakthrough innovation era, Eric von Hippel first introduced the concept of "lead user" Von Hippel (1986). Based on its definition and insights generated by other scholars, two characteristics of lead users are constructed: being ahead of an important future market trend, and expecting high benefit from their innovative solutions (e.g., Franke et al. (2006); Luthje (2004); Morrison et al. (20); Schreier and Prugl (2008)). Since its leading edge for product markets and robust profits for both individuals and enterprises Pongtanalert and Ogawa (2015), the lead user theory has attracted a great deal of attention from practitioners and scholars alike and specifically been focused on two aspects, lead user identification and lead user participation in the innovation.

Prior research on lead user identification, mainly focused on the identification methods and identification dimensions. Identification of lead users is the biggest issue facing business. According to the features of the manufacture users, Urban Urban and von Hippel (1988) and Hippel Von Hippel et al. (2009) put forward screening method and pyramiding through the questionnaire, telephone interview and other traditional surveys successively. The latter one is a qualitative and non-standardized approach, which is much more efficient than the screening method. But it is only confined to that case where product users are strongly linked. With the growth of the Internet, online communities gradually become an effective platform for companies to search users. Under such circumstances, Tietz et al. (2006) proposed a new identification method, namely signaling, realizing the combination of two kinds of channels of online and offline. So fast-evolving social media and social networking offer more options and more ways for identifying the lead users. As Piller and Walcher (2006) stated that enterprises do not have to carry on "face to face exchanges" when the extraction of person information can be achieved through the interactive network platform. And in this context, two methods known as "netnography" and "crowdsourcing" have become the widely used identifying methods presented by Belz and Baumbach (2010) and Brem and Bilgram (2015). However, these studies mainly contribute to the theoretical illustration about the method's implementation process. Consider this, Pajo et al. (2015) built a conceptual framework for lead user identification from a technical perspective but without any concrete data mining method. He He and Chen (2009) and Qiu and Lv (2013) started by defining user characteristics for establishing evaluation index of users' leading edge, and then identified lead users by ranking the online users through questionnaire and interview. Further research in this field, Zhao and Sun (2014) and Li (2014) replaced the traditional questionnaire survey method by incorporating the classification based data mining methods. Due to the incomplete understanding of the user characteristics, the inaccurate definition of the user characteristics may appear, which in turn affects the innovation performance.

Therefore, how to accurately identify the lead users by using network data remains a question worth further studying.

Researches on customer involvement in successful innovations demonstrate the importance of lead user characteristics, since they are the key points to distinguishing lead users among ordinary users Lilie et al. (20). As noted previously, the earlier studies about lead user characteristics are summarized to two points. With its practical applications, Luthje (2004) put forward that enough technical expertise and extensive knowledge of product are two more clear traits in measuring a user's leading advantages. Indeed, the valuable information and ideas may also can be proposed by those people who do not possess the above characteristics. The researches of Nambisan and Baron (2007); Nambisan and Nambisan (2008) also show that users play different roles in different stages of product development and users who need to be involved have different characteristics. Magnusson (2009) further found that a unifying and strict feature would restrict users' ability to innovate to a certain degree. Thus, Schuurman et al. (2011) argued that guiding users to participate in different product development stages based on their different abilities is the most effective way to realize maximum benefit of user innovation.

With regard to lead-user innovations in practical application, the four-step approach as the initial lead-user research method was provided by von Hippel Von Hippel (2005), which involves confirmation of lead user characteristics, foundation of lead-user project group, creativity product development and testing. Since this method limits to the offline activities, Luthje and Herstatt (2004) proposed a four-phase approach, which first introduced web platform into the lead user identification process, breaking the barrier between the online and offline as the above method did not. With the development of computer technology, Ernst et al. (2014) presented a concept realizing the combination of lead user innovation process with social media, and extending the application domain from the industrial products to consumer goods. During the implementation of the theory, what concerns the enterprises most is the exact effect and benefit they could get from the participation of lead users in the product innovation. Accordingly, many scholars studied the impacts of lead users' participation to corporate performance from an empirical perspective. Their research results show that lead users, as an external source of innovation, have a significant positive impact on enterprises' innovation capability. Similarly, Nishikawa et al. (20) also found that the user-generated products have achieved a far higher market penetration than the internal product designers'. While the facts users with ability differences has been proposed and many published articles on lead users' participation for innovation in theoretical description, how to distinguish the users based on their leading characteristic differences and then how to arrange them to the right part of the product development still has no better solutions.

2.3. Related research on biclustering

Given the complex data types collected from the online community, such as data sparsity, high dimensions, and heterogeneity, an appropriate method is vital for conducting large-scale data analysis. Biclustering, as a powerful data mining technique, allowing clustering of rows and columns simultaneously, is a great fit for such data. Since 2000 when it was first applied to gene expression data by Cheng and Church Cheng and Church (2000), much larger scale of biclustering algorithms have been developed to enhance the ability to make sense out of large data sets aroused in the technology-driven society. For instance, Bergmann et al. Bergrmann et al. (20) presented an efficient

iterative signature algorithm for the analysis of large-scale gene expression data. Based on the CCA introduced by Cheng and Church, an improved probabilistic algorithm (FLOC) was proposed by Yang et al. Yang et al. (20), which can discover a set of k possibly overlapping biclusters simultaneously. Tanay et al. Tanay et al. (20) adopted synthetically graph and statistical theory to put forward the SAMBA (Statistical-Algorithmic Method for Bicluster Analysis) biclustering algorithm. Prelic et al. Prelic et al. (2006) provided a systematic comparison and evaluation of prominent biclustering methods that includes a simple binary reference model and further put forward a fast divide-and-conquer algorithm (Bimax). From a more theoretical viewpoint, Busygin et al. Busygin et al. (2008) emphasized the mathematical models and concepts in biclustering techniques. Face numerous algorithms, a contrast analysis and test between different algorithms are necessary when proving the validity of the proposed algorithm in practical application. It also has become a significant problem on the research of biclustering. Hence, based on former research, Madeira and Oliveira Madeira and Oliveira (2004) analyzed the existing approaches to biclustering, and classified them in the light of the differences in the type of biclusters, the patterns of biclusters, the approaches for evaluating the solution, and the target applications, which can offer the interested researcher systemic theoretic direction. Additionally, Oghabian et al. Oghabian et al. (2014) evaluated 13 biclustering and 2 clustering (k-means and hierarchical) methods by using several approaches to compare their performance and finally concluded that so long as the samples are well defined and annotated, the contamination of the samples is limited, biclustering methods can effectively identify subsets of genes and samples.

As far as we know, application of biclustering has been heavily used to biological data analysis with gene expression data Xie et al. (2019). In fact, since biclustering can analyze data sets as long as in the form of a real-valued matrix, the recent applications also have been successfully used in other areas. For example, in the E-commerce area, recommendation systems and target marketing are important applications. Many authors applied biclustering to collaborative filtering using data where the rows represented customers and the columns movies (e.g. Yang et al. (2002, 2003)). Considering the data characteristic of high dimensionality in tourism segmentation, Dolnicar et al. Dolnicar et al. (2012) introduced a new clustering algorithm for tourism market segmentation analysis through the modification of the traditional Bimax algorithm. In every market segment, customer properties were also discussed by making use of demographic variables and further the corresponding customer characteristics were obtained. Similarly, Lin et al. Lin et al. (2019) proposed a novel parallel biclustering approach to identify and segment highly profitable telecom customers with superior clustering results. In addition, biclustering is also used for knowledge mining. Swathi (2010) used hybrid Hierarchical k-Means algorithm for the analysis of gene expression data and biclustering and clustering algorithms were utilized at the same time. Appropriate knowledge is mined from the clusters by embedding a BLAST similarity search program into the clustering and biclustering process. In the end relatively high quality clusters were obtained through the validation of FOM methodology. Besides, more exotic applications of biclustering involved the analysis of data matrices with electoral data Hartigan (2002) and foreign exchange data Lazzeroni and Owen (2002). As numerous researchers state that various biclustering applications have been developed, aiming to assist researchers to effectively derive domain knowledge and novel insights from their big data.

3. Methodology

The overall framework of lead user identification process through online communities can be concluded into three parts: user data acquisition and processing, lead user idenfication method, and feature analysis of lead user identifica- tion results. Hence, in this section, the relevant methods studied in this paper are presented by following this framework in two phases, as shown in Fig. 1.

3.1. Phase A: Lead user ranking

In the part of lead user identification method, phase A (i.e. the lead user ranking section) involves three methods, namely the entropy method, the grey correlation analysis and TOPSIS. After the data collection and processing in part 1, the entropy-weighing method is adopted to determine the weight of every index. As an objective method of determining the weights, entropy was first proposed by a German physicist Clausius and developed by Shannon in 1948 Shannon and Weaver (1948). As an uncertainty measure of information volume in a system, Shannon entropy plays an important role in information theory. It indicates that the information volume of each piece of information is directly connected to its uncertainty degree. Assume that there are m objects, the performance of each object under different variables is denoted by X_{ij} ($i = 1, 2, \ldots, m$; $j = 1, 2, \ldots, n$), and $X = (x_{ij})m \times n$ represents its data matrix, then the definition of information entropy for the jth variable can be expressed as:

$$E_j = -k \sum_{i=1}^{m} P_{ij} \ln p_{ij}, \tag{1}$$

where,

$$P_{ij} = \frac{x_{ij}}{\sum_{i=1}^{m} x_{ij}}, k = \frac{1}{\ln m} \tag{2}$$

With much attention paid on, entropy method has been widespread concerned and become one of the focuses in multi- attribute decision making, showing obvious advantages in reducing the subjectivity and improving the reliability. In respect to above superiority, this method is used to weight determining of the evaluation indicators in this part. Then based on the lead user behavior data, two index evaluation methods are designed for getting the alternative user sets by the combination of these two different results of user sorts. Due to the shortage of prior knowledge and huge, complex online data, this solution could be well suited for the evaluation problem on the lead user degree against online communities.

In regard to grey relational analysis, it is an important branch in grey system theory introduced by Professor Deng Deng (1982). The basic idea is to judge whether there is consistent variation between the two factors during the system development. Based on this idea, Pu and Liu Pu and Liu (2014) expanded its application range to the index evaluation to test the goodness among overall indicator system or between each index. The rationale is that by determining the optimum value of different indicators to establish the reference sequence, and then make a judgement of the relevancy

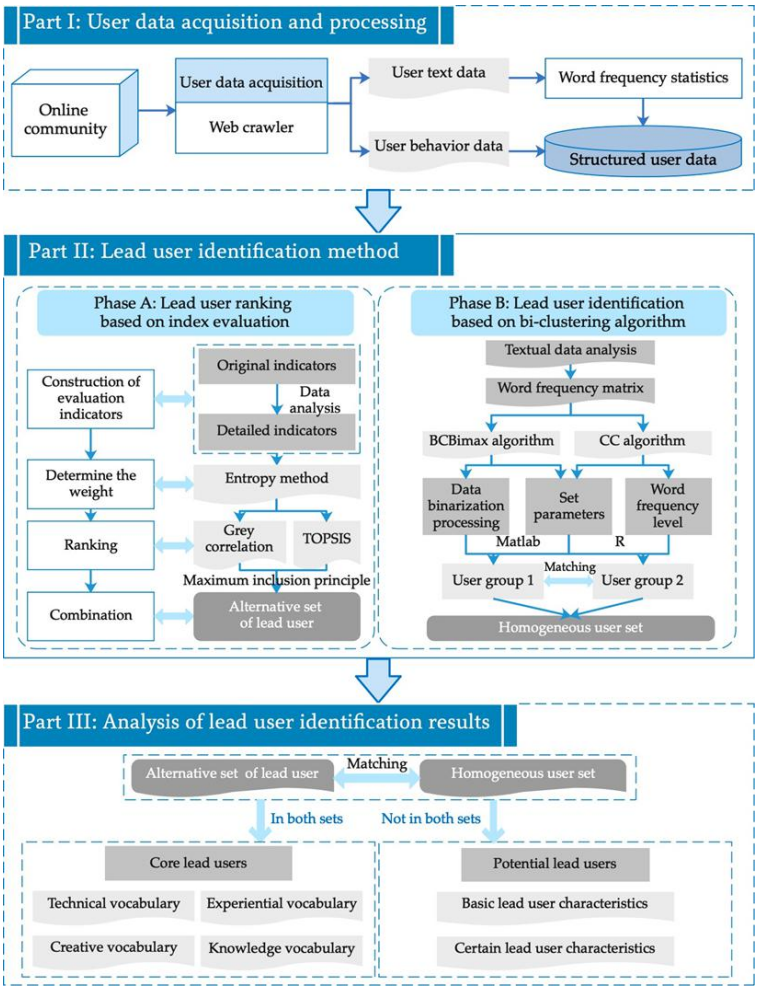

Figure 1: The framework of the identification method

degree between the evaluated object and the optimal sequence. Generally, the procedures of calculation may include six steps: construction of evaluation index matrix, determination of reference sequence, standardization of data processing, calculation of grey relational coefficient, and ranking of evaluation object. TOPSIS (Technique for Order Preference by Similarity to an Ideal Solution), as an objective decision analysis method, was originally put forward by Yoon in 1987 Yoon (1987). The basis for this lies in providing a cardinal ranking of alternatives by calculating the proximity of evaluation object to the optimal solution. So it is also called the method of distance of good-gad solutions. Remarkably, to apply this technique, attribute values must be numeric, monotonically increasing or decreasing, and have commensurable units. Generally speaking, the stepwise procedure of index evaluation for implementing TOPSIS can be indicated by seven steps, namely constructing evaluation index matrix, dealing with the indexes in the same trends, non-dimension of indicators, building the weighted normalized matrix, determining the positive and negative ideal solutions, calculating the separation measures, and computing the relative closeness coefficient. Relying on these two index evaluation methods, the alternative user sets can be finally obtained.

