The Review of Socionetwork Strategies (2020) 14:193–204
https://doi.org/10.1007/s12626-020-00063-4

ARTICLE

Improving Matching Process with Expanding and Classifying Criterial Keywords leveraging Word Embedding and Hierarchical Clustering Methods

Yutaka Iwakami[1] · Hironori Takuma[2] · Motoi Iwashita[3]

Received: 14 May 2020 / Accepted: 26 August 2020 / Published online: 14 September 2020

Abstract

Matching processes, such as the selection of producers of advertising content corresponding to specific products or the screening of job applicants based on predefined requirements, have become important operations required by enterprises. Such problems generally include several keywords representing the matching criteria, but it is difficult for enterprises to expand and classify criterial keywords properly to improve the matching performance. This study proposes solutions to this issue by extracting criterial keywords from social networking services (SNSs) based on word embedding and by classifying the obtained keywords via hierarchical clustering. This approach will enable enterprises to gather and prioritize criterial keywords more accurately to improve their matching processes.

Keywords Matching process · Word2Vec · Hierarchical clustering · NLP · SNS · Semantic analysis

1 Introduction

Over the recent years, matching processes have become an important component of commercial success. For example, it is vital for enterprises to advertise on the web to increase public recognition of their products and services. Product recommendations or introductions by individual content creators have also garnered attention of late. However, the number of available content creators is very high. It is, therefore,

✉ Yutaka Iwakami
yuta@norkresearch.co.jp

[1] Department of Research & Analysis for IT Industry, Nork Research Co., Ltd, 2-13-10 Shinjuku, Shinjuku-Ku, Tokyo 160-0022, Japan

[2] Department of Project Management, Chiba Institute of Technology, 2-17-1 Tsudanuma Narashino, Chiba 275-0016, Japan

[3] Department of Management Information Science, Chiba Institute of Technology, 2-17-1 Tsudanuma Narashino, Chiba 275-0016, Japan

critical for enterprises to identify the content producers that are most suitable to advertise their products and services. Several techniques have been proposed to resolve this issue [1–3]. On the other hand, employment is an essential requirement for enterprises. Owing to the rapid digitalization of society, job descriptions are becoming progressively more diverse and complicated making recruitment difficult to optimize. This also comprises a topic of active research [4, 5]. Matching of inquiries with solutions in customer support is yet another challenging matching problem [6].

In the aforementioned processes, enterprises usually provide a set of keywords as criteria for their requirements. For example, if a beer-brewer considers that “stout” well represents a characteristic of their product, they would wish to include the word as a criterion for selecting producers of advertising content. However, the producer’s keywords might not always capture the appealing facets of the product effectively. On the other hand, general consumers may have captured the essence of the product more effectively on their social networking services (SNS). Thus, the common challenge in such matching problems lies in identifying methods to expand criterial keywords properly and efficiently by extracting associated keywords from SNS.

The fundamental ideas underlying matching processes between enterprises and their counterparts are based on the identification of similarities between the criterial keywords. Co-occurrence network analysis is one of popular methods used to measure similarities among data, based on indices, such as the Jaccard coefficient. This method is known to be efficient for matching processes like product recommendation. Advanced studies on this method have also combined it with other analysis techniques, such as Analytic Hierarchy Process (AHP) [7, 8]. However, co-occurrence network analysis does not offer any means to expand criterial keywords by itself. Therefore, in this study, the authors propose a method to expand keywords by leveraging word embedding. Word2Vec [9] is a word embedding technique that converts each keyword into a vector and proceeds to calculate similarities between pairs of words by computing the inner product between the two corresponding vectors. Owing to this operational characteristic, Word2Vec can be applied to the identification and standardization of derivatives as well. For example, “Java7” and “Java9” can be identified to be different versions of the developing language, "Java" [10]. Estimation of similarity via Word2Vec can also be applied to extract keywords representing essential aspects of products [11] or detect human emotions inherent in social media content [12] or semantic analysis [13] etc. [14–16].

In this study, besides proposing a method to expand criterial keywords by leveraging Word2Vec, the authors demonstrate the use of hierarchical clustering to classify the obtained keywords. A combination of these two methods will enable enterprises to expand their criterial keywords effectively and to decide the priorities of expanded keywords. This would represent a significant improvement in their matching performances.

