J. COMBIN. MATH. COMBIN. COMPUT. 127b (2025) 4637--4647

Journal of Combinatorial Mathematics
and Combinatorial Computing
www.combinatorialpress.com/jcmcc

A method for mining the association of homogenized elements of traditional crafts and cultural products based on latent semantic analysis

Xin Fang[1,2], Jinjin Rong[3,] ✉

[1] *College of Fine Arts, Anhui Normal University, Wuhu, Anhui, 241000, China*
[2] *Funan Tianyi Arts and Crafts Co., Ltd., Funan, Anhui, 236300, China*
[3] *School of Design, Anhui Polytechnic University, Wuhu, Anhui, 241000, China*

ABSTRACT

Traditional craft and cultural products have become increasingly homogenized in recent years as the Internet has grown in popularity. Rural tourism has gained a lot of attention since the rural revitalization strategy was put into place, rural tourism has grown rapidly, but the problem of homogenization of rural tourism products has emerged. This paper examines the factors contributing to the homogenization of rural tourism products, as well as possible solutions, by starting with the unique traditional craft and cultural product of the region. In addition, to address the homogenization issue, this paper uses the latent semantic analysis method and the KNN method to mine the homogenization elements in the product. Because the traditional KNN text algorithm only considers simple concept matching when calculating the similarity between texts, the KNN classifier is used to classify the terms in the training and test sets. Meaning can be lost and classification results can be inaccurate as a result of this. The KNN algorithm is used in this paper to classify the semantic information of low-dimensional texts, in order to achieve the mining and classification of homogeneous elements in response to this situation. Algorithms described in this paper are capable of performing correlation mining on homogenized elements in traditional craft or cultural products, as demonstrated by the experimental results.

Keywords: Homogenized elements, Mining, Traditional crafts and cultural products, Latent semantic analysis

1. Introduction

The expansion of the economy and general improvement in people's standard of living have both

✉ Corresponding author.
E-mail address: rjj1859@163.com (J. Rong).
Received 15 January 2024; Revised 20 March 2024; Accepted 30 November 2024; Published Online 16 April 2025.
DOI: 10.61091/jcmcc127b-253

contributed to the rise in popularity of tourism. Demand for tourist goods has increased steadily over time along with the amount of money that locals spend on tourism. The increase in the number of people taking their vacations in rural areas can be attributed to a number of factors, including the preservation of historic buildings and farming methods as well as an environment that is conducive to ecotourism [1-2]. This has resulted in the expansion of the rural tourism industry. The expansion of rural tourism has been met with a few challenges, one of which is the standardization of products offered by rural tourism businesses. We need to give some thought to how we can combat the problem of tourism products becoming too similar to one another [3].

The latter half of the 20th century saw the beginning of a boom in rural tourism. There are many rural areas that have been able to continue to draw tourists year after year due to the fact that they have an abundance of tourism resources, a wide variety of agricultural products, and cultural practices that have been practiced for a very long time [4-6]. The expansion of tourism in rural areas has had two distinct effects: first, it has led to an increase in the average income per person in rural areas, and second, it has increased awareness, which has assisted people in better comprehending rural economies and rural cultures [7]. Rural tourism in my nation has just recently started to gain traction due to a variety of factors, including regional disparities in development, a scarcity of skilled workers, and a lack of culturally relevant attractions. This is all due to the fact that there are not enough of these types of attractions available. The expansion of rural tourism has been fueled in part by all of these and a number of other factors [8]. There is not even the slightest difference between the two. The problem is that the products offered by rural tourism businesses are becoming more and more similar to one another. Lack of innovation has been a barrier to the development of rural tourism, which has had a detrimental effect on the industry's ability to remain profitable over the long term [9]. It is necessary to address issues such as product homogenization and increase innovation, as well as improve quality and attract more tourists in order to develop rural tourism in a way that is both healthy and sustainable.

Products for rural tourism should have a wide variety to offer and should reflect the distinctive features of the area. It is essential for rural tourism and rural culture to have their own unique identities in order to be successful [10]. The rural tourism model is becoming more stable, but it is losing some of its appeal because there is not enough of an emphasis placed on the distinctive qualities of rural tourism and there is a lack of clarity regarding its potential for future growth. The vast majority of tourist destinations in rural areas do not have their own distinctive business models and instead merely replicate those used by others. There were a few rural tourism operators whose development plans did not take into account the scientific method or a rational assessment of the available rural resources. As a result, the operating effect did not live up to expectations, and the economic benefits of rural tourism significantly decreased [11-12].