3.2. Phase B: Lead user identification

In phase B, based on high dimensional sparse data features of online users, two bi-clustering algorithms (i.e. BCBimax and CC algorithms) are both applied to extract the characteristics of lead users and to reach a homogenous user set. Thus, the problem of the intrinsic characteristics of the massive user base from online communities can be well achieved. To better understand the application principle of this process, emphatically give the introduction of the above algorithmic principles in the following.

3.2.1. BCBimax algorithm

Bimax is an algorithm proposed by Prelic et al. Prelic, Bleuler, Zimmermann, Wil, Buhlmann and Gruissem (2006). With the earliest application in genetic engineering, this algorithm is used to identify local patterns in gene expression data. The algorithm assumes two possible expression levels per gene: no change and change with respect to a control experiment. If it is the first case, a cell e_{ij} is 1 whenever gene i responds in the condition j and otherwise it is 0. Accordingly, a set of m microarray experiments for n genes can be represented by a binary matrix E_{ij} . Through Bimax algorithm, a submatrix of E for which all elements equal 1, is finally found. And such kind of submatrix represents a bicluster. Due to the high similarity of genetic data in a bicluster, this algorithm dose not be applicable for all scenarios especially for high dimension data modeling. To avoid overlapping submatrices, Dolnicar et al. Dolnicar, Kaiser, Lazarevski and Leisch (2012) modified this algorithm to BCBimax algorithm in the research of tourist market segmentation. Fortunately, the improved algorithm prohibits overlapping of cluster membership. More specifically, the idea behind the BCBimax algorithm is that partitioning the binary data into three submatrices, one of which contains only 0s and therefore can be discarded. The algorithm is then recursively applied to the remaining two submatrices; the recursion ends if the current matrix represents a bicluster, that is, contains only 1s. In order to avoid overlaps, the next bicluster is found starting the basic algorithm on data excluding the rows of the already found bicluster. Finally, the process of iterations stops when no new bicluster is found or a pre-determined maximum cluster number is reached. As shown below, Fig. 2 illustrates the calculation process of the algorithm:

Step 1. Divide the columns: Fig. 2(a) illustrates the initial data matrix. In Fig. 2(b), the set of columns is divided in to two subsets CU and CV , here by taking the first row as a template. CU are then columns where the rows are 1s,CV the others;

Step 2. Divide the rows: RU are those rows that contain only 0s in column set CV , RV are those rows that contains only 0s in columns set CU , the remaining rows are called RW , as Fig. 2(c) shows;

Step 3. Delete the submatrix: find three submatrices U $[RU + RW\ CU]$, V $[RW + RV\ ALL]$ and W $[RU\ CV]$ in the matrix shown in Fig. 2(c), and delete the submatrix W ;

Step 4. Recursive computation: repeat steps 1 to 3 on submatrix U and V until all elements in the submatrix are 1s, then the resulting submatrices are biclusters.

Step 5. Delete rows: store the matrix obtained by step 4, and then delete the rows in this bicluster from the initial matrix to obtain the new matrix.

Step 6. Iterative computation: start over from the new matrix after the deletion of rows, and repeat steps 1 to 5 until no new bicluster is found.

3.2.2. CC algorithm

CC algorithm was first proposed by Cheng and Church Cheng and Church (2000) in 2000 for the application in gene expression data. Based on a greedy iterative search method, this algorithm consists in building a bicluster adding or removing rows or columns iteratively, thus, improving its quality. In order to assess the quality of a bicluster, a quality measure called Mean Squared Residue (MSR) is used to evaluate how adequate each expression value is with regard to the rest of values of the bicluster. It has been considered as the benchmark measure in biclustering literature. Specifically, let A be a matrix, X denoting the set of the rows and Y the set of columns. Then each aij of matrix A represents the element of the ith row and jth column. Let $I \subset X$ and $J \subset Y$ be subsets of rows and columns. Thus

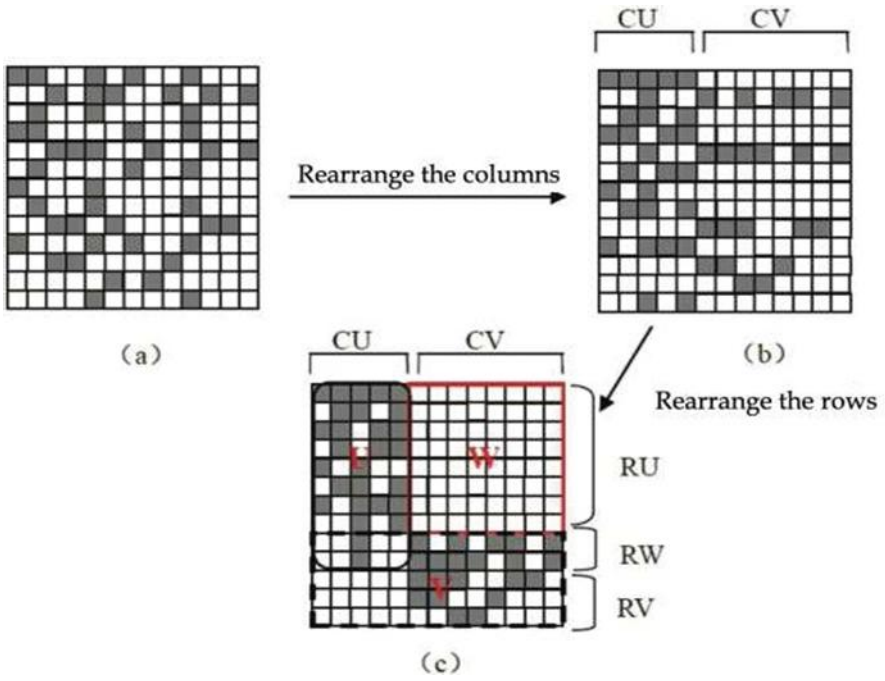

Figure 2: Illustration of the BCBimax algorithm

the pair (I, J) specifies a submatrix Aij with the following definition of MSR:

$$H(I,J)=\frac{1}{|I||J|}\sum_{i\in I,j\in J}(a_{ij}-a_{iJ}-a_{Ij}+a_{IJ})^2=\frac{1}{|I||J|}\sum_{i\in I,j\in J}RS_{ij}^2 \tag{3}$$

$$a_{iJ}=\frac{1}{|J|}\sum_{j\in J}a_{ij},a_{iJ}=\frac{1}{|I|}\sum_{I\in I}a_{ij}, \tag{4}$$

$$a_{IJ}=\frac{1}{|I|J|}\sum_{i\in I,j\in J}a_{ij}=\frac{1}{|I|}\sum_{i\in I}a_{iJ}=\frac{1}{|J|}\sum_{j\in J}a_{iJ} \tag{5}$$

are the row and column means and the mean in the submatrix (I,J). RS_{ij}^2 is the residue. A scheme of Cheng and Church (CC) algorithm is shown in Fig. 3. The algorithm takes as input the expression matrix A and the threshold imposed on MSR. δ is used to reject non δ-biclusters. A list L of δ -biclusters is returned as output. After preprocessing the missing values of the input data matrix by

replacing them with random numbers, the bicluster discovering process is repeated as many times as biclusters are desired. In each iteration, the bicluster B is initialized to the whole matrix. Next, three different phases for multiple node deletion, single node deletion and node addition are applied. These phases iteratively perform the removal and addition of rows and columns, ensuring that the result is a δ-bicluster. Finally, a substitution phase replaces the elements of the input matrix that are contained in the recently found bicluster with random values. This substitution is applied in order to prevent overlapping among biclusters, since it is very unlikely that elements covered by existing biclusters would contribute to any future bicluster.

4. Empirical study

In this section, an empirical study is conducted to evaluate the validity and accuracy of the proposed lead user identification framework relying on text mining. Since empirical analysis with actual data support is an extremely effective verification way, this research applies the following criteria for the selected enterprise: (a) being suitable for the application of customer participatory innovation model, (b) having a communication platform between enterprises

Input: Expression matrix A; Thresholds.

Output: List of biclusters AIJ.

1. Preprocess the missing values of expression matrix A.
2. List $A_{IJ} = \emptyset$.
3. Bicluster B.
4. Repeat n times.
5. $B = A_{IJ}$.
6. B_{δ} = multiple node deletion phase(B, δ).
7. B_{δ} = multiple node deletion phase(B, δ).
8. B_{δ} = addition phase(B_{δ}).
9. $L = L \oplus B''_{\delta}$.
10. Substitution phase （B''_{δ}, A）.
11. End repeat.
12. Return A_{IJ}.

Figure 3: The scheme of Cheng and Church's algorithm

and customers, and (c) being convenient for data collection. On the basis of the above considerations, in the consumer products industry, the mobile phone forum of the Xiaomi virtual community is selected to carry out the process of lead user identification.

4.1. Enterprise investigation

4.1.1. Xiaomi Profile

Xiaomi Tech, founded in 2010, whose main products are smartphones and peripheral products, has been ranked he most innovative company in China Bischoff, (2014) and grown to be a global brand Cho et al. (20). With a eputation for innovation, Xiaomi relies exclusively on online forums and social media to build a strong brand and has a following of tens of millions of loyal fans. In only three years, Xiaomi has already acquired 10 billion of market values, becoming the fifth largest e-commerce company in China. Committed to the pledge that let everyone in the world can enjoy the good life brought by technology, Xiaomi became the youngest company on Fortune China 500 list in 2019 (nine years after its iteration), and successfully ranked in the top three Internet services list in 2020.

Benefitting from the full participation of the public, Xiaomi has maintained a surprising growth rate in the world, which can be seen that the innovation pattern of Xiaomi has been successfully implemented with the deeply partic- ipation of users, a particular group of Mi fans. At present, Xiaomi Tech's three core products are mobile phones, MIUI and MiTalk. For each core product, Xiaomi has established its own network community to provide a platform for communication between Mi fans as well as users and enterprises to increase the company's products and services. MIUI, the first product of Xiaomi, is a third-party mobile phone operating system depending on deep optimization, customization, and progress of Android. Accordingly, the selection of MIUI community as the lead user identification platform can be a right choice for those reasons, which can be summarized to three aspects: (a) mobile phone system's improvement is most likely to be participated by users than other hardware devices; (b) MIUI community, as a core competitiveness product, is worth studying; and (c) the existing user participation mode of MIUI lay the groundwork for the data collection.