2 Example Case

The methods proposed in this study can be applied to any case where a product has been provided alongside a keyword that characterizes it, and the keyword needs to be expanded using keywords collected from SNS. As a simple analogy, let us reconsider the example of a beer-brewer. The brewer wishes to promote their primary product, which possesses the characteristic of “sharpness”. Thus, the brewer wishes to expand this expression using more appealing keywords collected from SNS (Twitter in this example) to provide more precise criteria to prospective producers of advertising content. In the following sections, the beer is denoted by “the product”, “sharpness” is denoted by “the criterial keyword”, and the proposed analytical procedures are explained using this analogy. However, all the methods described below are also applicable to a wide range of business as long as the purpose is to expand keywords characterizing a product and classify them based on the activities of consumers on SNS.

3 Proposed Procedures

The proposed procedures are classified into four categories — data collection, data preprocessing, keyword expansion, and keyword classification. In this section, each step is explained in detail. The statistical computing language and environment, “R”, is used as to retrieve and analyze data in this study.

3.1 Data Collection

In this step, recent tweets by a sufficient number of users are retrieved. The “rtweet” library of R is used as for tweet retrieval in this study. In terms of the frequency of occurrence of the product and the criterial keyword, users are classified into the two following categories:

Engaged users:Users whose tweets include both the criterial keyword and the product.

Non-engaged users:Users whose tweets include only the criterial keyword.

If users whose tweets include only the product were grouped together, their tweets would contain very diffused keywords about the product. The motivation of the brewer is to create highly appealing advertising content based on the criterial keyword. To prevent dilution due to keyword diffusion, tweets should be obtained from users who are adjudged to be familiar with the product based on their tweets. These are precisely the engaged users, as defined above.

However, some keywords used in the tweets of engaged users might be too close to the product, which can make it difficult for consumers to realize the features of the product in the context of the criterial keyword. Therefore, candidates for expanded keywords should be also collected from the tweets of users who only tweet about the criterial keyword. These are the non-engaged users, as defined above.

This concept of engaged and non-engaged users is also applicable in other cases, when an enterprise wants to identify keywords related to a product based on one of its specific features.

3.2 Data Preprocessing

In this step, text data obtained from tweets of the engaged and non-engaged users are decomposed into a set of words. As the data contain orthographic variants, morphological analysis is applied while splitting it into lemmatized words [17, 18]. "MeCab" is a popular tool, which can handle this process in Japanese and has been used in this study.

3.3 Keyword Expansion

In this step, the set of words is converted into vectors using Word2Vec. Word2Vec is a sort of applications of Neural network that represents a given set of words as vectors via contextual comprehension [19]. It embeds semantic relationships between words into the calculation of the corresponding vectors. For example, if the four words "King", "Man", "Woman", and "Queen" were converted into vectors via Word2Vec, the vector obtained using the following formula on the corresponding vectors "King"—"Man" + "Woman" would yield the vector corresponding to "Queen" [20]. Two algorithms are available to be used in Word2Vec — skip-gram and continuous bag of words also known as CBOW [21, 22]. The former network is supervised to predict the neighboring words of the current word, while the latter predicts the current word based on its neighbors. The purpose in this step is to expand the criterial keyword. Therefore, the skip-gram algorithm is implemented in this study using the "wordVectors" library of R Fig. 1.

The similarity between any pair of vectors can be measured by computing the inner product of the corresponding vectors determined by Word2Vec. The value of the inner product is proportional to the degree of the contextual relationship between the corresponding keywords. By leveraging this advantage, a list of keywords which are close to the criterial keyword are obtained. They comprise the "list of non-engaged keywords". In addition, the vectors corresponding to the product and the criterial keyword are added together to yield the compound vector. Then a list of its neighboring keywords is obtained. They comprise the "list of engaged keywords". Finally, keywords which satisfy the following two conditions are extracted Fig. 2.

- It exists in both the non-engaged list and the engaged list.
- It exhibits higher inner products with the criterial keyword in the engaged keywords list than that in the non-engaged list.