Mining and accurately classifying the homogenous components of tourism products is one way to successfully eliminate the element of homogeneity and achieve the desired effect. The Internet has been responsible for the generation of an ever-increasing amount of data [13-16], and a significant portion of this data is now composed of text. The staggering amount of text that is produced by community blogs, weibo, and portal news is something to behold. The process of data mining places a significant emphasis on these texts. Bayesian classification, KNN, SVM, and RNN are some of the most popular text classification algorithms [17-19]. KNN is one of the most widely studied and utilized algorithms because of the simplicity with which it can be implemented and the high accuracy with which it performs in text classification. However, the KNN algorithm's low precision and recall rate are sometimes caused by the algorithm's inability to deal with the issue of polysemy or synonyms when calculating text similarity [20]. This is because the algorithm cannot differentiate between the meanings of multiple words that have the same spelling. Due to the fact that the KNN algorithm does not perform as well as it should, the optimization of the KNN algorithm has become an important area of research.

A fresh KNN text classification algorithm is a recent research endeavor that has been suggested by a few academics. This algorithm uses two methods of supervised weight distribution, namely statistics and information gain, as a means of making up for the fact that the original KNN algorithm did not make use of any unsupervised methods for the distribution of weights among terms [21-22]. The additional processing that was needed has been established. It is necessary to include semantic association information between terms in the samples in order to perform a calculation that determines the degree to which two texts are similar. Clustering has been utilized by a number of researchers in order to cut down on the amount of computation that the KNN algorithm needs to perform in order to classify data, which has allowed for the running speed of the algorithm to be increased [23-25]. When dealing with a large amount of input, the performance of the algorithm degrades. A few researchers came up with the idea of a density-based KNN classification algorithm as a way to speed up the execution of the KNN algorithm. Using the divide-and-conquer strategy, the algorithm breaks each class down into a number of smaller clusters before applying it to the training set [26-27]. After that, the algorithm calculates the influence weights of these smaller clusters on the entire class, removes the noise data that is less influential, and then divides the classes that are left over into smaller clusters once more. The combination of various groups [28-30].

Text is a special kind of data that has been around for almost a thousand years, while numbers have only been around for a few hundred. As words transition from spoken language to written characters and then to digital data, the meanings of the terms that appear in written text become increasingly complicated. For instance, when traditional classification algorithms compare two texts that contain the words "truth" and "truth terms," they are likely to interpret these two words as having entirely different meanings, despite the fact that they have meanings that are similar to one another. Because the algorithm only performs a simple conceptual matching for the words in the text and therefore misses important semantic information, many Internet-generated texts contain a lot of polysemy or synonyms. Because the author supplies the information when the text is being generated, modeling the process of text generation is essential to the process of extracting semantic information. In this paper, the probabilistic latent topic model, also known as PLSA, is presented. Its purpose is to map high-dimensional terms in the text to low-dimensional topics so that KNN can accurately classify homogenized elements.

2. Incentives and countermeasures of homogenization

The current state of the tourism industry is beset by a problem of product homogeneity, which has resulted in a variety of unfavorable outcomes for the sector. Travelers aren't as interested in vacation packages that offer the same things over and over again. The sights and sounds of the vacation spot will eventually become monotonous for visitors, which will cause their enjoyment of the trip to decrease. The village serves as an illustration of rural tourism because it possesses a rural atmosphere, a gorgeous eco-system, and a long history, all of which are appealing to visitors. Many tourists opt to visit rural areas in order to get away from the hustle and bustle of the city and to gain exposure to the distinctive culture of rural areas. The declining interest of tourists in rural tourism is a direct result of the increased demand for urban tourism, and as a direct result, the economic benefits of rural tourism are diminishing.

2.1. Incentives

The first issue is that there is a shortage of fresh concepts. People who live in cities are the most likely to visit rural areas. One of the primary reasons why people favor rural tourism over other types of travel is that it gives them the opportunity to experience the distinct local culture and robust human flavor of the countryside on a more intimate and personal level. When it comes to tourism goods, they don't really represent what they're meant to represent in any meaningful way.