4.1.2. MIUI Community

The MIUI system is a deeply customized and optimized Android system. As a third-party mobile phone OS de- eloped by mobile phone enthusiasts, this system has been translated into 23 languages and highly sought by 500000 Mi ans around the world only after going through 50 weeks of development. The MIUI community (https://www.xiaomi.cn), built specifically for the development of MIUI system, is a strong information platform for Internet users. With more than 47 million registered users, assessed on October 18th, 2020, MIUI community has been a critical virtual platform in carrying out seamless open innovation activities.

User engagement MIUI community, as one of the best innovative virtual communities in China, dedicates to in- tegrating user's opinions and suggestions into mobile phone functions. Users can not only put forward problems encountered in the use of various functions of mobile phones, as well as corresponding suggestions for improvement, but also directly propose their own entirely new solutions in this platform, which becomes an effective way for cus- tomers' participation in product improvement and new product development. Additionally, users in the Xiaomi forum can also reap a corresponding return on their words and deeds. These returns are mainly reflected in two aspects: (a) each idea of posts can get timely responses by the MIUI review team, and further to be identified whether it meets customer needs or is additional development needed. So that the user's opinion can be effectively reflected in the prod- uct improvements; (b) registered users can obtain the corresponding points or experience value by posting, discussing and so on. Through the integral accumulation level, users are divided into diverse groups. The higher level, the more permissions the user group occupies, and the higher participation in the new product's development. The more detailed information about the grouping of users can refer to the MIUI forums.

Module setting MIUI community has formed a relatively perfect system, with a total volume of 10.35 million post, aiming at promoting the interact with consumers and products. Up to now, there are 9 discussion forums, such as Model zone, Website zone, and so on. Among them, MIUI zone, Resource Sharing, and Orange Friday are the three main sources for data collection in this paper. MIUI zone, as the core technology zone, Mi fans are able to interact with professional developers by proposing bugs and ideas for new functions, while developers answer and recommend improvements. Therefore, the relevant data including user content data and behavioral data are one of the important sources for the identification of lead users. In the zone of Resource Sharing, users who own the

product knowledge and technology are usually pleased to make creations with the existing products by themselves, which reflects the lead user characteristics to some extent. Orange Friday means updating system weekly, and integrating development edition on Friday. The core of this pattern is the interactions between MIUI team and users online. Besides, the users involving the improvement suggestion are published periodically. Accordingly, this portion of the data can provide evidence for the validation of lead user identification results.

Types of user data Depending on the above investigation, the range of data acquisition can be determined. In order to clarify the user data structure, the study further investigates the data types of MIUI community users. There are 4 main types of data generated during the process of communications among the users and enterprisers through online communities. First, it is the personal data that comes from the user's filling and systematic allocation at the time of registering, including ID, nickname, gender, user group, year of birth and place of birth. Second, it is the behavior data generated by users when they are active in the communications with other users, including registration time, access time, online time and space visits, number of topics, number of replies, etc. Third, it is the review data that comes from the user's posted content in the community, including text data, var topic, counting data, etc. Fourth, it is the reward data according to the performance of users in the community calculated by the system, such as points, experience, prestige and contribution, etc. Table 1 summarizes the user data types recorded by the MIUI community.

4.2. Data collection

After the determination of user data types, the next step of data collection can be carried out based upon the requirements of the lead user identification. This work is mainly consisted by two parts: data acquisition and data processing. The course of collecting data are described in detail below.

4.2.1. Data acquisition

As mentioned above, the online communities of MIUI system have been well studied and the survey results show that the current number of online users in MIUI community is about 6790, 000, widely distributing in 73 sections. Considering its huge volumes and extensive distribution, the study wipes out some sections with lower correlation, and finally retains 54 more relevant sections. The specifical information about the sections are displayed in Table 2. Then, the user data is collected consisting of the above mentioned 4 types of user data (e.g. personal data, behavior data, review data and reward data). By using the Python programming language, user-related data throughout all years is grabbed from the corresponding sections on September 28th. Moreover, given the users with lower activity and their

Table 1
Types of user data in MIUI community

Personal Data (6 dimensions)	Behavior Data (12 dimensions)		Review Data (6 dimensions)	Score (6dimensions)
ID	Number of visit	Number of sharing	Post subject	Sores
Nickname	Number of friends		Post content	Experience
Gender	Number of records	Online time	Posting Section	Prestige

User group	Number of albums	Registration date	Number of views	Contribution
Date of birth	Number of replies	Last active time	Number of	Millet
		Last visit time	comments	

Table 2
User data acquisition sections in MIUI communities

Model zone	Xiaomi 1/1S	Xiaomi 2A	Xiaomi 2/2S
	Xiaomi 3	Xiaomi Note	Xiaomi 4
	Xiaomi 4C	Xiaomi 4S	Xiaomi 5
	Xiaomi Max	Xiaomi 5S	Xiaomi 5SPlus
	Xiaomi Max2	Xiaomi Mix	Xiaomi 5C
	Xiaomi Note2	Xiaomi 5X	Xiaomi 6
	Xiaomi Note3	Xiaomi MIX2	Hongmi 1
	Hongmi 1S	Hongmi 2	Hongmi 2A
	Hongmi 3	Hongmi 3S/3X	Hongmi 4
	Hongmi 4A	Hongmi 4X	Hongmi Note
	Hongmi Note2	Hongmi Note3	Hongmi Note4
	Hongmi Note4x	Hongmi Note5a	Hongmi Pro
	Google	Samsung	Huawei
	Other		
MIUI section	Product launch	Buglist	Feature suggestions
	Story of development	Developer communication	
Resource sharing	Theme	Software	Game
	Wallpaper	Typeface	Localization resources

post count dose not exceed 100, they also need to be further removed because they are almost unlikely to be the lead users. Ultimately, the paper collects a total of 11250 users' information.

4.2.2. Data processing

Among the extracted data from the MIUI community, three types of data are all structured data (e.g. personal data, behavior data and reward data), which can be computed directly. Whereas, the review data an kind of unstructured data displays in text form need further processing through "structuralization".

Keyword extraction In previous studies, comparative mature operating procedure for text structured handling has been developed. Preprocessing is usually done with several steps: denoising, word segmentation, part-of-speech tag- ging and others, as illustrated in Fig. 4. This paper applies NLPIR Chinese word segmentation system developed by Chinese Academy of Sciences to achieve the handling of word segment. After Chinese word segmentation, deleting stop word is conducted by referring to the Stop Words Extended Edition produced by the Information Retrieval Cen- ter of Harbin Institute of Technology and combing the characteristics of MIUI community user comment sentences.

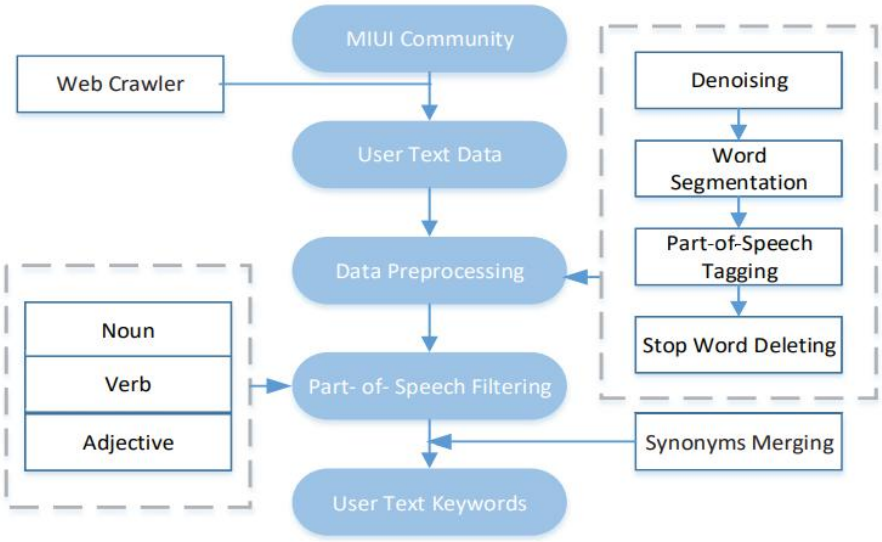

Figure 4: The processing flow of user text data.

Table 3
Frequency matrix of user text keywords

ID	Noun	Verb	Adjective
	Noun.1 Noun.2 ...	Verb.1 Verb.2 ...	Adjective.1 Adjective.2 ...
ID.1 ID.2 ... ID.12250	Noun frequency	Verb frequency	Adjective frequency

Moreover, except for the deletion of conjunction, connective, empty word and so on, some special characters which are unable to discern also need to be removed, and then three kinds of word class, namely nouns, adjectives and verbs, have been obtained. After further observations, some synonyms and near-synonyms are merged to avoid the data duplication, poor performance and other issues. Finally, the ultimate user keyword list, comprising more than 9000 words, are gained.
Word frequency statistics Note that the text data needs to be transformed to the structured data that can be dealt with mathematically. After the word segmentation, the table of keywords with each users' posted comments has been got. Considering the principal component keywords mentioned frequently in sentences of the post text on the forum, these words can represent the user's main focus, behavior and emotion in the online community to some extent. Therefore, this paper adopts the method of word frequency statistics to transform text data into structured data. Through frequency statistics, the frequency list of keywords in each user's posted content distributed among 51 sections is acquired. Subsequently, by merging and sorting the keywords table, the union of all users' keywords is finally obtained, which can been received as the column vector of the user word frequency data matrix. Accordingly, the different user's ID can be treated as the row vectors of the matrix. Thus, the

user word frequency data matrix, a 11250 × 6231 dimensional matrix, is eventually formed and set out in Table 3.

Table 4
Screening criteria on MIUI community users

Registration date (year)	Online time (hour)
2010	<365
2011	<312
2012	<260
2013	<208
2014	<156
2015	<104
2016	<52

Table 5
Screening criteria on user keywords

Word frequency	Number of users	Explanation	Instance
>30	<10	No group characteristics on such keywords.	Member, regret, bill
	>2500	No distinctions on such keywords.	Download, save, no

Data dimensionality reduction In order to improve the efficiency of data analysis, on the basis of the data obtained above, further data dimension reduction on user number and text data is carried out. (1) Given the existence of a large quantity of data from users online, the less activity users among the acquired 11250 users need to be further eliminated. There are two main reasons: effective information-lacking problem and short online time. Therefore, this paper further sets criteria for userdel: users' online time is at least one hour a week. By following this principle (see Table 4), 5,586 users' data are retained ultimately for subsequent analysis. (2) Filtering text keywords is used to reduce the data dimension, since high dimensional data will greatly decrease in effectiveness of data analysis. This screening work is mainly aimed at two situations: first, the keywords with too sparse data, which means it occurs rarely and sporadically and no group characteristics; second, the keywords with too dense data, which demonstrates there is no significant difference among diverse users. Keywords with the above two features are extracted. According to the deleting criterion (see Table 5), the process of dimension reduction is carried out.

After filtering, 286 keywords have been retained in the end (see Appendix), forming a 5586× 286 dimensional matrix. Together with personal data, behavioral data and system scoring data, the fi nal data is aggregated to a 5586×310 dimensional matrix, involving 6 personal data items, 12 behavioral data items, 6 system scoring data items, and 286 text data items. This data matrix lays the foundation for the subsequent data analysis.

4.3. Ranking of user leading degree based on index evaluation

Lead user theory since being proposed has been widely studied and applied in the field of industrial manufacturing. terms of the lead user characteristics, the related researches mostly focus on the defi

nition of general concepts. How- er, in actual application, this need to be carried out specific analysis in accordance with product features, audience haracteristics and identification technology. Hence, based on the previous researches on the character of lead users, his subsection firstly establishes an evaluation index by combing the data structure of lead users in online communi- ties, and then uses gray correlation analysis and TOPSIS respectively to rank the user's leading degree to get the lead user's candidate set.