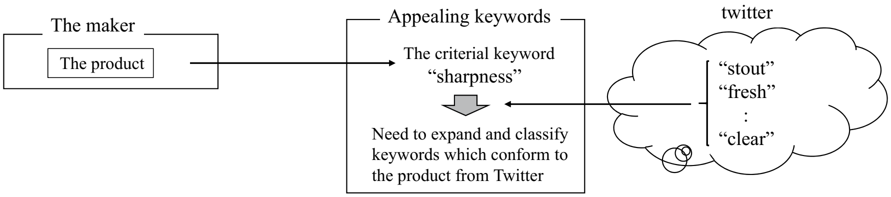


Fig. 1 Example case

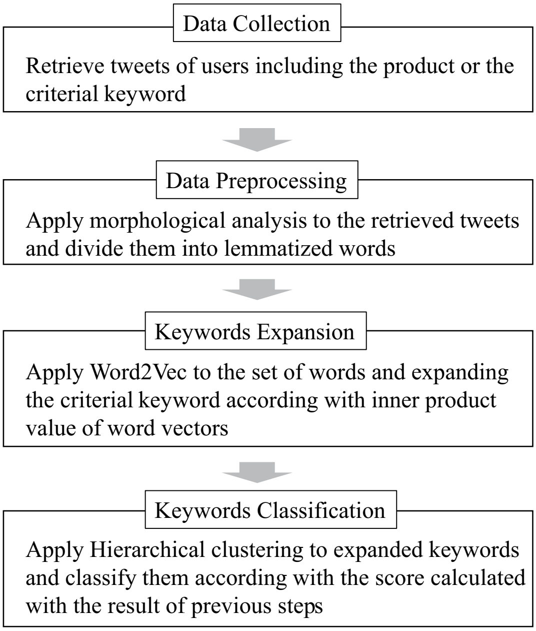

Data Collection
Retrieve tweets of users including the product or the criterial keyword
Data Preprocessing
Apply morphological analysis to the retrieved tweets and divide them into lemmatized words
Keywords Expansion
Apply Word2Vec to the set of words and expanding the criterial keyword according with inner product value of word vectors
Keywords Classification
Apply Hierarchical clustering to expanded keywords and classify them according with the score calculated with the result of previous steps

Fig. 2 Four steps of the proposed procedures

The extracted keywords are close to the criterial keyword and are also related to the product. Thus, they are considered to be expansions of the criterial keyword suitable for advertisement of the product. An outline of this step has been illustrated in Fig. 3.