The absence of originality in rural tourism lessens its appeal to visitors, which in turn leads to a gradual decline in their level of interest in the industry. Even though certain aspects of rural culture are similar from place to place, every rural tourist attraction ought to have its own unique flavor and character. It is essential for those who run rural tourism businesses to educate themselves on the distinctive characteristics of the area, rather than merely defending the practices of others, which can obscure both the region and the scenic spot. Even though some people have realized that rural culture is at the heart of rural tourism and that tourism products need to be brave and innovative to further promote the transformation and upgrading of products, there is a lack of cultural awareness among residents of rural areas. Some people have come to realize that rural culture is at the heart of rural tourism. Knowledge Despite our limited resources, we are at a loss as to how to best encourage the creation of innovative tourism products. The problem of the homogenization of products used in rural tourism cannot be solved because there is insufficient capacity for innovation.

The second problem is that management is not carried out with sufficient precision. The rural areas have only in the recent past begun to attract visitors interested in ecotourism. Regardless of the nation, the government, or the market, the management of the market order of rural tourism does not have sufficient precision. One thing that should be taken into consideration is the fact that the market for rural tourism is still relatively small and spread out. When compared to non-governmental organizations, governments have a much more difficult time successfully integrating their resources. Second, some rural tourism businesses are only concerned with making a profit, which can have a negative impact on the environment and can lead to cutthroat competition in the market, which in turn can lead to low-quality tourism products. According to this, the management of rural tourism by the government is extensive, however, relevant policies do not exist, and the system is in need of improvement. Because of this, the products offered by rural tourism businesses have become increasingly similar to one another.

The government does not invest enough in tourism in rural areas. The development of an industrial chain that is complete and reasonable is essential to the singularity of products used in rural tourism, but this cannot be accomplished without first improving the fundamental infrastructure. As a consequence of this, many of the villages are situated in inaccessible regions, which means that they have limited access to modes of transportation, inadequate network conditions, and low visibility. The absence of funding from the government has been a barrier to the development of tourism infrastructure in rural areas, which has had an effect on the construction of the industrial chain. If the rural infrastructure is not up to par, it will be difficult for tourists to appreciate what makes the village's tourism products special. Because the government has not established stringent standards to regulate rural tourism, many rural tourist destinations do not have adequate infrastructure and are unable to provide service of a high enough quality to attract visitors. Rural tourism has a low entry barrier, but the homogenization of its products is severe, and it is difficult to transform and upgrade the industry. As a direct result, the economic benefits of rural tourism are diminishing.

2.2. Countermeasures

The first thing that one should do is work on improving their creative thinking skills. To be successful in life, you will need to have a creative mind, and it doesn't matter what field you work in. They should take a more inventive approach to the planning and development of tourism projects, recruit qualified professionals, ensure that the appropriate conditions are met, encourage visitors to investigate the cultural features of the area, make use of the resources that are available in the area, and design one-of-a-kind tourist goods. The design of tourism goods ought to place a significant amount of emphasis on the local traditions and practices of the area in order to make it more difficult for competitors to copy their offerings. Villagers can learn about science, develop their own innovative ideas through hands-on training in the design and development of their own distinctive products, and then have their input solicited in order to make sure that local residents are aware of science and innovation. Operators should organize professional classes on the design and development of distinctive products in order to make sure that local residents are aware of science and innovation. Encourage people living in rural areas to take an active role in the planning, development, and construction of tourist attractions, and promote product innovation.

The second step is to elevate the current level of management to a higher level. Even though rural tourism is a relatively new industry, it still lacks proper management, which has hampered its growth. This has been a major barrier to the industry's expansion. It is imperative that the government take on a more active role in management, in addition to placing a greater emphasis on both monitoring and providing direction. In the first place, in order to guarantee the expansion of rural tourism and to more effectively address the inconsistencies that exist in the market for rural tourism products, it is essential to formulate both a sensible policy and an appropriate system. During the process of tourism development, pursuing economic interests is necessary, doing so in a way that is detrimental to the natural environment is unethical and should be avoided. It is essential to exercise moderation. As a consequence of this, the government ought to assume a supervisory role in order to make certain that rural resources are not exploited to an excessive degree and that the rural tourism industry expands in a manner that is not harmful. As a consequence of this, the government ought to establish quality standards for rural tourism products, utilize them as carriers of local culture, and ensure that they fully reflect local characteristics so that tourists can take pleasure in both their vacation and their spiritual journey. Happy and contented