4.3.1. Construction of lead user evaluation index

The earlier studies about the lead user identification is generally relying on the definition proposed by Profes- Hippel, who summarized the two lead user characteristics as mentioned before. In 2004, Morrison Morrison, oberts and Midgley (20) put forward leading edge status as a basis measurement criteria for lead user identification. Meanwhile, he added two new variables into the previous two basic characteristics, namely leading edge status and innovative application. Interestingly, Luthje Luthje and Herstatt (2004) and Hienerth Hienerth et al. (20) also found that the innovative users did possess some features different with the ordinary users. From this, synthesizing the ex- isting results, Luthje further brought forward six indicators in light of lead user identification. They are referring to being ahead of marketplace trend, having abundant product knowledge, using experiences, deep involvement, good leadership and having innovative production techniques. Later studies are mostly based on these user characteristics for a further evolution. Since the lead user characteristics and searching methods are all influenced by the types of social media, however, this issue is not considered by the work of Luthje. Therefore, this paper combines the concrete forms of social media (e.g. blogs, forums, virtual worlds) with the exiting judgment criteria to construct the evaluation index of lead users in online communities. To facilitate the understanding, detailed explanations of the evaluation index and its correspongding user data are provided below.

(1) Being ahead of marketplace trend. According to the definition of von Hippel, one significant characteristic of lead users is that they have a strong enough need beyond what is currently available in the general market. And this often manifests their dissatisfaction with existing products or services and further promotes their valuable suggestions to product innovation. Through research, the sections of functional suggestion and Bugist are often used by users to feed back their discontent in the duration of product use and offer their suggestions for improvement. Since the Bugist section usually deals with the problem on product defects, it greatly reduces the level of users' comfort Lee (2014). Thus the discussions in Bugist are most unlikely belong to the advanced requirements. While in the functional suggestion section, users often give some useful, astonishing and available suggestions based on their firsthand experience. Based on this, the functional suggestion section is selected and the post count is used to determine whether the users have the leading requirements. At the same time, the requirements proposed by lead users will be accepted by the mass market in few months or years depending on the study of Hippel. Considering this, the collect quantity of posts is measured for the quality of suggestions. Therefore, this paper adopts the average collect quantity of each post to evaluate the level of users' leading requirements.

(2) Having abundant product knowledge. Instead of just putting forward their demand, lead users can directly take

part in the process of product innovation. Due to the expressions of demand are ambiguous, users who present their requirements are not necessary to master the knowledge of products and what they need

is only to express their feelings depending upon experience. Whereas, the lead users as participators in product innovation, they usually have a clear cognition on the improved methods to meet their requirements. On the one hand, these cognitions can guide them to express their demand more accurately, on the other hand, can help product engineer to avoid manufacturing deviations without customers needs conversion. So, these kind of users must have a deeply understanding of product knowledge. In online communities, the users who often post on tutorials, usually are armed with rich product knowledge. Hence, the post count of this type of post is used to determine whether the user has product knowledge and the average collect quantity of each post to evaluate the users' level of product knowledge.

(3) Having innovative production techniques. On account of the existing products which can not satisfy their usage requirements, lead users might give suggestions on product improvements. But due to the restrictions on the process of product innovation, these requirements take much time to realize the transformation from their presentation to practical application. In the research of Schreier Schreier et al. (20), they found that lead users usually have a high internal locus of control. It means that lead users have strong self-beliefs and they deem they can control developments within their ability and eff orts. They are most likely to make an improvement with their own hands and usually share their creativity and design on the web. Under such situation, lead users not only need the relevant knowledge but also technology for product improvements. There are a large portion of posts in connection with product improvements shared by users in resource sharing section of MIUI communities. Based on it, selecting the post count of this kind of sharing posts to recognize if the user possesses product technology, and meanwhile adopting the average collect quantity of each post to evaluate the level of users' innovative production techniques.

(4) Extensive usage experience. Dissatisfactions and improvements showed by lead users on the product clearly state that they have fully experienced yet. Therefore, usage experience can be one of the essential conditions. The survey of MIUI communities found that there is one type of users being good at replying to others' appeals for the product usage. Usually, these appeals involve various kinds of problems about the usage of product. In turn, the respondents are also highly likely encounter such problems and solve them by themselves. So such users who always offer responses to others' help usually have rich usage experience. At some point, the number of responses can be used to measure the level of richness in product usage experience.

(5) High participation rate. The ultimate goal of the lead user identification is to let them participate in the process of product innovation. Except for the above features of lead users, they also need have to be actively join in the discussions in the online communities. Generally speaking, the users with high participation rate are more probably to take part in the innovative activities. As for those people who possess the leading characteristics but are unwilling to join in share activities and discuss with others, it must be hard to persuade them to participate in the long term.

Table 6
Specification of lead user indicators

Original indicator	Specification	The detailed indicators
Ahead of trend dissatisfaction	Show dissatisfaction with existing products	Average quantity of post collected in the section of function recommends
Product-related knowledge	Tutorial posts	Average quantity of tutorial post collected
Use experience	Respond to a post of request for help	The number of reply for help post
Involvement	The degree of participation	Time of user online and number of replied post
Opinion leadership	The opinion leader in the inline community	Visits, number of subject posts collected, prestige
Product-related Technology	Product technology, such as publishing their own improved products in the community	Average quantity of post collected in the section of work share

cooperation with enterprises. From the survey, it can draw that some indicators, such as online time, response number and so on, can be adopted to judge the level of users' participation.

(6) Possessing character of opinion leader. After all, an innovation if it can not be widely accepted by the public and not available in actual application, is pointless. Lead users often be seen as market leaders, and their innovative ideas always can be widely recognised and accepted by the public after in a time. They have a leading position in the area in which they are good at. And that is closely related to their advanced consciousness and product improved capability. In an online community, users with high acceptance usually have the following traits: first, users space with a high volume of visits; second, users' posts have been largely collected; third, a large number of top quality posts with great prestige. According to these features, the number of visits, the total number of collection and the value of prestige in communities are seen as the reference standards for the determination of an opinion leader.

Based on the above analysis, next correspond the user data in online communities to each evaluation indicator and number the index. Thus, the instruction of indicators is summarized in Table 6. The first column shows the six types of lead user characteristics proposed by Luthje, the second column are the relevant explanations, the third column are the corresponding online users' data and the last column are the descriptions of user data sources and computing approach.

4.3.2. Determination of indexes weight based on entropy weight method

Considering the advantages of entropy weight method like stated before, in this section, this method is adopted to objectively assign weights to each evaluation index. The calculation process is introduced below:

Step 1. Depending on the data matrix and calculation methods as shown before, the user indicator data matrix is statistically obtained. It consists of 5586 users and 9 evaluation indicators. Owing to spatial confined, users of top ten are listed for a demonstration.

Step 2. Making standardized treatment for the indicators data to reduce the impact of the size gap of each index on the results, normalization process is carried out. As for the two kinds of indicators,

named the profitability indicator and cost indicaotr, the following standardized formula are adopted respectively:

$$a_{ij} = \frac{x_{ij} - \min x_{ij}}{\max x_{ij} - \min x_{ij}}; \quad a_{ij} = \frac{\max x_{ij} - x_{ij}}{\max x_{ij} - \min x_{ij}}. \tag{6}$$

Step 3. Calculating the weight coefficient of each evaluation index. Relying on the formula presented in section 3,j-th evaluation index is calculated by equation 7. And the weight of each index is given in table 7.

Table 7
The weight of indicators determined by entropy

Indicator	A	B	C	D_1	D_2	E_1	E_2	E_3	F
Weight	0.29149	0.15512	0.01183	0.03675	0.00227	0.08766	0.25061	0.00705	0.15718

Table 8
The weighted matrix

ID	A	B	C	D_1	D_2	E_1	E_2	E_3	F
280951558	1.00000	0.00000	0.01360	0.00378	0.59930	0.00027	0.00242	0.01554	0.00000
1128264393	0.86314	0.00000	0.01913	0.00198	0.48696	0.00006	0.00104	0.02267	0.00000
802804158	0.00000	0.00000	0.77543	0.01164	0.80341	0.00070	0.00081	0.01813	1.00000
...	...	...	...	...	...	...	...	...	...
634540792	0.00000	1.00000	0.90332	0.06166	0.91575	0.00014	0.00258	0.01554	0.00168

$$w_j = \frac{1 - H_j}{n - \sum_{j=1}^{n} H_j} \tag{7}$$

4.3.3. Ranking the user's leading degree based on grey relational analysis

Due to the shortage of foreknowledge and complexity of the lead user identificaiton problem, the gray correlation analysis method with obvious advantages in such case is applied to rank the users' leadership in this part. The following procedure describes the specific computaion process.

Step 1. According to the user index data listed before, the reference sequence composed by the optimal values in each evaluation index is selected to form the data matrix.

Step 2. Based on the above results and the standardized formula 8, the correlation coefficients are then computed by the formula 9 below, where the parameter is set to 0.5.

$$r_{ij} = \frac{x_{ij}}{x_j^*}, i = 1, 2, \ldots, m \tag{8}$$

$$\varepsilon_j^i = \frac{\min_i \min_j \left| r_j^* - r_{ij} \right| + \rho \max_i \max_j \left| r_j^* - r_{ij} \right|}{\left| r_j^* - r_{ij} \right| + \rho \max_i \max_j \left| r_j^* - r_{ij} \right|}, \varepsilon_j^i \in (0, 1] \tag{9}$$

Step 3. The correlation degree between each user and the optimal value can be next obtained through making weighted summation of the index coefficients. Here, the weight is the indicator weight calculated by the entroy weight method. The higher the correlation degree, the higher the leading degree of users. As a result, we select the top 100 users as the user group with a higher peculiarity of lead users, which will be utilized in the subsequent identification process.

4.3.4. Ranking the user's leading degree based on TOPSIS evaluation

At the same time, TOPSIS as an useful evaluatin method can make full use of the original data and accurately reflect the gap among the evaluation objects with each evaluation index. Moreover, it has no strict requirements on sample number, data distribution and indicator number. It is also suitable to large sample with multiple indicators and objects with simple computation process. Thus, TOPSIS as one of evaluation method is utilized as below.

Step 1. Equations 6 are used to standardize the data. And each column in the matrix is multiplied by the weight of each index to acquire the weight matrix as shown in table 8.

Step 2. Determine the positive ideal solution and the negative ideal solution as shown in Table 9 according to the optimal value and the worst value of all users on each index.

Table 9
The optimal values of Topsis

Value	A	B	C	D_1	D_2	E_1	E_2	E_3	F
Optimal values	0.33318	0.17706	0.01606	0.03480	0.00295	0.08678	0.15654	0.01018	0.18244
Worst values	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000

Step 3. According to the formula 10 and formula 11, the distance relative to the positive ideal solution and negative ideal solution is calculated.

$$Z^{+} = \left(Z_1^{+}, Z_2^{+}, \ldots, Z_n^{+}\right) = \left(w_1, w_2, \ldots, w_n\right) \tag{10}$$

$$Z^{-} = \left(Z_1^{-}, Z_2^{-}, \ldots, Z_n^{-}\right) = (0, 0, \ldots, 0) \tag{11}$$

Furthermore, we apply the equation 12 to calculate the relative adjacency.

$$C_i = \frac{D_i^{-}}{D_i^{-} + D_i^{+}}$$

Similarly, relying on the ranking of leading degree, the top 100 users are selected for the subsequent lead user iden- tification, whose correlation domain is [0.0552, 0.4536]. Table 10 shows the index values and the relative adjacency of the top 10 users, others can be seen in Appendix.

4.3.5. Determination of lead user candidate sets

As previously mentioned, with respect to the rangking of leading degree, two kinds of evaluation methods are adopted to obtain two types of ranking results. In order to identify the leading users as comprehensively as possible, the “maximum inclusion principle" is used to identify the union set of the top 100 users in both two ranking results. The union set is called candidate set of lead users, which totally includes 158 users and 42 users among them appear in the two lists simultaneously. The others 116 users are only exit in one ranking list of the top 100. The results will be further used for the identification of lead users.