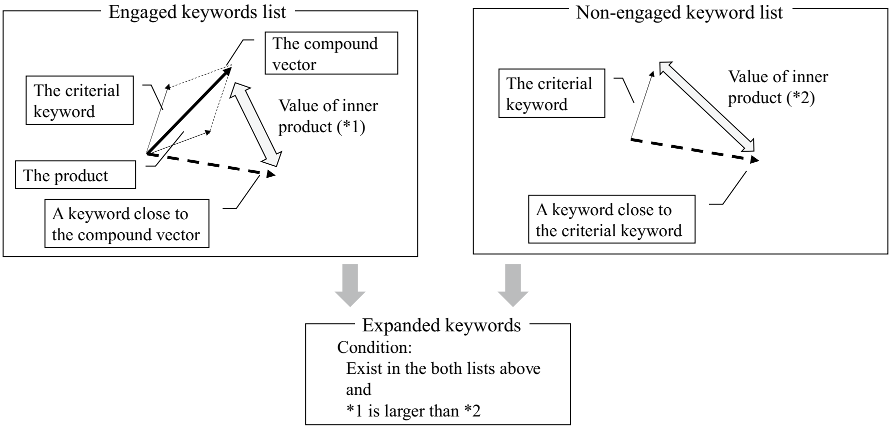


Fig. 3 The outline of keywords expanding

3.4 Keyword Classification

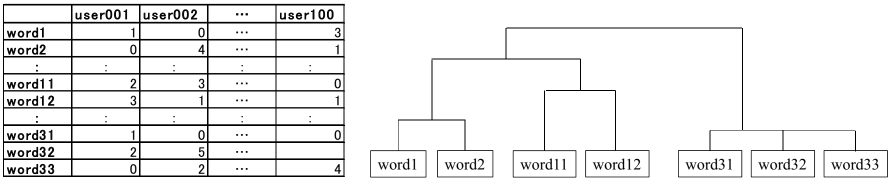
In this step, the expanded keywords obtained in the previous step are prioritized. Although the expanded keywords can be ranked based on their inner products with the criterial keyword, it is necessary to consider the similarity between them as well. For example, suppose there were four expanded keywords A, B, C, and D, such that A and B are often seen in the tweets of some subgroup of users, and that C and D are often seen in the tweets of another subgroup. To obtain a comprehensive collection of expanded keywords, representative keywords should not only be chosen between A and B but between C and D. However, if the keywords were selected only based on their inner products with the criterial keyword, C and D might both be excluded. To prevent this improper exclusion, expanded keywords must be first classified based on their similarity to each other.

To this end, the authors apply hierarchical clustering, which is a method that classifies data based on the pair-wise distances between them after regarding them as points in an n-dimensional space [23–25]. k-means clustering is another method for this kind of classification, which is also attracting theme [26, 27]. In k-means clustering, the number of clusters is required to be determined in advance. In real businesses, the number of expanded keywords which can be configured generally depends on related systems or marketing expenses. The number of clusters in hierarchical clustering does not need to be configured in advance. Because of this reason, hierarchical clustering is adopted in this study.

While various techniques may be used to measure the pair-wise similarities between the expanded keywords, the authors adopt the number of times each keyword has appeared in each user's tweet as the metric. This is to characterize the measured similarity between each pair of words based on whether are closely related in the tweets of engaged users or those of non-engaged ones.

Before applying the hierarchical clustering method, the frequency of each expanded keyword corresponding to each user is calculated based on the tweet data obtained during the Data Processing step. In other words, if the number of expanded keywords is N and the number of users is M, a dataset of N rows and M column is obtained, as depicted on the left side of Fig. 4. Then, hierarchical clustering is applied to the dataset and the hierarchical classification among the expanded keywords is obtained in the form of dendrogram, as depicted on the right side of Fig. 4.

To obtain a comprehensive collection of expanded keywords, one keyword should be chosen from each cluster. Thus, one keyword should be chosen from all possible candidates belonging to the same cluster based on some sort of



	user001	user002	···	user100
word1	1	0	···	3
word2	0	4	···	1
:	:	:	:	:
word11	2	3	···	0
word12	3	1	···	1
:	:	:	:	:
word31	1	0	···	0
word32	2	5	···	
word33	0	2	···	4

Fig. 4 The outline of keywords classifying

prioritization, because the number of expanded keywords that can be configured is often limited by associated systems or marketing expenses.

Therefore, the only remaining issue is to identify a method to prioritize the keywords. Two factors indicate the importance of a keyword in this model — the value of the inner product with the compound vector, which represents the proximity of the keyword to the product and the criterial keyword, and the frequency of occurrence of the keyword, as more frequently used keywords are generally more appealing.

Deviation from the mean is selected as a metric for the first factor. Deviations are more useful to compare similar data than the values themselves. The final values are multiplied by 100 for readability, as the absolute values of the inner products are less than 1.

If the frequency of appearance of a word was taken to be the metric for the second factor, the influence of temporarily trending keywords on SNS might be inflated. For this reason, the log value (base 10) of the frequency of occurrence is adopted as the metric in this study.

Finally, the two factors are combined. Because the appeal of a keyword can be considered to be roughly proportional to its frequency of occurrence, the product of the two factors is used as the metric. Therefore, the authors propose the following metric for prioritization of the expanded words.

$$priority(word) = (inn(word) - mean) * 100 * log(freq(word)) \tag{1}$$

inn: inner product of an expanded keywords with the compound vector.

mean:mean of the *inn* of all expanded keywords.

freq: sum of the frequencies of occurrence of each expanded keyword in the tweets of each user.

Based on this metric, the priority list of expanded keywords belonging to the same cluster can be ascertained.