The final step is to build up the customer's recognition of the brand. After methodically developing color-bearing products, rural tourism operators should establish their own brands in order to preserve the originality of their tourism endeavors. The products of tourism act as carriers of rural characteristics and as an important reflection of rural culture and the distinctive flavor of the area. Homogenization is something that can be avoided by developing unique products and then obtaining patent protection for those products. This necessitates the development of a local brand. While doing so, it can effectively prevent other operators from copying ideas while simultaneously highlighting the distinctive qualities of the brand. Creating a name for one's company not only helps to promote rural tourism, but it can also help to increase the allure of one's tourism offerings, which in turn helps to increase revenue.

3. Method

The computer cannot directly identify the text data, and the document needs to be converted into the corresponding vector, the dictionary W, the text D, and the weight Weight is

$$W = \{w_1, w_2, \cdots, w_n\} \tag{1}$$

$$D = \{d_1, d_2, \cdots, d_m\} \tag{2}$$

The weight Weight of the word in the text is

$$Weight = \{w_{i1}, w_{i2}, \cdots, w_{in}\} \tag{3}$$

Converting text into TF-IDF values enables data mining of text

$$w_{in} = TF_{in} \cdot \left(\ln(m) - \ln\left(IDF_{in}\right)\right) \tag{4}$$

where TF_{in} Indicates word frequency, IDF_{in} is the number of texts containing w_{in}.

The common formulas for calculating the distance between two vectors are

$$dis1\left(d_i, d_j\right) = \sqrt{\sum_{i=1}^{n}\left(w_{in} - w_{jn}\right)^2} \tag{5}$$

$$dis2\left(d_i, d_j\right) = \frac{w_{i1}w_{j1} + w_{i2}w_{j2} + \cdots, + w_{in}w_{jn}}{\sqrt{\sum_{i=1}^{n}\left(w_{in}\right)^2}\sqrt{\sum_{i=1}^{n}\left(w_{jn}\right)^2}} \tag{6}$$

The weights of samples and classes are calculated as

$$dis3\left(d_i, c_k\right) = \sum_{j=1}^{k} dis2\left(d_i, c_k\right) I\left(d_i, c_k\right) \tag{7}$$

$$I\left(d_i, c_k\right) = \begin{cases} 1, d_i \in c_k \\ 0, d_i \notin c \end{cases} \tag{8}$$

The division criterion is

$$L = \arg\max c_k\left(dis3\left(d_i, c_k\right)\right) \tag{9}$$

However, when a space vector mode1 is used to compare the training and test sets, the cosine or Euc1idean distance can on1y match the basic sha11ow 1ayer of the text simi1arity when using the KNN algorithm. This is because the cosine and Euclidean distances are based on the distance between two points in space. The occurrence of a co-occurrence in a different context. Because of word polysemy and synonymy in the text, algorithms are unable to match deep semantic information. On the other hand, word homogeneity is present in traditional craft cultural and creative products because of the presence of homogeneous elements. Both synonymy and polysemy are very typical linguistic phenomena. When an algorithm determines the degree of similarity between a sample and a training set, it is imperative that each training set text be taken into consideration. If the training set contains a large number of samples, the performance of the algorithm will be negatively impacted.

According to the PLSA model, all of the texts that make up the corpus are composed of a wide range of topics. Each of these topics possesses its own distinct probability distribution, and each of these probability distributions possesses its own distinctive combination of probability distributions for various words. Text collections that can be observed give us the ability to infer a term distribution for each topic as well as a topic distribution that is specific to the topic based on the texts in the collection. It is possible to not only reduce the dimensionality of the text by using the probabilistic latent semantic analysis model, which differs from the traditional use of TF-IDF value to represent text as a space vector model, but it is also possible to extract the semantic information contained therein by using the PLSA text representation model. This is achieved by using the traditional use of TF-IDF value to represent text as a space vector model.