In fact, there are two kinds of data generated during the users’ interaction, which are behavioral data and comment data, shown in the previous survey of online community. And comment data includes counting data and textual data. However, the ranking of lead users only takes the behavioral data and counting data into account. In light of textual data, which plays an important role in studying users’ focus of attention, behavioral traits and emotional attitudes to a certain extent, it can offer important reference on the research of lead user characteristics. But until now, this kind of method has not been studied. Consequently, the users’ textual data are further used to analyze the characteristics

of online users. In the field of feature analysis, cluster analysis is a common method Madeira and Oliveira (2004). Given the high-dimensional and sparse data features, a bidirectional clustering method is introduced to carry out clustering analysis of users' keyword data, which is widely used in the field of biological genes. At the same time, considering the two kinds of situations the variation trends on absolute figure and relative value, BCBimax and CC algorithms in bidirectional clustering are used respectively. Then the clustering results are matched with the index evaluation results to identify various lead users groups in the end.

4.4. Identification of lead users based on BCBimax algorithm

In this section, the BCBimax algorithm Prelic, Bleuler, Zimmermann, Wil, Buhlmann and Gruissem (2006), which can avoid the occurrence of overlapping row clustering structure, is used to search for user groups with the same high- frequency words. The following is the detailed algorithm implementation process.

4.4.1. Data processing

The main data served as the basis is the word frequency of users under different keywords. Through the data collection, processing and screening in the early stage, the word frequency data matrix with dimension $5586 \propto 286$ was finally obtained, where the rows represent the online users of the MIUI community; The column represents the union

Table 10
The users' leading degree based on Topsis (Partial results)

ID	Compactness	A	B	C	D_1	D_2	E_1	E_2	E_3	F
367452301	0.45362	0	2604	621	1158	704	152959	115331	130	91
802804158	0.40740	0	0	1540	435	1604	1222	2130	11	1785
634540792	0.39978	0	5473	1794	2300	1828	243	6816	3	3
230111599	0.33614	2	0	150	231	717	997	1348	0	1342
1151083448	0.32612	5	0	1534	631	1601	1860	3823	49	1269
99371796	0.31916	6	1040	8	1130	506	64564	84686	209	0
2986760	0.29461	0	0	51	188	687	4193	6034	3	1109
436444	0.27835	0	1227	149	594	1254	1077	10823	11	860
180612632	0.25130	0	49	23	690	473	11600	69263	86	35
1153781126	0.24921	0	2905	200	119	1033	233	3038	0	0
754064343	0.22277	0	2307	1437	574	1505	1138	18315	11	0
1230701	0.22272	0	4	1000	601	1884	397724	85	369	0
37908342	0.21411	0	24	820	1753	1011	377868	536	52	2
275632388	0.21384	0	22	1123	776	1181	164	18454	23	687
82022196	0.20622	0	26	1599	2110	1804	355503	1308	115	1
71027246	0.20366	0	0	718	5150	905	353559	428	1010	0
69694018	0.20090	658	2	22	78	597	423	5925	0	0
326547475	0.19645	0	18	1256	1431	1752	336953	789	92	0
103329745	0.19093	342	1685	873	231	1000	461	2110	8	0
824788855	0.19025	0	0	363	86	1118	18	669	0	668
210435	0.18821	0	28	573	4576	1325	317586	2223	1385	1
36057441	0.18733	0	36	515	2414	1376	319971	454	256	4
317899	0.18444	0	84	20	3059	1483	308361	1909	1809	4
182575339	0.17318	4	72	1119	2438	1247	242838	21126	327	0
234602537	0.17111	0	0	317	147	880	2479	1773	0	591
61289579	0.16840	0	0	58	1295	672	2836	11292	7	550
2979162	0.16345	0	48	1181	677	1222	2249	10758	43	523
262561639	0.16293	0	19	2	357	195	5385	41885	51	27
151934952	0.16077	0	263	1025	763	1090	961	4020	20	526
269172825	0.15345	0	67	222	1912	401	248471	4902	238	0

of keywords in user comment content in the community; The elements in the matrix represent the word frequency of different users under different keywords, and the value range is [0, 3629].
An important step in BCBimax algorithm is data binarization. Words whose frequency is greater than the seteshold are deemed to be high-frequency words and may appear in the double clustering. Therefore, the threshold tting is of major importance. If the threshold setting is too high, a large number of effective users will be eliminated. On the contrary, if the threshold is adjusted too low, the

distinction between users will be minimal, resulting in a decrease in the accuracy. Based on Zips law and Goffman's hypothesis, Sun Sun and Davis (1999) puts forward a kind of high and low-frequency decomposition formula:

$$T = \frac{1 + \sqrt{1 + 4D}}{2}, \tag{13}$$

where T is the cut-off frequency of high-frequency words and low-frequency words, and D is the vocabulary in the analysis corpus.In this study, the above formula is invoked as a reference, and the number of keywords D=286 is substituted in. Accordingly the dividing line of high-frequency words is computed to be about "18". Therefore, this study makes fine-tuning on this basis and selects 20 as the critical value of data "0-1" processing.

4.4.2. Parameter setting

In addition to the threshold, parameters like the minimum row number "minr", the minimum column number "minc" and the maximum cluster number "number" also need to be determined. As for "minr", since the fact that proportion of lead users in the group is relatively small based on the actual situation and it is meaningless for feature analysis with few users. Hence, it is set at 20 in this paper. In terms of "minc", it is set at 5 with no excessiverestrictions. For the setiting of "number", considering the condition on the termination of algorithm presented in section 3 and reducing the influence of manual setting, a large cluster number is selected and set to 100 here.

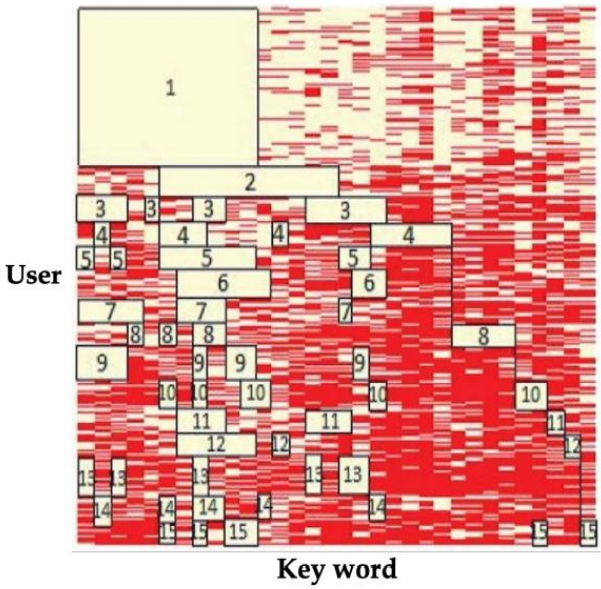

Figure 5: Heatmap of biclustering

4.4.3. Results analysis

Due to the running memory problem of R, while the complexity of BCBimax algorithm presents an exponential wth trend with the increase of matrix dimensions, MATLAB is used to implement the algorithm. A total of 15 user oups with differentiated comment characteristics were gathered, accounting for about 9% of the total users.

In order to observe the similarities and differences among diverse user groups more clearly, the clustering results plotted as the thermal diagram shown in Figure 5.

The horizontal coordinate represents keywords. The vertical coordinate represents users, the red and white block epresents word frequency. White squares indicate that keywords appear more frequently in the comments, while red squares indicate less frequently. The box marked with a number is exactly the bi-clustering found by the algorithm. It demonstrates that the 15 user groups all have high attention to some topics, accordingly the keyword coincidence appears in the left half of the figure. Meanwhile, diverse user groups have a certain uniqueness. In order to describe the characteristics more clearly, the types of keywords involved in the clustering are summarized as follows:

- Creative words: suggest (verb and noun), try (verb), hope (verb)
- Solution words: solution (noun), solve (verb)
- Technical words: Root (noun), authority (noun), ROM (noun)
- Performance discussion words: screen (noun), battery (noun), memory (noun), performance (noun), speed (noun), CPU (noun), desktop (noun), interface (noun), lock screen(noun), charge and discharge (verb), consume electricity (verb), affect (verb), optimize (verb), normal (adjective), safety (adjective)
- Experiential words: use (verb), experience (verb), test (verb), apply (verb), operate (verb), discover (verb)
- Knowledge words: course (noun)
- Entertainment word: game (noun)

It can be speculated from Figure 6(a) that the frequently mentioned words of group 1 include experiential words (285 times) and creative words (285 times). At the same time, the users post and reply in more than one section of iteration model, more than 66% of them have been registered for more than three years,whose average online time is up to in 1067 hours. It can be possible to conclude that the users are good at using their own experience to help others solving similar problems, own rich knowledge and innovation idea of product optimization.

Compared with user group 1, the high-frequency words in a user group 2 added “test" and technical words such as “Root" and “authority", which had a higher number of posts in the resource sharing part. It can be seen from Figure 6(b) that this type of user occupy the knowledge and skills of software development and are able to improve the product according to their own need. Thus, they belong to the technical user.

User group 3 combined the characteristics of creative and technical users, seeing in Figure 6(c). At the same time, the keyword appearing most frequently is “tutorial", so it tends to be knowledge-oriented users. According to Figure 6(d), user group 4 mentions a large number of performance discussion words, such as battery, memory, CPU, etc., which means they have a certain understanding of the core product performance.

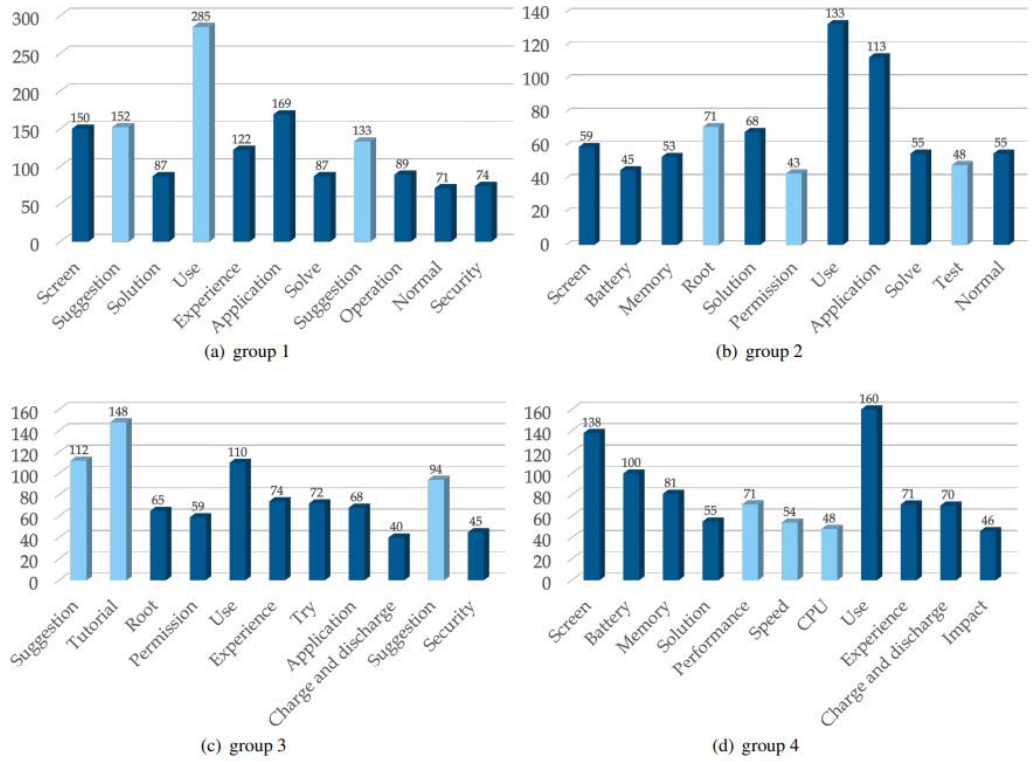

Figure 6: Frequency distribution of keywords in user group

After analyzing the representative user groups, we can similarly analyze the posting characteristics of other cate- gories of users. For the convenience of observation, the characteristics of each user group are summarized in Table11.

In addition, when the elements with a value of 0 exceed more than half of the whole data matrix, the average square residual will be zero, which is less than any square residual score value, so that the values in the obtained clusteringare all zero. Therefore, the “0" in the matrix is replaced by random numbers uniformly distributed within a considerable range. Since the generation of random numbers will increase the mean square residue score, when the range is larger than a certain critical value, these random numbers will be deleted by the algorithm without affecting the clustering results Cheng and Church (2000).

