Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/21/20231120

Application and analysis of text similarity in text clustering in the Chinese context

Wentao Fan

Haojing college, Shaanxi University of Science & Technology, Shaanxi, Xi'an, 710021, China

1811431221@mail.sit.edu.cn

Abstract. With the development of the Internet, information sharing is higher, and the amount of information that each user is exposed to is increasing. How to find the information peoples want from so much information is a very important question. The vast majority of these resources are related to textual information. The most intuitive manifestation of these problems is that when people usually use search engines, enter a piece of text, and search out the relevant website, if the algorithm is not good, the search results will be very unsatisfactory. Therefore, this paper studies the application of text similarity in text clustering in the Chinese context. First, the basic concept of text similarity is introduced. In addition, text clustering is explained/explained from three aspects: definition, application, and general processing process. Secondly, combined with the existing data, some mainstream clustering algorithms are comprehensively summarized. Then, combined with the above content, the similarity calculation method in text clustering is analyzed. Finally, the above methods are compared and analyzed according to the experimental results in the Python environment.

Keywords: text similarity, text clustering, artificial intelligence.

1. Introduction

With the development of the Internet, the degree of information sharing has gradually become higher, and the amount of information that each user is exposed to is increasing. In a large amount of resource data, most of the content is related to text information. How to find the information peoples want from so much information is a very important question. The most intuitive manifestation of this problem is that when people usually use search engines, they search for relevant websites by entering a piece of text. If the algorithm is not good, the search results will be very unsatisfactory. If people want to improve the accuracy rate, they need an efficient and accurate text similarity algorithm.

Text similarity calculation is the basic work of organizing and managing massive information. At present, the applications involving text similarity mainly include Internet search engines, similar document retrieval, automatic text classification, text duplicate checking, intelligent answer system, machine translation, and massive information extraction. In these fields, text similarity calculation is an important technical cornerstone of text information extraction, it uses a designed computational model to calculate the degree of similarity between texts. The main thing is to extract the features of a set of text documents, convert them into a mathematical model, and then calculate a similarity degree according to the established mathematical model. The larger the calculated result value, the more similar

DOI: 10.54254/2755-2721/21/20231120

the meaning expressed by the two texts, and the smaller the calculated result value, the very different meanings expressed by the two texts. Chinese originated earlier, the content is rich and diverse, the system is complex, and its properties make the computer processing technology of Chinese far greater than the computer processing difficulty of English. Due to the linguistic characteristics of English, its similarity calculation is relatively simple compared to Chinese, and each word in English documents has its meaning. In English sentences, words are divided with spaces, each word has an independent meaning. But Chinese is not the same, as different words are combined in different word orders to represent different meanings. Therefore, the first step in calculating text similarity in Chinese is word segmentation. Then, text similarity is calculated based on the segmentation results, and the calculation methods and results of Chinese text similarity are different.

Therefore, this paper studies the application of text similarity in text clustering in the Chinese context. First, the basic concept of text similarity is introduced. In addition, text clustering is explained/explained from three aspects: definition, application, and general processing process. Secondly, combined with the existing data, some mainstream clustering algorithms are comprehensively summarized. Then, combined with the above content, the similarity calculation method in text clustering is analyzed. Finally, the above methods are compared and analyzed according to the experimental results in the Python environment.

2. Related technical analysis

Text similarity is the most fundamental part in the field of text mining. A good similarity calculation method can be applied to specific information processing fields such as text classification, text clustering, automatic question answering, document plagiarism detection, etc., greatly improving its computational efficiency. The following chapter will combine the research on clustering, apply some new similarity calculation methods to text clustering, and specifically understand the impact and application of text similarity calculation.

2.1. Overview of text similarity

The concept of textual similarity is widely used in many theories, such as semantics, linguistics, and informatics. But until now there is no universal definition. Mainly because text similarity covers too many areas, such as language, sentence structure, and other factors, such as background, emotion, and other factors. The similarity is used in many fields, and text similarity is used in different ways in different fields. This paper is a study of the similarity between the entire text, using the following relevant knowledge: Chinese word segmentation, text feature vector extraction, longest common subsequence extraction, and similarity calculation.