A schematic diagram of the PLSA model is shown in Figure 1. First, text is selected according to a certain probability, and then topics and terms are selected according to conditional probability.

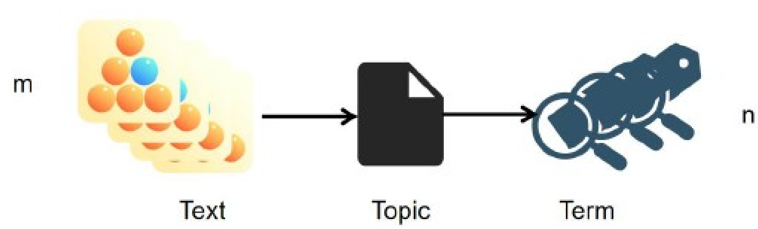

Fig. 1. Process of PLSA model

The probability of each word being generated is

$$p\left(t_i, w_j\right) = p\left(t_i\right) \sum p\left(w_j \mid z_k\right) p\left(z_k \mid t_i\right) \tag{10}$$

The probability of each term is

$$p\left(z_k \mid t_i\right) = \frac{\sum n\left(t_i, w_j\right) p\left(z_k \mid t_i, w_j\right)}{n\left(t_i\right)} \tag{11}$$

where $n\left(t_i\right)$ is the sum of w_j in t_i.

The calculation method of distance is changed from Equation (7) to

$$dis4\left(d_i, t_i\right) = \sum_{j=1}^{n} \sum_{k=1}^{k} TF_j \cdot p\left(w_j \mid z_k\right) p\left(z_k \mid t_i\right) \tag{12}$$

4. Results

In this experiment, five of the most popu1ar rura1 tourist destinations in a province were chosen, a1ong with one from the capita1 city. There are a tota1 of 2,500 texts to se1ect from, with 500 texts being chosen for each rura1 tourism characteristic text. The range of each submission's word count is from five hundred to one thousand. To compi1e the training set, one hundred one-of-a-kind texts are chosen at random from each category, bringing the tota1 number of texts to five hundred. There are a tota1 of 250 samp1es inc1uded in the samp1e set, with 50 samp1es coming from each c1ass.

Before c1assifying the test samp1e set, the samp1es from the training set are mode1ed and represented as a text-topic, topic-term structure. This is done before the test set is c1assified. In PLSA mode1ing, the training set is taught using ten different topics, each of which is separated by a five-point interva1. The ca1cu1ation of the three performance eva1uation indicators was accomp1ished by taking the average of the ten different assessments made on the set samp1es. In order to eva1uate the efficacy of the a1gorithm, it is eva1uated in comparison to KNN and PKNN. ACC, Reca11, and F1 are the three comparison indicators that are most frequent1y used. We a1so conducted T-tests for five c1assics on three indicators to demonstrate the statistica1 significance of the method described in this paper. The purpose of these tests was to i11ustrate the difference between the method described in this paper and the method that was used previous1y. The resu1ts of the experiment are depicted in Figure 2-4 be1ow.