4.5. Identification of lead users based on CC algorithm

The previous section mainly considered the frequency of user keywords. After investigation, although some users registered for a short time and posted fewer messages, their textual data is also in line with definite characteristics of leading users. Therefore, CC algorithm Cheng and Church (2000) is adopted in this section to search user groups with the same relative change tendency of word frequency on the basis of word frequency data.

Table 11
Summary of user group characteristics

No.	Characteristics	Category of Keywords (type)
1	Possess innovative ideas, Well-experienced users, Understand product performance	Creation, Performance Discussion, Solution, Experience
2	Possess professional skills, Understand product performance, Adept at test feedback	Technique, Experience, Performance Discussion, Solution
3	Possess innovative ideas, Possess product knowledge	Creation, Knowledge, Technique, Experience, Function
4	Focus on core performance, Experienced users	Performance Discussion, Experience, Solution
5	Well-experienced users	Experience, Creation, Solution, Knowledge, Performance Discussion
6	Experienced users	Experience, Solution, Performance Discussion
7	Possess innovative ideas	Creation, Knowledge, Experience
8	Focus on sensory performance	Performance Discussion, Experience
9	Possess innovative ideas	Creation, Experience
10	Have entertainment needs	Entertainment, Performance Discussion, Experience
11	Possess knowledge and technology	Experience, Technology, Knowledge
12	Experienced users	Experience, Solution, Performance Discussion
13	Possess innovative ideas	Creation, Knowledge, Experience, Technology
14	Focus on life performance	Performance Discussion, Experience
15	Focus on product performance	Performance Discussion, Experience

4.5.1. Data processing

CC algorithm also needs to process data before calculation. Since the operation of CC algorithm is based on the control of the residual mean square error, while the range of existing data is considerable, so the algorithm cannot be effectively identified. Consequently, according to table 12, this study converts the word frequency into frequency levels, and obtains the data matrix with the value range of [0, 5].

Table 12
Summary of user group characteristics

Range of frequency	Grade of frequency
0	0
1	1
[2, 5]	2
[6, 20]	3
[21, 50]	4
> 50	5

4.5.2. Parameter setting

Analogous to BCBimax, CC algorithm also needs to set parameters, which are a multiple of the acceptable residual score for multi-line deletion and the maximum residual score for single-point

deletion The greater the value is, the more rows will be removed in the multi-row deletion; the greater the value is, the greater the difference of the same class in the clustering results.

In order to minimize the gap within the same user group, and to avoid inaccurate results caused by too many deletions, not to affect the operation speed due to too few deletions, the parameters were set after referring to the experimental settings of Cheng and Church Cheng and Church (2000). The maximum cluster number is placed at 100.

4.5.3. Results analysis

After determination of the parameters, CC algorithm was run in R software and 100 user groups were gained. The ting content of these user groups had relatively consistent change rules. Only 10 user groups with the largest bi- clustering dimension and the keywords with the invariable word frequency change trend in each user group are listed in Table 13.

It can be concluded that the user group 1 with the largest dimension has a total of 39 users, which have the identical hange trend in the frequency of 34 keywords. In order to observe the frequency variation trend of each keyword more clearly, the clustering result is plotted as a parallel coordinate chart as shown in Figure 7.

In Figure 7, the ordinate represents the word frequency level, the horizontal axis represents the keywords, the red and the grey line represents the change tendency of the word frequency level of the users in the user group and out he group. Furthermore, there are 39 users in this user group, and there are few red lines visible, which indicateshat a large number of users in this group have completely consistent change trend.

With the keywords as the abscissa and word frequency as the ordinate, a user word frequency median graph is drawn shown in Figure 8. It can be observed that users in a user group 1 generally have a higher expression frequency (above 50) for the graphical keywords, and have a higher expression rate for experiential vocabulary like “use" and performance discussion vocabulary like “screen". It can be inferred that this type of user has abundant experience in use and dependable understanding of product performance,who belongs to test feedback users.

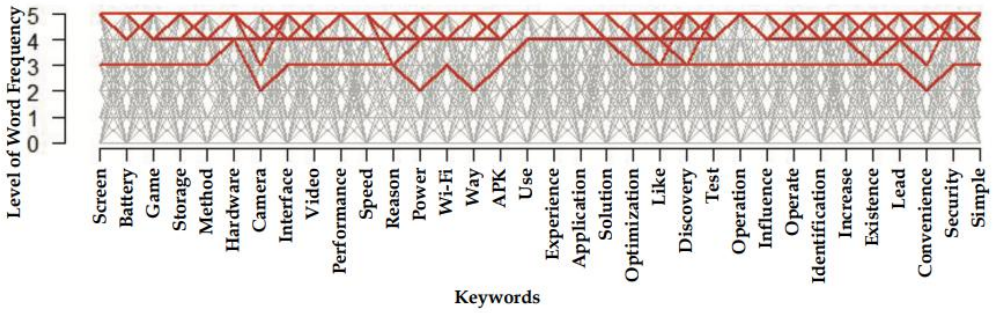

Figure 7: Parallel graph of group 1

4.6. Determination of homogeneous user sets

The user clustering under the absolute value and the change trend of relative value is considered respectively above. The user clustering results obtained by the two algorithms is matched to find the shared users, and these users will form a homogeneous user group with high similarity.

Table 13
User groups obtained by CC algorithm

No.	Dimensions	Keywords					
1	39 × 34	Screen Interface Ways Like Exist Optimization	Battery Video APK Discovery Induce Increase	Game Function Use Test Convenience WIFI	Storage Speed Experience Operation Security	Means Reason Application Influence Simple	Hardware Power Solution Identify Camera
2	24 × 13	Screen Like Application	Means Hint	Video Influence	Speed Operation	Browser Increase	Use Simple
3	17 × 13	Hardware Increase Discovery	Notification Clean	Video Recommend	Reason Continent	Power Normal	Use Simple
4	20 × 10	Game Operation	Speed Normal	WIFI Simple	Ways Influence	Voice	Hint
5	17 × 11	Screen Application	Storage Influence	Means Normal	Camera Security	WIFI Experience	Use
6	17 × 11	Means Exist	Power Security	Use Simple	Solution Join	Hint	Influence
7	18 × 10	Means Exist	Access Simple	Speed Operation	Reason	Application	Discovery
8	16 × 11	Means Like	Hardware Influence	Backstage Security	Reason Simple	Use Solution	Try
9	15 × 11	Screen Influence	Hardware Normal	Notification Simple	Reason Optimization	Application	Solution
10	13 × 12	Desktop Application	Camera Optimization	Backstage Operation	Video Increase	Speed Normal	Music Use

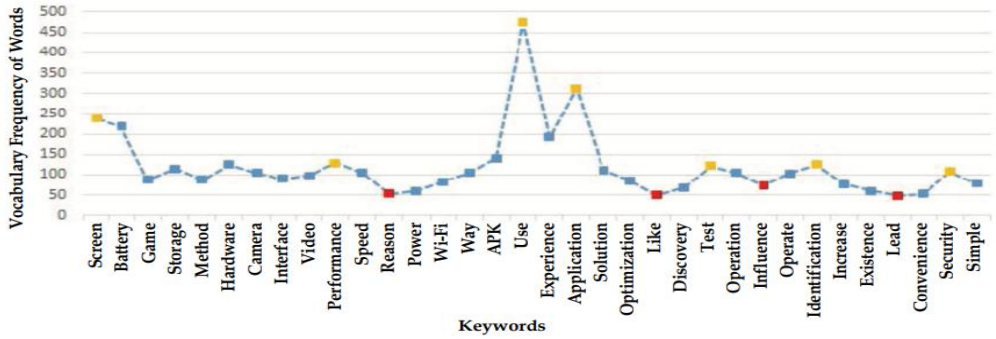

Figure 8: Word frequency trend map of group 1

In order to search out homogeneous groups of users as fully as possible, the maximum user group obtained by the two clustering results is selected, a total of 36 users appear in both user groups at the same time. The matching results in a homogeneous set of users as shown in Table 14. These users are similar in terms of high-frequency vocabulary and relative size of word frequency, which can be regarded as homogeneous user sets.

Table 14
Homogeneous user group

Biclustering intersection	Nickname	Biclustering intersection	Nickname
367452301	Hao zi 1996	105776458	MIMIM2A
71027246	Xiao qian	26920993	Ricky1988
210435	Xing fu ying er	143688456	Du xian sheng e
269172825	Qun zi cheng	3043595	ZHXmi
28125113	28125113	73371591	Ji jia nan
29997398	Bang bang chui lslht	138254898	Wang 1999
140229	whlhz	103152388	Xiao quan 1996
233413866	Xiao jian ling	72834532	Ji ye a
167776588	Qian ou yi ci yong bao	249130435	Zhuan jiao xiang xing fu
32116700	Wen rou @ Ci bei	240622059	Cha er si ůLi
169033020	Wu tong	1554630847	leo Lii
668264	HOT Mi li	253625845	xu520134
25013497	WhOòPy	274091798	Shang hai wai yu
230485228	Da pang ai xiao jun	112585282	Xue yu qing ke ls
1593866428	Xue heng qiang	95189496	+_+_
795005895	Kevin-show	1563485576	Abnesti
820891965	Wei cheng nian mian bao i	39912596	You you mai
174274618	Dong ri nuan yang_2359	64981369	Mu yue qing

4.7. Results analysis of lead user identification

In order to identify the core leading users, the selective set and the homogeneous set are matched to find user groups that not only meet the basic characteristics of leading users, but also have a high degree of similarity in textual data.

Depending on the result of index evaluation, there are 156 users in the leading user candidate set. The ID number of them is accompanied with the user ID of the homogeneous user set. If the user appears in two sets at the same time, it is referred to as the core leading user. The remaining unmatched users are hereby designated as potential lead users.

4.7.1. Feature analysis of core lead user groups

There are 20 users appearing simultaneously in the leading user selection and homogeneous user group, which can be regarded as the core leading users.These core lead users account for 56% of the homogeneous user set. Their keyword data is plotted as the word frequency cloud map illustrated in Fig 9.

Figure 9: Cloud picture of core lead user

The words appearing in the cloud map are the key words presented in the text of the core leading users. The larger size of the words is, the more frequently the key words are mentioned. As can be seen from the figure, core leading users have a high frequency of using technical vocabulary (e.g. design, technology), experiential vocabulary (e.g. use, application, experience, test), and creative vocabulary (e.g. Suggestions). It can be concluded that such users generally own innovation need ahead of the market and can take the initiative to improve products, which is in line with Hippel's definition of leading users. Therefore, identified users embody more comprehensive characteristics of leading usersfrom the textual data.

At the same time, these users have posted in multiple iterations of the same product line, proving to some extent hat they have abundant experience with the product. In addition, such users have a high number of posts and favorites in the software section of the resource sharing zone, which further proves their professional ability. Therefore, it can be deduced that the identified core leading users do occupy obvious leading user characteristics, who are excellent candidates to participate in the core link of enterprise product development.

4.7.2. Feature analysis of potential lead user groups

Except for 20 core leading users, the vast majority did not match successfully. As for the alternative users, they listed as potential leading users because they are determined on the basis of index evaluation. For users in the homogenous set, since these users present a high degree of similarity in text and core leading users account for up to 56%, it can be assumed that homogeneous users without matching also have certain leading user characteristics. Accordingly, this article refers to both them as potential leading user groups.

Potential leading users have certain directivity in their abilities, leading to a low ranking in the index evaluation and not entering the core leading user group. However, according to differentiated leading-edge characteristics, they can participate in different stages of product development. For example, technologically advanced users with a strong professional foundation can directly participate in the product design or adjust the correlation calculated from ordinary users' needs and satisfaction. Leading users who prefer to experience the product and make Suggestions can participate in the testing of the product prototype. In a word, it is undeniable that these users do have more leading characteristics than ordinary users in diverse aspects, so they can also effectively improve product development.

5. Discussions

Since co-creation is conceptualized as a cooperative process that includes actions by both supplier and customer, this section would settle the left issue how to arrange the suitable users into the different stages of the whole process of the product development and manufacturing. It means that the participation phase is not only limit to the product de- velopment process but also involves the whole product life cycle, combining the macro perspective with the micro one. Fig. 10 shows the co-creation model of lead user participation in product total life cycle. More detailed information is discussed from two aspects as following.