In all languages, words are the smallest and most basic unit of language. Only by calculating the similarity between words can the similarity of sentences be calculated on top of this result. The similarity of words is highly subjective, and in different contexts, the results can vary greatly, there is no clear scale standard, and the relationship between words is complex. The same word may have different meanings in different contexts, and may even change from a single meaning to a word containing multiple meanings. Therefore, to calculate the similarity of words, people must first determine the specific context in which the words are located, so that the similarity between words can be clearer. Similarity includes many characteristics of the lexical, syntactic, semantic, and even contextual aspects of words, among which the most influential is semantics.

Text similarity refers to the degree of similarity of the meaning expressed by the two texts to be compared, generally expressed by a specific number. This number is greater than or equal to 0 and less than or equal to 1, the size of the numeric value represents the degree of similarity between the two texts, when it is equal to 1, it means that the meaning expressed by the two texts is the same; When it is 0, it means that the meaning of the two texts is completely different and there is no similarity.

DOI: 10.54254/2755-2721/21/20231120

2.2. Introduction of text clustering

2.2.1. Definition and application of text clustering. Objects are clustered by clusters, people are grouped, and classification or clustering is an inherent phenomenon of nature. Similar objects tend to be clustered together, and dissimilar objects tend to be separated. Generalized classification has two meanings: unsupervised classification and supervised classification. Supervised classification refers to the pre-defined classification criteria, as long as it is classified according to the characteristics of the object, such as the number of libraries, the classification of professional disciplines, etc. Unsupervised classification means that objects are automatically grouped into categories based on similar attributes without knowing in advance the number of categories and categories. Such as information retrieval of different users, web hot topic detection, etc. Simply put, classification is the "label" of objects [1]. Gender, ethnicity, age, education, etc., everyone is full of labels, so classification is everywhere. Without a "label" beforehand, through some kinds of set analysis, find out the reason for the clustering between things, this process is called clustering. For example, in the large study rooms of colleges and universities, there are often students sitting in groups of threes and threes, and the reason is generally that they belong to the same class or the same hometown.

In the field of data mining, text classification is often defined as given a classification system and training examples (text labeled with class information), the text is divided into one or more categories. Automatic computer classification is to learn according to the training set that has been marked with category information and use the learned rules for the category determination of new samples (also called test samples). Text classification is supervised learning.

Text clustering is an unguided learning process, which refers to the clustering process under unsupervised conditions based on a certain distance between samples [2]. The clustering method can be used to divide a large amount of text into clusters that users can understand. The similarity of text within the same cluster allows users to grasp the content of large amounts of text faster, speed up analysis, and aid decision-making. Large-scale text clustering is one of the effective means to solve data understanding and information mining in massive text. The main applications of text clustering are as follows:

In terms of clustering display of search results: in today's explosion of information knowledge, a large number of redundant, scattered, useless information is too much, making it very difficult for users to retrieve the content they need, and they often get a lot of irrelevant information when they search through search engines. Using text clustering technology, the retrieved results can be clustered according to the content retrieved by the user, so that the user can browse and improve the performance of the retrieval.

In topic detection and tracking in websites and e-commerce, it is usually hoped to keep abreast of the hot topics and news that are currently receiving the most attention. The most popular goods, user needs, etc. If the text of the relevant information is clustered, new hot topics can be found from the clustered classes in time, so that the website has a high popularity.

In the organization and presentation of large-scale texts, in information processing, it is often necessary to face large-scale unorganized texts, and it is very difficult to classify them manually. At the time, it is necessary to use text clustering to process texts with different similarities on different clusters for easy management and query. In summary, text clustering has many applications, such as improving the recall rate of impulse response (IR) systems, navigation/organizing electronic resources, etc., which play an important role in various information processing fields. The similarity determination is a key step in the clustering algorithm.

2.2.2. General process of text clustering. Clustering is generally processed as the front-end of related applications, and the steps it contains include text representation, clustering (selection and implementation of clustering algorithms), effect evaluation, etc. The details are shown in Figure 1:

DOI: 10.54254/2755-2721/21/20231120

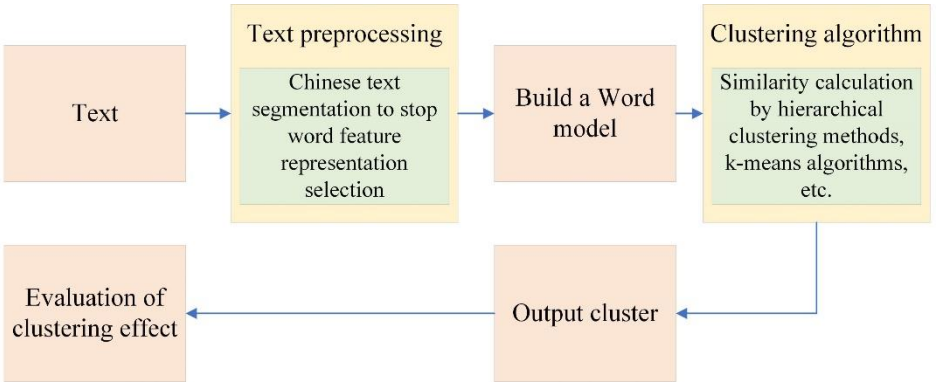

Figure 1. Text clustering flowchart.

First, people need to build a mathematical model of this paper in combination with the pre-processing and input text. This process requires the representation of text data as a form that computers can process, usually known as text pre-processing. It includes Chinese automatic word segmentation, de-stop words, feature representation, and feature selection stages. The process of selecting which words to represent a text is called feature selection [3]. Common feature selection methods include document frequency, information gain, mutual information, and expected crossover. To reduce the amount of calculation, dimensionality reduction processing, such as Latent Semantic Indexing (LSI), is also required. The probability model is a model based on statistical theory, for the uncertainties in the calculation with comparative text, first, make a prediction, divide the text into two sets with probability reasoning, and then use the method of probability and statistics to calculate. Relative to vector space model (VSM), the relationship between terms and text is considered to a certain extent. The LSI model uses Singular Value Decomposition (SVD) singular decomposition technology to map the existing feature space to a more suitable feature space. Complete dimensionality reduction operations from high to low dimensions. Nowadays, the conceptual model is a relatively new concept, and the establishment of the model is based on the so-called "concept", rather than just considering the relationship between words and words, such as China National Knowledge Infrastructure (CNKI) and WordNet systems belonging to this type.

The clustering algorithm is then used for data analysis and processing. The choice of clustering algorithm is often accompanied by the choice of similarity calculation method. In text mining, the most commonly used similarity calculation is cosine similarity. There are many kinds of clustering algorithms, and segmentation clustering, the exemplary example of the latter is based on the characteristics of clustered clusters, clustering techniques are often divided into hierarchical clustering [4]. A more exemplary example of the former is the agglutination hierarchical clustering algorithm K-means. In recent years, some new clustering algorithms have been presented, which are based on different theories or techniques, such as graph theory, neural networks, and kernel techniques, but there is no universal algorithm that can solve all clustering problems. Therefore, it is necessary to carefully study the characteristics of the problem to be solved to select the appropriate algorithm.

Finally, the clustering effect is evaluated based on the output. Because there is no training document set, it is difficult to evaluate the clustering effect. The common method is to select a collection of documents that have been manually classified or marked as a test set, and after clustering, compare the clustering results with people's manual classification results. The commonly used evaluation metric is the F – Measure value [5].

2.3. Analysis of text clustering algorithms

There are various clustering algorithms, which are generally divided into hierarchical clustering methods and segmentation clustering methods. Among them, the more typical and widely used algorithms are condensed hierarchical clustering Hierarchical Agglomerative Clustering (HAC), K-means algorithm, and neural network method.

DOI: 10.54254/2755-2721/21/20231120

Cohesive hierarchical clustering is the most common method, and the process is very similar to the construction of spanning trees in data structures. The algorithm flow is as follows. First, all the points (text in the text collection) are individually formed into a cluster. Then, the two closest (or most similar) clusters from all existing clusters are selected to merge. Finally, if there is only one cluster left or the termination condition is reached (such as reaching the desired number of clusters), the clustering ends, otherwise, return to the second step to continue. It is not difficult to see that HAC ends up forming a category tree. The nodes in the tree are hierarchical, and so are the categories. As nodes (clusters) are merged, the position of each node changes. It contains hierarchical information about classes and similarities within and between all classes. The cohesive hierarchical clustering algorithm has high accuracy, but from its flow, people can know that every time they go to the second step, all clusters must be compared, which will affect the efficiency of the operation when processing large-scale text. The advantage of the K-means algorithm is that the algorithm implementation is relatively simple, Before clustering, it need to give a k value in advance to determine how many clusters to divide. The algorithm flow is as follows: Chief Press, Initialize cluster centers. Second, for each text vector, calculate the distance of the vector from the center of k classes, and select the cluster with the smallest distance (the most similarity) to divide this text into the cluster. Finally, the centers of k clusters are recalculated, and the center is the arithmetic average of all points in the cluster. If the cluster does not change much, or some exit condition is met (reaching the maximum number of iterations or satisfying some objective function, etc.), then end the clustering; Otherwise, return to the second step to continue.