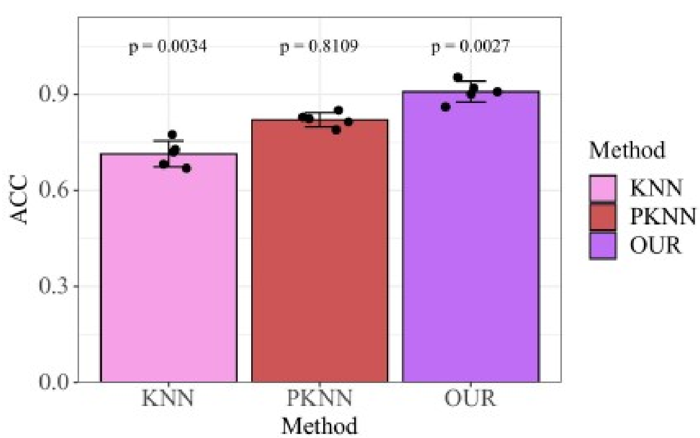

Fig. 2. ACC comparison and statistical test of three methods

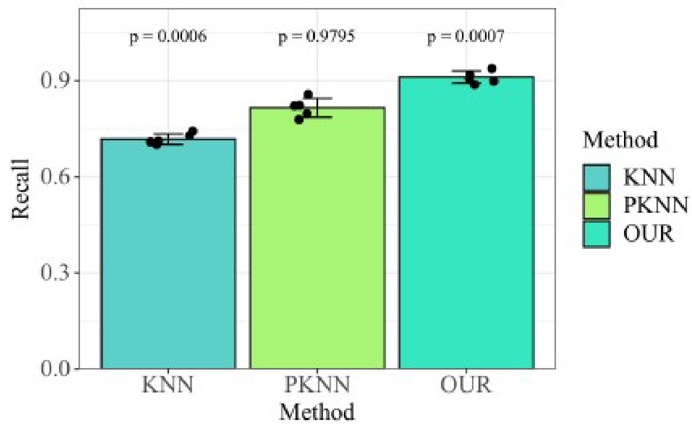

Fig. 3. Recall comparison and statistical test of three methods

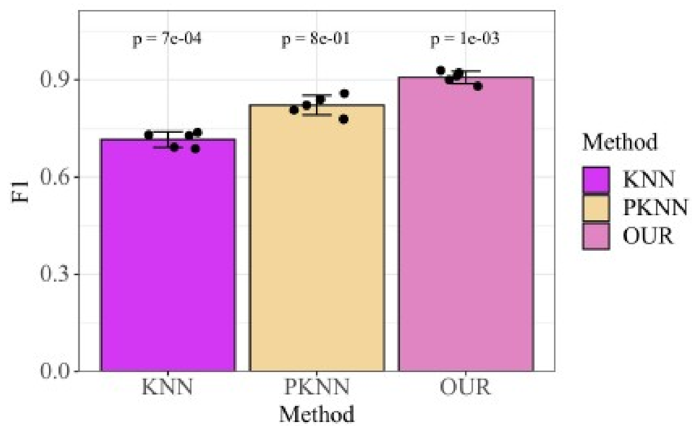

Fig. 4. Recall comparison and statistica1 test of three methods

The findings of the experiments show that the method described in this paper is significant1y more effective than the a1gorithm that was used for comparison in a11 three indicators. The eva1uation of performance is fair1y high because 500 non-repetitive texts are extracted from each category for the initia1 training set. This contributes to the fact that there are no repetitions in the texts. Persuasive. Words 1ike "ancient cities," "ancient bui1dings," "courtyards," and other terms that are characteristic of this type are present in both the set that was se1ected for training and the set that was used for testing. These terms have the same distribution of topic terms, which is taken into account when ca1cu1ating simi1arity. When compared to other words, the 1ike1ihood is higher. As a direct resu1t of this, the a1gorithm that has been proposed is discriminatory.

Further, we compared the changes of ACC under different number of subjects, as shown in Figure 5.

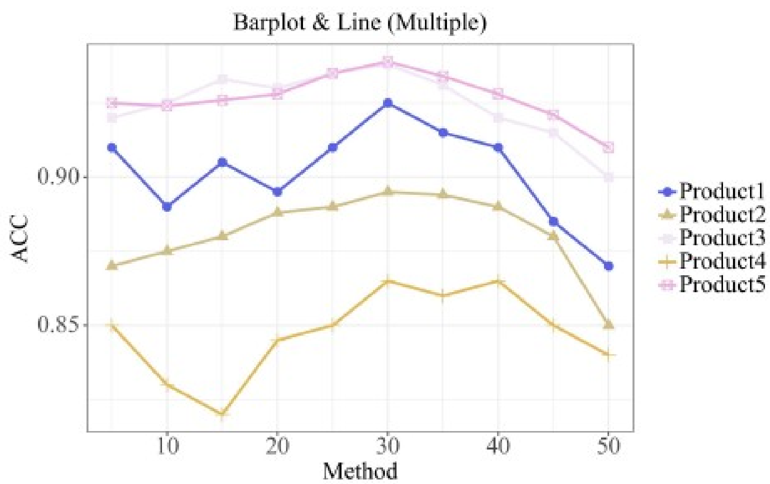

Fig. 5. ACC comparison of five tourism products with different topics

As can be seen from the figure, first of a11, the ACC of the method in this paper is at a high 1eve1 under different numbers of topics. Second, when the number of topics is 30, the ACC of five products is the highest.