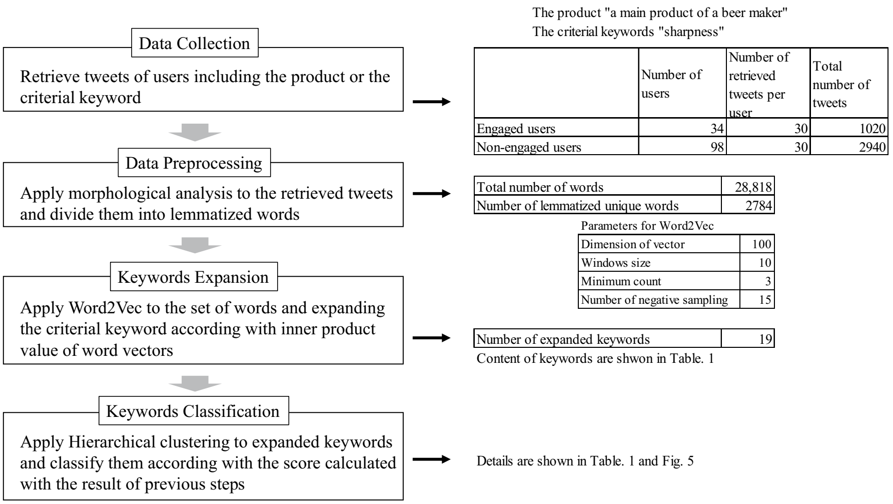
4 Experimental Evaluation

In this section, the authors apply the proposed procedure on an existing beer-brewer who characterized his primary product using the word “sharpness”. Figure 5 depicts the actual amount of data corresponding to each step of the proposed method. 1020 and 2940 tweets were retrieved from engaged and non-engaged users, respectively.

The number of tweets retrieved per user was 30. In the case of products like beer, appropriate expanded keywords may vary depending on the season. In this example, for simplicity, one month is taken to be the term of tweet retrieval. This parameter should be changed depending on the characteristic of each product.

The parameters for Word2Vec are used to configure the “wordVectors” library in R, which follows the recommended data size, etc.

As a result, 19 expanded keywords were obtained.


Data Collection
Retrieve tweets of users including the product or the criterial keyword
Data Preprocessing
Apply morphological analysis to the retrieved tweets and divide them into lemmatized words
Keywords Expansion
Apply Word2Vec to the set of words and expanding the criterial keyword according with inner product value of word vectors
Keywords Classification
Apply Hierarchical clustering to expanded keywords and classify them according with the score calculated with the result of previous steps

The product "a main product of a beer maker"
The criterial keywords "sharpness"

	Number of users	Number of retrieved tweets per user	Total number of tweets
Engaged users	34	30	1020
Non-engaged users	98	30	2940

Total number of words	28,818
Number of lemmatized unique words	2784

Parameters for Word2Vec

Dimension of vector	100
Windows size	10
Minimum count	3
Number of negative sampling	15

Number of expanded keywords	19

Content of keywords are shwon in Table. 1

Details are shown in Table. 1 and Fig. 5

Fig. 5 Actual amount of data in an example case

Table 1 presents the 19 expanded keywords and associated important factors like eng (word), noeng (word), freq (word) and the priority index discussed in the Sects. 3 and 4.

Table. 1 The content of expanded keywords and important factors

	Keyword	*inn* (word)	*freq* (word)	Priority index
word 1	Wheat	0.165	3	0.86
word2	Rich	0.165	9	1.70
word3	Guzzle	0.168	3	1.02
word4	Chilled	0.133	5	−1.01
word5	Drinkable	0.166	18	2.41
word6	Lager	0.171	4	1.46
word7	Alcohol percent	0.110	3	−1.76
word8	Craft beer	0.181	3	1.63
word9	Dry	0.188	6	3.20
word10	Belgium	0.154	10	0.71
word11	Fruity	0.165	3	0.83
word12	Cheers	0.172	7	2.07
word13	Bitterness	0.167	8	1.77
word14	Refreshing	0.133	6	−1.09
word15	Brisk	0.128	5	−1.34
word16	Strongest	0.094	9	−5.06
word17	Solid	0.141	13	−0.74
word18	Thick	0.097	8	−4.58
word19	Taste	0.098	187	−11.26

Dry” is observed to be contextually similar to “sharpness” in the expression of beer, and it exhibited a high priority index in the analysis. On the other hand, the index of “taste” was observed to be low in spite of its high frequency. This can be attributed to its low *inn* (word) value, signifying that “taste” does not share a strong relationship with “sharpness” or with the product, as it is a very general word. Thus, the result of the analysis was observed to correspond with daily impressions of consumers about beer.