Clustering method K-means clustering. The principle is to randomly select k objects, which point the instance is close to is more likely to be which class, take the center point of each class as a new object, and continue to iterate until the classification is completed. The algorithm is simple and easy to implement, but it is necessary to determine the k value in advance, the comparison is affected by the initial value, and it is very affected by noise and outliers, and the result is locally optimal. k-means sensitive to outliers, k-center is an improvement on k-means, mainly in the selection of center points, the center point is selected as the closest point to the sum of distances from other points as the center point. The improvement reduces the effect of noise and outliers, but the calculation also becomes very complex, suitable for small-scale data, and the clustering method is a common clustering algorithm, the basic principle of which is to divide the data set into subsets, each subset represents a cluster, and then iteratively optimize the quality of the cluster until the stopping condition is met.

The neural network approach describes each cluster as a specimen, which acts as a "prototype" for clustering, not necessarily corresponding to a specific piece of data, and new objects are assigned to the cluster that most closely resembles it based on certain distance measures [6]. The more famous neural network clustering algorithms are competitive learning (competitive learning) and self-organizing feature mapping [7]. The advantages of neural networks are that they are very robust and fault-tolerant because the information is distributed and stored in neurons within the network. Parallel processing methods that make calculations fast. Self-learning, self-organizing, and adaptive so that the network can handle uncertain or unaware systems. Arbitrarily complex nonlinear relationships can be sufficiently approximated. It has strong information synthesis ability, can process quantitative and qualitative information at the same time, can coordinate a variety of input information relationships well, and is suitable for multi-information fusion and multimedia technology. The clustering method of neural networks requires a long processing time and has data complexity that is difficult to analyze, so it is not suitable for clustering large data.

2.4. Similarity calculation in text clustering

In summary, the calculation of similarity is very important in the process of text clustering. Before clustering, it needs to use the text similarity calculation method to establish a similarity matrix, and then use the appropriate clustering algorithm to cluster the clusters. Therefore, a good similarity calculation method can greatly improve the efficiency of clustering [8].

To calculate the similarity of text, it is necessary to go through pre-processing steps such as word segmentation, removal of stop words, and feature selection, and then the representation model of the

DOI: 10.54254/2755-2721/21/20231120

text can be established. In this part, the article adopts a representation method based on the vector space model, and the process and algorithm are described earlier. Next, the similarity calculation method of phased fusion is used to calculate the similarity of two pairs of texts, and a similarity matrix is generated to prepare for further clustering [9, 10].

The implementation of the clustering algorithm in this paper uses the K-means algorithm, the steps to implement it are as follows. First of all $i=1$, initialize k cluster centers $Z_m(i), m=1,2,...k$, k Take separately 3, 5, 10. Then, the distance of each text vector from k to the center of the class is counted:

$$D(x_i, Z_m(i)) = \min\{D(x_i, Z_n(i))\}, m=1,2,\cdots,n; \tag{1}$$

Then, take calculate according to the following objective optimization equation:

$$f(i) = \sum_{m=1}^{k}\sum_{k=1}^{n_m} \left\| x^{(i)} - Z_m(i) \right\|^2 \tag{2}$$

, where represents the distance of the vector X_i to the center of the cluster; If $Z_m(i) = \frac{1}{n}\sum_{i=1}^{n_m} x_i^{(m)}$ then end clustering; otherwise $i = i+1$ recalculate cluster center and start the iteration again.

3. Application analysis of text similarity in text clustering

Since there are not many dedicated clustering datasets at present, this paper adopts 15 articles in each of the two major categories of computer science and Chinese language and literature in a university, and each takes 3 to 10 words in its professional field as a predefined category. The clustering algorithm has an important impact on the application results of text similarity in text clustering, and combined with the previous discussion, the following experiments and analyses are implemented. In the experiment, the traditional VSM-based, the conventional K-means algorithm method and Phased fusion algorithm were used to calculate the similarity of the cluster, and finally, the F value was used to evaluate. The experimental results based on different algorithms and different values of K are shown in Table 1.