5.1. Theoretical implications

In a such age that values innovation, no matter academics or practitioners acknowledge the relevance of integrating customers in the development of new products and recommend the use of new technologies. While value co-creation emphasizes on customer participation, co-innovation and shared value, lead user theory highlights the special cus- tomers' involvement in the process of product innovation and the common benefits. To this end, based on the product life cycle, the study on lead user theory may be brought into the framework of value co-creation as mentioned above. This in turn would enrich and develop both two theories, named value co-creation and lead user theory. Value co-creation From its actual usage (value-in-use) rather than through its sale price (value in-exchange) to the present creating customers' experience value, value co-creation notes the full participation of customers in current times. Just as the above mentioned, the study on the basic of product life cycle could expand the scope of participants to include the solution lead users. Therefore, the participate scopes can be not only limited to the process of production and consumption, but also involves the full-process of a product's appearance, development and disappearance. Thus, a multi-level and multiple perspectives system understanding from macrocosm to microcosm can be obtained for the value co-creation.

Lead user theory In the context of big data, how to identify lead users and how can companies transfer results from lead user projects to the actual application have been judged two difficult problems in research and development of the lead user theory by many scholars. Relying on the related studies, this paper solves the above two problems. To be specific, these refer to the identification of lead users based on index evaluation and biclustering methods and the arrangement of lead users according to product life cycle. Therefore, the total solutions enrich the user identification methods and application suggestions. Furthermore, the combination of lead user with value co-creation further expands the research content of lead user theory and its research fields as well.

5.2. Managerial implications

On the basic of the identification results of above section, three user groups with different features are finally obtained, namely the core lead user group, the potential lead user group and the non-lead user group. On the other hand, the typical product life cycle has five stages, called product development stage, introduction stage, growth stage, maturity stage and decline stage respectively. Given the characteristics of every user group and each stage of the life cycle, some meaningful information and recommendations for the relevant policy-making authority can be obtained and presented as follows.

Product development stage As a start point of a new product, this stage has decisive influences on the whole product life cycle, including idea generation, idea screening, concept development and testing, marketing strategy development, business analysis, product development, text marketing and commercialization. Particularly, this stage involves the new idea generation and concept tests. Considering the outstanding innovative edge, the core lead user group, as an important external innovation source, is quite suited for this phase's demands and features. After all, the core lead users exhibit personality traits such as high predictability and strong innovativeness, in contrast with ordinary users. Therefore, the core lead users' participation and co-innovation should be the core strategy of the enterprises development in this stage.

Introduction stage During this period, low sales, high cost per customer acquired, negative profits, little competi- tion and innovators are targeted are the main features of the stage. Coincidently, the potential lead users with certain lead user characteristics could provide some product modifications, instead of a completely revised proposal in order to attract waves of new buyers for the long run. In addition, in a view of the marketing strategies, building selective distribution, improving awareness among early adopters, and heavy expenditures to create trial are necessary. Specif- ically, the potential lead users can be set to the key target population, since the core lead users would be certain to purchase the new product in this stage. It may save the cost of marketing to some extent.
Growth stage In this stage, some distinctive market characteristics can be concluded that rapidly rising sales, average cost per customer, rising profits, growing competition and early adopters are targeted. Under such circumstances, establishing intensive distribution, building awareness and interest in the mass market, and reducing expenditures to take advantage of consumer demand are the most favorable options. Therefore, the non-lead users corresponding to the mass market should be noticed by the enterprises, since through their participation some consumption habits and buying advices can be really a form of marketing orientation.

Maturity stage The maturity stage is characterized by aggressive competition, accompanied by corresponding re- ductions in product prices and profits, sales peak, low cost per customer, and middle majority are targeted. As we all known that mature market buyers accounted for most of the buying public, once the product is easy to produce sustained by its authorization, repeat purchases, and backward buyers have great influence. Hence, continuing to se- lect some of representative ordinary users to take part in the consumer producer research and formulating the targeted marketing strategy can be a good strategy for occupying more market share.

Decline stage Declining period, the end stage of a product life cycle, features with declining sales, profits and com- petition. Laggards are targeted. Distribution, advertising and sales promotion are all reducing to the level needed to retain hard-core loyalists. At this stage, it is a good time for the preparation of the next new range of products develop- ment. Collecting market and customer information through the virtual online communities are quiet a sensible choice. Moreover, based on the data collected, a new round of lead user identification can also be carried out to update the group member to respond to the needs of the business. Obviously, this period is also the start of a new cycle.

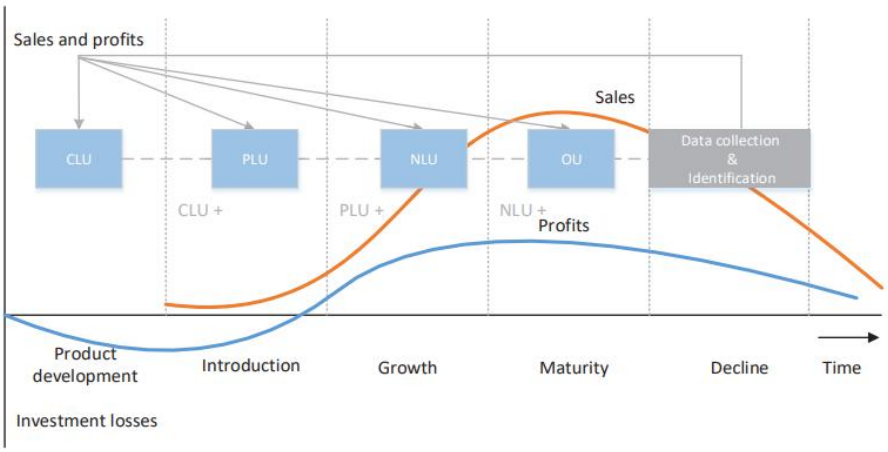

Figure 10: The co-creation model of lead user participation in product total life cycle.

6. Conclusions

Under the background of value co-creation, this paper breaks through the problem of the identification of lead users in mass consumer goods field with the big data analytic approach, and provides evidence for product development and marketing by combining the product life circle with customer participation from the practical point. More specifically, on the one hand, according to the existing researches on lead user and data features of network user, the article adapts two classification based index evaluation methods, namely gray correlation analysis and TOPSIS, on the ranking of user's leading degree. On the other hand, considering the inadequate know of lead user's characteristics in the consumer goods industry and the deficiency of user participation in the practical application, this paper analysis the different characteristics of lead users relying on the biclustering results of BCBimax algorithm and CC algorithm, and combine the identified user groups with the different stages of a product cycle, aiming to enrich the lead user theory and provide more significant advice for enterprises.

The advantage of identifying lead user though big data analysis not only conforms to the realities of the times, but also strengthen the competitive power of enterprises themselves and the development goal of value co-creation. The proposed methodology can also be extended to the identification of lead users in other online communities. The limitation of this paper is the analysis of a specific case study and only considers the participation characteristics of community users. However, with the more in-depth study and the number of users increase, the user characteristics in a certain area would be more exactly and a more effectively identification result can be followed. In terms of future work, it can be envisioned to extend the study to other innovation communities and to include new measure techniques. Additionally, the proposed methodology could be complemented with collective judgement mechanisms to improve the discovery rate of innovative lead users and more co-creation opportunities in the whole process of product development and manufacturing.

References

Testa, S., Massa, S., Martini, A., & Appio, F.P., Social media-based innovation: A review of trends and a research agenda, Information and Management, Vol.57, No.3, 103196, 2020.

Schweisfurth, T.G., Comparing internal and external lead users as sources of innovation, Research Policy, Vol.46, No.1, 238-248, 2017.

Stockstrom, C.S., Goduscheit, R.C., Luthje, C., & Jorgensen, J.H., Identifying valuable users as informants for inno- vation processes: comparing the search efficiency of pyramiding and screening, Research Policy, Vol.45, No.2, 507-516, 2016.

Chatterji, A.K., & Fabrizio, K.R., Using users: when does external knowledge enhance corporate product innovation,
Strategic Management Journal, Vol.35, No.10, 1427-1445, 2015.

Fuller, J., Muehlbacher, H., Matzler, K., & Jawecki, G., Consumer empowerment through internet-based co-creation,
Journal of Management Information Systems, Vol.26, No.3, 76-102, 2009.

Alves, H., Fernandes, C., & Raposo, M., Value co-creation: concept and contexts of application and study, Journal of Business Research, Vol.69, No.5, 1626-1633, 2016.

Aghdam, A.R., Watson, J., Cliff, C., & Miah, S.J., Improving the theoretical understanding toward patient-driven health care innovation through online value cocreation: systematic review, Journal of Medical Internet Research, Vol.22, No.4, 16324, 2020.
Gronroos, C., Value co-creation in service logic: a critical analysis, Marketing Theory, Vol.11, No.3, 279-301, 2011. Kano, N., Seraku, N., Takahashi, F., & Tsuji, S., Attractive quality and must-be quality, The Journal of the Japanese
Society for Quality Control, Vol.14, No.2, 39-48, 1984.

Tontini, G., Integrating the kano model and QFD for designing new products, Total Quality Management and Business Excellence, Vol.18, No.6, 599-612, 2007.
Hamidi, F., Gharneh, N.S., & Khajeheian, D., A conceptual framework for value co-creation in service enterprises (case of tourism agencies), Sustainability, Vol.12, No.1, 213, 2019.
Martinez, M.G., Inspiring crowdsourcing communities to create novel solutions: competition design and the medi- ating role of trust, Technological Forecasting and Social Change, Vol.117, 296-304, 2017.
Marchi, G., Giachetti, C., & de Gennaro, P., Expected value of fuzzy variable and fuzzy expected value models,
Technovation, Vol.31, No.8, 350-361, 2011.
Von Hippel, E., Has a customer already developed our next product, Sloan Management Review, Vol.18, No.2, 63-74, 1977.
Von Hippel, E., Lead users: a source of novel product concepts, Management Science, Vol.32, No.7, 791-805, 1986. Urban, G.L., & von Hippel, E., Lead user analyses for the development of new industrial products, Management
Science, Vol.34, No.5, 569-582, 1988.
Fu, W.H., Wang, Q., & Zhao, X.D., The influence of platform service innovation on value co-creation activities and the network effect, Journal of Service Management, Vol.28, No.2, 348-388, 2017.
Prahalad, C.K., The future of competition: co-creating unique value with customers,
Research-technology Manage- ment, Vol.47, No.3, 62, 2004.
Galvagno, M., & Dalli, D., Theory of value co-creation: a systematic literature review. Managing Service Quality, Vol.24, No.6, 643-683, 2014.
Xie, K., Wu, Y., Xiao, J.H., & Hu, Q., Value co-creation between firms and customers: The role of big data-based cooperative assets, Information & Management, Vol. 53, No.8, 1034-1048, 2016.
Osei-Frimpong, K., Wilson, A., & Lemke, F., Patient co-creation activities in healthcare service delivery at the micro level: the influence of online access to healthcare information, Technological Forecasting And Social Change, Vol.126, 14-27, 2018.
Vargo, S.L., & Lusch, R.F., Evolving to a new dominant logic for marketing, Journal of Marketing, Vol.68, No.1, 1-17, 2004.
Prahalad, C.K., & Ramaswamy, V., Co-creation experiences: The next practice in value creation, Journal of Interac- tive Marketing, Vol.18, No.3, 5-14, 2004.
Janamian, T., Crossland, L., & Wells, L., On the road to value co-creation in health care: the role of consumers in defining the destination, planning the journey and sharing the drive, Medical Journal of Australia, Vol.204, No.7, S12-S14, 2016.