As discussed in the Sect. 3.2, prioritization of expanded keywords is required in a real-life business. Based on the proposed methods, expanded keywords were classified via hierarchical clustering in this example, as depicted in Fig. 6.

A large cluster including 10 keywords (word12, word9, word8, word14, word13, word10, word15, word6, word1, word7, word11) was observed on the right half of the dendrogram. If one keyword is to be selected for business matching, the keyword “dry” (word9) should be selected, as it exhibits the highest priority index within the cluster, as evident from Table 1. However, the effectiveness of the proposed priority index should also be verified. For this purpose, the number of actual tweets including both the product and each expanded keyword in

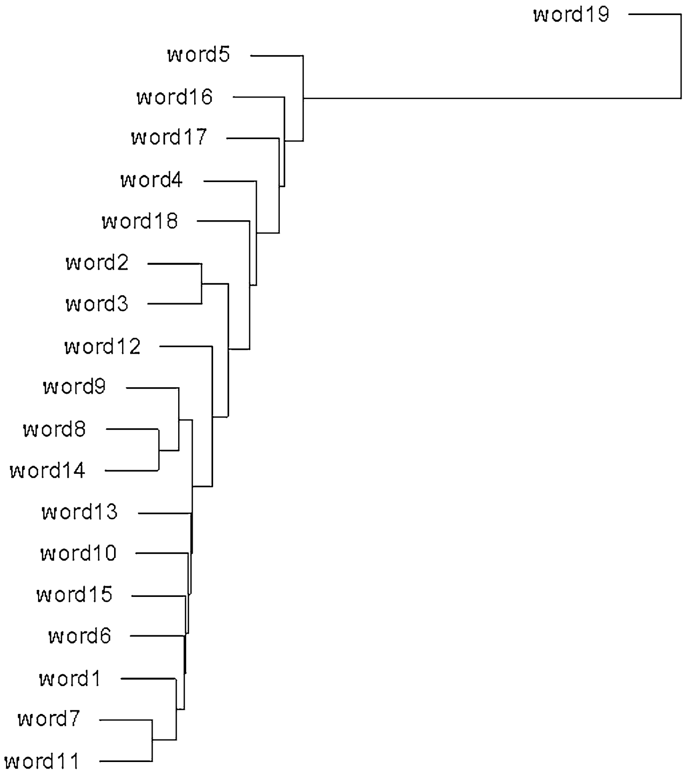


Fig. 6 Classification of expanded keywords

the cluster was computed and compared to its corresponding value in the priority index. The results have been depicted in Fig. 7.

A trend is clearly evident — keywords with higher priority index exhibit higher frequency of appearance alongside the product. That implies that the priority index proposed in this study reflects the true relationship between the product and each expanded keyword. Therefore, enterprises can use the proposed method in its entirety to actually expand, classify, and prioritize keywords suitable for their products.

5 Conclusion and Future Work

Effective selection of criterial keywords is essential in business matching. For this purpose, it is useful to expand the criterial keywords by applying word2Vec and classify the keywords via hierarchical clustering. Further, the keywords must be prioritized based on the indices derived from the results of the two aforementioned methods.

Based on verification using actual data, we confirmed the effectiveness of the proposed procedure. However, in future works, similar experimental verification for other products is required. The definition of priority index should be also be made more sophisticated in future works, for instance, by considering time series of keywords, etc.

As a further step, relationship or influence between expanded keywords should be considered. This approach is similar to the identification of KPIs/KGIs in project management [28–30]. For example, an appeal using two keywords with lower priority indices might produce a better effect than that using one keyword with a higher priority index. Probabilistic inference methods, such as Bayesian Network [31, 32] or Decision Tree [33, 34], would be suitable for this investigation.