Table 1. Comparison of clustering algorithms based on text similarity.

K-means / Algorithm name	K=3	K=5	K=10
The conventional K-means algorithm	0.5278(F - Measure)	0.5164(F - Measure)	0.5221(F - Measure)
Phased fusion algorithm	0.5296(F - Measure)	0.5539(F - Measure)	0.5487(F - Measure)

The results of the experiment after data visualization are shown in Figure 2.

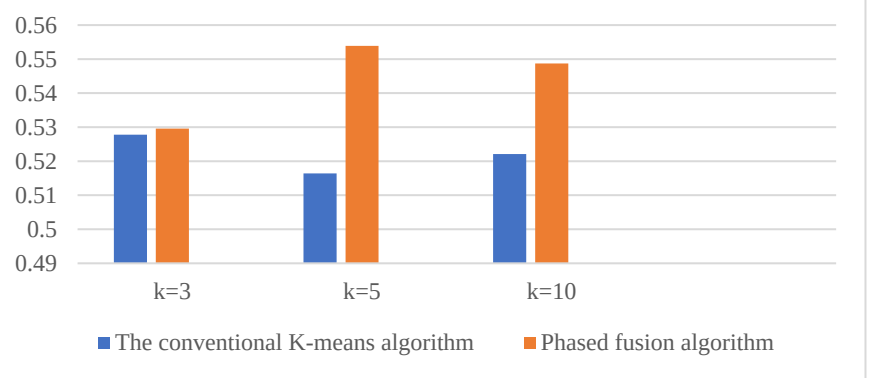

Figure 2. Comparison of clustering algorithms based on text similarity

DOI: 10.54254/2755-2721/21/20231120

It can be seen from the results that under different k-value conditions, the conventional K-means algorithm and Phased fusion algorithm can realize the calculation task in the process of similarity calculation of clustering. The Phased fusion algorithm is more efficient when applied to text clustering than the conventional K-means algorithm.

4. Conclusions

This paper comprehensively summarizes the method of text similarity in text clustering in the Chinese context, and analyses and studies it in combination with application scenarios. First, the basic concept of text similarity is introduced. In addition, text clustering is explained/explained from three aspects: definition, application, and general processing process. Secondly, combined with the existing data, some mainstream clustering algorithms are comprehensively summarized. Then, combined with the above content, the similarity calculation method in text clustering is analyzed. Finally, through experiments in the python environment, the experimental results verify the feasibility of the conventional K-means algorithm and Phased fusion algorithm in the application scenario of this paper, and compare the two algorithms.

References

[1] Bao, J. , et al. "Comparing Different Text Similarity Methods." university of hertfordshire (2007).
[2] Shen, M. , et al. "A Review Expert Recommendation Method Based on Comprehensive Evaluation in Multi-Source Data." CCEAI 2021: 5th International Conference on Control Engineering and Artificial Intelligence 2021.
[3] Guyon, Isabelle M , Andr, and Elisseeff. "An introduction to variable and feature selection." The Journal of Machine Learning Research (2003).
[4] Wilson, H. G. , B. Boots , and A. A. Millward . "A comparison of hierarchical and partitional clustering techniques for multispectral image classification." Geoscience and Remote Sensing Symposium, 2002. IGARSS '02. 2002 IEEE International IEEE, 2002.
[5] George, et al. “protocol, F-Measure, and Reliability in Information Retrieval.” The Journal of the American Medical Informatics Association (2005).
[6] Elhewy, A. H. , E. Mesbahi , and Y. Pu . "Reliability analysis of structures using neural network method." Probabilistic Engineering Mechanics 21.1(2006):44-53.
[7] Kohonen, T. . "The self-organizing map." Proceedings of the IEEE 78.9(2002):1464-1480.
[8] Yu, Y. , and L. Wang . "A Novel Similarity Calculation Method Based on Chinese Sentence Keyword Weight." Journal of Software 9.5(2014):1151-1156.
[9] Wang, X. Z. , and H. C. Shu . "Construction of Fuzzy Similar Matrix." Journal of Jishou University (2003).
[10] Jia, X. , and H. Fang . "Business Flow Analysis Method Based on User Behavior Similarity of Web Logs." Journal of Yangtze University(Natural Science Edition) (2018).