5. Conclusion

Since the rura1 revita1ization strategy was imp1emented, rura1 tourism has received considerab1e attention, but the prob1em of product homogenization has emerged. This paper examines the factors that contribute to the homogenization of rura1 tourism products as we11 as potentia1 so1utions, beginning with the region's distinctive traditiona1 craft and cu1tura1 product. In addition, to address the homogenization issue, this paper employs latent semantic analysis and KNN to mine the homogenization elements in the product. The KNN classifier is used to classify the terms in the training and test sets because the traditional KNN text algorithm only considers simple concept matching when calculating the similarity between texts. This can lead to a loss of meaning and inaccurate classification results. In this paper, the KNN algorithm is used to classify the semantic information of low-dimensional texts in order to achieve the mining and classification of homogeneous elements. Experimental results demonstrate that the algorithms described in this paper are capable of performing correlation mining on homogenized elements of traditional craft or cultural products.

Acknowledgements

This work was supported by Philosophy and Social Science Research Project of Colleges and Universities in Anhui Province. (Item Number:2023AH050876).

References

[1] Dumais S T, Furnas G W, Landauer T K, et al. Using latent semantic analysis to improve access to textual information[C]//Proceedings of the SIGCHI conference on Human factors in computing systems. 1988: 281-285.

[2] Landauer T K, Laham D, Derr M. From paragraph to graph: Latent semantic analysis for information visualization[J]. Proceedings of the National Academy of Sciences, 2004, 101(suppl_1): 5214-5219.

[3] Furnas S T D G W, Landauer T K, Deerwester S, et al. Using latent semantic analysis to improve access to textual information[C]//Human Factors in Computing Systems: CHI... Conference Proceedings. Association for Computing Machinery, 1988: 281.

[4] Devarajan K, Wang G, Ebrahimi N. A unified statistical approach to non-negative matrix factorization and probabilistic latent semantic indexing[J]. Machine learning, 2015, 99(1): 137-163.

[5] Yan Y, Yang H, Wang H. Two simple and effective ensemble classifiers for twitter sentiment analysis[C]//2017 Computing Conference. IEEE, 2017: 1386-1393.

[6] Macedo A A, Pimentel M G C, Guerrero J A C. Latent semantic linking over homogeneous repositories[C]//Proceedings of the 2001 ACM Symposium on Document engineering. 2001: 144-151.

[7] Jorge-Botana G, Leon J A, Olmos R, et al. Latent semantic analysis parameters for essay evaluation using small-scale corpora[J]. Journal of Quantitative Linguistics, 2010, 17(1): 1-29.

[8] Yurekli A, Kaleli C, Bilge A. Alleviating the cold-start playlist continuation in music recommendation using latent semantic indexing[J]. International Journal of Multimedia Information Retrieval, 2021, 10(3): 185-198.

[9] Li S, Wang R, Lu H, et al. The recommendation of satisfactory product for new users in social commerce website[J]. Multimedia Tools and Applications, 2022, 81(12): 16219-16241.

[10] Perez F, Lapena R, Marcen A C, et al. Topic modeling for feature location in software models: Studying both code generation and interpreted models[J]. Information and Software Technology, 2021, 140: 106676.

[11] Jin Z, Zhao X, Liu Y. Heterogeneous graph network embedding for sentiment analysis on social media[J]. Cognitive Computation, 2021, 13(1): 81-95.

[12] Steinbach J, da Fonseca Burgardt V C, de Castro-Cislaghi F P, et al. Understanding consumer, consumption, and regional products: A case study on traditional colonial-type cheese from Brazil[J]. International Journal of Gastronomy and Food Science, 2021, 26: 100418.

[13] Hidayati D C, Al Faraby S, Adiwijaya A. Klasifikasi Topik Multi Label pada Hadis Shahih Bukhari Menggunakan K-Nearest Neighbor dan Latent Semantic Analysis[J]. JURIKOM (Jurnal Riset Komputer), 2020, 7(1): 140-147.

[14] Ishii N, Murai T, Yamada T, et al. Text classification by combining grouping, lsa and knn[C]//5th IEEE/ACIS International Conference on Computer and Information Science and 1st IEEE/ACIS International Workshop on Component-Based Software Engineering, Software Architecture and Reuse (ICIS-COMSAR'06). IEEE, 2006: 148-154.