Acknowledgements Use the plural heading even if you have single acknowledgement

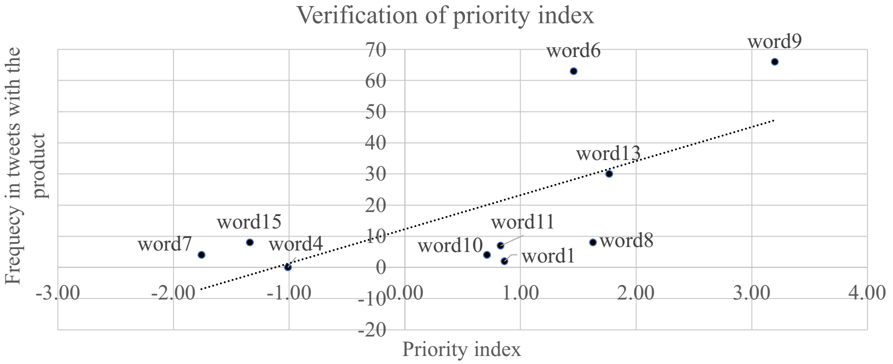


Fig. 7 Verification of priority index

Compliance with ehtical standards

Conflict of Interest Statement On behalf of all authors, the corresponding author states that there is no conflict of interest to declare.

Appendix

Appendixes, if needed, appear before the acknowledgments.

References

1. Iwashita, M. (2019). A proposal of matching algorithm for new type of advertisement business model. *Procedia Computer Science, 159*, 1966–1975.
2. Haan, W., & Kaltenbrunner, G. (2009). Anticipated growth and business cycles in matching models. *Journal of Monetary Economics, 56*(3), 309–327.
3. Iwashita, M., Tanimoto, S., & Tsuchiya, K. (2018). Framework of highly secure transaction management for affiliate services of video advertising. *Procedia Computer Science, 126*, 1802–1809.
4. Hall, R., & Schulhofer-Wohl, S. (2018). Measuring job-finding rates and matching efficiency with heterogeneous job-seekers. *American Economic Journal, 10*(1), 1–32.
5. Higashi, Y. (2018). Spatial spillovers in job matching: Evidence from the Japanese local labor markets. *Journal of the Japanese and International Economics, 50*, 1–15.
6. Iwashita, M., Shimogawa, S., & Nishimatsu, K. (2011). Semantic analysis and classification method for customer enquiries in telecommunication services. *Engineering Applications of Artificial Intelligence, 24*(8), 1521–1531.
7. Garg, M., & Kumar, M. (2018). Identifying influential segments from word co-occurrence networks using AHP. *Cognitve Systems Research, 47*, 28–41.
8. Angelo, L., Stefan, P., Fratocchi, L., Marzola. A. (2018) An AHP-based method for choosing the best 3D scanner for cultural heritage applications. *Journal of Cultural Heritage 34*, 109–115.
9. Mkolov, T., Chen, K., Corrado, G., Dean, J. (2013) Efficient estimation of word representations in vector space. *Computation and Language*
10. Fukui, K., Miyazaki, T., Ohira, M. (2019) Suggesting questions that match each user's expertise in community question and answering services, 20th IEEE/ACIS international conference on software engineering, artificial intelligence, networking and parallel/distributed computing (SNPD) https://doi.org/10.1109/SNPD.2019.8935747
11. Jing, X., Wang, P., & Rayz, J. (2018). Discovering attribute-specific features from online reviews: what is the gap between automated tools and human cognition? *Software Science and Computational Intelligence*. https://doi.org/10.4018/IJSSCI.201804010.
12. Jan, R., Khan, A. (2020). Emotion mining using semantic similarity. *Natural Language Processing*. https://doi.org/10.4018/978-1-7998-0951-7.ch053.
13. Kim, S., Park, H., Lee, J (2020) Word2vec-based latent semantic analysis (W2V-LSA) for topic modeling: A study on blockchain technology trend analysis. *Expert Systems with Applications152*, 113401
14. Jatnika, D., Biijaksana, M., & Suryani, A. (2019). Word2Vec model analysis for semantic similarities in english words. *Procedia Computer Science, 157*, 160–167.
15. Kai, H., Qing, L., Kunlun, Qi., Siluo, Y., Jin, M., Xiaokang, F., Jie, Z., Huayi, W., Ya, G., and Qibing, Z. (2019) Understanding the topic evolution of scientific literatures like an evolving city: Using Google Word2Vec model and spatial autocorrelation analysis. *Information Processing and Management*, *56* (4), 1185–1203
16. Wolf, L., Hanani, Y., Bar, K., Dershowitz N. (2014) Joint word2vec Networks for Bilingual Semantic Representations. *IJCLA 5*, (1): 27–42
17. Goel, A., Ganesh, L., Kaur, A. (2019) Sustainability integration in the management of construction projects: A morphological analysis of over two decades' research literature. *Journal of Cleaner Production*, *236*, 117676