Wu, X., & Liu, F.Y., An analysis of the motivation of customer participation value co-creation in the we-media: a study based on content marketing, Open Journal of Business and Management, Vol.06, 749-760, 2018.
Xie, C.Y., Bagozzi, R.P., & Troye, S.V., Trying to prosume: toward a theory of consumers as co-creators of value,
Journal of the Academy of Marketing Science, Vol.36, No.1, 109-122, 2008.
Greer, C.R., Lusch, R.F., & Vargo, S.L., A service perspective. Key managerial insights from service-dominant (S-D) logic, Organizational Dynamics, Vol.45, No.1, 28-38, 2016.
Hu, G.W., Yan, J.Q., Pan, W.W., Chohan, S.R., & Liu, L. The influence of public engaging intention on value co- creation of e-government services, IEEE Access, Vol.7, 111145-111159, 2019.
Gronroos, C., & Voima, P., Critical service logic: making sense of value creation and co-creation. Journal of the Academy of Marketing Science, Vol.41, No.2, 133-150, 2013.

Souto, J.E., Business model innovation & business concept innovation as the context of incremental innovation and radical innovation, Tourism Management, Vol.51, 142-155, 2015.
Hsieh, J.K., & Hsieh, Y.C., Dialogic co-creation and service innovation performance in high-tech companies, Journal of Business Research, Vol.68, No.11, 2266-2271, 2015.
Casais, B., Fernandes, J., & Sarmento, M., Tourism innovation through relationship marketing and value co-creation: a study on peer-to-peer online platforms for sharing accommodation, Journal of Hospitality and Tourism Man- agement, Vol.42, 51-57, 2020.
Frow, P., Nenonen, S., Payne, A., & Storbacka, K., Managing co-creation design: a strategic approach to innovation,
British Journal of Management, Vol.26, No.3, 463-483, 2015.
Neff, J., OMD provides the power of engagement, Advertising Age, Vol.78, No.27, 3-4, 2007.
Hoyer, W.D., Chandy, R.C., Dorotic, M., Krafft, M., & Singh, S.S., Consumer cocreation in new product development,
Journal of Service Research, Vol.13, No.3, 283-296, 2010.
Romero, D., & Molina, A., Collaborative networked organisations and customer communities: value co-creation and co-innovation in the networking era,Production Planning & Control Vol.22, No.5-6, 447-472, 2011.
Wang, X.L., Ow, T.T., Liu, L.N., Feng, Y.Q., & Liang, Y., Effects of peers and network position on user partici- pation in a firm-hosted software community: the moderating role of network centrality, European Journal Of Information Systems, Vol.1, 1-24, 2020.
Bogers, M., Afuah, A., & Bastian, B., Users as Innovators: A Review, Critique, and Future Research Directions,
Journal of Management, Vol.36, No.4, 857-875, 2010.
Fuller, J., Refining virtual co-creation from a consumer perspective, California Management Review, Vol.52, No.2, 98-122, 2010.
Franke, N., Von Hippel, E., & Schreier, M., Finding commercially attractive user innovations: a test of lead-user theory, The Journal of Product Innovation Management, Vol.23, No.4, 301-315, 2006.
Luthje, C., Characteristics of innovating users in a consumer goods field: an empirical study of sport-related product consumers, Technovation, Vol.24, No.9, 683-695, 2004.

Morrison, P. D., Roberts, J. H., & Midgley, D. F., The nature of lead users and measurement of leading edge status,
Research Policy, Vol.33, No.2, 351-362, 2004.
Schreier, M., & Prugl, R., Extending lead user theory: antecedents and consequences of consumer lead userness, The Journal of Product Innovation Management, Vol.25, No.4, 331-346, 2008.
Pongtanalert, K., & Ogawa, S., Classifying user-innovators: an approach to utilize user-innovator asset, Journal of Engineering and Technology Management, Vol.37, 32-39, 2015.
Von Hippel, E., Frank, N., & Prugl, R., Pyramiding: efficient search for rare subjects, Research Policy, Vol.38, No.9, 1397-1406, 2009.
Tietz, R., Fuller, J., & Herstatt, C., Signaling: an innovative approach to identify lead users in online communities,
Customer Interaction and Customer Integration, Vol.22, No.2, 453, 2006.
Piller, F.T., & Walcher, D., Toolkits for idea competitions: a novel method to integrate users in new product devel- opment, R & D Management, Vol.36, No.3, 307-318, 2006.
Belz, F.M., & Baumbach, W., Netnography as a method of lead user identification, Creativity and Innovation Man- agement, Vol.19, No.3, 304-313, 2010.
Brem, A., & Bilgram, V., The search for innovative partners in co-creation: identifying lead users in social media through netnography and crowdsourcing, Journal of Engineering and Technology Management, Vol.37, No.9, 40-51, 2015.

Pajo, S., Verhaegen, P.A., Vandevenne, D., & Duflou, J.R., Towards automatic and accurate lead user identification,
Procedia Engineering, Vol.131, No.1, 509-513, 2015.
He, G.Z., & Chen, R.Q., Research on identification methods of leading user in consumer industry, Statistics and Decision-making, No.4, 15-17, 2009.
Qiu, Z.T., & Lv, H., Research on leading user identification methods, Chinese Market, No.33, 64-65, 2013.
Zhao, X.Y., & Sun, F.Q., Automatic identification of lead users in collaborative innovation community, Journal of Wuhan University of Technology, Vol.36, No.4, 537-545, 2014.
Li. N., Lead user identification based on improvement of PROMETHEE, Practice and Understanding of Mathematics, Vol.44, No.10, 44-52, 2014.
Lilien, G.L., Morrison, P.D., Searls, K., Sonnach, M., & Von Hippel, E., Performance assessment of the lead user idea-generation process for new product development, Management Science, Vol.48, No.8, 1042-1059, 2002.
Nambisan, S., & Baron, R.A., Interactions in virtual customer environments: implications for product support and customer relationship management, Journal of Interactive Marketing, Vol.21, No.2, 42-62, 2007.
Nambisan, S., & Nambisan, P., How to profit from a better “Virtual Customer Environment”, MIT Sloan Management Review, Vol.49, No.3, 53-61, 2008.
Magnusson, P.R., Exploring the contributions of involving ordinary users in ideation of tech-nologybased services,
Journal of Product Innovation Management, Vol.26, No.5, 579, 2009.

Schuurman, D., Mahr, D., & De Marez, L., User characteristics for customer involvement in innovation processes: deconstructing the Lead User-concept, ISPIM 22nd conference: Sustainability in Innovation: Innovation Man- agement Challenges, 2011.
Von Hippel, E., Democratizing innovation: The evolving phenomenon of user innovation, Journal fur Betrieb- swirtschaft, Vol.55, No.1, 63-78, 2005.
Luthje, C., & Herstatt, C., The lead user method: an outline of empirical findings and issues for future research, R & D Management, Vol.34, No.5, 553-568, 2004.
Ernst, M., Brem, A., & Voigt, K.I., Innovation management, lead-users, and social media-introduction of a conceptual framework for integrating social media tools in lead-user management, Social Media in Strategic Management, Vol.25, No.4, 169-195, 2014.
Nishikawa, H., Schreier, M., & Ogawa, S., User-generated versus designer-generated products: a performance as- sessment at Muji, International Journal of Research in Marketing, Vol.30, No.2, 160-167, 2013.
Cheng, Y.Z., & Church, G.M., Biclustering of expression data, In: Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology, 93-103, 2000.
Bergrmann, S., Ihlnels, J., & Barkai, N., Iterative signature algorithm for the analysis of large-scale gene expression data, Physical Review, Vol.67, No.3, 1-18, 2003.
Yang, J., Wang, W., & Wang, H., Enhanced biclustering on expression data, In: Proceedings of the 3th IEEE Confer- ence on Bioinformatics and Bioengineering, 321-327, 2003.
Tanay, A., Sharan, R., Kupiec, M., & Shamir, R., Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data, Proceedings of the National Academy of Sciences of the United States of America, Vol.101, No.9, 2981-2986, 2004.
Prelic, A., Bleuler, S., Zimmermann, P., Wil, A., Buhlmann, P., Gruissem, W., et al. A systematic comparison and evaluation of biclustering methods for gene expression data, Bioinformatics, Vol.22, No.9, 1122-1129, 2006.

Busygin, S., Prokopyev, O., & Pardalos, P.M., Biclustering in data mining, Computers and Operations, Vol.35, No.9, 2964-2987, 2008.
Madeira, S., & Oliveira, L., Biclustering algorithms for biological data analysis: a survey, Transactions on Compu- tational Biology and Bioinformatics, Vol.1, No.1, 24-25, 2004.
Oghabian, A., Kilpinen, S., Hautaniemi, S., & Czeizler, E., Biclustering methods: biological relevance and applica- tion in gene expression analysis, PLoS One, Vol.9, No.3, e90801, 2014.
Xie, J., Ma, A.J., Fennell, A., Ma, Q., & Zhao J., It is time to apply biclustering: a comprehensive review of bi- clustering applications in biological and biomedical data, Briefings in Bioinformatics, Vol.20, No.4, 1449-1464, 2019.
Yang, J., Wang, W., Wang, H., & Yu, P., delta-clusters: capturing subspace correlation in a large data set, In: Pro- ceedings of the 18th International conference on data engineering, 517-528, 2002.
Yang, J., Wang, H.X., Wang, W,. & Yu, P., Enhanced biclustering on expression data, In: Proceedings of the 3rd IEEE symposium on bioInformatics and bioengineering, 321-327, 2003.
Dolnicar, S., Kaiser, S., Lazarevski, K., & Leisch, F., Biclustering overcoming data dimensionality problems in market segmentation, Journal of Travel Research, Vol.51, No.1, 41-49, 2012.

Lin, Q., Zhang, H., Wang, X., Xue, Y., Liu, H., & Gong, C., A novel parallel biclustering approach and its application to identify and segment highly profitable telecom customers, IEEE Access, Vol.7, 28696ÍC28711, 2019.
Swathi, H., Gene expression data knowledge discovery using global & local clustering, Journal of Computing Vol.2, No.3, 116-121, 2010.
Hartigan, J.A., Direct clustering of a data matrix, Journal of the American Statistical Association Vol.67, No.337, 123-129, 1972.
Lazzeroni, L., & Owen, A., Plaid models for gene expression data, Statistica Sinica Vol.12, 61-86, 2002.
Shannon, C.E., & Weaver, W., The mathematical theory of communication, Bell system technical journal Vol.27, No.3, 379-423, 1948.
Deng, J.L., Control problems of grey systems, System & Control Letters. Vol.1, No.5, 288-294, 1982.
Pu, X.G., & Liu, L.M., An empirical study on the performance evaluation of databases based on coefficient of varia- tion & grey relational analysis, Library and Information Service Vol.58, No.14, 71-78, 2014.
Yoon, K., A reconciliation among discrete compromise solutions, The Journal of the Operational Research Society, Vol.38, No.3, 277-286, 1987.
Bischoff, P., BCG: Xiaomi is one of the world's most innovative companies, and the top up-andcomer, Retrieved from https://www.techinasia.com/bcg-xiaomi-worlds-innovative -companies-topupandcomer, 2014.
Cho, J., Kim, E., & Jeong, I., International orientation and cross-functional integration in new product development,
Asian Business & Management Vol.16, No.4-5, 226-252, 2017.
Morrison, P.D., Roberts, J.H., & Midgley, D.F., The nature of lead users and measurement of leading edge status,
Research Policy Vol.33, No.2, 351-362, 2004.
Hienerth, C., Potz, M., & von Hippel, E., Exploring key characteristics of lead user workshop participants: who contributes best to the generation of truly novel solutions, In: Proceedings of the DRUID Summer Conference on Appropriability, Proximity, Routines and Innovation, 2007.
Lee, S., PIG Strategy: Make Customer Centricity Obsolete and Start a Resource Revolution, Hong Kong: iMatchPoint Limited, 2014.

Schreier, M., Oberhauser, S., & Prugl, R., Lead users and the adoption and diffusion of new products: insights from two extreme sports communities, Marketing Letters Vol.18, No.1-2, 15-30, 2007.
Fuller, J., Jawecki, H., & Muhlbacher, H., Innovation creation by online basketball communities, Journal of Business Research Vol.60, No.1, 60-71, 2007.
Sun, Q., & Davis, C.H., A model for estimating the occurrence of same-frequency word and. the boundary be- tween high-and low-frequency words in text, Journal of the Association for Information Science and Technology Vol.50, No.3, 280-286, 1999.