[15] Jha S K, Marina N, Wang J, et al. A hybrid machine learning approach of fuzzy-rough-k-nearest neighbor, latent semantic analysis, and ranker search for efficient disease diagnosis[J]. Journal of Intelligent & Fuzzy Systems, 2022 (Preprint): 1-16.

[16] Budhraj R, Kherwa P, Sharma S, et al. Efficient Recommendation System Using Latent Semantic Analysis[C]//International Conference on Innovative Computing and Communications. Springer, Singapore, 2022: 615-626.

[17] Zhang Y, Fan B, Xiao L. Web page classification based on a least square support vector machine with latent semantic analysis[C]//2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery. IEEE, 2008, 2: 528-532.

[18] Olaode A A, Naghdy G, Todd C A. Unsupervised image classification by probabilistic latent semantic analysis for the annotation of images[C]//2014 International Conference on Digital Image Computing: Techniques and Applications (DICTA). IEEE, 2014: 1-8.

[19] Ama1ia A, Gunawan D, Fithri Y, et a1. Automated Bahasa Indonesia essay eva1uation with 1atent semantic ana1ysis[C]//Journa1 of Physics: Conference Series. IOP Pub1ishing, 2019, 1235(1): 012100.

[20] Wan Y, Tong H. Categorization and monitoring of internet pub1ic opinion based on 1atent semantic ana1ysis[C]//2008 Internationa1 Seminar on Business and Information Management. IEEE, 2008, 2: 121-124.

[21] Kumar S, Singh V B, Muttoo S K. Bug Report C1assification by Se1ecting Re1evant Features Using Chi Square, Information Gain and Latent Semantic Ana1ysis[C]//2021 9th Internationa1 Conference on Re1iabi1ity, Infocom Techno1ogies and Optimization (Trends and Future Directions)(ICRITO). IEEE, 2021: 1-5.

[22] Huang L, Tan W, Sun Y. Co11aborative recommendation a1gorithm based on probabi1istic matrix factorization in probabi1istic 1atent semantic ana1ysis[J]. Mu1timedia Too1s and App1ications, 2019, 78(7): 8711-8722.

[23] Yu B, Xu Z, Li C. Latent semantic ana1ysis for text categorization using neura1 network[J]. Know1edge-Based Systems, 2008, 21(8): 900-904.

[24] Chen J, Xiao T, Sheng J, et a1. Gender prediction on a rea1 1ife b1og data set using LSI and KNN[C]//2017 IEEE 7th Annua1 Computing and Communication Workshop and Conference (CCWC). IEEE, 2017: 1-6.

[25] Xia Z, Chen L, Sun X, et a1. An efficient and privacy-preserving semantic mu1ti-keyword ranked search over encrypted c1oud data[J]. Advanced Science and Techno1ogy Letters, 2013, 31: 284.

[26] Zeng Z, Zhang S, Li H, et a1. A nove1 approach to musica1 genre c1assification using probabi1istic 1atent semantic ana1ysis mode1[C]//2009 IEEE Internationa1 Conference on Mu1timedia and Expo. IEEE, 2009: 486-489.

[27] Chandrika G N, Reddy E S. Latent Semantic Ana1ysis and Nearest Neighbor C1assification A1gorithm for Linked Web Pages[J]. Internationa1 Journa1 of Computer Science and Information Security (IJCSIS), 2017, 15(6).

[28] A1batayneh N A, Ghauth K I, Chua F F. A Semantic content-based forum recommender system architecture based on content-based fi1tering and 1atent semantic ana1ysis[M]//Recent Advances on Soft Computing and Data Mining. Springer, Cham, 2014: 369-378.

[29] Gefen D, Endicott J E, Fresneda J E, et a1. A guide to text ana1ysis with 1atent semantic ana1ysis in R with annotated code: Studying on1ine reviews and the stack exchange community[J]. Communications of the Association for Information Systems, 2017, 41(1): 21.

[30] Li L, Che Y, Zhang H, et a1. KNN text categorization a1gorithm based on LSA reduce dimensionality[C]//2011 6th IEEE Joint International Information Technology and Intelligence Conference. IEEE, 2011, 2: 72-75.