18. Lee, H., Park, G., Kim, H. (2018) Effective integration of morphological analysis and named entity recognition based on a recurrent neural network. *Pattern Recognition Letters*, *112*, 361–365
19. Mikolov, T., Chen, K., Corrado, G., Dean, J. (2013) Efficient estimation of word representations in vector space. arXiv preprint. arXiv: 1301.3781
20. Church, K. (2017). Word2Vec. *Natural Language Engineering, 23*(1), 155–162.
21. Jianqiang, L., Jing, L., Xianghua, F., Masud, M., Zhexue, H. (2016) Learning distributed word representation with multi-contextual mixed embedding. *Knowledge-Based Systems*, *106*, 220–230
22. Carrasco, R., & Sicilia, M. (2018). Unsupervised intrusion detection through skip-gram models of network behavior. *Computers and Security, 78*, 187–197.
23. Lior, R., Maimon, O. (2005) Clustering methods - Data mining and knowledge discovery, handbook, (Springer US), 321−352
24. Chakraborty, S., Paul, D., & Das, S. (2020). Hierarchical clustering with optimal transport. *Statistics and Probability Letters, 163*, 108781.
25. Xu, Q., Zhang, Q., Liu, J., Luo, B. (2020) Efficient synthetical clustering validity indexes for hierarchical clustering. *Expert Systems with Applications*, *151*, 113367
26. Kim, Hy., Kim, Ha., Cho, S. (2020) Improving spherical k-means for document clustering: Fast initialization, sparse centroid projection, and efficient cluster labeling. *Expert Systems with Applications*, *150*, 113288
27. Bai, L., Liang, J., & Cao, F. (2020). A multiple k-means clustering ensemble algorithm to find nonlinearly separable clusters. *Information Fusion, 61*, 36–47.
28. Takuma, H. (2018) Consideration of feasibility to support function for value indicator management by mathematical analysis for implementation of IoPM. *J. Intern. Assoc. of P2M 13(1)*, 249−259
29. Takuma, H., Hiyama, M. (2015) Discussion of the value indicators for associating projects with programs. *Journal International of Association. of P2M. 10(1)*: 23−34
30. Takuma, H., Iwakami, Y. (2018) Extraction of fundamental KPIs in new product development using Bayesian network analysis. *Proceedings of the 6th Asian Conf. on Innovative Energy and Environmental Chemical Engineering* 163−169
31. Yedidia, J., Freeman, W., Weiss, Y. (2019) Understanding belief propagation and its generalizations. *Mitsubishi Electric Research Laboratories* TR2001–22. Accessed May 30, 2019
32. Sanchez, F., Bonjour, E., Micaelli, J., & Monitcolo, D. (2020). An approach based on bayesian network for improving project management maturity: an application to reduce cost overrun risks in engineering projects. *Computers in Industry, 119*, 103227.
33. Yan, J., Zhang, Z., Lin, K., Yang, F., Luo, X. (2020) A hybrid scheme-based one-vs-all decision trees for multi-class classification tasks. *Knowledge-Based Systems*, *198,* 105922
34. Barsacchi, M., Bechini, A., Marcelloni, F.: An analysis of boosted ensembles of binary fuzzy decision trees. *Expert Systems with Applications*, *154*, 113436